Lecture 17: Logarithmic Models and Testing Joint Hypotheses

ECON 480 - Econometrics - Fall 2018

Ryan Safner

November 28, 2018

Logarithmic Models

Linear-Log Model

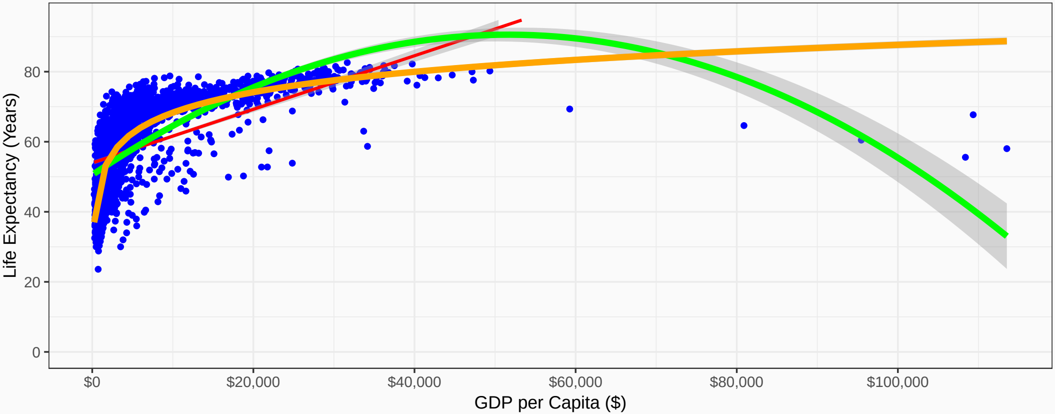
Log-Linear Model

Log-Log Model

Comparing Across Units

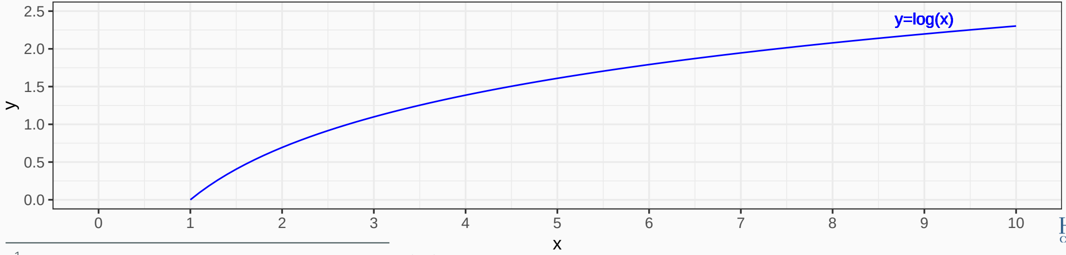
Joint-Hypothesis Testing

Nonlinearities? Example



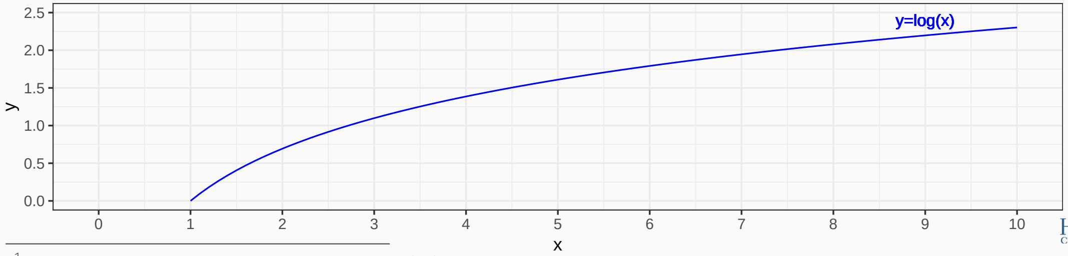
Logarithmic Models

- Another model specification for nonlinear data is the logarithmic model[1]



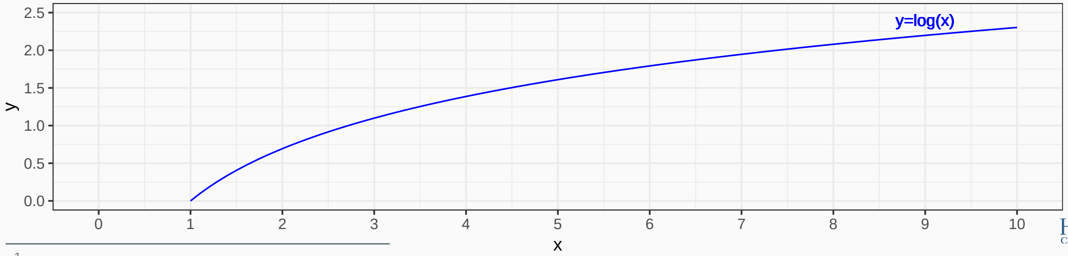
[1] Note, this should not be confused with a **logistic model**, which is a model for dependent dummy variables.

- Another model specification for nonlinear data is the logarithmic model[1]
  - We transform either X, Y, or *both* by taking the (natural) logarithm



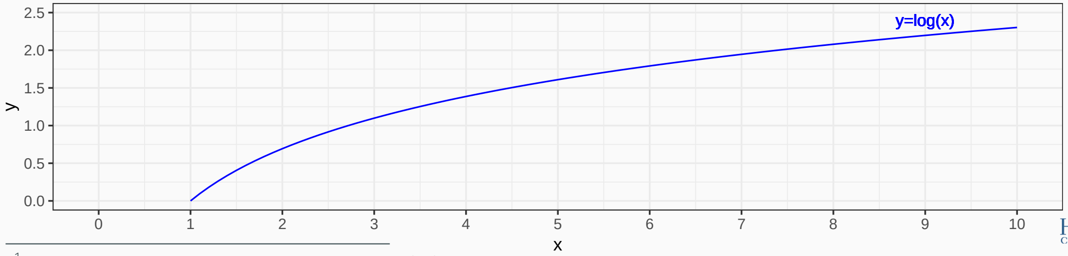
[1] Note, this should not be confused with a **logistic model**, which is a model for dependent dummy variables.

- Another model specification for nonlinear data is the logarithmic model[1]
    - We transform either X, Y, or *both* by taking the (natural) logarithm
- Logarithmic model has two additional advantages



[1] Note, this should not be confused with a **logistic model**, which is a model for dependent dummy variables.

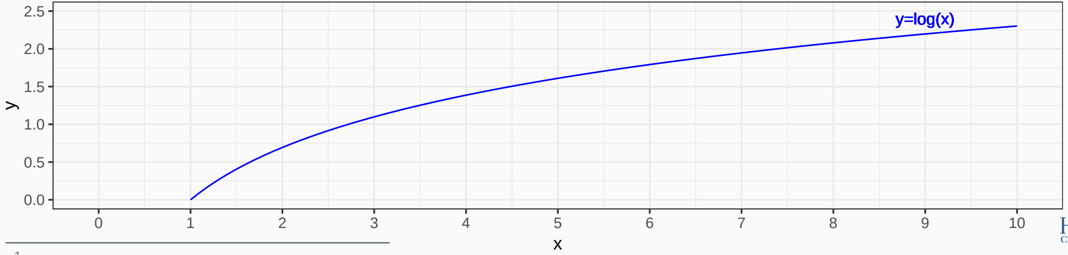
- Another model specification for nonlinear data is the logarithmic model[1]
  - We transform either X, Y, or *both* by taking the (natural) logarithm
- Logarithmic model has two additional advantages
  - We can easily interpret coefficients as **percentage changes** or **elasticities**



[1] Note, this should not be confused with a **logistic model**, which is a model for dependent dummy variables.

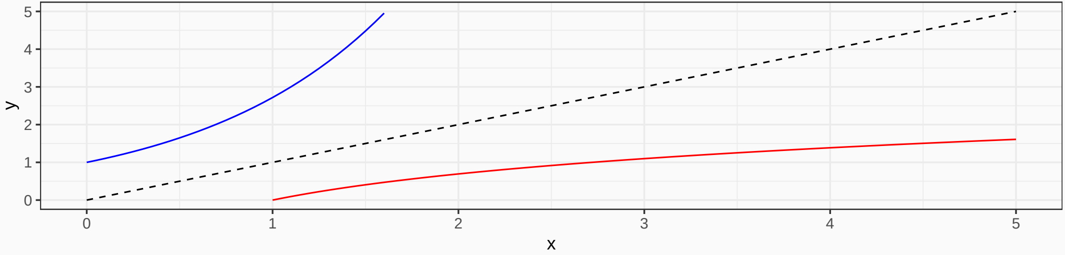
Logarithmic Models

- Another model specification for nonlinear data is the logarithmic model[1]
    - We transform either X, Y, or *both* by taking the (natural) logarithm
- Logarithmic model has two additional advantages
    - We can easily interpret coefficients as **percentage changes** or **elasticities**
    - Useful economic shape: diminishing returns (production functions, utility functions, etc)

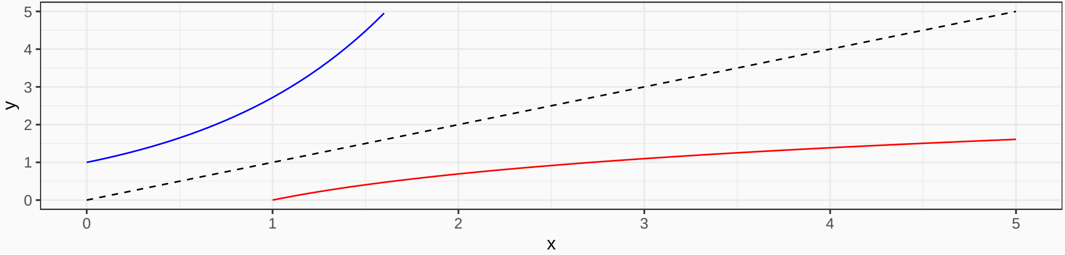


[1] Note, this should not be confused with a **logistic model**, which is a model for dependent dummy variables.

- The **exponential function,** e^x or $exp(x)$, where base $e = 2.71828\ldots$



- The **exponential function,** e^x or $exp(x)$, where base $e = 2.71828...$
- **Natural logarithm** is the inverse, $y = ln(x)$



- Exponents are defined as

$$b^n = \underbrace{b \times b \times \ldots \times b}_{n}$$

where base b is multiplied by itself n times

- Exponents are defined as

$$b^n = \underbrace{b \times b \times \ldots \times b}_{n}$$

  where base b is multiplied by itself n times
- e.g. $2^3 = \underbrace{2 \times 2 \times 2}_{n=3} = 8$

- Exponents are defined as

$$b^n = \underbrace{b \times b \times \ldots \times b}_{n}$$

where base b is multiplied by itself n times

- e.g. $2^3 = \underbrace{2 \times 2 \times 2}_{n=3} = 8$
- Logarithms the inverse, defined as the exponents in the expressions above

$$\text{If } b^n = y \text{, then } log_b(y) = n$$

n is the number you must raise b to in order to get y

- Exponents are defined as

$$b^n = \underbrace{b \times b \times \ldots \times b}_{n}$$

where base b is multiplied by itself n times

- e.g. $2^3 = \underbrace{2 \times 2 \times 2}_{n=3} = 8$
- Logarithms the inverse, defined as the exponents in the expressions above

$$\text{If } b^n = y \text{, then } log_b(y) = n$$

n is the number you must raise b to in order to get y

- e.g. $log_2(6) = 3$

- Exponents are defined as

$$b^n = \underbrace{b \times b \times \ldots \times b}_{n}$$

where base b is multiplied by itself n times

- e.g. $2^3 = \underbrace{2 \times 2 \times 2}_{n=3} = 8$
- Logarithms the inverse, defined as the exponents in the expressions above

$$\text{If } b^n = y \text{, then } log_b(y) = n$$

n is the number you must raise b to in order to get y

- e.g. $log_2(6) = 3$
- Logarithms can have any base, but common to use the natural logarithm (ln) with base e=2.71828...

$$\text{If } e^n = y \text{, then } ln(y) = n$$

HOOD COLLEGE

The Natural Logarithm: Properties

- Natural logs have a lot of useful properties:

The Natural Logarithm: Properties

- Natural logs have a lot of useful properties:
  - $ln(\frac{1}{x}) = -ln(x)$

- Natural logs have a lot of useful properties:
  - $ln(\frac{1}{x}) = -ln(x)$
  - $ln(ab) = ln(a) + ln(b)$

The Natural Logarithm: Properties

- Natural logs have a lot of useful properties:
  - $ln(\frac{1}{x}) = -ln(x)$
  - $ln(ab) = ln(a) + ln(b)$
  - $ln(\frac{x}{a}) = ln(x) - ln(a)$

- Natural logs have a lot of useful properties:
    - $ln(\frac{1}{x}) = -ln(x)$
    - $ln(ab) = ln(a) + ln(b)$
    - $ln(\frac{x}{a}) = ln(x) - ln(a)$
    - $ln(x^a) = a\,ln(x)$

- Natural logs have a lot of useful properties:
    - $ln(\frac{1}{x}) = -ln(x)$
    - $ln(ab) = ln(a) + ln(b)$
    - $ln(\frac{x}{a}) = ln(x) - ln(a)$
    - $ln(x^a) = a\,ln(x)$
    - $\frac{d\,ln\,x}{d\,x} = \frac{1}{x}$

- Most useful property: for small change in x, Δx:

$$\underbrace{ln(x+\Delta x) - ln(x)}_{\text{Difference in logs}} \approx \underbrace{\frac{\Delta x}{x}}_{\text{Relative change}}$$

- Most useful property: for small change in x, Δx:

$$\underbrace{ln(x + \Delta x) - ln(x)}_{\text{Difference in logs}} \approx \underbrace{\frac{\Delta x}{x}}_{\text{Relative change}}$$

Example

- Most useful property: for small change in x, Δx:

$$\underbrace{ln(x+\Delta x) - ln(x)}_{\text{Difference in logs}} \approx \underbrace{\frac{\Delta x}{x}}_{\text{Relative change}}$$

Example

- Let $x = 100$ and $\Delta x = 1$, relative change is:

$$\frac{\Delta x}{x} = \frac{(101 - 100)}{100} = 0.01 \text{ or } 1\%$$

- Most useful property: for small change in x, Δx:

$$\underbrace{ln(x+\Delta x) - ln(x)}_{\text{Difference in logs}} \approx \underbrace{\frac{\Delta x}{x}}_{\text{Relative change}}$$

Example

- Let $x = 100$ and $\Delta x = 1$, relative change is:

$$\frac{\Delta x}{x} = \frac{(101 - 100)}{100} = 0.01 \text{ or } 1\%$$

- The logged difference:

$$ln(101) - ln(100) = 0.00995 \approx 1\%$$

- Most useful property: for small change in x, Δx:

$$\underbrace{ln(x + \Delta x) - ln(x)}_{\text{Difference in logs}} \approx \underbrace{\frac{\Delta x}{x}}_{\text{Relative change}}$$

Example

- Let $x = 100$ and $\Delta x = 1$, relative change is:

$$\frac{\Delta x}{x} = \frac{(101 - 100)}{100} = 0.01 \text{ or } 1\%$$

- The logged difference:

$$ln(101) - ln(100) = 0.00995 \approx 1\%$$

- This allows us to very easily interpret coefficients as *percent changes* or **elasticities**

- **Difference (Δ):** the difference between two values of x, x_1 and x_2

$$\Delta x = x_2 - x_1$$

- **Difference (Δ):** the difference between two values of x, x_1 and x_2

$$\Delta x = x_2 - x_1$$

- **Relative Difference**: the difference expressed in terms of the original value

$$\frac{\Delta x}{x} = \frac{x_2 - x_1}{x_1}$$

this becomes a proportion ($\pm$ between 0 and 1)

- **Difference (Δ):** the difference between two values of x, x_1 and x_2

$$\Delta x = x_2 - x_1$$

- **Relative Difference**: the difference expressed in terms of the original value

$$\frac{\Delta x}{x} = \frac{x_2 - x_1}{x_1}$$

  this becomes a proportion ($\pm$ between 0 and 1)

- **Percentage Change** or **Growth Rate**: relative difference expressed as a *percentage* ($\pm$ between 0 and 100%)

$$\begin{aligned} \%\Delta &= \frac{\Delta x}{x_1} \times 100\% \\ &= \frac{x_2 - x_1}{x_1} \times 100\% \end{aligned}$$

Example

A country's GDP is \$100 in 2017, and \$120 in 2018. Calculate the country's GDP growth rate for 2018.

Example

A country's GDP is \$100 in 2017, and \$120 in 2018. Calculate the country's GDP growth rate for 2018.

$$\text{GDP Growth Rate}_{2018} = \frac{GDP_{2018} - GDP_{2017}}{GDP_{2017}} \times 100\%$$

Example
A country's GDP is \$100 in 2017, and \$120 in 2018. Calculate the country's GDP growth rate for 2018.

$$\begin{aligned}\text{GDP Growth Rate}_{2018} &= \frac{GDP_{2018} - GDP_{2017}}{GDP_{2017}} \times 100\% \\ &= \frac{105 - 100}{100} \times 100\%\end{aligned}$$

Example

A country's GDP is \$100 in 2017, and \$120 in 2018. Calculate the country's GDP growth rate for 2018.

$$\begin{aligned}
\text{GDP Growth Rate}_{2018} &= \frac{GDP_{2018} - GDP_{2017}}{GDP_{2017}} \times 100\% \\
&= \frac{105 - 100}{100} \times 100\% \\
&= \frac{5}{100} \times 100\%
\end{aligned}$$

Relative Change and Percentage Change: Example

Example

A country's GDP is \$100 in 2017, and \$120 in 2018. Calculate the country's GDP growth rate for 2018.

$$
\begin{aligned}
\text{GDP Growth Rate}_{2018} &= \frac{GDP_{2018} - GDP_{2017}}{GDP_{2017}} \times 100\% \\
&= \frac{105 - 100}{100} \times 100\% \\
&= \frac{5}{100} \times 100\% \\
&= 0.05 \times 100\%
\end{aligned}
$$

Example

A country's GDP is \$100 in 2017, and \$120 in 2018. Calculate the country's GDP growth rate for 2018.

$$
\begin{aligned}
\text{GDP Growth Rate}_{2018} &= \frac{GDP_{2018} - GDP_{2017}}{GDP_{2017}} \times 100\% \\
&= \frac{105 - 100}{100} \times 100\% \\
&= \frac{5}{100} \times 100\% \\
&= 0.05 \times 100\% \\
&= 5\%
\end{aligned}
$$

HOOD COLLEGE

- An elasticity between two variables, $E_{y,x}$ describes the *responsiveness* of one variable to a change in another.

Elasticity

- An elasticity between two variables, $E_{y,x}$ describes the *responsiveness* of one variable to a change in another.
- Measured in percentages: a 1% change in x will cause a E% change in y

Elasticity

- An elasticity between two variables, $E_{y,x}$ describes the *responsiveness* of one variable to a change in another.
- Measured in percentages: a 1% change in x will cause a E% change in y

$$E_{y,x} = \frac{\%\Delta y}{\%\Delta x}$$

- An elasticity between two variables, $E_{y,x}$ describes the *responsiveness* of one variable to a change in another.
- Measured in percentages: a 1% change in x will cause a E% change in y

$$\begin{aligned} E_{y,x} &= \frac{\%\Delta y}{\%\Delta x} \\ &= \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}} \end{aligned}$$

Elasticity

- An elasticity between two variables, $E_{y,x}$ describes the *responsiveness* of one variable to a change in another.
- Measured in percentages: a 1% change in x will cause a E% change in y

$$E_{y,x} = \frac{\%\Delta y}{\%\Delta x} = \frac{\frac{\Delta y}{y}}{\frac{\Delta x}{x}}$$

- Numerator is relative change in Y, Denominator is relative change in X

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

Math FYI: Cobb Douglas Functions and Logs

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

- Taking logs, relationship becomes linear:

$$ln(Y) = ln(A) + \alpha ln(L) + \beta ln(K)$$

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

- Taking logs, relationship becomes linear:

$$ln(Y) = ln(A) + \alpha ln(L) + \beta ln(K)$$

- With data on (Y, L, K) and linear regression, can estimate α and β

Math FYI: Cobb Douglas Functions and Logs

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

- Taking logs, relationship becomes linear:

$$ln(Y) = ln(A) + \alpha ln(L) + \beta ln(K)$$

- With data on (Y, L, K) and linear regression, can estimate α and β
  - α: elasticity of Y with respect to L

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

- Taking logs, relationship becomes linear:

$$ln(Y) = ln(A) + \alpha ln(L) + \beta ln(K)$$

- With data on (Y, L, K) and linear regression, can estimate α and β
  - α: elasticity of Y with respect to L
    - A 1% change in L will lead to an α% change in Y

Math FYI: Cobb Douglas Functions and Logs

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

- Taking logs, relationship becomes linear:

$$ln(Y) = ln(A) + \alpha ln(L) + \beta ln(K)$$

- With data on (Y, L, K) and linear regression, can estimate α and β
    - α: elasticity of Y with respect to L
        - A 1% change in L will lead to an α% change in Y
    - β: elasticity of Y with respect to K

- One of the (many) reasons why economists love Cobb-Douglas functions:

$$Y = AL^{\alpha}K^{\beta}$$

- Taking logs, relationship becomes linear:

$$ln(Y) = ln(A) + \alpha ln(L) + \beta ln(K)$$

- With data on (Y, L, K) and linear regression, can estimate α and β
  - α: elasticity of Y with respect to L
    - A 1% change in L will lead to an α% change in Y
  - β: elasticity of Y with respect to K
    - A 1% change in K will lead to a β% change in Y

Example

$$\widehat{Wages_{it}} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

Example

$$\widehat{Wages_{it}} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Does this make sense?

Example

$$\widehat{Wages_{it}} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Does this make sense?
- Wages increase by $\hat{\beta}_1$ for every 1 unit increase in Inflation

Example

$$\widehat{Wages_{it}} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Does this make sense?
- Wages increase by $\hat{\beta}_1$ for every 1 unit increase in Inflation
- Suppose $\hat{\beta}_1 = 1.25$: for every 1 unit increase in Inflation, *everyone's* (CEOs, janitors, etc) wages increase by $1.25.

Example

$$\widehat{Wages}_{it} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Does this make sense?
- Wages increase by $\hat{\beta}_1$ for every 1 unit increase in Inflation
- Suppose $\hat{\beta}_1 = 1.25$: for every 1 unit increase in Inflation, *everyone's* (CEOs, janitors, etc) wages increase by $1.25.
- What we really want is to estimate the *percentage* increase in people's wages

Example

$$\widehat{ln(Wages_{it})} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

Example

$$ln(\widehat{Wages_{it}}) = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Use ln(wages) for us to see the *percentage* change in wages from inflation

Example

$$\textcolor{magenta}{ln}(\widehat{Wages_{it}}) = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Use ln(wages) for us to see the *percentage* change in wages from inflation
- If $\hat{\beta}_1 = 1.25$, wages increase by 1.25% for every 1 unit increase in inflation

Example

$$\widehat{ln(Wages_{it})} = \beta_0 + \beta_1 Inflation_t + \epsilon_t$$

- Use ln(wages) for us to see the *percentage* change in wages from inflation
- If $\hat{\beta}_1 = 1.25$, wages increase by 1.25% for every 1 unit increase in inflation
- Different *levels* of wages between CEO & janitor, but increase at same *rate*

- The `log()` function can easily take the log

- The `log()` function can easily take the log

```
gapminder<-gapminder %>%
  mutate(l.gdp=log(gdpPercap))
```

- The `log()` function can easily take the log

```
gapminder<-gapminder %>%
  mutate(l.gdp=log(gdpPercap))
```

- Note, `log()` by default is the **natural logarithm** $ln()$, i.e. base **e**
  - Can change base with e.g. `log(x,base=5)`

- The `log()` function can easily take the log

```
gapminder<-gapminder %>%
  mutate(l.gdp=log(gdpPercap))
```

- Note, `log()` by default is the **natural logarithm** $ln()$, i.e. base **e**
    - Can change base with e.g. `log(x,base=5)`
    - Some common built-in logs: `log10`, `log2`

- Three types of log regression models, depending on which variables we log

- Three types of log regression models, depending on which variables we log

    1. Linear-log model:

$$Y = \beta_0 + \beta_1 ln(X)$$

- Three types of log regression models, depending on which variables we log

1. Linear-log model:

$$Y = \beta_0 + \beta_1 ln(X)$$

2. Log-linear model:

$$ln(Y) = \beta_0 + \beta_1 X$$

3.Log-log model:

$$ln(Y) = \beta_0 + \beta_1 ln(X)$$

Linear-Log Model

- Linear-log model has an independent variable (X) that is logged

$$Y = \beta_0 + \beta_1 ln(X)$$

- Linear-log model has an independent variable (X) that is logged

$$Y = \beta_0 + \beta_1 ln(X)$$

$$\beta_1 = \frac{\Delta Y}{\left(\frac{\Delta X}{X}\right)}$$

- Linear-log model has an independent variable (X) that is logged

$$Y = \beta_0 + \beta_1 ln(X)$$

$$\beta_1 = \frac{\Delta Y}{\left(\frac{\Delta X}{X}\right)}$$

- $\hat{\beta}_1$: **a 1% change in $X \rightarrow \frac{\beta_1}{100}$ unit change in Y**

```r
lin_log_reg<-lm(lifeExp~l.gdp, data = gapminder)
summary(lin_log_reg)
```

```
##
## Call:
## lm(formula = lifeExp ~ l.gdp, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.778  -4.204   1.212   4.658  19.285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.1009     1.2277  -7.413 1.93e-13 ***
## l.gdp         8.4051     0.1488  56.500  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.62 on 1702 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.652
## F-statistic:  3192 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Life Expectancy}}_i = -9.10 + 9.41 ln(GDP)_i$$

Linear-Log Model In R

```
lin_log_reg<-lm(lifeExp~l.gdp, data = gapminder)
summary(lin_log_reg)
```

```
##
## Call:
## lm(formula = lifeExp ~ l.gdp, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.778  -4.204   1.212   4.658  19.285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.1009     1.2277  -7.413 1.93e-13 ***
## l.gdp         8.4051     0.1488  56.500  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.62 on 1702 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.652
## F-statistic:  3192 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Life Expectancy}}_i = -9.10 + 9.41 ln(GDP)_i$$

- $\hat{\beta}_1$: a **1% change in GDP** $\rightarrow$ a $\frac{9.41}{100} =$ **0.0941 year increase** in Life Expectancy
- A **25% fall in GDP** $\rightarrow$ a $(25 \times 0.0941) =$ **2.353 year decrease** in Life Expectancy

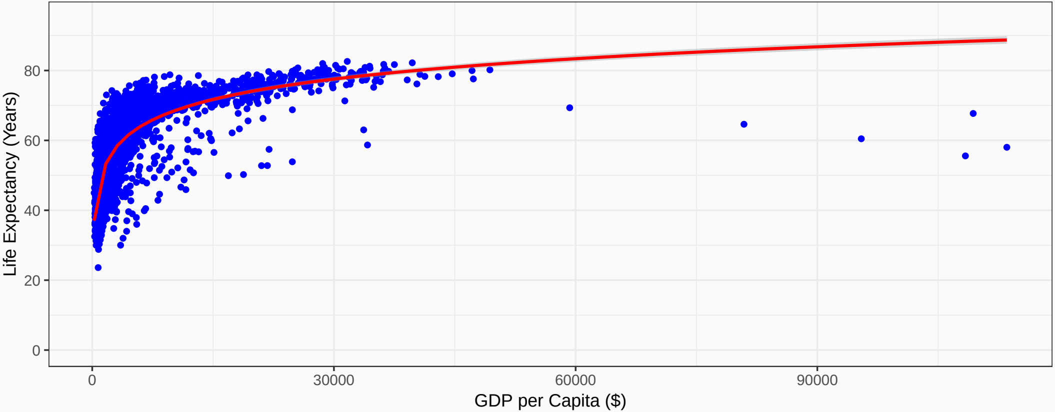
```r
lin_log_reg<-lm(lifeExp~l.gdp, data = gapminder)
summary(lin_log_reg)
```

```
##
## Call:
## lm(formula = lifeExp ~ l.gdp, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.778  -4.204   1.212   4.658  19.285
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.1009     1.2277  -7.413 1.93e-13 ***
## l.gdp         8.4051     0.1488  56.500  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.62 on 1702 degrees of freedom
## Multiple R-squared:  0.6522, Adjusted R-squared:  0.652
## F-statistic:  3192 on 1 and 1702 DF,  p-value: < 2.2e-16
```

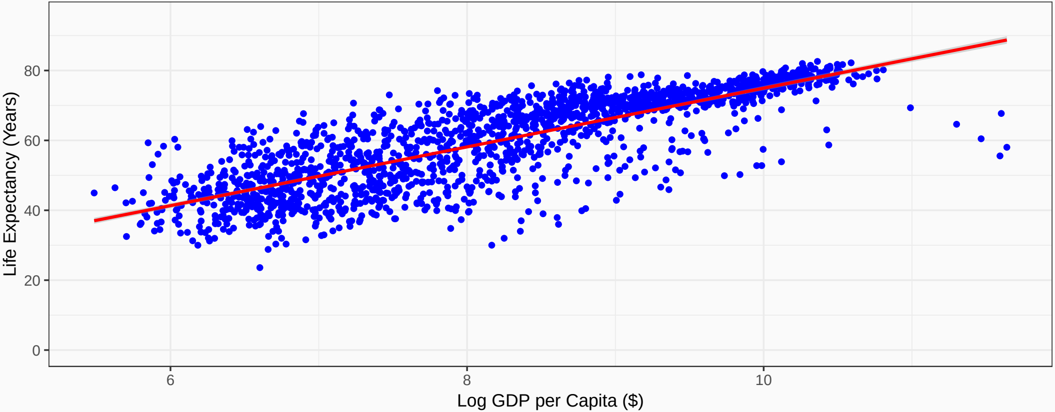
$$\widehat{\text{Life Expectancy}_i} = -9.10 + 9.41 ln(GDP)_i$$

- $\hat{\beta}_1$: a **1% change in GDP** $\rightarrow$ a $\frac{9.41}{100}$ = **0.0941 year increase** in Life Expectancy
- A **25% fall in GDP** $\rightarrow$ a $(25 \times 0.0941) =$ **2.353 year decrease** in Life Expectancy
- A **100% rise in GDP** $\rightarrow$ a $(100 \times 0.0941) =$ **9.041 year increase** in Life Expectancy

Linear-Log Model Graph



Linear-Log Model Graph II



Log-Linear Model

- Log-linear model has the dependent variable (Y) logged

$$ln(Y) = \beta_0 + \beta_1 X$$

- Log-linear model has the dependent variable (Y) logged

$$ln(Y) = \beta_0 + \beta_1 X$$

$$\beta_1 = \frac{\left(\frac{\Delta Y}{Y}\right)}{\Delta X}$$

- **Log-linear model** has the dependent variable (Y) logged

$$ln(Y) = \beta_0 + \beta_1 X$$
$$\beta_1 = \frac{\left(\frac{\Delta Y}{Y}\right)}{\Delta X}$$

- $\hat{\beta}_1$: a **1 unit change in** $X \rightarrow \beta_1 \times 100\%$ **change in** Y

- We will again have very large/small coefficients if we deal with GDP directly, again let's transform `gdpPercap` into $1,000s, call it `gdp.t`

- We will again have very large/small coefficients if we deal with GDP directly, again let's transform `gdpPercap` into $1,000s, call it `gdp.t`

```
gapminder <- gapminder %>%
  mutate(gdp.t=gdpPercap/1000)
```

```r
log_lin_reg<-lm(l.life~gdp.t, data = gapminder)
summary(log_lin_reg)
```

```
##
## Call:
## lm(formula = l.life ~ gdp.t, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37201 -0.12789  0.04738  0.14988  0.30925
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.9666387  0.0058346  679.85   <2e-16 ***
## gdp.t       0.0129170  0.0004777   27.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1943 on 1702 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.3001
## F-statistic: 731.1 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$ln\left(\widehat{\text{Life Expectancy}}\right)_i = 3.967 + 0.013GDP_i$$

Log-Linear Model In R

```r
log_lin_reg<-lm(l.life~gdp.t, data = gapminder)
summary(log_lin_reg)
```

```
##
## Call:
## lm(formula = l.life ~ gdp.t, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37201 -0.12789  0.04738  0.14988  0.30925
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.9666387  0.0058346  679.85   <2e-16 ***
## gdp.t       0.0129170  0.0004777   27.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1943 on 1702 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.3001
## F-statistic: 731.1 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$\widehat{ln(\text{Life Expectancy})_i} = 3.967 + 0.013GDP_i$$

- $\hat{\beta}_1$: a **\$1 (thousand) change in (thousands of) GDP** $\rightarrow$ a $0.013 \times 100\%$ = **1.3% increase** in Life Expectancy
- A **\$25K fall in GDP** $\rightarrow$ a $(25 \times 1.3\%)$ = **32.5% decrease** in Life Expectancy

Log-Linear Model In R

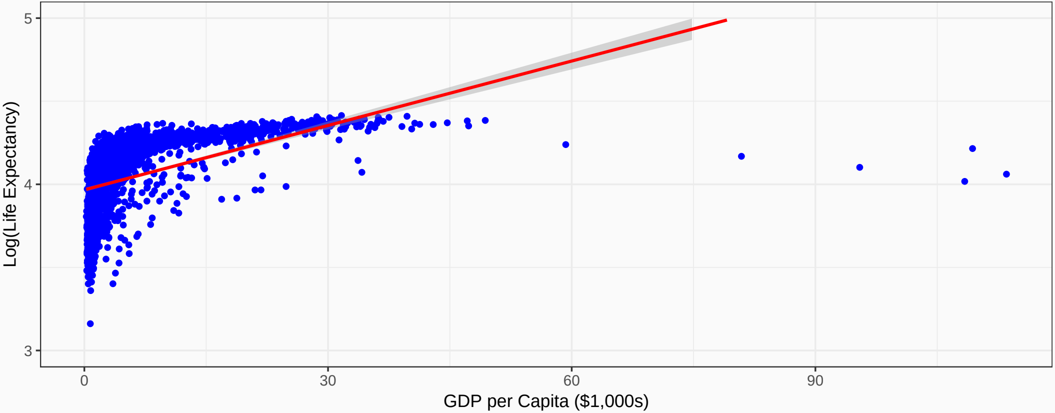
```r
log_lin_reg<-lm(l.life~gdp.t, data = gapminder)
summary(log_lin_reg)
```

```
##
## Call:
## lm(formula = l.life ~ gdp.t, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37201 -0.12789  0.04738  0.14988  0.30925
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.9666387  0.0058346  679.85   <2e-16 ***
## gdp.t       0.0129170  0.0004777   27.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1943 on 1702 degrees of freedom
## Multiple R-squared:  0.3005, Adjusted R-squared:  0.3001
## F-statistic: 731.1 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$ln\widehat{(\text{Life Expectancy})}_i = 3.967 + 0.013GDP_i$$

- $\hat{\beta}_1$: a **\$1 (thousand) change in (thousands of) GDP** $\rightarrow$ a $0.013 \times 100\%$ = **1.3% increase** in Life Expectancy
- A **\$25K fall in GDP** $\rightarrow$ a $(25 \times 1.3\%)$ = **32.5% decrease** in Life Expectancy
- A **\$100K rise in GDP** $\rightarrow$ a $(100 \times 1.3\%)$ = **130% increase** in Life Expectancy

Log-Linear Model Graph



Log-Log Model

Log-Log Model

- Log-log model has both variables (X and Y) logged

$$ln(Y) = \beta_0 + \beta_1 ln(X)$$

- Log-log model has both variables (X and Y) logged

$$ln(Y) = \beta_0 + \beta_1 ln(X)$$

$$\beta_1 = \frac{\left(\frac{\Delta Y}{Y}\right)}{\left(\frac{\Delta X}{X}\right)}$$

- Log-log model has both variables (X and Y) logged

$$ln(Y) = \beta_0 + \beta_1 ln(X)$$

$$\beta_1 = \frac{\left(\frac{\Delta Y}{Y}\right)}{\left(\frac{\Delta X}{X}\right)}$$

- $\hat{\beta}_1$: a **1% change in** $X \rightarrow \beta_1\%$ **change in** Y
- $\hat{\beta}_1$ is the **elasticity** of Y with respect to X

Log-Log Model In R

```r
log_log_reg<-lm(l.life~l.gdp, data = gapminder)
summary(log_log_reg)
```

```
##
## Call:
## lm(formula = l.life ~ l.gdp, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67059 -0.06453  0.01978  0.09086  0.36156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.864177   0.023283  123.02   <2e-16 ***
## l.gdp       0.146549   0.002821   51.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1445 on 1702 degrees of freedom
## Multiple R-squared:  0.6132, Adjusted R-squared:  0.613
## F-statistic:  2698 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$\widehat{ln(\text{Life Expectancy})_i} = 3.967 + 0.013 GDP_i$$

```r
log_log_reg<-lm(l.life~l.gdp, data = gapminder)
summary(log_log_reg)
```

```
##
## Call:
## lm(formula = l.life ~ l.gdp, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67059 -0.06453  0.01978  0.09086  0.36156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.864177   0.023283  123.02   <2e-16 ***
## l.gdp       0.146549   0.002821   51.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1445 on 1702 degrees of freedom
## Multiple R-squared:  0.6132, Adjusted R-squared:  0.613
## F-statistic:  2698 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$\widehat{ln(\text{Life Expectancy})_i} = 3.967 + 0.013GDP_i$$

- $\hat{\beta}_1$: a **\$1% change in GDP** $\rightarrow$ a **0.147% increase** in Life Expectancy
- A **25% fall in GDP** $\rightarrow$ a $(25 \times 0.147\%)$ = **3.675% decrease** in Life Expectancy

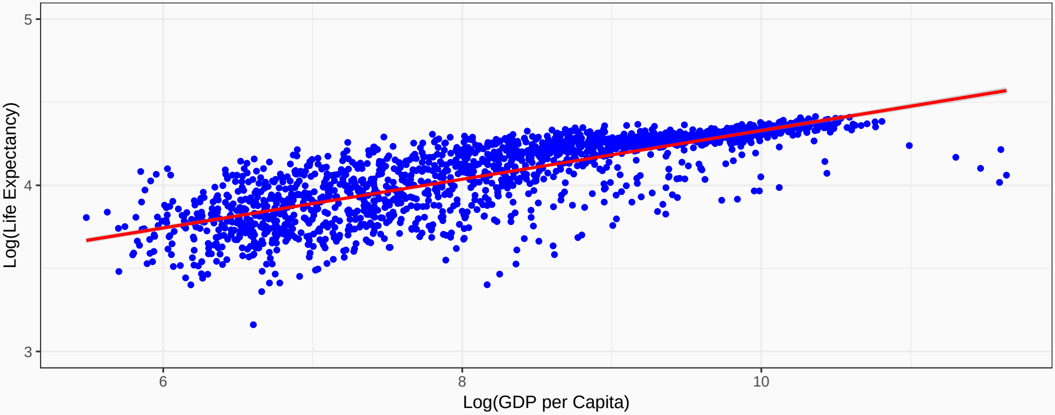
```r
log_log_reg<-lm(l.life~l.gdp, data = gapminder)
summary(log_log_reg)
```

```
##
## Call:
## lm(formula = l.life ~ l.gdp, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67059 -0.06453  0.01978  0.09086  0.36156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.864177   0.023283  123.02   <2e-16 ***
## l.gdp       0.146549   0.002821   51.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1445 on 1702 degrees of freedom
## Multiple R-squared:  0.6132, Adjusted R-squared:  0.613
## F-statistic:  2698 on 1 and 1702 DF,  p-value: < 2.2e-16
```

$$ln\widehat{(\text{Life Expectancy})_i} = 3.967 + 0.013GDP_i$$

- $\hat{\beta}_1$: a **\$1% change in GDP** $\rightarrow$ a **0.147% increase** in Life Expectancy
- A **25% fall in GDP** $\rightarrow$ a $(25 \times 0.147\%)$ = **3.675% decrease** in Life Expectancy
- A **100% rise in GDP** $\rightarrow$ a $(100 \times 0.147\%)$ = **14.7% increase** in Life Expectancy

Log-Log Model Graph



Model	Equation	Interpretation
Linear-Log	$Y = \beta_0 + \beta_1 ln(X)$	1% change in $X \rightarrow \frac{\hat{\beta}_1}{100}$ **unit** change in Y
Log-Linear	$ln(Y) = \beta_0 + \beta_1 X$	1 **unit** change in $X \rightarrow \hat{\beta}_1 \times 100$% change in Y
Log-Log	$ln(Y) = \beta_0 + \beta_1 ln(X)$	1% change in $X \rightarrow \hat{\beta}_1$% change in Y

- Hint: the variable that gets logged changes in percent terms, the variable not logged changes in **unit** terms

	Dependent variable:		
	lifeExp	l.life	
	Linear-Log	Log-Linear	Log-Log
	(1)	(2)	(3)
l.gdp	8.405***		0.147***
	(0.149)		(0.003)
gdp.t		0.013***	
		(0.0005)	
Constant	−9.101***	3.967***	2.864***
	(1.228)	(0.006)	(0.023)
Observations	1,704	1,704	1,704
R^2	0.652	0.300	0.613
Adjusted R^2	0.652	0.300	0.613
Residual Std. Error (df = 1702)	7.620	0.194	0.145
F Statistic (df = 1; 1702)	3,192.273***	731.139***	2,698.233***
Note:			*$p<0.1$; **$p<0.05$; ***$p<0.01$

- Models are very different units, how to choose?

Comparing Models I

	Dependent variable:		
	lifeExp	l.life	
	Linear-Log	Log-Linear	Log-Log
	(1)	(2)	(3)
l.gdp	8.405***		0.147***
	(0.149)		(0.003)
gdp.t		0.013***	
		(0.0005)	
Constant	−9.101***	3.967***	2.864***
	(1.228)	(0.006)	(0.023)
Observations	1,704	1,704	1,704
R^2	0.652	0.300	0.613
Adjusted R^2	0.652	0.300	0.613
Residual Std. Error (df = 1702)	7.620	0.194	0.145
F Statistic (df = 1; 1702)	3,192.273***	731.139***	2,698.233***
Note:			*p<0.1; **p<0.05; ***p<0.01

- Models are very different units, how to choose?
- Compare R^2's

	Dependent variable:		
	lifeExp	l.life	
	Linear-Log	Log-Linear	Log-Log
	(1)	(2)	(3)
l.gdp	8.405***		0.147***
	(0.149)		(0.003)
gdp.t		0.013***	
		(0.0005)	
Constant	−9.101***	3.967***	2.864***
	(1.228)	(0.006)	(0.023)
Observations	1,704	1,704	1,704
R^2	0.652	0.300	0.613
Adjusted R^2	0.652	0.300	0.613
Residual Std. Error (df = 1702)	7.620	0.194	0.145
F Statistic (df = 1; 1702)	3,192.273***	731.139***	2,698.233***
Note:			*p<0.1; **p<0.05; ***p<0.01

- Models are very different units, how to choose?
- Compare R^2's
- Compare graphs

Comparing Models I

	Dependent variable:		
	lifeExp	l.life	
	Linear-Log	Log-Linear	Log-Log
	(1)	(2)	(3)
l.gdp	8.405***		0.147***
	(0.149)		(0.003)
gdp.t		0.013***	
		(0.0005)	
Constant	−9.101***	3.967***	2.864***
	(1.228)	(0.006)	(0.023)
Observations	1,704	1,704	1,704
R^2	0.652	0.300	0.613
Adjusted R^2	0.652	0.300	0.613
Residual Std. Error (df = 1702)	7.620	0.194	0.145
F Statistic (df = 1; 1702)	3,192.273***	731.139***	2,698.233***
Note:			*p<0.1; **p<0.05; ***p<0.01

- Models are very different units, how to choose?
- Compare R^2's
- Compare graphs
- Compare intution

- In practice, the following types of variables are logged (ln):

When to Log?

- In practice, the following types of variables are logged (ln):
    - Variables that must always be positive (prices, sales, market values)

- In practice, the following types of variables are logged (ln):
  - Variables that must always be positive (prices, sales, market values)
  - Very large numbers (population, GDP)

- In practice, the following types of variables are logged (ln):
  - Variables that must always be positive (prices, sales, market values)
  - Very large numbers (population, GDP)
  - Variables expressed as percentages or percentage changes (inflation, population growth, GDP growth, labor force participation rate, unemployment rate)

- In practice, the following types of variables are logged (ln):
  - Variables that must always be positive (prices, sales, market values)
  - Very large numbers (population, GDP)
  - Variables expressed as percentages or percentage changes (inflation, population growth, GDP growth, labor force participation rate, unemployment rate)
  - Variables that have nonlinear scatterplots -Never use logs for:

When to Log?

- In practice, the following types of variables are logged (ln):
  - Variables that must always be positive (prices, sales, market values)
  - Very large numbers (population, GDP)
  - Variables expressed as percentages or percentage changes (inflation, population growth, GDP growth, labor force participation rate, unemployment rate)
  - Variables that have nonlinear scatterplots -Never use logs for:
  - Variables that are less than one, decimals, 0, or negative (temperature)

When to Log?

- In practice, the following types of variables are logged (ln):
  - Variables that must always be positive (prices, sales, market values)
  - Very large numbers (population, GDP)
  - Variables expressed as percentages or percentage changes (inflation, population growth, GDP growth, labor force participation rate, unemployment rate)
  - Variables that have nonlinear scatterplots -Never use logs for:
  - Variables that are less than one, decimals, 0, or negative (temperature)
  - Categorical variables

- In practice, the following types of variables are logged (ln):
    - Variables that must always be positive (prices, sales, market values)
    - Very large numbers (population, GDP)
    - Variables expressed as percentages or percentage changes (inflation, population growth, GDP growth, labor force participation rate, unemployment rate)
    - Variables that have nonlinear scatterplots -Never use logs for:
    - Variables that are less than one, decimals, 0, or negative (temperature)
    - Categorical variables
    - Time variables (year, week, day)

Comparing Across Units

Comparing Coefficients of Different Units

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- We often want to compare coefficients to see which variable X_1 or X_2 has a bigger effect on Y

Comparing Coefficients of Different Units

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- We often want to compare coefficients to see which variable X_1 or X_2 has a bigger effect on Y
- What if X_1 and X_2 are different scales?

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- We often want to compare coefficients to see which variable X_1 or X_2 has a bigger effect on Y
- What if X_1 and X_2 are different scales?

Example

$$\widehat{\text{Salary}_i} = \beta_0 + \beta_1 \text{Batting average}_i + \beta_2 \text{Home runs}_i$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- We often want to compare coefficients to see which variable X_1 or X_2 has a bigger effect on Y
- What if X_1 and X_2 are different scales?

Example

$$\widehat{\text{Salary}_i} = \beta_0 + \beta_1 \text{Batting average}_i + \beta_2 \text{Home runs}_i$$

$$\widehat{\text{Salary}_i} = -2{,}869{,}439.40 + 12{,}417{,}629.72 \text{Batting average}_i + 129{,}627.36 \text{Home runs}_i$$

- An easy way is to standardize the variables (i.e. take the Z-score)

$$X^{std} = \frac{X - \overline{X}}{sd(X)}$$

Example

Variable	Mean	Std. dev.
Salary	\$2,024,616	\$2,764,512
Batting average	0.267	0.031
Home runs	12.11	10.31

$$\widehat{\text{Salary}_i} = -2{,}869{,}439.40 + 12{,}417{,}629.72\text{Batting average}_i + 129{,}627.36\text{Home runs}_i$$
$$\widehat{\text{Salary}_i}^{std} = 0.00 + 0.14\text{Batting average}_i^{std} + 0.48\text{Home runs}_i^{std}$$

- Marginal effect on Y (in *standard deviations* of Y) from 1 *standard deviation* change in X

Example

Variable	Mean	Std. dev.
Salary	\$2,024,616	\$2,764,512
Batting average	0.267	0.031
Home runs	12.11	10.31

$$\widehat{\text{Salary}_i} = -2{,}869{,}439.40 + 12{,}417{,}629.72\text{Batting average}_i + 129{,}627.36\text{Home runs}_i$$

$$\widehat{\text{Salary}_i}^{std} = 0.00 + 0.14\text{Batting average}_i^{std} + 0.48\text{Home runs}_i^{std}$$

- Marginal effect on Y (in *standard deviations* of Y) from 1 *standard deviation* change in X
- e.g. for 1 standard deviation increase in Batting Average, Salary increases by 0.14 standard deviations

Standardizing Variables: Example

Example

Variable	Mean	Std. dev.
Salary	\$2,024,616	\$2,764,512
Batting average	0.267	0.031
Home runs	12.11	10.31

$$\widehat{\text{Salary}_i} = -2{,}869{,}439.40 + 12{,}417{,}629.72\text{Batting average}_i + 129{,}627.36\text{Home runs}_i$$

$$\widehat{\text{Salary}_i}^{std} = 0.00 + 0.14\text{Batting average}_i^{std} + 0.48\text{Home runs}_i^{std}$$

- Marginal effect on Y (in *standard deviations* of Y) from 1 *standard deviation* change in X
- e.g. for 1 standard deviation increase in Batting Average, Salary increases by 0.14 standard deviations
  - $0.14 \times 2{,}764{,}512 = \$387{,}032$

- Use the `scale()` command inside `dplyr`'s `mutate()` function to standardize a variable

```
gapminder<-gapminder %>%
  mutate(s.life=scale(lifeExp),
         s.gdp=scale(gdpPercap))

summary(lm(s.life~s.gdp, data=gapminder))
```

```
##
## Call:
## lm(formula = s.life ~ s.gdp, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

Joint-Hypothesis Testing

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i$$

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i$$

- Maybe region doesn't affect wages *at all*?

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i$$

- Maybe region doesn't affect wages *at all*?
- H_0: $\beta_2 = 0, \beta_3 = 0, \beta_4 = 0$

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i$$

- Maybe region doesn't affect wages *at all*?
- H_0: $\beta_2 = 0$, $\beta_3 = 0$, $\beta_4 = 0$
- This is a joint hypothesis to test

- A joint hypothesis tests against the null hypothesis of a value for *multiple* parameters:

$$H_0 : \beta_1 = \beta_2 = 0$$

  the hypotheses that multiple regressors are equal to zero (have no causal effect on the outcome)

- A joint hypothesis tests against the null hypothesis of a value for *multiple* parameters:

$$H_0 : \beta_1 = \beta_2 = 0$$

the hypotheses that multiple regressors are equal to zero (have no causal effect on the outcome)

- Our alternative hypothesis is that:

$$H_1 : \text{ either } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both}$$

or simply, that H_0 is not true

- Three possible cases:

- Three possible cases:
    1. H_0: $\beta_1 = \beta_2 = 0$

Joint Hypothesis Tests

- Three possible cases:
    1. H_0: $\beta_1 = \beta_2 = 0$
        - Testing if multiple variables don't matter

- Three possible cases:
  1. H_0: $\beta_1 = \beta_2 = 0$
     - Testing if multiple variables don't matter
     - Useful under high multicollinearity

- Three possible cases:

  1. H_0: $\beta_1 = \beta_2 = 0$
     - Testing if multiple variables don't matter
     - Useful under high multicollinearity
     - H_A: at least one parameter is not 0

- Three possible cases:
    1. H_0: $\beta_1 = \beta_2 = 0$
        - Testing if multiple variables don't matter
        - Useful under high multicollinearity
        - H_A: at least one parameter is not 0
    2. H_0: $\beta_1 = \beta_2$

- Three possible cases:
    1. H_0: $\beta_1 = \beta_2 = 0$
        - Testing if multiple variables don't matter
        - Useful under high multicollinearity
        - H_A: at least one parameter is not 0
    2. H_0: $\beta_1 = \beta_2$
        - Testing if two variables matter the same

- Three possible cases:
  1. H_0: $\beta_1 = \beta_2 = 0$
     - Testing if multiple variables don't matter
     - Useful under high multicollinearity
     - H_A: at least one parameter is not 0
  2. H_0: $\beta_1 = \beta_2$
     - Testing if two variables matter the same
     - Variables must be in the same units

- Three possible cases:

1. H_0: $\beta_1 = \beta_2 = 0$
    - Testing if multiple variables don't matter
    - Useful under high multicollinearity
    - H_A: at least one parameter is not 0
2. H_0: $\beta_1 = \beta_2$
    - Testing if two variables matter the same
    - Variables must be in the same units
    - H_A: $\beta_1 (\neq, <, \text{ or } >) \beta_2$

HOOD COLLEGE

- Three possible cases:
  1. H_0: $\beta_1 = \beta_2 = 0$
     - Testing if multiple variables don't matter
     - Useful under high multicollinearity
     - H_A: at least one parameter is not 0
  2. H_0: $\beta_1 = \beta_2$
     - Testing if two variables matter the same
     - Variables must be in the same units
     - H_A: $\beta_1 (\neq, <, \text{ or } >) \beta_2$
  3. H_0: all β's$= 0$

- Three possible cases:

1. H_0: $\beta_1 = \beta_2 = 0$
   - Testing if multiple variables don't matter
   - Useful under high multicollinearity
   - H_A: at least one parameter is not 0
2. H_0: $\beta_1 = \beta_2$
   - Testing if two variables matter the same
   - Variables must be in the same units
   - H_A: $\beta_1(\neq, <, \text{ or } >)\beta_2$
3. H_0: all β's$= 0$
   - The "**Overall *F*-test**"

- Three possible cases:

1. H_0: $\beta_1 = \beta_2 = 0$
    - Testing if multiple variables don't matter
    - Useful under high multicollinearity
    - H_A: at least one parameter is not 0
2. H_0: $\beta_1 = \beta_2$
    - Testing if two variables matter the same
    - Variables must be in the same units
    - H_A: $\beta_1 (\neq, <, \text{ or } >) \beta_2$
3. H_0: all β's$= 0$
    - The "**Overall** *F*-**test**"
    - Testing if regression explains *no* variation in Y

Joint Hypothesis Tests: The F-Statistic

- The *F-statistic* is used to test joint hypotheses about regression coefficients with an *F-test*

Joint Hypothesis Tests: The F-Statistic

- The *F-statistic* is used to test joint hypotheses about regression coefficients with an *F-test*
- This involves comparing two models:

- The *F*-statistic is used to test joint hypotheses about regression coefficients with an *F*-test
- This involves comparing two models:

  1. **Unrestricted model**: regression with all coefficients

- The *F*-statistic is used to test joint hypotheses about regression coefficients with an *F*-test
- This involves comparing two models:

1. **Unrestricted model**: regression with all coefficients
2. **Restricted model**: regression under null hypothesis (coefficients equal hypothesized values)

- The *F-statistic* is used to test joint hypotheses about regression coefficients with an *F-test*
- This involves comparing two models:
    1. **Unrestricted model**: regression with all coefficients
    2. **Restricted model**: regression under null hypothesis (coefficients equal hypothesized values)
- F is an **analysis of variance (ANOVA)**, essentially tests whether R^2 increases statistically significantly as we go from the restricted model$\longrightarrow$unrestricted model

Joint Hypothesis Tests: The *F*-Statistic

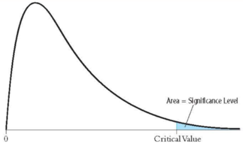
- The *F*-statistic is used to test joint hypotheses about regression coefficients with an *F*-test
- This involves comparing two models:

  1. **Unrestricted model**: regression with all coefficients
  2. **Restricted model**: regression under null hypothesis (coefficients equal hypothesized values)

- F is an **analysis of variance (ANOVA)**, essentially tests whether R^2 increases statistically significantly as we go from the restricted model$\longrightarrow$unrestricted model
- F has its own distribution, with *two* sets of degrees of freedom

Joint Hypothesis Tests: The F-Statistic II

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

Example

$$\widehat{Wages}_i = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- Let H_0: $\beta_2 = \beta_3 = \beta_4 = 0$

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- Let H_0: $\beta_2 = \beta_3 = \beta_4 = 0$
- Let H_1: H_0 is not true (at least one beta$\neq$ 0

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- Let H_0: $\beta_2 = \beta_3 = \beta_4 = 0$
- Let H_1: H_0 is not true (at least one beta$\neq 0$
- **Unrestricted Model**:

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- Let H_0: $\beta_2 = \beta_3 = \beta_4 = 0$
- Let H_1: H_0 is not true (at least one beta$\neq$ 0
- **Unrestricted Model**:

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- **Restricted Model**:

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \epsilon_i$$

Example

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- Let H_0: $\beta_2 = \beta_3 = \beta_4 = 0$
- Let H_1: H_0 is not true (at least one beta$\neq$ 0
- **Unrestricted Model**:

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \hat{\beta}_2 Northeast_i + \hat{\beta}_3 Northcen_i + \hat{\beta}_4 South_i + \epsilon_i$$

- **Restricted Model**:

$$\widehat{Wages_i} = \hat{\beta}_0 + \hat{\beta}_1 Male + \epsilon_i$$

- F: does going from restricted to unrestricted statistically significantly improve R^2?

Calculating the F-statistic

$$F_{q,n-k-1} = \frac{\dfrac{(R_u^2 - R_r^2)}{q}}{\dfrac{(1 - R_u^2)}{(n-k-1)}}$$

$$F_{q,n-k-1} = \frac{\frac{(R_u^2 - R_r^2)}{q}}{\frac{(1 - R_u^2)}{(n-k-1)}}$$

- R_u^2 : the R^2 from the unrestricted model

Calculating the F-statistic

$$F_{q,n-k-1} = \frac{\frac{(R^2_u - R^2_r)}{q}}{\frac{(1 - R^2_u)}{(n-k-1)}}$$

- R^2_u : the R^2 from the unrestricted model
- R^2_r: the R^2 from the restricted model
- q: number of restrictions

$$F_{q,n-k-1} = \frac{\frac{(R^2_u - R^2_r)}{q}}{\frac{(1-R^2_u)}{(n-k-1)}}$$

- R^2_u : the R^2 from the unrestricted model
- R^2_r: the R^2 from the restricted model
- q: number of restrictions
- k: number of variables in the unrestricted model

$$F_{q,n-k-1} = \frac{\frac{(R^2_u - R^2_r)}{q}}{\frac{(1 - R^2_u)}{(n - k - 1)}}$$

- R^2_u : the R^2 from the unrestricted model
- R^2_r: the R^2 from the restricted model
- q: number of restrictions
- k: number of variables in the unrestricted model
- F has two sets of degrees of freedom, q for numerator, $n - k - 1$ for denominator

$$F_{q,n-k-1} = \frac{\frac{(R_u^2 - R_r^2)}{q}}{\frac{(1 - R_u^2)}{(n-k-1)}}$$

- The bigger the difference between $(R_u^2 - R_r^2)$, the greater the improvement in fit by adding variables, the larger the F

$$F_{q,n-k-1} = \frac{\frac{(R_u^2 - R_r^2)}{q}}{\frac{(1 - R_u^2)}{(n-k-1)}}$$

- The bigger the difference between $(R_u^2 - R_r^2)$, the greater the improvement in fit by adding variables, the larger the F
- This formula is actually a bit simplified, assumes **homoskedastic** errors

$$F_{q,n-k-1} = \frac{\frac{\left(R_u^2 - R_r^2\right)}{q}}{\frac{\left(1 - R_u^2\right)}{(n-k-1)}}$$

- The bigger the difference between $\left(R_u^2 - R_r^2\right)$, the greater the improvement in fit by adding variables, the larger the F
- This formula is actually a bit simplified, assumes **homoskedastic** errors
    - Heteroskedasticity-robust formula a lot more complicated

$$F_{q,n-k-1} = \frac{\frac{\left(R_u^2 - R_r^2\right)}{q}}{\frac{\left(1 - R_u^2\right)}{(n-k-1)}}$$

- The bigger the difference between $\left(R_u^2 - R_r^2\right)$, the greater the improvement in fit by adding variables, the larger the F
- This formula is actually a bit simplified, assumes **homoskedastic** errors
    - Heteroskedasticity-robust formula a lot more complicated
- This formula is just to give you an *intuition* of what F is doing

```r
# Load WAGE1 as wages
library("foreign") # to load .dta Stata files
wages<-read.dta("../Data/WAGE1.dta")

unrestricted<-lm(wage~female+northcen+west+south, data=wages)
restricted<-lm(wage~female, data=wages)
```

The F-test Example II

```r
library("car") # load car package for additional regression tools
linearHypothesis(unrestricted, c("northcen", "west", "south")) # test that northcen=west=south=0
```

```
## Linear hypothesis test
##
## Hypothesis:
## northcen = 0
## west = 0
## south = 0
##
## Model 1: restricted model
## Model 2: wage ~ female + northcen + west + south
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    524 6332.2
## 2    521 6174.8  3    157.36 4.4258 0.004377 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing a Joint Hypothesis: Are Two Coefficients Equal

- Testing whether two coefficients equal one another

Example

$$\widehat{wage}_i = \beta_0 + \beta_1 \text{Adolescent height}_i + \beta_2 \text{Adult height}_i + \beta_3 \text{Male}_i$$

- Testing whether two coefficients equal one another

Example

$$\widehat{wage_i} = \beta_0 + \beta_1 \text{Adolescent height}_i + \beta_2 \text{Adult height}_i + \beta_3 \text{Male}_i$$

- Does height as an adolescent have the same effect on wages as height as an adult?

$$H_0 : \beta_1 = \beta_2$$

- Testing whether two coefficients equal one another

Example

$$\widehat{wage_i} = \beta_0 + \beta_1 \text{Adolescent height}_i + \beta_2 \text{Adult height}_i + \beta_3 \text{Male}_i$$

- Does height as an adolescent have the same effect on wages as height as an adult?

$$H_0 : \beta_1 = \beta_2$$

- What is the restricted regression?

Testing a Joint Hypothesis: Are Two Coefficients Equal

- Testing whether two coefficients equal one another

Example

$$\widehat{wage}_i = \beta_0 + \beta_1 \text{Adolescent height}_i + \beta_2 \text{Adult height}_i + \beta_3 \text{Male}_i$$

- Does height as an adolescent have the same effect on wages as height as an adult?

$$H_0 : \beta_1 = \beta_2$$

- What is the restricted regression?

$$\widehat{wage}_i = \beta_0 + \beta_1 (\text{Adolescent height}_i + \text{Adult height}_i) + \beta_3 \text{Male}_i$$

HOOD COLLEGE

Testing a Joint Hypothesis: Are Two Coefficients Equal: Example

```r
# load HeightWages
height<-read.csv("../Data/HeightWages.csv")

# make a "heights" variable as the sum of adolescent and adult height
height <- height %>%
  mutate(heights=height81+height85)

h.unrestricted<-lm(wage96~height81+height85+male, data=height)
h.restricted<-lm(wage96~heights+male, data=height)
```

Testing a Joint Hypothesis: Are Two Coefficients Equal: Example II

```r
linearHypothesis(h.unrestricted, "height81=height85") # F-test
```

```
## Linear hypothesis test
##
## Hypothesis:
## height81 - height85 = 0
##
## Model 1: restricted model
## Model 2: wage96 ~ height81 + height85 + male
##
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   6591 5128243
## 2   6590 5127284  1     959.2 1.2328 0.2669
```

- No. Insufficient evidence to reject H_0

Testing a Joint Hypothesis: The "Overall F-test"

- The "overall F-test" tests against H_0: *all* regression coefficients $= 0$

Testing a Joint Hypothesis: The "Overall F-test"

- The "overall F-test" tests against H_0: *all* regression coefficients $= 0$
- The $R^2_{restricted} = 0$

- The "overall F-test" tests against H_0: *all* regression coefficients $= 0$
- The $R^2_{restricted} = 0$
- Tests if R^2 of a model is statistically significantly greater than 0

Testing a Joint Hypothesis: The "Overall F-test"

- The "overall F-test" tests against H_0: *all* regression coefficients $= 0$
- The $R^2_{restricted} = 0$
- Tests if R^2 of a model is statistically significantly greater than 0
- R calculates automatically for every regression run (bottom line of output)