ECO 344 - Applied Econometrics I
Chapter 6

Muhammad Salman Khalid

School of Economics & Social Sciences

October 28, 2026



Outline

1. Data Scaling
2. More Functional Forms
3. More on Goodness of Fit & Selection of Regressors
4. Prediction & Residual Analysis

Data Scaling

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.
- Mostly data scaling is used for

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.
- Mostly data scaling is used for
  - Cosmetics (showing numbers that look good).

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.
- Mostly data scaling is used for
  - Cosmetics (showing numbers that look good).
  - Changing units to appeal to different audiences.

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.
- Mostly data scaling is used for
  - Cosmetics (showing numbers that look good).
  - Changing units to appeal to different audiences.
- We have covered in earlier chapters how the β coefficients change under data scaling.

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.
- Mostly data scaling is used for
  - Cosmetics (showing numbers that look good).
  - Changing units to appeal to different audiences.
- We have covered in earlier chapters how the β coefficients change under data scaling.
- However, we have not discussed what happens to the standard errors, t-test, F-test, R^2 and so on.

Effect of Data Scaling on OLS Statistics

- Data scaling is when variables are multiplied or divided by some constant.
- Mostly data scaling is used for
  - Cosmetics (showing numbers that look good).
  - Changing units to appeal to different audiences.
- We have covered in earlier chapters how the β coefficients change under data scaling.
- However, we have not discussed what happens to the standard errors, t-test, F-test, R^2 and so on.
- The good news is that nothing of substance changes. Scaling is only a change of units.

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- We want to scale the dependent variable by a factor c. The model becomes

$$cy = c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + c\mu$$

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- We want to scale the dependent variable by a factor c. The model becomes

$$cy = c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + c\mu$$

- What is the effect on the OLS statistics?

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- We want to scale the dependent variable by a factor c. The model becomes

$$cy = c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + c\mu$$

- What is the effect on the OLS statistics?
  - All the estimators are scaled by the same factor c. $\hat{\beta}_j^{new} = c\hat{\beta}_j$ for all $j = 0, 1, 2, \ldots, k$.

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- We want to scale the dependent variable by a factor c. The model becomes

$$cy = c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + c\mu$$

- What is the effect on the OLS statistics?
  - All the estimators are scaled by the same factor c. $\hat{\beta}_j^{new} = c\hat{\beta}_j$ for all $j = 0, 1, 2, \ldots, k$.
  - The standard errors are also scaled by c. $se(\hat{\beta}_j^{new}) = c \cdot se(\hat{\beta}_j)$.

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- We want to scale the dependent variable by a factor c. The model becomes

$$cy = c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + c\mu$$

- What is the effect on the OLS statistics?
  - All the estimators are scaled by the same factor c. $\hat{\beta}_j^{new} = c\hat{\beta}_j$ for all $j = 0, 1, 2, \ldots, k$.
  - The standard errors are also scaled by c. $se(\hat{\beta}_j^{new}) = c \cdot se(\hat{\beta}_j)$.
  - The t-statistics do not change. Why?

Data Scaling of the Dependent Variable

- Let we have the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- We want to scale the dependent variable by a factor c. The model becomes

$$cy = c\beta_0 + c\beta_1 x_1 + c\beta_2 x_2 + \cdots + c\beta_k x_k + c\mu$$

- What is the effect on the OLS statistics?
  - All the estimators are scaled by the same factor c. $\hat{\beta}_j^{new} = c\hat{\beta}_j$ for all $j = 0, 1, 2, \ldots, k$.
  - The standard errors are also scaled by c. $se(\hat{\beta}_j^{new}) = c \cdot se(\hat{\beta}_j)$.
  - The t-statistics do not change. Why?
  - Because $t = \hat{\beta}_j / se(\hat{\beta}_j)$ and both the numerator and the denominator are multiplied by c.

Data Scaling of the Dependent Variable

- The residuals are scaled by c, so the sums of squares are scaled by c^2.

$$SST^{new} = c^2 SST, \quad SSE^{new} = c^2 SSE, \quad SSR^{new} = c^2 SSR$$

Data Scaling of the Dependent Variable

- The residuals are scaled by c, so the sums of squares are scaled by c^2.

$$SST^{new} = c^2 SST, \quad SSE^{new} = c^2 SSE, \quad SSR^{new} = c^2 SSR$$

- The standard error of the regression is scaled by c. $\hat{\sigma}^{new} = c\hat{\sigma}$.

Data Scaling of the Dependent Variable

- The residuals are scaled by c, so the sums of squares are scaled by c^2.

$$SST^{new} = c^2 SST, \quad SSE^{new} = c^2 SSE, \quad SSR^{new} = c^2 SSR$$

- The standard error of the regression is scaled by c. $\hat{\sigma}^{new} = c\hat{\sigma}$.
- R^2 does not change, because $R^2 = 1 - \frac{SSR}{SST}$ and c^2 cancels out.

Data Scaling of the Dependent Variable

- The residuals are scaled by c, so the sums of squares are scaled by c^2.

$$SST^{new} = c^2 SST, \quad SSE^{new} = c^2 SSE, \quad SSR^{new} = c^2 SSR$$

- The standard error of the regression is scaled by c. $\hat{\sigma}^{new} = c\hat{\sigma}$.
- R^2 does not change, because $R^2 = 1 - \frac{SSR}{SST}$ and c^2 cancels out.
- The F-statistic does not change. Why?

Data Scaling of the Dependent Variable

- The residuals are scaled by c, so the sums of squares are scaled by c^2.

$$SST^{new} = c^2 SST, \quad SSE^{new} = c^2 SSE, \quad SSR^{new} = c^2 SSR$$

- The standard error of the regression is scaled by c. $\hat{\sigma}^{new} = c\hat{\sigma}$.
- R^2 does not change, because $R^2 = 1 - \frac{SSR}{SST}$ and c^2 cancels out.
- The F-statistic does not change. Why?
- Because $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$ and R^2 has not changed.

Data Scaling of an Independent Variable

- Let we have the same model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Data Scaling of an Independent Variable

- Let we have the same model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Now we scale one of the independent variables by a factor c, so $x_1^{new} = cx_1$. The model becomes

$$y = \beta_0 + \frac{\beta_1}{c}(cx_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Data Scaling of an Independent Variable

- Let we have the same model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Now we scale one of the independent variables by a factor c, so $x_1^{new} = cx_1$. The model becomes

$$y = \beta_0 + \frac{\beta_1}{c}(cx_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- What is the effect on the OLS statistics?

Data Scaling of an Independent Variable

- Let we have the same model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Now we scale one of the independent variables by a factor c, so $x_1^{new} = cx_1$. The model becomes

$$y = \beta_0 + \frac{\beta_1}{c}(cx_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- What is the effect on the OLS statistics?
  - Only the coefficient of the scaled variable changes. It is scaled by the inverse factor. $\hat{\beta}_1^{new} = \frac{1}{c}\hat{\beta}_1$.

Data Scaling of an Independent Variable

- Let we have the same model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Now we scale one of the independent variables by a factor c, so $x_1^{new} = cx_1$. The model becomes

$$y = \beta_0 + \frac{\beta_1}{c}(cx_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- What is the effect on the OLS statistics?
  - Only the coefficient of the scaled variable changes. It is scaled by the inverse factor. $\hat{\beta}_1^{new} = \frac{1}{c}\hat{\beta}_1$.
  - Likewise, only its standard error changes. $se(\hat{\beta}_1^{new}) = \frac{1}{c} se(\hat{\beta}_1)$.

Data Scaling of an Independent Variable

- Let we have the same model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Now we scale one of the independent variables by a factor c, so $x_1^{new} = cx_1$. The model becomes

$$y = \beta_0 + \frac{\beta_1}{c}(cx_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- What is the effect on the OLS statistics?
  - Only the coefficient of the scaled variable changes. It is scaled by the inverse factor. $\hat{\beta}_1^{new} = \frac{1}{c}\hat{\beta}_1$.
  - Likewise, only its standard error changes. $se(\hat{\beta}_1^{new}) = \frac{1}{c} se(\hat{\beta}_1)$.
  - The intercept and all the other coefficients and standard errors stay the same.

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?
- Because the coefficient and the standard error are both multiplied by $\frac{1}{c}$.

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?
- Because the coefficient and the standard error are both multiplied by $\frac{1}{c}$.
- The residuals do not change, so SST, SSE, SSR and $\hat{\sigma}$ do not change.

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?
- Because the coefficient and the standard error are both multiplied by $\frac{1}{c}$.
- The residuals do not change, so SST, SSE, SSR and $\hat{\sigma}$ do not change.
- R^2 does not change.

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?
- Because the coefficient and the standard error are both multiplied by $\frac{1}{c}$.
- The residuals do not change, so SST, SSE, SSR and $\hat{\sigma}$ do not change.
- R^2 does not change.
- The F-statistic does not change. Why?

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?
- Because the coefficient and the standard error are both multiplied by $\frac{1}{c}$.
- The residuals do not change, so SST, SSE, SSR and $\hat{\sigma}$ do not change.
- R^2 does not change.
- The F-statistic does not change. Why?
- Because R^2 has not changed.

Data Scaling of an Independent Variable

- The t-statistic of the scaled variable does not change. Why?
- Because the coefficient and the standard error are both multiplied by $\frac{1}{c}$.
- The residuals do not change, so SST, SSE, SSR and $\hat{\sigma}$ do not change.
- R^2 does not change.
- The F-statistic does not change. Why?
- Because R^2 has not changed.
- In short, scaling x rescales its own slope and standard error. Scaling y rescales everything in units of y.

Data Scaling and Logarithms

- What if the variable we scale is already in logs?

Data Scaling and Logarithms

- What if the variable we scale is already in logs?
- Scaling the dependent variable: $\log(cy) = \log(c) + \log(y)$

$$\log(cy) = [\beta_0 + \log(c)] + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

Data Scaling and Logarithms

- What if the variable we scale is already in logs?
- Scaling the dependent variable: $\log(cy) = \log(c) + \log(y)$

$$\log(cy) = [\beta_0 + \log(c)] + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- Only the intercept changes. It shifts by $\log(c)$.

Data Scaling and Logarithms

- What if the variable we scale is already in logs?
- Scaling the dependent variable: $\log(cy) = \log(c) + \log(y)$

$$\log(cy) = [\beta_0 + \log(c)] + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- Only the intercept changes. It shifts by $\log(c)$.
- Scaling an independent variable: $\log(cx_1) = \log(c) + \log(x_1)$

$$y = [\beta_0 - \beta_1 \log(c)] + \beta_1 \log(cx_1) + \cdots + \mu$$

Data Scaling and Logarithms

- What if the variable we scale is already in logs?
- Scaling the dependent variable: $\log(cy) = \log(c) + \log(y)$

$$\log(cy) = [\beta_0 + \log(c)] + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- Only the intercept changes. It shifts by $\log(c)$.
- Scaling an independent variable: $\log(cx_1) = \log(c) + \log(x_1)$

$$y = [\beta_0 - \beta_1 \log(c)] + \beta_1 \log(cx_1) + \cdots + \mu$$

- Again only the intercept changes. The slope coefficients and all their standard errors stay the same.

Data Scaling and Logarithms

- What if the variable we scale is already in logs?
- Scaling the dependent variable: $\log(cy) = \log(c) + \log(y)$

$$\log(cy) = [\beta_0 + \log(c)] + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$$

- Only the intercept changes. It shifts by $\log(c)$.
- Scaling an independent variable: $\log(cx_1) = \log(c) + \log(x_1)$

$$y = [\beta_0 - \beta_1 \log(c)] + \beta_1 \log(cx_1) + \cdots + \mu$$

- Again only the intercept changes. The slope coefficients and all their standard errors stay the same.
- This is why the units of a logged variable never matter for the slopes.

Example

- Let's go back to the example that we covered in one of the labs: "The Effect of Smoking on Birth Weight". We are interested in the following model:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + \mu$$

where *bwght* is birth weight in ounces, *cigs* is the number of cigarettes smoked per day during pregnancy and *faminc* is family income.

Example

- Let's go back to the example that we covered in one of the labs: "The Effect of Smoking on Birth Weight". We are interested in the following model:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + \mu$$

  where *bwght* is birth weight in ounces, *cigs* is the number of cigarettes smoked per day during pregnancy and *faminc* is family income.
- For an American audience ounces are the natural unit for birth weight.

Example

- Let's go back to the example that we covered in one of the labs: "The Effect of Smoking on Birth Weight". We are interested in the following model:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + \mu$$

  where *bwght* is birth weight in ounces, *cigs* is the number of cigarettes smoked per day during pregnancy and *faminc* is family income.
- For an American audience ounces are the natural unit for birth weight.
- For a Pakistani audience pounds or kilograms make more sense. (1 lb = 16 ounces)

Example

- Let's go back to the example that we covered in one of the labs: "The Effect of Smoking on Birth Weight". We are interested in the following model:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + \mu$$

  where *bwght* is birth weight in ounces, *cigs* is the number of cigarettes smoked per day during pregnancy and *faminc* is family income.
- For an American audience ounces are the natural unit for birth weight.
- For a Pakistani audience pounds or kilograms make more sense. (1 lb = 16 ounces)
- Similarly, for occasional smokers the number of cigarettes makes sense, but for chain smokers the number of packs makes more sense. (1 pack = 20 cigarettes)

Example

- Let's go back to the example that we covered in one of the labs: "The Effect of Smoking on Birth Weight". We are interested in the following model:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + \mu$$

  where *bwght* is birth weight in ounces, *cigs* is the number of cigarettes smoked per day during pregnancy and *faminc* is family income.
- For an American audience ounces are the natural unit for birth weight.
- For a Pakistani audience pounds or kilograms make more sense. (1 lb = 16 ounces)
- Similarly, for occasional smokers the number of cigarettes makes sense, but for chain smokers the number of packs makes more sense. (1 pack = 20 cigarettes)
- The next two slides show the regression with different units of birth weight and cigarettes.

Example

TABLE 6.1 Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables			
cigs	–.4634 (.0916)	–.0289 (.0057)	—
packs	—	—	–9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

Example

Dependent	*bwght*	*bwghtlbs*	*bwght*
Independent	*cigs*	*cigs*	*packs*
cigs	-0.4634 (0.0916)	$-0.4634/16 = -0.0289$ $(0.0916/16 = 0.0057)$	-
packs	-	-	$-0.4634 \times 20 = -9.268$ $(0.0916 \times 20 = 1.832)$
faminc	0.0927 (0.0292)	$0.0927/16 = 0.0058$ $(0.0292/16 = 0.0018)$	0.0927 (0.0292)
intercept	116.974 (1.049)	$116.974/16 = 7.3109$ $(1.049/16 = 0.0656)$	116.974 (1.049)
SSR	557,485.51	$557{,}485.51/16^2 = 2{,}177.68$	557,485.51
SER	20.063	$20.063/16 = 1.2539$	20.063
R^2	0.0298	0.0298	0.0298

Standard errors in parentheses. $n = 1{,}388$ in all three columns.

Reading the Table

- Column 2 divides everything in units of y by 16. The coefficients, standard errors, SER are divided by 16 and SSR is divided by 16^2.

Reading the Table

- Column 2 divides everything in units of y by 16. The coefficients, standard errors, *SER* are divided by 16 and *SSR* is divided by 16^2.
- Column 3 multiplies the *cigs* coefficient and its standard error by 20, because one pack is 20 cigarettes. Everything else is unchanged.

Reading the Table

- Column 2 divides everything in units of y by 16. The coefficients, standard errors, *SER* are divided by 16 and *SSR* is divided by 16^2.
- Column 3 multiplies the *cigs* coefficient and its standard error by 20, because one pack is 20 cigarettes. Everything else is unchanged.
- The t-statistic of *cigs* is the same in all three columns.

$$\frac{-0.4634}{0.0916} \approx \frac{-0.0289}{0.0057} \approx \frac{-9.268}{1.832} \approx -5.06$$

Reading the Table

- Column 2 divides everything in units of y by 16. The coefficients, standard errors, *SER* are divided by 16 and *SSR* is divided by 16^2.
- Column 3 multiplies the *cigs* coefficient and its standard error by 20, because one pack is 20 cigarettes. Everything else is unchanged.
- The t-statistic of *cigs* is the same in all three columns.

$$\frac{-0.4634}{0.0916} \approx \frac{-0.0289}{0.0057} \approx \frac{-9.268}{1.832} \approx -5.06$$

- R^2 and the F-statistic are the same too.

Reading the Table

- Column 2 divides everything in units of y by 16. The coefficients, standard errors, SER are divided by 16 and SSR is divided by 16^2.
- Column 3 multiplies the *cigs* coefficient and its standard error by 20, because one pack is 20 cigarettes. Everything else is unchanged.
- The t-statistic of *cigs* is the same in all three columns.

$$\frac{-0.4634}{0.0916} \approx \frac{-0.0289}{0.0057} \approx \frac{-9.268}{1.832} \approx -5.06$$

- R^2 and the F-statistic are the same too.
- Conclusion: the choice of units is a presentation choice. It never changes what the data say.

Standardized Coefficients

- We read about standardization in Chapter 2 and Chapter 3.

$$z_x = \frac{x - \bar{x}}{\sigma_x}$$

where z_x is the standardized version of x and σ_x is the (sample) standard deviation of x.

Standardized Coefficients

- We read about standardization in Chapter 2 and Chapter 3.

$$z_x = \frac{x - \bar{x}}{\sigma_x}$$

  where z_x is the standardized version of x and σ_x is the (sample) standard deviation of x.

- In cases where units are arbitrary or hard to interpret, we use the standardized versions of the variables.

Standardized Coefficients

- We read about standardization in Chapter 2 and Chapter 3.

$$z_x = \frac{x - \bar{x}}{\sigma_x}$$

  where z_x is the standardized version of x and σ_x is the (sample) standard deviation of x.
- In cases where units are arbitrary or hard to interpret, we use the standardized versions of the variables.
- For example, SAT score as an independent variable (proxy for ability) is hard to interpret, as a different test will have a different scale.

Standardized Coefficients

- We read about standardization in Chapter 2 and Chapter 3.

$$z_x = \frac{x - \bar{x}}{\sigma_x}$$

  where z_x is the standardized version of x and σ_x is the (sample) standard deviation of x.
- In cases where units are arbitrary or hard to interpret, we use the standardized versions of the variables.
- For example, SAT score as an independent variable (proxy for ability) is hard to interpret, as a different test will have a different scale.
- However, we can use the standardized SAT score. This tells us the effect of a 1 standard deviation change in SAT score on the dependent variable.

Standardized Coefficients

- We read about standardization in Chapter 2 and Chapter 3.

$$z_x = \frac{x - \bar{x}}{\sigma_x}$$

  where z_x is the standardized version of x and σ_x is the (sample) standard deviation of x.
- In cases where units are arbitrary or hard to interpret, we use the standardized versions of the variables.
- For example, SAT score as an independent variable (proxy for ability) is hard to interpret, as a different test will have a different scale.
- However, we can use the standardized SAT score. This tells us the effect of a 1 standard deviation change in SAT score on the dependent variable.
- It also lets us compare which independent variable matters more, since all of them are now measured in standard deviations.

Standardized Coefficients

- The regression with standardized variables looks like

$$\frac{y-\bar{y}}{\sigma_y} = b_0 + b_1\frac{x_1-\bar{x}_1}{\sigma_{x_1}} + b_2\frac{x_2-\bar{x}_2}{\sigma_{x_2}} + \cdots + b_k\frac{x_k-\bar{x}_k}{\sigma_{x_k}} + \text{error}$$

Standardized Coefficients

- The regression with standardized variables looks like

$$\frac{y-\bar{y}}{\sigma_y} = b_0 + b_1\frac{x_1-\bar{x}_1}{\sigma_{x_1}} + b_2\frac{x_2-\bar{x}_2}{\sigma_{x_2}} + \cdots + b_k\frac{x_k-\bar{x}_k}{\sigma_{x_k}} + \text{error}$$

- Or

$$z_y = b_0 + b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \text{error}$$

Standardized Coefficients

- The regression with standardized variables looks like

$$\frac{y-\bar{y}}{\sigma_y} = b_0 + b_1\frac{x_1-\bar{x}_1}{\sigma_{x_1}} + b_2\frac{x_2-\bar{x}_2}{\sigma_{x_2}} + \cdots + b_k\frac{x_k-\bar{x}_k}{\sigma_{x_k}} + \text{error}$$

- Or

$$z_y = b_0 + b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \text{error}$$

- We denote the coefficients by b_j and not β_j. b_j is called the **standardized coefficient**.

Standardized Coefficients

- The regression with standardized variables looks like

$$\frac{y-\bar{y}}{\sigma_y} = b_0 + b_1\frac{x_1-\bar{x}_1}{\sigma_{x_1}} + b_2\frac{x_2-\bar{x}_2}{\sigma_{x_2}} + \cdots + b_k\frac{x_k-\bar{x}_k}{\sigma_{x_k}} + \text{error}$$

- Or

$$z_y = b_0 + b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \text{error}$$

- We denote the coefficients by b_j and not β_j. b_j is called the **standardized coefficient**.
- Interpretation: a one standard deviation increase in x_j leads to a b_j standard deviation change in y, holding the other variables constant.

Standardized Coefficients

- How can we get the standardized coefficients without running the regression on standardized variables?

Standardized Coefficients

- How can we get the standardized coefficients without running the regression on standardized variables?
- We start with the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Standardized Coefficients

- How can we get the standardized coefficients without running the regression on standardized variables?
- We start with the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Subtract $\bar{y}$ from both sides and substitute $\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \cdots + \beta_k \bar{x}_k$ (the residuals average to zero).

$$y - \bar{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu - (\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_k \bar{x}_k)$$

Standardized Coefficients

- How can we get the standardized coefficients without running the regression on standardized variables?
- We start with the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- Subtract $\bar{y}$ from both sides and substitute $\bar{y} = \beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \cdots + \beta_k \bar{x}_k$ (the residuals average to zero).

$$y - \bar{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu - (\beta_0 + \beta_1 \bar{x}_1 + \cdots + \beta_k \bar{x}_k)$$

- Re-arranging:

$$y - \bar{y} = \beta_1 (x_1 - \bar{x}_1) + \beta_2 (x_2 - \bar{x}_2) + \cdots + \beta_k (x_k - \bar{x}_k) + \mu$$

Standardized Coefficients

$$y - \bar{y} = \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_k(x_k - \bar{x}_k) + \mu$$

- Divide both sides by σ_y.

$$\frac{y - \bar{y}}{\sigma_y} = \frac{1}{\sigma_y}\left(\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_k(x_k - \bar{x}_k) + \mu\right)$$

Standardized Coefficients

$$y - \bar{y} = \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_k(x_k - \bar{x}_k) + \mu$$

- Divide both sides by σ_y.

$$\frac{y - \bar{y}}{\sigma_y} = \frac{1}{\sigma_y}\left(\beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_k(x_k - \bar{x}_k) + \mu\right)$$

- Multiply and divide each term on the R.H.S. by its own σ_{x_j} for $j = 1, 2, \ldots, k$.

$$\frac{y - \bar{y}}{\sigma_y} = \frac{\sigma_{x_1}}{\sigma_y}\beta_1\frac{(x_1 - \bar{x}_1)}{\sigma_{x_1}} + \cdots + \frac{\sigma_{x_k}}{\sigma_y}\beta_k\frac{(x_k - \bar{x}_k)}{\sigma_{x_k}} + \frac{\mu}{\sigma_y}$$

Standardized Coefficients

- So we get

$$z_y = b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \frac{\mu}{\sigma_y}$$

where

$$b_j = \frac{\sigma_{x_j}}{\sigma_y}\hat{\beta}_j \quad \text{for } j = 1, 2, \ldots, k \qquad \text{and} \qquad b_0 = 0$$

Standardized Coefficients

- So we get

$$z_y = b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \frac{\mu}{\sigma_y}$$

where

$$b_j = \frac{\sigma_{x_j}}{\sigma_y}\hat{\beta}_j \quad \text{for } j = 1, 2, \ldots, k \qquad \text{and} \qquad b_0 = 0$$

- The standardized coefficient is the usual slope coefficient multiplied by the ratio of the standard deviation of that independent variable to the standard deviation of the dependent variable.

Standardized Coefficients

- So we get

$$z_y = b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \frac{\mu}{\sigma_y}$$

where

$$b_j = \frac{\sigma_{x_j}}{\sigma_y}\hat{\beta}_j \quad \text{for } j = 1, 2, \ldots, k \qquad \text{and} \qquad b_0 = 0$$

- The standardized coefficient is the usual slope coefficient multiplied by the ratio of the standard deviation of that independent variable to the standard deviation of the dependent variable.
- The intercept is zero because every standardized variable has mean zero.

Standardized Coefficients

- So we get

$$z_y = b_1 z_1 + b_2 z_2 + \cdots + b_k z_k + \frac{\mu}{\sigma_y}$$

where

$$b_j = \frac{\sigma_{x_j}}{\sigma_y}\hat{\beta}_j \quad \text{for } j = 1, 2, \ldots, k \qquad \text{and} \qquad b_0 = 0$$

- The standardized coefficient is the usual slope coefficient multiplied by the ratio of the standard deviation of that independent variable to the standard deviation of the dependent variable.
- The intercept is zero because every standardized variable has mean zero.
- The t-statistics and R^2 are the same as in the original regression. Standardizing is just a change of units.

Standardized Coefficients - Example

- Example 6.1 in the book looks at the effect of pollution on housing prices, with *nox* (nitrogen oxide in the air) and *rooms* among the regressors.

Standardized Coefficients - Example

- Example 6.1 in the book looks at the effect of pollution on housing prices, with *nox* (nitrogen oxide in the air) and *rooms* among the regressors.
- The standardized coefficient on *nox* is about -0.34. A one standard deviation increase in *nox* lowers the house price by about 0.34 standard deviations.

Standardized Coefficients - Example

- Example 6.1 in the book looks at the effect of pollution on housing prices, with *nox* (nitrogen oxide in the air) and *rooms* among the regressors.
- The standardized coefficient on *nox* is about -0.34. A one standard deviation increase in *nox* lowers the house price by about 0.34 standard deviations.
- The standardized coefficient on *rooms* is about 0.51. A one standard deviation increase in the number of rooms raises the house price by about 0.51 standard deviations.

Standardized Coefficients - Example

- Example 6.1 in the book looks at the effect of pollution on housing prices, with *nox* (nitrogen oxide in the air) and *rooms* among the regressors.
- The standardized coefficient on *nox* is about -0.34. A one standard deviation increase in *nox* lowers the house price by about 0.34 standard deviations.
- The standardized coefficient on *rooms* is about 0.51. A one standard deviation increase in the number of rooms raises the house price by about 0.51 standard deviations.
- Since both are in standard deviation units, we can compare them. In this model *rooms* matters more than *nox*.

Standardized Coefficients - Example

- Example 6.1 in the book looks at the effect of pollution on housing prices, with *nox* (nitrogen oxide in the air) and *rooms* among the regressors.
- The standardized coefficient on *nox* is about -0.34. A one standard deviation increase in *nox* lowers the house price by about 0.34 standard deviations.
- The standardized coefficient on *rooms* is about 0.51. A one standard deviation increase in the number of rooms raises the house price by about 0.51 standard deviations.
- Since both are in standard deviation units, we can compare them. In this model *rooms* matters more than *nox*.
- Go through Example 6.1 in the book.

More Functional Forms

Logarithmic Functional Forms

- We have covered using logarithmic variables in the regression and how it changes the interpretation of the coefficients.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Logarithmic Functional Forms

- We have covered using logarithmic variables in the regression and how it changes the interpretation of the coefficients.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 unit change in x_1 leads to approximately a $100\beta_1$ percent change in y.

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Logarithmic Functional Forms

- We have covered using logarithmic variables in the regression and how it changes the interpretation of the coefficients.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 unit change in x_1 leads to approximately a $100\beta_1$ percent change in y.

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 percent change in x_1 leads to a $\frac{\beta_1}{100}$ unit change in y.

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

Logarithmic Functional Forms

- We have covered using logarithmic variables in the regression and how it changes the interpretation of the coefficients.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 unit change in x_1 leads to approximately a $100\beta_1$ percent change in y.

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 percent change in x_1 leads to a $\frac{\beta_1}{100}$ unit change in y.

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 percent change in x_1 leads to a β_1 percent change in y. (β_1 is an elasticity.)

Logarithmic Functional Forms

- We have covered using logarithmic variables in the regression and how it changes the interpretation of the coefficients.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 unit change in x_1 leads to approximately a $100\beta_1$ percent change in y.

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 percent change in x_1 leads to a $\frac{\beta_1}{100}$ unit change in y.

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \cdots + \beta_k x_k + \mu$$

- A 1 percent change in x_1 leads to a β_1 percent change in y. (β_1 is an elasticity.)
- There are problems with using logarithms that we should be aware of, but first let's look at the benefits.

Benefits of Logarithmic Transformation

- It gives us a nice interpretation of slope estimators (in percentage changes).

Benefits of Logarithmic Transformation

- It gives us a nice interpretation of slope estimators (in percentage changes).
- It can make many skewed distributions closer to normal. (Easier to satisfy M.L.R - VI.)

Benefits of Logarithmic Transformation

- It gives us a nice interpretation of slope estimators (in percentage changes).
- It can make many skewed distributions closer to normal. (Easier to satisfy M.L.R - VI.)
- It reduces the sensitivity of the estimates to outliers, as the log of a large number is a small number.

Benefits of Logarithmic Transformation

- It gives us a nice interpretation of slope estimators (in percentage changes).
- It can make many skewed distributions closer to normal. (Easier to satisfy M.L.R - VI.)
- It reduces the sensitivity of the estimates to outliers, as the log of a large number is a small number.
- Strictly positive variables with large values and outliers are generally transformed to logs (wage, population, sales, firm size, house price and so on).

Benefits of Logarithmic Transformation

- It gives us a nice interpretation of slope estimators (in percentage changes).
- It can make many skewed distributions closer to normal. (Easier to satisfy M.L.R - VI.)
- It reduces the sensitivity of the estimates to outliers, as the log of a large number is a small number.
- Strictly positive variables with large values and outliers are generally transformed to logs (wage, population, sales, firm size, house price and so on).
- Rule of thumb: do not log variables measured in years (education, experience, age) or variables already in percentages or proportions (unemployment rate, interest rate).

Limitations of Logarithmic Transformations

1. For large changes the log approximation is not exact.

Limitations of Logarithmic Transformations

1. For large changes the log approximation is not exact.
   - We derived the interpretation using differentiation, which is valid for small changes.

Limitations of Logarithmic Transformations

1. For large changes the log approximation is not exact.
   - We derived the interpretation using differentiation, which is valid for small changes.
   - For a change Δx_1 the exact percentage change in y is

$$\%\Delta y = 100\,[\exp(\beta_1 \Delta x_1) - 1]$$

Limitations of Logarithmic Transformations

1. For large changes the log approximation is not exact.
   - We derived the interpretation using differentiation, which is valid for small changes.
   - For a change Δx_1 the exact percentage change in y is

$$\%\Delta y = 100\,[\exp(\beta_1 \Delta x_1) - 1]$$

   - For $\beta_1 = 0.2$ and $\Delta x_1 = 1$ the approximation says 20% but the exact change is 22.1%. For $\Delta x_1 = -1$ it is -18.1%, not -20%. (We will cover this in Lab 8.)

Limitations of Logarithmic Transformations

1. For large changes the log approximation is not exact.
   - We derived the interpretation using differentiation, which is valid for small changes.
   - For a change Δx_1 the exact percentage change in y is

$$\%\Delta y = 100\,[\exp(\beta_1 \Delta x_1) - 1]$$

   - For $\beta_1 = 0.2$ and $\Delta x_1 = 1$ the approximation says 20% but the exact change is 22.1%. For $\Delta x_1 = -1$ it is -18.1%, not -20%. (We will cover this in Lab 8.)
2. We cannot take the log of zero or of a negative number. Any variable that takes zero or negative values cannot be transformed to logs. (log(0) is undefined.)

Limitations of Logarithmic Transformations

1. For large changes the log approximation is not exact.
   - We derived the interpretation using differentiation, which is valid for small changes.
   - For a change Δx_1 the exact percentage change in y is

$$\%\Delta y = 100\left[\exp(\beta_1 \Delta x_1) - 1\right]$$

   - For $\beta_1 = 0.2$ and $\Delta x_1 = 1$ the approximation says 20% but the exact change is 22.1%. For $\Delta x_1 = -1$ it is -18.1%, not -20%. (We will cover this in Lab 8.)
2. We cannot take the log of zero or of a negative number. Any variable that takes zero or negative values cannot be transformed to logs. (log(0) is undefined.)
3. A log transformation is not a linear rescaling. Unlike data scaling, the coefficients, standard errors, sums of squares, R^2, t-tests and F-tests all change and we cannot recover them without re-running the regression.

Percentage Change VS Percentage Point Change

- You also need to distinguish between a percentage change and a percentage point change.

Percentage Change VS Percentage Point Change

- You also need to distinguish between a percentage change and a percentage point change.
- Consider the following model:

$$\log(wage) = \beta_0 + \beta_1 \log(sales) + \mu$$

Here β_1 is the percentage change in *wage* due to a 1 percent change in *sales*.

Percentage Change VS Percentage Point Change

- You also need to distinguish between a percentage change and a percentage point change.
- Consider the following model:

$$\log(wage) = \beta_0 + \beta_1 \log(sales) + \mu$$

Here β_1 is the percentage change in *wage* due to a 1 percent change in *sales*.

- Now consider the following model:

$$marks = \beta_0 + \beta_1 attendance + \mu$$

where both *marks* and *attendance* are in percent. Here β_1 is the percentage point change in *marks* due to a 1 percentage point change in *attendance*.

Percentage Change VS Percentage Point Change

- You also need to distinguish between a percentage change and a percentage point change.
- Consider the following model:

$$\log(wage) = \beta_0 + \beta_1 \log(sales) + \mu$$

Here β_1 is the percentage change in *wage* due to a 1 percent change in *sales*.

- Now consider the following model:

$$marks = \beta_0 + \beta_1 attendance + \mu$$

where both *marks* and *attendance* are in percent. Here β_1 is the percentage point change in *marks* due to a 1 percentage point change in *attendance*.

- What's the difference? Moving from 10% to 11% is a 1 percentage point change but a 10 percent change.

Percentage Change VS Percentage Point Change

- You also need to distinguish between a percentage change and a percentage point change.
- Consider the following model:

$$\log(wage) = \beta_0 + \beta_1 \log(sales) + \mu$$

Here β_1 is the percentage change in *wage* due to a 1 percent change in *sales*.

- Now consider the following model:

$$marks = \beta_0 + \beta_1 attendance + \mu$$

where both *marks* and *attendance* are in percent. Here β_1 is the percentage point change in *marks* due to a 1 percentage point change in *attendance*.

- What's the difference? Moving from 10% to 11% is a 1 percentage point change but a 10 percent change.
- If y is in logs and x is a percentage, then a 1 percentage point increase in x changes y by $100\beta_1$ percent.

Predicting y when the Dependent Variable is $\log(y)$

- In any model where the dependent variable is in log form, the fitted value is $\widehat{\log(y)}$ and not $\hat{y}$.

Predicting y when the Dependent Variable is $\log(y)$

- In any model where the dependent variable is in log form, the fitted value is $\widehat{\log(y)}$ and not $\hat{y}$.
- You might think it is easy to convert back, as $e^{\log(c)} = c$. However, simply taking the exponent of the fitted value does not work. It systematically underestimates the expected value of y.

$$\exp\left(\widehat{\log(y)}\right) = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k} \quad \text{underestimates} \quad E[y|x]$$

Predicting y when the Dependent Variable is $\log(y)$

- In any model where the dependent variable is in log form, the fitted value is $\widehat{\log(y)}$ and not $\hat{y}$.
- You might think it is easy to convert back, as $e^{\log(c)} = c$. However, simply taking the exponent of the fitted value does not work. It systematically underestimates the expected value of y.

$$\exp\left(\widehat{\log(y)}\right) = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k} \quad \text{underestimates} \quad E[y|x]$$

- It can be shown that under M.L.R - I through M.L.R - VI (normal errors)

$$\hat{y} = e^{\hat{\sigma}^2/2} \times \exp\left(\widehat{\log(y)}\right)$$

Predicting y when the Dependent Variable is $\log(y)$

- In any model where the dependent variable is in log form, the fitted value is $\widehat{\log(y)}$ and not $\hat{y}$.
- You might think it is easy to convert back, as $e^{\log(c)} = c$. However, simply taking the exponent of the fitted value does not work. It systematically underestimates the expected value of y.

$$\exp\left(\widehat{\log(y)}\right) = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k} \quad \text{underestimates} \quad E[y|x]$$

- It can be shown that under M.L.R - I through M.L.R - VI (normal errors)

$$\hat{y} = e^{\hat{\sigma}^2/2} \times \exp\left(\widehat{\log(y)}\right)$$

- If we drop the normality assumption then

$$\hat{y} = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y)}\right), \quad \text{where} \quad \hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\mu}_i}$$

Predicting y when the Dependent Variable is $\log(y)$

Steps to predict y when the dependent variable is $\log(y)$

1. Obtain the fitted values $\widehat{\log(y)}$ and the residuals $\hat{\mu}_i$ from the regression of $\log(y)$ on $x_1, x_2, \ldots, x_k$.

Predicting y when the Dependent Variable is $\log(y)$

Steps to predict y when the dependent variable is $\log(y)$

1. Obtain the fitted values $\widehat{\log(y)}$ and the residuals $\hat{\mu}_i$ from the regression of $\log(y)$ on $x_1, x_2, \ldots, x_k$.
2. Obtain $\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\mu}_i}$. (It is always greater than 1.)

Predicting y when the Dependent Variable is $\log(y)$

Steps to predict y when the dependent variable is $\log(y)$

1. Obtain the fitted values $\widehat{\log(y)}$ and the residuals $\hat{\mu}_i$ from the regression of $\log(y)$ on $x_1, x_2, \ldots, x_k$.
2. Obtain $\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\mu}_i}$. (It is always greater than 1.)
3. For any given point $c_1, c_2, \ldots, c_k$ get the predicted value of $\log(y)$.

Predicting y when the Dependent Variable is $\log(y)$

Steps to predict y when the dependent variable is $\log(y)$

1. Obtain the fitted values $\widehat{\log(y)}$ and the residuals $\hat{\mu}_i$ from the regression of $\log(y)$ on $x_1, x_2, \ldots, x_k$.
2. Obtain $\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\mu}_i}$. (It is always greater than 1.)
3. For any given point $c_1, c_2, \ldots, c_k$ get the predicted value of $\log(y)$.
4. Obtain the predicted value of y using $\hat{y} = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y)}\right)$.

Predicting y when the Dependent Variable is $\log(y)$

Steps to predict y when the dependent variable is $\log(y)$

1. Obtain the fitted values $\widehat{\log(y)}$ and the residuals $\hat{\mu}_i$ from the regression of $\log(y)$ on $x_1, x_2, \ldots, x_k$.
2. Obtain $\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\mu}_i}$. (It is always greater than 1.)
3. For any given point $c_1, c_2, \ldots, c_k$ get the predicted value of $\log(y)$.
4. Obtain the predicted value of y using $\hat{y} = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y)}\right)$.
   - There are other methods to estimate α_0 as well.

Predicting y when the Dependent Variable is $\log(y)$

Steps to predict y when the dependent variable is $\log(y)$

1. Obtain the fitted values $\widehat{\log(y)}$ and the residuals $\hat{\mu}_i$ from the regression of $\log(y)$ on $x_1, x_2, \ldots, x_k$.
2. Obtain $\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\mu}_i}$. (It is always greater than 1.)
3. For any given point $c_1, c_2, \ldots, c_k$ get the predicted value of $\log(y)$.
4. Obtain the predicted value of y using $\hat{y} = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y)}\right)$.
   - There are other methods to estimate α_0 as well.
   - Go through Example 6.7 in the book.

Goodness of Fit when the Dependent Variable is $\log(y)$

- The R^2 from a $\log(y)$ regression measures the variation explained in $\log(y)$, not in y.

Goodness of Fit when the Dependent Variable is $\log(y)$

- The R^2 from a $\log(y)$ regression measures the variation explained in $\log(y)$, not in y.
- So it cannot be compared with the R^2 of a model for y in levels.

Goodness of Fit when the Dependent Variable is $\log(y)$

- The R^2 from a $\log(y)$ regression measures the variation explained in $\log(y)$, not in y.
- So it cannot be compared with the R^2 of a model for y in levels.
- To make them comparable, compute the predictions of y for each observation

$$\hat{m}_i = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y_i)}\right)$$

Goodness of Fit when the Dependent Variable is $\log(y)$

- The R^2 from a $\log(y)$ regression measures the variation explained in $\log(y)$, not in y.
- So it cannot be compared with the R^2 of a model for y in levels.
- To make them comparable, compute the predictions of y for each observation

$$\hat{m}_i = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y_i)}\right)$$

- Then take the squared correlation between the actual and predicted values

$$R^2 = [Corr(y_i, \hat{m}_i)]^2$$

Goodness of Fit when the Dependent Variable is $\log(y)$

- The R^2 from a $\log(y)$ regression measures the variation explained in $\log(y)$, not in y.
- So it cannot be compared with the R^2 of a model for y in levels.
- To make them comparable, compute the predictions of y for each observation

$$\hat{m}_i = \hat{\alpha}_0 \times \exp\left(\widehat{\log(y_i)}\right)$$

- Then take the squared correlation between the actual and predicted values

$$R^2 = [Corr(y_i, \hat{m}_i)]^2$$

- This number can be compared with the R^2 of the level model. (In a level model it equals the usual R^2.)

Model with Quadratics

- To capture non-linear relationships between the dependent and independent variables, we can include the quadratic form of an independent variable.

Model with Quadratics

- To capture non-linear relationships between the dependent and independent variables, we can include the quadratic form of an independent variable.
- Recall the wage, education and experience model. There might be a non-linear relationship between wage and experience.

Model with Quadratics

- To capture non-linear relationships between the dependent and independent variables, we can include the quadratic form of an independent variable.
- Recall the wage, education and experience model. There might be a non-linear relationship between wage and experience.
  - The effect of experience on wage for inexperienced workers might be different from the effect for experienced workers.

Model with Quadratics

- To capture non-linear relationships between the dependent and independent variables, we can include the quadratic form of an independent variable.
- Recall the wage, education and experience model. There might be a non-linear relationship between wage and experience.
  - The effect of experience on wage for inexperienced workers might be different from the effect for experienced workers.
- To address this, we add the quadratic form of experience to our model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \mu$$

Model with Quadratics

- To capture non-linear relationships between the dependent and independent variables, we can include the quadratic form of an independent variable.
- Recall the wage, education and experience model. There might be a non-linear relationship between wage and experience.
  - The effect of experience on wage for inexperienced workers might be different from the effect for experienced workers.
- To address this, we add the quadratic form of experience to our model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \mu$$

- How do we interpret the effect of experience on wage now? We cannot hold $exper^2$ fixed while changing $exper$, so we differentiate:

$$\frac{\Delta wage}{\Delta exper} \approx \beta_2 + 2\beta_3 exper$$

Model with Quadratics

- To capture non-linear relationships between the dependent and independent variables, we can include the quadratic form of an independent variable.
- Recall the wage, education and experience model. There might be a non-linear relationship between wage and experience.
  - The effect of experience on wage for inexperienced workers might be different from the effect for experienced workers.
- To address this, we add the quadratic form of experience to our model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \mu$$

- How do we interpret the effect of experience on wage now? We cannot hold $exper^2$ fixed while changing $exper$, so we differentiate:

$$\frac{\Delta wage}{\Delta exper} \approx \beta_2 + 2\beta_3 exper$$

- The effect of experience on wage now depends on the level of experience itself.

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

- Based on the model, what is the effect of experience on wage?

$$\frac{\Delta wage}{\Delta exper} \approx 0.27 - 0.008\, exper$$

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

- Based on the model, what is the effect of experience on wage?

$$\frac{\Delta wage}{\Delta exper} \approx 0.27 - 0.008\, exper$$

- The effect of one more year of experience on wage is

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

- Based on the model, what is the effect of experience on wage?

$$\frac{\Delta wage}{\Delta exper} \approx 0.27 - 0.008\, exper$$

- The effect of one more year of experience on wage is
  - For an individual with 0 years of experience the marginal effect is 0.27 USD.

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

- Based on the model, what is the effect of experience on wage?

$$\frac{\Delta wage}{\Delta exper} \approx 0.27 - 0.008\, exper$$

- The effect of one more year of experience on wage is
  - For an individual with 0 years of experience the marginal effect is 0.27 USD.
  - For an individual with 10 years of experience the marginal effect is 0.19 USD.

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

- Based on the model, what is the effect of experience on wage?

$$\frac{\Delta wage}{\Delta exper} \approx 0.27 - 0.008\, exper$$

- The effect of one more year of experience on wage is
  - For an individual with 0 years of experience the marginal effect is 0.27 USD.
  - For an individual with 10 years of experience the marginal effect is 0.19 USD.
  - For an individual with 20 years of experience the marginal effect is 0.11 USD.

Model with Quadratics

- Let we have the following estimates after running the model

$$\widehat{wage} = -3.96 + 0.6\, educ + 0.27\, exper - 0.004\, exper^2$$

- Based on the model, what is the effect of experience on wage?

$$\frac{\Delta wage}{\Delta exper} \approx 0.27 - 0.008\, exper$$

- The effect of one more year of experience on wage is
  - For an individual with 0 years of experience the marginal effect is 0.27 USD.
  - For an individual with 10 years of experience the marginal effect is 0.19 USD.
  - For an individual with 20 years of experience the marginal effect is 0.11 USD.
  - At what level of experience is the effect zero? At $exper = \frac{0.27}{0.008} = 33.75$ years. Beyond that, the effect is negative.

Model with Quadratics

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + \mu$$

- In the model above, there are four possibilities based on the signs of the coefficients of x_1 and x_1^2.

Model with Quadratics

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + \mu$$

- In the model above, there are four possibilities based on the signs of the coefficients of x_1 and x_1^2.
  1. $\beta_1 > 0$, $\beta_2 > 0$ - The effect of x_1 on y is positive and increasing with x_1. It grows without bound as x_1 gets large.

Model with Quadratics

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + \mu$$

- In the model above, there are four possibilities based on the signs of the coefficients of x_1 and x_1^2.
  1. $\beta_1 > 0$, $\beta_2 > 0$ - The effect of x_1 on y is positive and increasing with x_1. It grows without bound as x_1 gets large.
  2. $\beta_1 > 0$, $\beta_2 < 0$ - The effect is positive but decreasing with x_1. It reaches zero and then turns negative. The turning point is $x_1^* = \frac{\beta_1}{2|\beta_2|}$.

Model with Quadratics

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + \mu$$

- In the model above, there are four possibilities based on the signs of the coefficients of x_1 and x_1^2.
  1. $\beta_1 > 0$, $\beta_2 > 0$ - The effect of x_1 on y is positive and increasing with x_1. It grows without bound as x_1 gets large.
  2. $\beta_1 > 0$, $\beta_2 < 0$ - The effect is positive but decreasing with x_1. It reaches zero and then turns negative. The turning point is $x_1^* = \frac{\beta_1}{2|\beta_2|}$.
  3. $\beta_1 < 0$, $\beta_2 < 0$ - The effect of x_1 on y is negative and decreasing with x_1. It goes to $-\infty$ as x_1 gets large.

Model with Quadratics

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \cdots + \beta_k x_{k-1} + \mu$$

- In the model above, there are four possibilities based on the signs of the coefficients of x_1 and x_1^2.
  1. $\beta_1 > 0$, $\beta_2 > 0$ - The effect of x_1 on y is positive and increasing with x_1. It grows without bound as x_1 gets large.
  2. $\beta_1 > 0$, $\beta_2 < 0$ - The effect is positive but decreasing with x_1. It reaches zero and then turns negative. The turning point is $x_1^* = \frac{\beta_1}{2|\beta_2|}$.
  3. $\beta_1 < 0$, $\beta_2 < 0$ - The effect of x_1 on y is negative and decreasing with x_1. It goes to $-\infty$ as x_1 gets large.
  4. $\beta_1 < 0$, $\beta_2 > 0$ - The effect is negative but increasing with x_1. It reaches zero and then turns positive. The turning point is $x_1^* = \frac{|\beta_1|}{2\beta_2}$.

Model with Quadratics

- The book estimates the relationship between wage and experience (Figure 6.1) as

$$\widehat{wage} = 3.73 + 0.298\, exper - 0.0061\, exper^2$$

Model with Quadratics

- The book estimates the relationship between wage and experience (Figure 6.1) as

$$\widehat{wage} = 3.73 + 0.298\, exper - 0.0061\, exper^2$$

- Here $\beta_1 > 0$ and $\beta_2 < 0$, so the effect of experience is positive but diminishing.

Model with Quadratics

- The book estimates the relationship between wage and experience (Figure 6.1) as

$$\widehat{wage} = 3.73 + 0.298\, exper - 0.0061\, exper^2$$

- Here $\beta_1 > 0$ and $\beta_2 < 0$, so the effect of experience is positive but diminishing.
- The turning point is

$$exper^* = \frac{0.298}{2 \times 0.0061} \approx 24.4 \text{ years}$$

Model with Quadratics

- The book estimates the relationship between wage and experience (Figure 6.1) as

$$\widehat{wage} = 3.73 + 0.298\, exper - 0.0061\, exper^2$$

- Here $\beta_1 > 0$ and $\beta_2 < 0$, so the effect of experience is positive but diminishing.
- The turning point is

$$exper^* = \frac{0.298}{2 \times 0.0061} \approx 24.4 \text{ years}$$

- Up to 24.4 years more experience raises the predicted wage. After that, more experience lowers it.

Model with Quadratics

- The book estimates the relationship between wage and experience (Figure 6.1) as

$$\widehat{wage} = 3.73 + 0.298\, exper - 0.0061\, exper^2$$

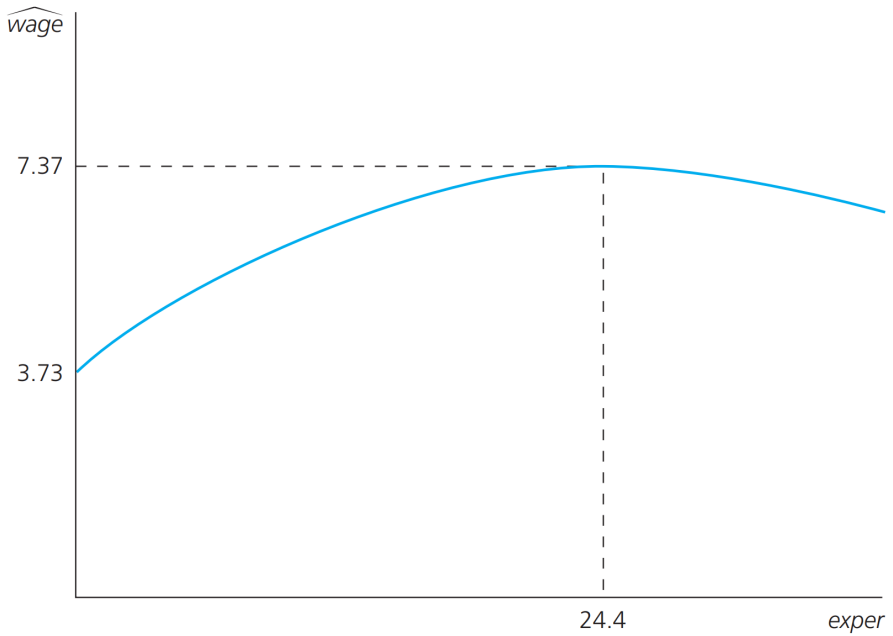
- Here $\beta_1 > 0$ and $\beta_2 < 0$, so the effect of experience is positive but diminishing.
- The turning point is

$$exper^* = \frac{0.298}{2 \times 0.0061} \approx 24.4 \text{ years}$$

- Up to 24.4 years more experience raises the predicted wage. After that, more experience lowers it.
- Always check how many observations lie beyond the turning point. If almost none do, the downward part of the curve is not something the data can speak about.

Model with Quadratics

FIGURE 6.1 Quadratic relationship between $\widehat{wage}$ and *exper*.



Model with Quadratics

Average Partial Effect

- As the effect of x_1 on y depends on x_1 itself, what is the simplest way to summarize the effect of x_1 on y?

$$\frac{\Delta y}{\Delta x_1} \approx \beta_1 + 2\beta_2 x_1$$

Model with Quadratics

Average Partial Effect

- As the effect of x_1 on y depends on x_1 itself, what is the simplest way to summarize the effect of x_1 on y?

$$\frac{\Delta y}{\Delta x_1} \approx \beta_1 + 2\beta_2 x_1$$

- One approach is to use the **average partial effect**.

Model with Quadratics

Average Partial Effect

- As the effect of x_1 on y depends on x_1 itself, what is the simplest way to summarize the effect of x_1 on y?

$$\frac{\Delta y}{\Delta x_1} \approx \beta_1 + 2\beta_2 x_1$$

- One approach is to use the **average partial effect**.
- The average partial effect is the effect of x_1 on y at the average value of x_1. It is a single number and does not depend on x_1 explicitly.

$$\text{Average Partial Effect} = \beta_1 + 2\beta_2 \bar{x}_1$$

Model with Quadratics

Average Partial Effect

- As the effect of x_1 on y depends on x_1 itself, what is the simplest way to summarize the effect of x_1 on y?

$$\frac{\Delta y}{\Delta x_1} \approx \beta_1 + 2\beta_2 x_1$$

- One approach is to use the **average partial effect**.
- The average partial effect is the effect of x_1 on y at the average value of x_1. It is a single number and does not depend on x_1 explicitly.

$$\text{Average Partial Effect} = \beta_1 + 2\beta_2 \bar{x}_1$$

- In our wage example, with average experience of 17 years:

$$0.27 - 0.008 \times \overline{exper} = 0.27 - 0.008 \times 17 = 0.134$$

Model with Quadratics

Average Partial Effect

- As the effect of x_1 on y depends on x_1 itself, what is the simplest way to summarize the effect of x_1 on y?

$$\frac{\Delta y}{\Delta x_1} \approx \beta_1 + 2\beta_2 x_1$$

- One approach is to use the **average partial effect**.
- The average partial effect is the effect of x_1 on y at the average value of x_1. It is a single number and does not depend on x_1 explicitly.

$$\text{Average Partial Effect} = \beta_1 + 2\beta_2 \bar{x}_1$$

- In our wage example, with average experience of 17 years:

$$0.27 - 0.008 \times \overline{exper} = 0.27 - 0.008 \times 17 = 0.134$$

- So on average, one more year of experience raises the wage by about 0.134 USD.

Interaction Term

- The effect of x_1 on y can also depend on other variables.

Interaction Term

- The effect of x_1 on y can also depend on other variables.
- For example, the effect of an extra bedroom on house price may also depend on the square footage of the house.

Interaction Term

- The effect of x_1 on y can also depend on other variables.
- For example, the effect of an extra bedroom on house price may also depend on the square footage of the house.
- To capture such effects we use interaction terms.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 (sqrft \times bdrms) + \mu$$

Interaction Term

- The effect of x_1 on y can also depend on other variables.
- For example, the effect of an extra bedroom on house price may also depend on the square footage of the house.
- To capture such effects we use interaction terms.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 (sqrft \times bdrms) + \mu$$

- What is the effect of an extra bedroom on the price of a house?

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

Interaction Term

- The effect of x_1 on y can also depend on other variables.
- For example, the effect of an extra bedroom on house price may also depend on the square footage of the house.
- To capture such effects we use interaction terms.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 (sqrft \times bdrms) + \mu$$

- What is the effect of an extra bedroom on the price of a house?

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

- If $\beta_3 > 0$, adding an extra bedroom to a larger house has a higher effect on price.

Interaction Term

- The effect of x_1 on y can also depend on other variables.
- For example, the effect of an extra bedroom on house price may also depend on the square footage of the house.
- To capture such effects we use interaction terms.

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 (sqrft \times bdrms) + \mu$$

- What is the effect of an extra bedroom on the price of a house?

$$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

- If $\beta_3 > 0$, adding an extra bedroom to a larger house has a higher effect on price.
- Similarly, the effect of square footage depends on the number of bedrooms.

$$\frac{\Delta price}{\Delta sqrft} = \beta_1 + \beta_3 bdrms$$

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.
- The effect of x_1 on y is $\beta_1 + \beta_3 x_2$. What is the interpretation of β_1?

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.
- The effect of x_1 on y is $\beta_1 + \beta_3 x_2$. What is the interpretation of β_1?
- It is the effect of x_1 on y when x_2 is zero.

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.
- The effect of x_1 on y is $\beta_1 + \beta_3 x_2$. What is the interpretation of β_1?
- It is the effect of x_1 on y when x_2 is zero.
- But think about the house price example. Does it make sense to say that β_2 is the increase in price from an extra bedroom in a house with zero square feet?

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.
- The effect of x_1 on y is $\beta_1 + \beta_3 x_2$. What is the interpretation of β_1?
- It is the effect of x_1 on y when x_2 is zero.
- But think about the house price example. Does it make sense to say that β_2 is the increase in price from an extra bedroom in a house with zero square feet?
- It is better to report the **average partial effect**. For example, $\beta_2 + \beta_3 \overline{sqrft}$ is the increase in price from an extra bedroom in a house of average size.

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.
- The effect of x_1 on y is $\beta_1 + \beta_3 x_2$. What is the interpretation of β_1?
- It is the effect of x_1 on y when x_2 is zero.
- But think about the house price example. Does it make sense to say that β_2 is the increase in price from an extra bedroom in a house with zero square feet?
- It is better to report the **average partial effect**. For example, $\beta_2 + \beta_3 \overline{sqrft}$ is the increase in price from an extra bedroom in a house of average size.
- A neat trick is to center the variable. Use $(sqrft - \overline{sqrft})$ in the interaction instead of *sqrft*. Then the coefficient on *bdrms* is itself the effect at the average size.

Interaction Term

- Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$.
- The effect of x_1 on y is $\beta_1 + \beta_3 x_2$. What is the interpretation of β_1?
- It is the effect of x_1 on y when x_2 is zero.
- But think about the house price example. Does it make sense to say that β_2 is the increase in price from an extra bedroom in a house with zero square feet?
- It is better to report the **average partial effect**. For example, $\beta_2 + \beta_3 \overline{sqrft}$ is the increase in price from an extra bedroom in a house of average size.
- A neat trick is to center the variable. Use $(sqrft - \overline{sqrft})$ in the interaction instead of *sqrft*. Then the coefficient on *bdrms* is itself the effect at the average size.
- Always think about the interpretation of the coefficients when using interaction terms, as it depends on the context.

Interaction Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$$

- In the model above, there are four possibilities based on the signs of β_1 and β_3.

Interaction Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$$

- In the model above, there are four possibilities based on the signs of β_1 and β_3.
  1. $\beta_1 > 0$, $\beta_3 > 0$ - The effect of x_1 on y is positive and increasing with x_2.

Interaction Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$$

- In the model above, there are four possibilities based on the signs of β_1 and β_3.
  1. $\beta_1 > 0$, $\beta_3 > 0$ - The effect of x_1 on y is positive and increasing with x_2.
  2. $\beta_1 > 0$, $\beta_3 < 0$ - The effect is positive but decreasing with x_2. It reaches zero and then turns negative when $x_2 = \frac{\beta_1}{|\beta_3|}$.

Interaction Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$$

- In the model above, there are four possibilities based on the signs of β_1 and β_3.
  1. $\beta_1 > 0$, $\beta_3 > 0$ - The effect of x_1 on y is positive and increasing with x_2.
  2. $\beta_1 > 0$, $\beta_3 < 0$ - The effect is positive but decreasing with x_2. It reaches zero and then turns negative when $x_2 = \frac{\beta_1}{|\beta_3|}$.
  3. $\beta_1 < 0$, $\beta_3 < 0$ - The effect of x_1 on y is negative and decreasing with x_2.

Interaction Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$$

- In the model above, there are four possibilities based on the signs of β_1 and β_3.
  1. $\beta_1 > 0$, $\beta_3 > 0$ - The effect of x_1 on y is positive and increasing with x_2.
  2. $\beta_1 > 0$, $\beta_3 < 0$ - The effect is positive but decreasing with x_2. It reaches zero and then turns negative when $x_2 = \frac{\beta_1}{|\beta_3|}$.
  3. $\beta_1 < 0$, $\beta_3 < 0$ - The effect of x_1 on y is negative and decreasing with x_2.
  4. $\beta_1 < 0$, $\beta_3 > 0$ - The effect is negative but increasing with x_2. It reaches zero and then turns positive when $x_2 = \frac{|\beta_1|}{\beta_3}$.

Interaction Term

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \mu$$

- In the model above, there are four possibilities based on the signs of β_1 and β_3.
  1. $\beta_1 > 0$, $\beta_3 > 0$ - The effect of x_1 on y is positive and increasing with x_2.
  2. $\beta_1 > 0$, $\beta_3 < 0$ - The effect is positive but decreasing with x_2. It reaches zero and then turns negative when $x_2 = \frac{\beta_1}{|\beta_3|}$.
  3. $\beta_1 < 0$, $\beta_3 < 0$ - The effect of x_1 on y is negative and decreasing with x_2.
  4. $\beta_1 < 0$, $\beta_3 > 0$ - The effect is negative but increasing with x_2. It reaches zero and then turns positive when $x_2 = \frac{|\beta_1|}{\beta_3}$.
- For practice, go over Example 6.3 in the book.

More on Goodness of Fit & Selection of Regressors

Adjusted R-Squared

- We have studied that R^2 never falls when we add independent variables, regardless of their relevance.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Adjusted R-Squared

- We have studied that R^2 never falls when we add independent variables, regardless of their relevance.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- The adjusted R^2 addresses this problem by adjusting for degrees of freedom. We write it as $\bar{R}^2$.

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

Note that there is no equivalent formula in terms of SSE.

Adjusted R-Squared

- We have studied that R^2 never falls when we add independent variables, regardless of their relevance.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- The adjusted R^2 addresses this problem by adjusting for degrees of freedom. We write it as $\bar{R}^2$.

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

  Note that there is no equivalent formula in terms of SSE.

- There is now a penalty for adding an independent variable. $\bar{R}^2$ rises only if SSR falls by enough to offset the lost degree of freedom.

Adjusted R-Squared

- We have studied that R^2 never falls when we add independent variables, regardless of their relevance.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- The adjusted R^2 addresses this problem by adjusting for degrees of freedom. We write it as $\bar{R}^2$.

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

  Note that there is no equivalent formula in terms of SSE.
- There is now a penalty for adding an independent variable. $\bar{R}^2$ rises only if SSR falls by enough to offset the lost degree of freedom.
- In fact, $\bar{R}^2$ rises when a variable is added if and only if the absolute value of its t-statistic is greater than 1.

Adjusted R-Squared

- In terms of R^2,

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Adjusted R-Squared

- In terms of R^2,

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

- The interpretation of $\bar{R}^2$ is similar to R^2. It is the proportion of the variation in y explained by the independent variables, adjusted for degrees of freedom.

Adjusted R-Squared

- In terms of R^2,

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

- The interpretation of $\bar{R}^2$ is similar to R^2. It is the proportion of the variation in y explained by the independent variables, adjusted for degrees of freedom.
- $\bar{R}^2$ can increase or decrease when an independent variable is added, depending on how much the variable reduces SSR.

Adjusted R-Squared

- In terms of R^2,

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

- The interpretation of $\bar{R}^2$ is similar to R^2. It is the proportion of the variation in y explained by the independent variables, adjusted for degrees of freedom.
- $\bar{R}^2$ can increase or decrease when an independent variable is added, depending on how much the variable reduces SSR.
- $\bar{R}^2$ is always smaller than R^2 when $k \geq 1$, and it can be negative. A negative value indicates a very poor fit relative to the degrees of freedom.

Choice Between Non-Nested Models

- So far, we have learned to use the F-test to choose between two nested models.

Choice Between Non-Nested Models

- So far, we have learned to use the F-test to choose between two nested models.
- Models are nested when the restricted model is a special case of the unrestricted one. Non-nested models are those where neither is a special case of the other.

Choice Between Non-Nested Models

- So far, we have learned to use the F-test to choose between two nested models.
- Models are nested when the restricted model is a special case of the unrestricted one. Non-nested models are those where neither is a special case of the other.
- Say we have the following two models and want to know which is better.

$$rd = \beta_0 + \beta_1 \log(sales) + \mu$$
$$rd = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \mu$$

Choice Between Non-Nested Models

- So far, we have learned to use the F-test to choose between two nested models.
- Models are nested when the restricted model is a special case of the unrestricted one. Non-nested models are those where neither is a special case of the other.
- Say we have the following two models and want to know which is better.

$$rd = \beta_0 + \beta_1 \log(sales) + \mu$$
$$rd = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \mu$$

- Why would the usual F-test not work for non-nested models?

Choice Between Non-Nested Models

- So far, we have learned to use the F-test to choose between two nested models.
- Models are nested when the restricted model is a special case of the unrestricted one. Non-nested models are those where neither is a special case of the other.
- Say we have the following two models and want to know which is better.

$$rd = \beta_0 + \beta_1 \log(sales) + \mu$$
$$rd = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \mu$$

- Why would the usual F-test not work for non-nested models?
- Because the F-test works by imposing restrictions on a model. Here neither model is a restricted version of the other.

Choice Between Non-Nested Models

- So far, we have learned to use the F-test to choose between two nested models.
- Models are nested when the restricted model is a special case of the unrestricted one. Non-nested models are those where neither is a special case of the other.
- Say we have the following two models and want to know which is better.

$$rd = \beta_0 + \beta_1 \log(sales) + \mu$$
$$rd = \beta_0 + \beta_1 sales + \beta_2 sales^2 + \mu$$

- Why would the usual F-test not work for non-nested models?
- Because the F-test works by imposing restrictions on a model. Here neither model is a restricted version of the other.
- We can compare the adjusted R^2 of the two models instead. The model with the higher $\bar{R}^2$ fits better after the penalty. This works only if both models have the same dependent variable. We cannot compare $\bar{R}^2$ of a model for y with a model for $\log(y)$.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.
- Adjusted R^2 can be one input when deciding whether to keep a variable, but theory should come first.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.
- Adjusted R^2 can be one input when deciding whether to keep a variable, but theory should come first.
- In some cases, it is best not to control for a particular variable.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.
- Adjusted R^2 can be one input when deciding whether to keep a variable, but theory should come first.
- In some cases, it is best not to control for a particular variable.
  - Say you are interested in the ceteris paribus effect of x_1 on y.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.
- Adjusted R^2 can be one input when deciding whether to keep a variable, but theory should come first.
- In some cases, it is best not to control for a particular variable.
  - Say you are interested in the ceteris paribus effect of x_1 on y.
  - And x_1 affects y through w.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.
- Adjusted R^2 can be one input when deciding whether to keep a variable, but theory should come first.
- In some cases, it is best not to control for a particular variable.
  - Say you are interested in the ceteris paribus effect of x_1 on y.
  - And x_1 affects y through w.
  - If you control for w, you will find an insignificant effect of x_1 on y, as ceteris paribus now keeps w constant too.

Controlling for too Many Variables

- As we know, an omitted variable can lead to biased estimates.
- The worry about bias can sometimes lead to **over controlling** the model.
- Over controlling is when we control for too many variables, including some that should not be held fixed.
- Adjusted R^2 can be one input when deciding whether to keep a variable, but theory should come first.
- In some cases, it is best not to control for a particular variable.
  - Say you are interested in the ceteris paribus effect of x_1 on y.
  - And x_1 affects y through w.
  - If you control for w, you will find an insignificant effect of x_1 on y, as ceteris paribus now keeps w constant too.
  - Example: if we want the effect of beer taxes on traffic fatalities, we should not control for beer consumption. The tax works through consumption.

Adding Regressors to Reduce the Error Variance

- We know we should add variables that affect y and are correlated with the independent variable of interest, to avoid omitted variable bias.

Adding Regressors to Reduce the Error Variance

- We know we should add variables that affect y and are correlated with the independent variable of interest, to avoid omitted variable bias.
- What about variables that affect y but are uncorrelated with the independent variables?

Adding Regressors to Reduce the Error Variance

- We know we should add variables that affect y and are correlated with the independent variable of interest, to avoid omitted variable bias.
- What about variables that affect y but are uncorrelated with the independent variables?
- They do not cause bias when left out. But we should still include them if they significantly decrease the error variance (SSR).

Adding Regressors to Reduce the Error Variance

- We know we should add variables that affect y and are correlated with the independent variable of interest, to avoid omitted variable bias.
- What about variables that affect y but are uncorrelated with the independent variables?
- They do not cause bias when left out. But we should still include them if they significantly decrease the error variance (SSR).
- Since they are uncorrelated with the other regressors, they do not add to multicollinearity. VIF does not rise.

Adding Regressors to Reduce the Error Variance

- We know we should add variables that affect y and are correlated with the independent variable of interest, to avoid omitted variable bias.
- What about variables that affect y but are uncorrelated with the independent variables?
- They do not cause bias when left out. But we should still include them if they significantly decrease the error variance (SSR).
- Since they are uncorrelated with the other regressors, they do not add to multicollinearity. VIF does not rise.
- A smaller error variance gives us a lower sampling variance and thus lower standard errors for the estimators, leading to more precise estimates.

Adding Regressors to Reduce the Error Variance

- We know we should add variables that affect y and are correlated with the independent variable of interest, to avoid omitted variable bias.
- What about variables that affect y but are uncorrelated with the independent variables?
- They do not cause bias when left out. But we should still include them if they significantly decrease the error variance (SSR).
- Since they are uncorrelated with the other regressors, they do not add to multicollinearity. VIF does not rise.
- A smaller error variance gives us a lower sampling variance and thus lower standard errors for the estimators, leading to more precise estimates.
- This is especially useful in experiments where the treatment is randomly assigned.

Prediction & Residual Analysis

Confidence Interval of $\hat{y}$

- As we know, we can find a confidence interval by

$$\hat{c} - t_{\frac{\alpha}{2}} \times se(\hat{c}) \leq c \leq \hat{c} + t_{\frac{\alpha}{2}} \times se(\hat{c})$$

where c is the population parameter and $\hat{c}$ is the sample estimator. ($t_{\alpha/2}$ has $n - k - 1$ degrees of freedom and is close to 1.96 for large n.)

Confidence Interval of $\hat{y}$

- As we know, we can find a confidence interval by

$$\hat{c} - t_{\frac{\alpha}{2}} \times se(\hat{c}) \leq c \leq \hat{c} + t_{\frac{\alpha}{2}} \times se(\hat{c})$$

  where c is the population parameter and $\hat{c}$ is the sample estimator. ($t_{\alpha/2}$ has $n - k - 1$ degrees of freedom and is close to 1.96 for large n.)
- So we need the point estimate and its standard error to construct a confidence interval.

Confidence Interval of $\hat{y}$

- As we know, we can find a confidence interval by

$$\hat{c} - t_{\frac{\alpha}{2}} \times se(\hat{c}) \leq c \leq \hat{c} + t_{\frac{\alpha}{2}} \times se(\hat{c})$$

  where c is the population parameter and $\hat{c}$ is the sample estimator. ($t_{\alpha/2}$ has $n - k - 1$ degrees of freedom and is close to 1.96 for large n.)
- So we need the point estimate and its standard error to construct a confidence interval.
- For the predicted value the point estimate is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Confidence Interval of $\hat{y}$

- As we know, we can find a confidence interval by

$$\hat{c} - t_{\frac{\alpha}{2}} \times se(\hat{c}) \leq c \leq \hat{c} + t_{\frac{\alpha}{2}} \times se(\hat{c})$$

  where c is the population parameter and $\hat{c}$ is the sample estimator. ($t_{\alpha/2}$ has $n - k - 1$ degrees of freedom and is close to 1.96 for large n.)
- So we need the point estimate and its standard error to construct a confidence interval.
- For the predicted value the point estimate is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- What about the standard error of $\hat{y}$?

Confidence Interval of $\hat{y}$

- As we know, we can find a confidence interval by

$$\hat{c} - t_{\frac{\alpha}{2}} \times se(\hat{c}) \leq c \leq \hat{c} + t_{\frac{\alpha}{2}} \times se(\hat{c})$$

  where c is the population parameter and $\hat{c}$ is the sample estimator. ($t_{\alpha/2}$ has $n - k - 1$ degrees of freedom and is close to 1.96 for large n.)
- So we need the point estimate and its standard error to construct a confidence interval.
- For the predicted value the point estimate is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- What about the standard error of $\hat{y}$?
- We could use the properties of variance, but $Var(\hat{y})$ involves $Cov(\hat{\beta}_i, \hat{\beta}_j)$ for every pair, and regression output does not report them.

Confidence Interval of $\hat{y}$

- As we know, we can find a confidence interval by

$$\hat{c} - t_{\frac{\alpha}{2}} \times se(\hat{c}) \leq c \leq \hat{c} + t_{\frac{\alpha}{2}} \times se(\hat{c})$$

  where c is the population parameter and $\hat{c}$ is the sample estimator. ($t_{\alpha/2}$ has $n - k - 1$ degrees of freedom and is close to 1.96 for large n.)
- So we need the point estimate and its standard error to construct a confidence interval.
- For the predicted value the point estimate is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- What about the standard error of $\hat{y}$?
- We could use the properties of variance, but $Var(\hat{y})$ involves $Cov(\hat{\beta}_i, \hat{\beta}_j)$ for every pair, and regression output does not report them.
- So we need a trick.

Confidence Interval of $\hat{y}$ for the Expected Value of y

- Take a point $(c_1, c_2, \ldots, c_k)$. In the population, the expected value of y at that point is

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

Confidence Interval of $\hat{y}$ for the Expected Value of y

- Take a point $(c_1, c_2, \ldots, c_k)$. In the population, the expected value of y at that point is

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

- Using the sample, the estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

Confidence Interval of $\hat{y}$ for the Expected Value of y

- Take a point $(c_1, c_2, \ldots, c_k)$. In the population, the expected value of y at that point is

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

- Using the sample, the estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

- To construct a confidence interval we need $se(\hat{\theta}_0)$. From the equation above, $\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$.

Confidence Interval of $\hat{y}$ for the Expected Value of y

- Take a point $(c_1, c_2, \ldots, c_k)$. In the population, the expected value of y at that point is

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

- Using the sample, the estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

- To construct a confidence interval we need $se(\hat{\theta}_0)$. From the equation above, $\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$.
- Plug this into our regression function:

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \cdots + \beta_k(x_k - c_k) + \mu$$

Confidence Interval of $\hat{y}$ for the Expected Value of y

- Take a point $(c_1, c_2, \ldots, c_k)$. In the population, the expected value of y at that point is

$$\theta_0 = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k$$

- Using the sample, the estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

- To construct a confidence interval we need $se(\hat{\theta}_0)$. From the equation above, $\beta_0 = \theta_0 - \beta_1 c_1 - \cdots - \beta_k c_k$.
- Plug this into our regression function:

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \cdots + \beta_k(x_k - c_k) + \mu$$

- So the intercept in this regression is θ_0, and its standard error is $se(\hat{\theta}_0)$.

Confidence Interval of $\hat{y}$ for the Expected Value of y

Steps for the confidence interval of $\hat{y}$ at a specific point:

1. For a specific point $(c_1, c_2, \ldots, c_k)$, find the predicted value $\hat{\theta}_0$.

Confidence Interval of $\hat{y}$ for the Expected Value of y

Steps for the confidence interval of $\hat{y}$ at a specific point:

1. For a specific point $(c_1, c_2, \ldots, c_k)$, find the predicted value $\hat{\theta}_0$.
2. Construct new independent variables $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.

Confidence Interval of $\hat{y}$ for the Expected Value of y

Steps for the confidence interval of $\hat{y}$ at a specific point:

1. For a specific point $(c_1, c_2, \ldots, c_k)$, find the predicted value $\hat{\theta}_0$.
2. Construct new independent variables $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
3. Run the regression of y on $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.

Confidence Interval of $\hat{y}$ for the Expected Value of y

Steps for the confidence interval of $\hat{y}$ at a specific point:

1. For a specific point $(c_1, c_2, \ldots, c_k)$, find the predicted value $\hat{\theta}_0$.
2. Construct new independent variables $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
3. Run the regression of y on $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
4. Find the standard error of the intercept, $se(\hat{\theta}_0)$.

Confidence Interval of $\hat{y}$ for the Expected Value of y

Steps for the confidence interval of $\hat{y}$ at a specific point:

1. For a specific point $(c_1, c_2, \ldots, c_k)$, find the predicted value $\hat{\theta}_0$.
2. Construct new independent variables $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
3. Run the regression of y on $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
4. Find the standard error of the intercept, $se(\hat{\theta}_0)$.
5. Construct the confidence interval:

$$\hat{\theta}_0 - t_{\frac{\alpha}{2}} \times se(\hat{\theta}_0) \leq \theta_0 \leq \hat{\theta}_0 + t_{\frac{\alpha}{2}} \times se(\hat{\theta}_0)$$

Confidence Interval of $\hat{y}$ for the Expected Value of y

Steps for the confidence interval of $\hat{y}$ at a specific point:

1. For a specific point $(c_1, c_2, \ldots, c_k)$, find the predicted value $\hat{\theta}_0$.
2. Construct new independent variables $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
3. Run the regression of y on $(x_1 - c_1), (x_2 - c_2), \ldots, (x_k - c_k)$.
4. Find the standard error of the intercept, $se(\hat{\theta}_0)$.
5. Construct the confidence interval:

$$\hat{\theta}_0 - t_{\frac{\alpha}{2}} \times se(\hat{\theta}_0) \leq \theta_0 \leq \hat{\theta}_0 + t_{\frac{\alpha}{2}} \times se(\hat{\theta}_0)$$

6. This is not a general formula. You have to repeat it for every point at which you want a confidence interval.

Confidence Interval of $\hat{y}$ - Example

- Using the college GPA data, the following equation is estimated.

$$\widehat{colGPA} = 1.493 + 0.00149\, sat - 0.01386\, hsperc - 0.06088\, hsize + 0.00546\, hsize^2$$

Confidence Interval of $\hat{y}$ - Example

- Using the college GPA data, the following equation is estimated.

$$\widehat{colGPA} = 1.493 + 0.00149\, sat - 0.01386\, hsperc - 0.06088\, hsize + 0.00546\, hsize^2$$

- We want the confidence interval for the predicted value at $sat = 1200$, $hsperc = 30$, $hsize = 5$.

Confidence Interval of $\hat{y}$ - Example

- Using the college GPA data, the following equation is estimated.

$$\widehat{colGPA} = 1.493 + 0.00149\, sat - 0.01386\, hsperc - 0.06088\, hsize + 0.00546\, hsize^2$$

- We want the confidence interval for the predicted value at $sat = 1200$, $hsperc = 30$, $hsize = 5$.
  1. Predicted value:

$$\widehat{colGPA} = 1.493 + 0.00149(1200) - 0.01386(30) - 0.06088(5) + 0.00546(5^2) \approx 2.70$$

Confidence Interval of $\hat{y}$ - Example

- Using the college GPA data, the following equation is estimated.

$$\widehat{colGPA} = 1.493 + 0.00149\, sat - 0.01386\, hsperc - 0.06088\, hsize + 0.00546\, hsize^2$$

- We want the confidence interval for the predicted value at $sat = 1200$, $hsperc = 30$, $hsize = 5$.
  1. Predicted value:

$$\widehat{colGPA} = 1.493 + 0.00149(1200) - 0.01386(30) - 0.06088(5) + 0.00546(5^2) \approx 2.70$$

  2. Construct $sat_0 = sat - 1200$, $hsperc_0 = hsperc - 30$, $hsize_0 = hsize - 5$ and $hsizesq_0 = hsize^2 - 25$. (For the squared term we subtract 5^2, not $(hsize - 5)^2$.)

Confidence Interval of $\hat{y}$ - Example

- Using the college GPA data, the following equation is estimated.

$$\widehat{colGPA} = 1.493 + 0.00149\, sat - 0.01386\, hsperc - 0.06088\, hsize + 0.00546\, hsize^2$$

- We want the confidence interval for the predicted value at $sat = 1200$, $hsperc = 30$, $hsize = 5$.
  1. Predicted value:

$$\widehat{colGPA} = 1.493 + 0.00149(1200) - 0.01386(30) - 0.06088(5) + 0.00546(5^2) \approx 2.70$$

  2. Construct $sat_0 = sat - 1200$, $hsperc_0 = hsperc - 30$, $hsize_0 = hsize - 5$ and $hsizesq_0 = hsize^2 - 25$. (For the squared term we subtract 5^2, not $(hsize - 5)^2$.)
  3. Regress *colGPA* on these four variables. The intercept is 2.70 and its standard error is 0.020.

Confidence Interval of $\hat{y}$ - Example

- Using the college GPA data, the following equation is estimated.

$$\widehat{colGPA} = 1.493 + 0.00149\, sat - 0.01386\, hsperc - 0.06088\, hsize + 0.00546\, hsize^2$$

- We want the confidence interval for the predicted value at $sat = 1200$, $hsperc = 30$, $hsize = 5$.
  1. Predicted value:

$$\widehat{colGPA} = 1.493 + 0.00149(1200) - 0.01386(30) - 0.06088(5) + 0.00546(5^2) \approx 2.70$$

  2. Construct $sat_0 = sat - 1200$, $hsperc_0 = hsperc - 30$, $hsize_0 = hsize - 5$ and $hsizesq_0 = hsize^2 - 25$. (For the squared term we subtract 5^2, not $(hsize - 5)^2$.)
  3. Regress *colGPA* on these four variables. The intercept is 2.70 and its standard error is 0.020.
  4. The 95% confidence interval is

$$2.70 - 1.96 \times 0.020 \leq E[colGPA] \leq 2.70 + 1.96 \times 0.020 \Rightarrow [2.66,\ 2.74]$$

Prediction Interval of y^0 for an Individual Outcome

- The interval we just built is for the **average** value of y for everyone with the characteristics $(c_1, \ldots, c_k)$. It assumes the average error is zero.

Prediction Interval of y^0 for an Individual Outcome

- The interval we just built is for the **average** value of y for everyone with the characteristics $(c_1, \ldots, c_k)$. It assumes the average error is zero.
- But often we want to predict the outcome of one particular individual, say a student who is not in our sample.

Prediction Interval of y^0 for an Individual Outcome

- The interval we just built is for the **average** value of y for everyone with the characteristics $(c_1, \ldots, c_k)$. It assumes the average error is zero.
- But often we want to predict the outcome of one particular individual, say a student who is not in our sample.
- The predicted value is still the best guess, but now we also face the unknown error of that individual.

Prediction Interval of y^0 for an Individual Outcome

- The interval we just built is for the **average** value of y for everyone with the characteristics $(c_1, \ldots, c_k)$. It assumes the average error is zero.
- But often we want to predict the outcome of one particular individual, say a student who is not in our sample.
- The predicted value is still the best guess, but now we also face the unknown error of that individual.
- So for a single new observation we must include an additional source of variation, the variation in the unobserved error.

Prediction Interval of y^0 for an Individual Outcome

- The interval we just built is for the **average** value of y for everyone with the characteristics $(c_1, \ldots, c_k)$. It assumes the average error is zero.
- But often we want to predict the outcome of one particular individual, say a student who is not in our sample.
- The predicted value is still the best guess, but now we also face the unknown error of that individual.
- So for a single new observation we must include an additional source of variation, the variation in the unobserved error.
- This gives a **prediction interval**, which is wider than the confidence interval for the average.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.
  - Let μ^0 be the unobserved error of that individual.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.
  - Let μ^0 be the unobserved error of that individual.
  - So $y^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + \mu^0$.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.
  - Let μ^0 be the unobserved error of that individual.
  - So $y^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + \mu^0$.
- The predicted value is $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.
  - Let μ^0 be the unobserved error of that individual.
  - So $y^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + \mu^0$.
- The predicted value is $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$.
- The prediction error is $\hat{e}^0 = y^0 - \hat{y}^0$.

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.
  - Let μ^0 be the unobserved error of that individual.
  - So $y^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + \mu^0$.
- The predicted value is $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$.
- The prediction error is $\hat{e}^0 = y^0 - \hat{y}^0$.
- What is the expected value of the prediction error?

$$E[\hat{e}^0] = E[y^0] - E[\hat{y}^0] = 0$$

Prediction Interval of y^0 for an Individual Outcome

- Consider the following:
  - Let y^0 be the unobserved value we want to predict.
  - Let $x_1^0, x_2^0, \ldots, x_k^0$ be the point at which we predict.
  - Let μ^0 be the unobserved error of that individual.
  - So $y^0 = \beta_0 + \beta_1 x_1^0 + \cdots + \beta_k x_k^0 + \mu^0$.
- The predicted value is $\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0$.
- The prediction error is $\hat{e}^0 = y^0 - \hat{y}^0$.
- What is the expected value of the prediction error?

$$E[\hat{e}^0] = E[y^0] - E[\hat{y}^0] = 0$$

- This is because $E[\hat{\beta}_j] = \beta_j$ and $E[\mu^0] = 0$.

Prediction Interval of y^0 for an Individual Outcome

- What is the variance of the prediction error?

$$\hat{e}^0 = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^0 + \cdots + (\beta_k - \hat{\beta}_k)x_k^0 + \mu^0$$

Prediction Interval of y^0 for an Individual Outcome

- What is the variance of the prediction error?

$$\hat{e}^0 = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^0 + \cdots + (\beta_k - \hat{\beta}_k)x_k^0 + \mu^0$$

- The error μ^0 is not part of our sample, so it is uncorrelated with the $\hat{\beta}_j$. Therefore

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(\mu^0) = Var(\hat{y}^0) + \sigma^2$$

Prediction Interval of y^0 for an Individual Outcome

- What is the variance of the prediction error?

$$\hat{e}^0 = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^0 + \cdots + (\beta_k - \hat{\beta}_k)x_k^0 + \mu^0$$

- The error μ^0 is not part of our sample, so it is uncorrelated with the $\hat{\beta}_j$. Therefore

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(\mu^0) = Var(\hat{y}^0) + \sigma^2$$

- The standard error of the prediction error is

$$se(\hat{e}^0) = \sqrt{[se(\hat{y}^0)]^2 + \hat{\sigma}^2}$$

Prediction Interval of y^0 for an Individual Outcome

- What is the variance of the prediction error?

$$\hat{e}^0 = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^0 + \cdots + (\beta_k - \hat{\beta}_k)x_k^0 + \mu^0$$

- The error μ^0 is not part of our sample, so it is uncorrelated with the $\hat{\beta}_j$. Therefore

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(\mu^0) = Var(\hat{y}^0) + \sigma^2$$

- The standard error of the prediction error is

$$se(\hat{e}^0) = \sqrt{[se(\hat{y}^0)]^2 + \hat{\sigma}^2}$$

- The prediction interval is

$$\hat{y}^0 - t_{\frac{\alpha}{2}} \times se(\hat{e}^0) \leq y^0 \leq \hat{y}^0 + t_{\frac{\alpha}{2}} \times se(\hat{e}^0)$$

Prediction Interval of y^0 for an Individual Outcome

- What is the variance of the prediction error?

$$\hat{e}^0 = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_1^0 + \cdots + (\beta_k - \hat{\beta}_k)x_k^0 + \mu^0$$

- The error μ^0 is not part of our sample, so it is uncorrelated with the $\hat{\beta}_j$. Therefore

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(\mu^0) = Var(\hat{y}^0) + \sigma^2$$

- The standard error of the prediction error is

$$se(\hat{e}^0) = \sqrt{[se(\hat{y}^0)]^2 + \hat{\sigma}^2}$$

- The prediction interval is

$$\hat{y}^0 - t_{\frac{\alpha}{2}} \times se(\hat{e}^0) \leq y^0 \leq \hat{y}^0 + t_{\frac{\alpha}{2}} \times se(\hat{e}^0)$$

- $se(\hat{y}^0)$ is found using the same steps as before, and $\hat{\sigma}$ is the standard error of the regression.

Prediction Interval - Example

- Back to the college GPA example. At $sat = 1200$, $hsperc = 30$, $hsize = 5$ we had $\hat{y}^0 \approx 2.70$ and $se(\hat{y}^0) = 0.020$.

Prediction Interval - Example

- Back to the college GPA example. At $sat = 1200$, $hsperc = 30$, $hsize = 5$ we had $\hat{y}^0 \approx 2.70$ and $se(\hat{y}^0) = 0.020$.
- The standard error of the regression is $\hat{\sigma} \approx 0.560$. So

$$se(\hat{e}^0) = \sqrt{0.020^2 + 0.560^2} \approx 0.560$$

Prediction Interval - Example

- Back to the college GPA example. At $sat = 1200$, $hsperc = 30$, $hsize = 5$ we had $\hat{y}^0 \approx 2.70$ and $se(\hat{y}^0) = 0.020$.
- The standard error of the regression is $\hat{\sigma} \approx 0.560$. So

$$se(\hat{e}^0) = \sqrt{0.020^2 + 0.560^2} \approx 0.560$$

- The 95% prediction interval for one student is

$$2.70 \pm 1.96 \times 0.560 \Rightarrow [1.60,\ 3.80]$$

Prediction Interval - Example

- Back to the college GPA example. At $sat = 1200$, $hsperc = 30$, $hsize = 5$ we had $\hat{y}^0 \approx 2.70$ and $se(\hat{y}^0) = 0.020$.
- The standard error of the regression is $\hat{\sigma} \approx 0.560$. So

$$se(\hat{e}^0) = \sqrt{0.020^2 + 0.560^2} \approx 0.560$$

- The 95% prediction interval for one student is

$$2.70 \pm 1.96 \times 0.560 \Rightarrow [1.60,\ 3.80]$$

- Compare with the confidence interval for the average student, $[2.66,\ 2.74]$. It is far narrower.

Prediction Interval - Example

- Back to the college GPA example. At $sat = 1200$, $hsperc = 30$, $hsize = 5$ we had $\hat{y}^0 \approx 2.70$ and $se(\hat{y}^0) = 0.020$.
- The standard error of the regression is $\hat{\sigma} \approx 0.560$. So

$$se(\hat{e}^0) = \sqrt{0.020^2 + 0.560^2} \approx 0.560$$

- The 95% prediction interval for one student is

$$2.70 \pm 1.96 \times 0.560 \Rightarrow [1.60,\ 3.80]$$

- Compare with the confidence interval for the average student, $[2.66,\ 2.74]$. It is far narrower.
- Notice that $se(\hat{e}^0)$ is almost entirely $\hat{\sigma}$. With a large sample, the estimation error becomes tiny, but the individual's own error never goes away.

Prediction Interval - Example

- Back to the college GPA example. At $sat = 1200$, $hsperc = 30$, $hsize = 5$ we had $\hat{y}^0 \approx 2.70$ and $se(\hat{y}^0) = 0.020$.
- The standard error of the regression is $\hat{\sigma} \approx 0.560$. So

$$se(\hat{e}^0) = \sqrt{0.020^2 + 0.560^2} \approx 0.560$$

- The 95% prediction interval for one student is

$$2.70 \pm 1.96 \times 0.560 \Rightarrow [1.60,\ 3.80]$$

- Compare with the confidence interval for the average student, $[2.66,\ 2.74]$. It is far narrower.
- Notice that $se(\hat{e}^0)$ is almost entirely $\hat{\sigma}$. With a large sample, the estimation error becomes tiny, but the individual's own error never goes away.
- Go over Example 6.6 in the book.

Residual Analysis

- Sometimes it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value. That is, we examine the residuals of individual observations. This is called **residual analysis**.

Residual Analysis

- Sometimes it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value. That is, we examine the residuals of individual observations. This is called **residual analysis**.
- If a set of points have large residuals, we can get insights on how to improve the model. For example, if houses above some price have larger residuals, maybe they are different from the rest (luxury houses), and we may need another model or a variable that describes them well.

Residual Analysis

- Sometimes it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value. That is, we examine the residuals of individual observations. This is called **residual analysis**.
- If a set of points have large residuals, we can get insights on how to improve the model. For example, if houses above some price have larger residuals, maybe they are different from the rest (luxury houses), and we may need another model or a variable that describes them well.
- Similarly, if the price in the housing data is the asking price, a large positive residual may mean the house will be difficult to sell, or the seller is demanding above the market price.

Residual Analysis

- Sometimes it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value. That is, we examine the residuals of individual observations. This is called **residual analysis**.
- If a set of points have large residuals, we can get insights on how to improve the model. For example, if houses above some price have larger residuals, maybe they are different from the rest (luxury houses), and we may need another model or a variable that describes them well.
- Similarly, if the price in the housing data is the asking price, a large positive residual may mean the house will be difficult to sell, or the seller is demanding above the market price.
- When the dependent variable is $\log(y)$, a residual of -0.30 means that the actual value is roughly 30 percent below the predicted value.

Residual Analysis

- Sometimes it is useful to examine individual observations to see whether the actual value of the dependent variable is above or below the predicted value. That is, we examine the residuals of individual observations. This is called **residual analysis**.
- If a set of points have large residuals, we can get insights on how to improve the model. For example, if houses above some price have larger residuals, maybe they are different from the rest (luxury houses), and we may need another model or a variable that describes them well.
- Similarly, if the price in the housing data is the asking price, a large positive residual may mean the house will be difficult to sell, or the seller is demanding above the market price.
- When the dependent variable is $\log(y)$, a residual of -0.30 means that the actual value is roughly 30 percent below the predicted value.
- Therefore, residual analysis can generate different insights depending on the context.