3.9 — Logarithmic Regression

ECON 480 • Econometrics • Fall 2020

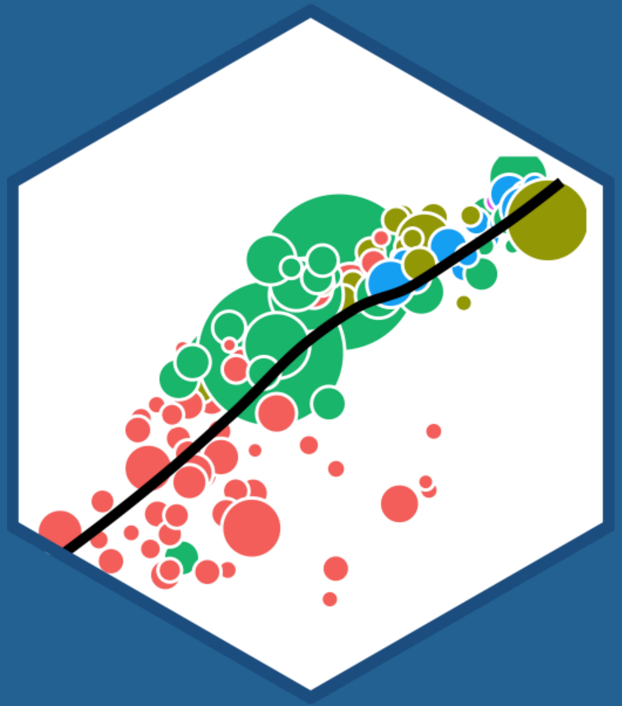
Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF20

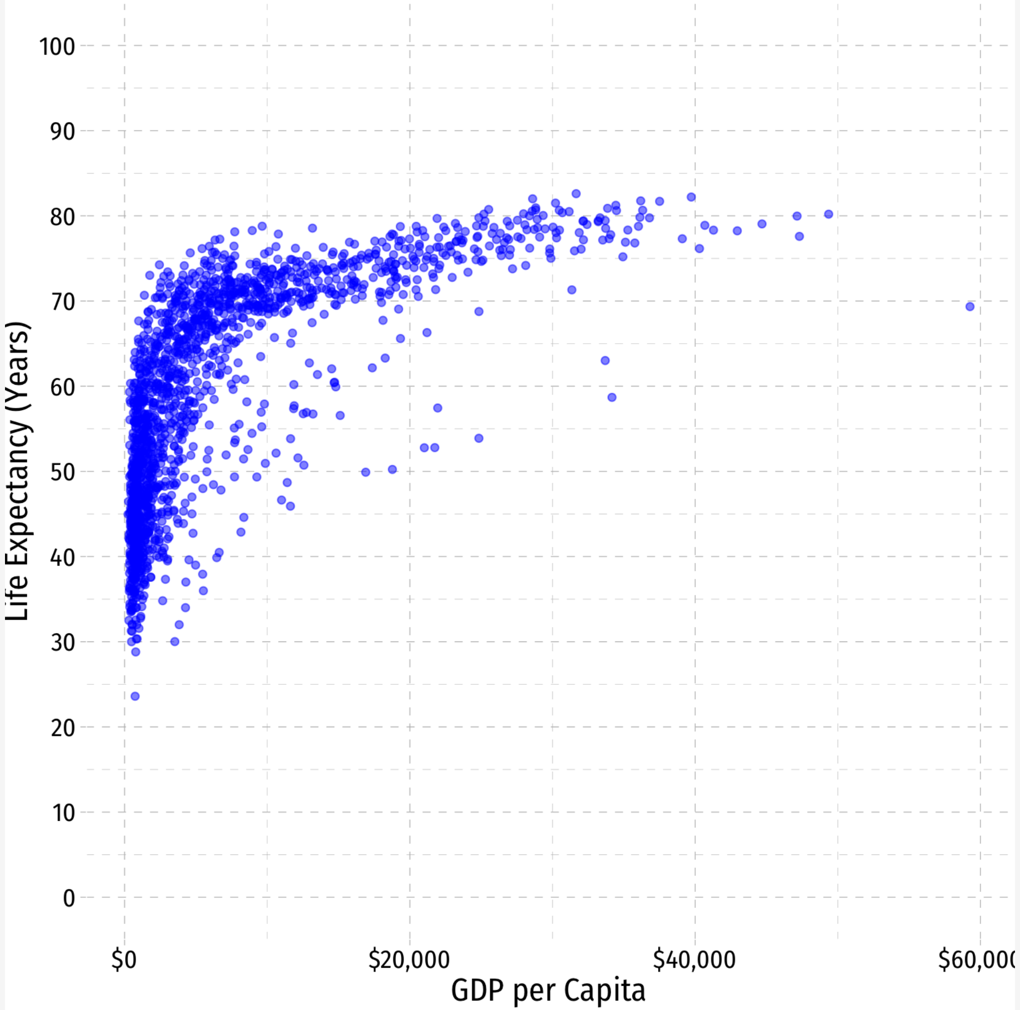
metricsF20.classes.ryansafner.com



Outline

Nonlinearities

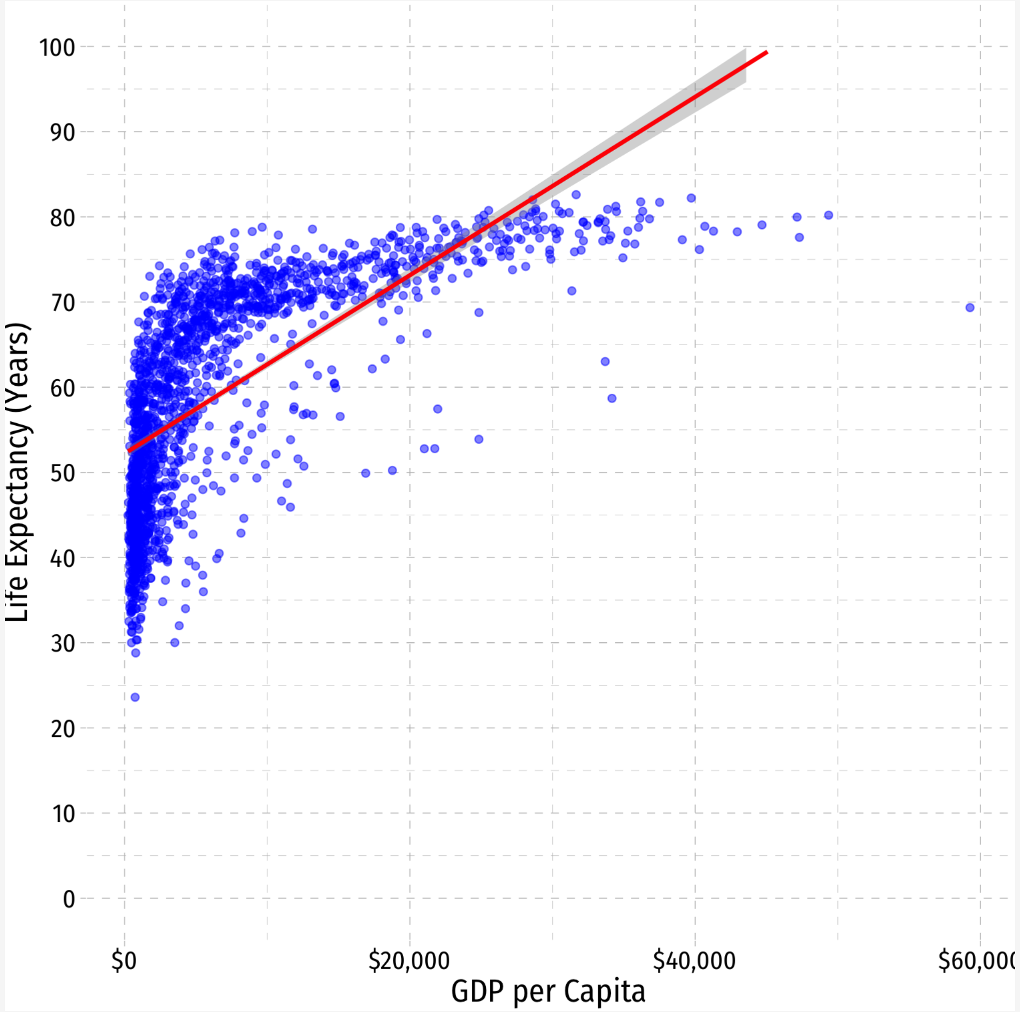
- Consider the `gapminder` example



Nonlinearities

- Consider the `gapminder` example

$$\color{red}{\widehat{\text{Life Expectancy}_i}=\hat{\beta_0}+\hat{\beta_1}\text{GDP per capita}_i}$$

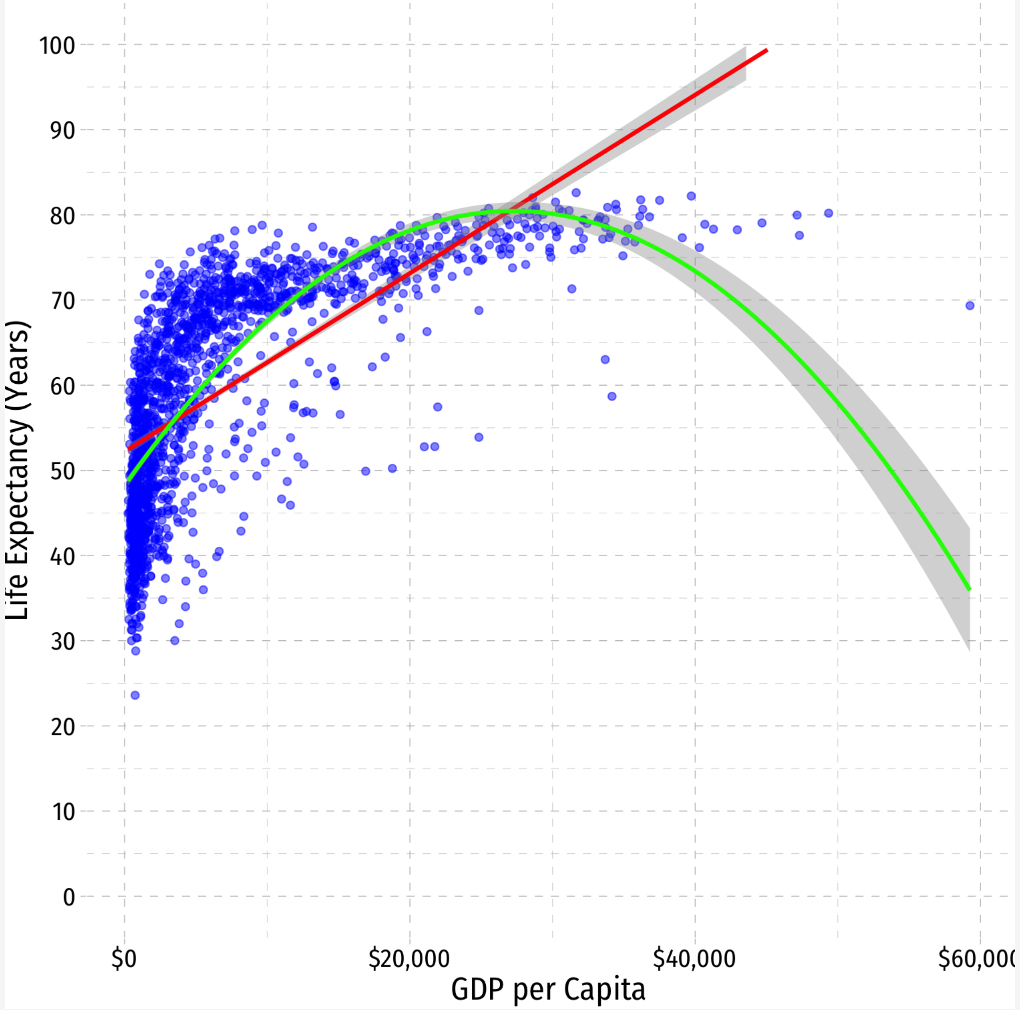


Nonlinearities

- Consider the `gapminder` example

$$\color{red}{\widehat{\text{Life Expectancy}_i}=\hat{\beta_0}+\hat{\beta_1}\text{GDP per capita}_i}$$

$$\color{green}{\widehat{\text{Life Expectancy}_i}=\hat{\beta_0}+\hat{\beta_1}\text{GDP per capita}_i+\hat{\beta_2}\text{GDP per capita}_i^2}$$



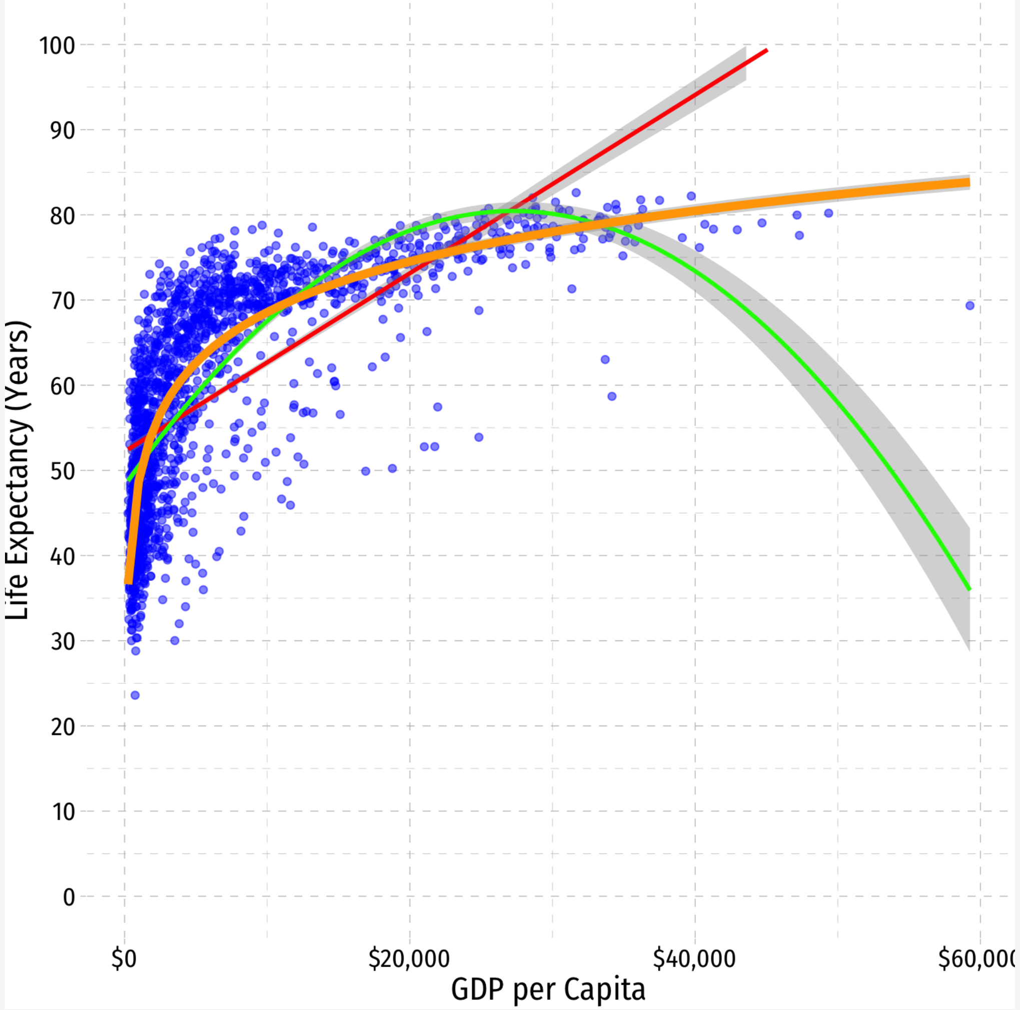
Nonlinearities

- Consider the `gapminder` example

$$\color{red}{\widehat{\text{Life Expectancy}_i}=\hat{\beta_0}+\hat{\beta_1}\text{GDP per capita}_i}$$

$$\color{green}{\widehat{\text{Life Expectancy}_i}=\hat{\beta_0}+\hat{\beta_1}\text{GDP per capita}_i+\hat{\beta_2}\text{GDP per capita}_i^2}$$

$$\color{orange}{\widehat{\text{Life Expectancy}_i}=\hat{\beta_0}+\hat{\beta_1}\ln \text{GDP per capita}_i}$$



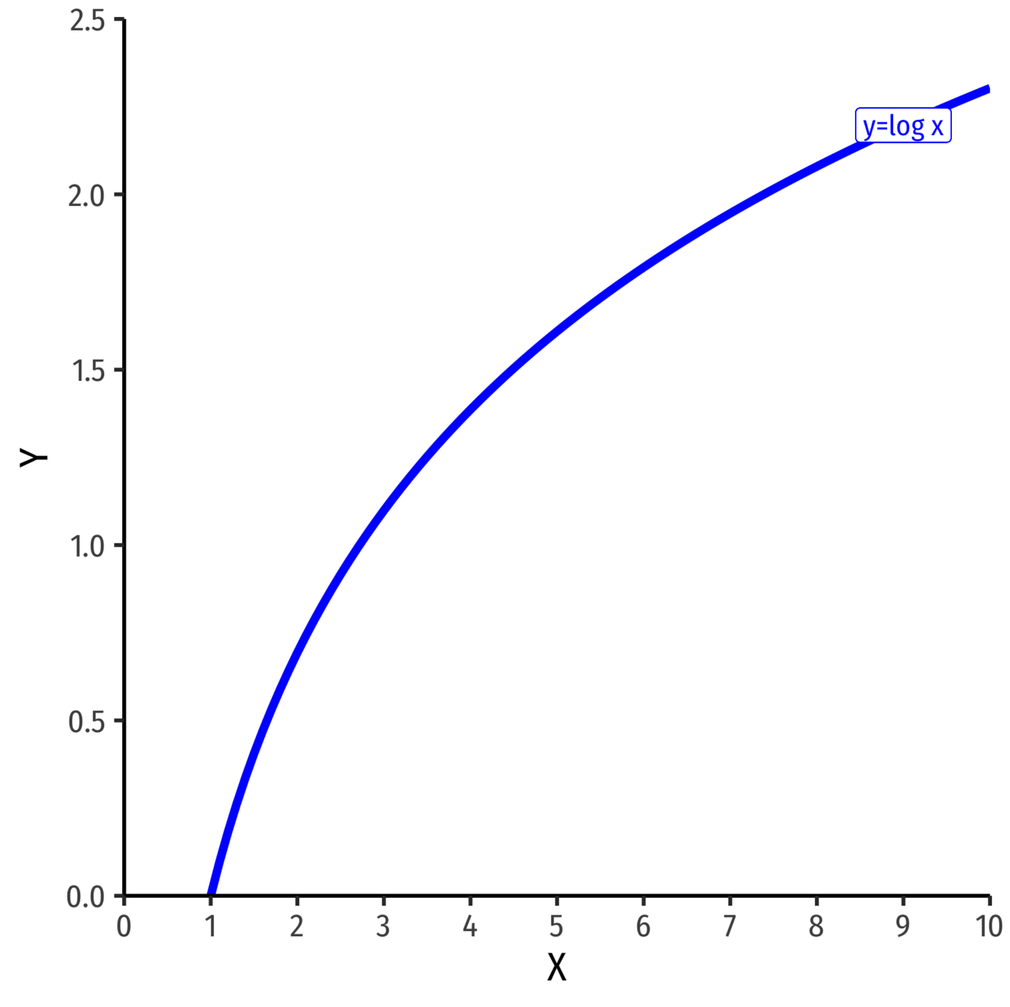
Natural Logarithms

Logarithmic Models

- Another useful model for nonlinear data is the **logarithmic model**†
    - We transform either \(X\), \(Y\), or *both* by taking the **(natural) logarithm**

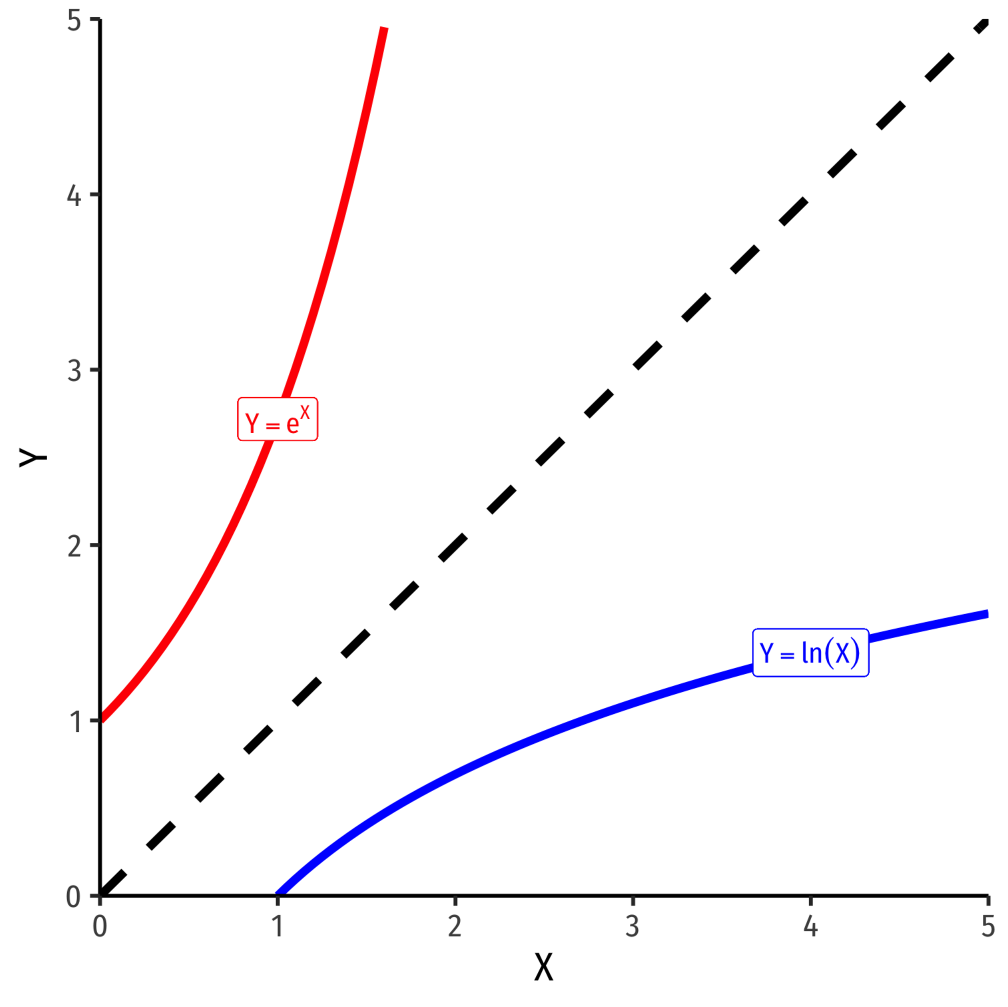
- Logarithmic model has two additional advantages

    1. We can easily interpret coefficients as **percentage changes** or **elasticities**
    2. Useful economic shape: diminishing returns (production functions, utility functions, etc)



† Don't confuse this with a **logistic (logit) model** for *dependent* dummy variables.

The Natural Logarithm



- The exponential function, $Y=e^X$ or $Y=\exp(X)$, where base $e=2.71828...$
- Natural logarithm is the inverse, $Y=\ln(X)$

The Natural Logarithm: Review I

- **Exponents** are defined as $$\color{#6A5ACD}{b}^{\color{#e64173}{n}}=\underbrace{\color{#6A5ACD}{b} \times \color{#6A5ACD}{b} \times \cdots \times \color{#6A5ACD}{b}}_{\color{#e64173}{n} \text{ times}}$$
    - where base \(\color{#6A5ACD}{b}\) is multiplied by itself \(\color{#e64173}{n}\) times

- **Example**: \(\color{#6A5ACD}{2}^{\color{#e64173}{3}}=\underbrace{\color{#6A5ACD}{2} \times \color{#6A5ACD}{2} \times \color{#6A5ACD}{2}}_{\color{#e64173}{n=3}}=\color{#314f4f}{8}\)

- **Logarithms** are the inverse, defined as the exponents in the expressions above $$\text{If } \color{#6A5ACD}{b}^{\color{#e64173}{n}}=\color{#314f4f}{y}\text{, then }log_{\color{#6A5ACD}{b}}(\color{#314f4f}{y})=\color{#e64173}{n}$$
    - \(\color{#e64173}{n}\) is the number you must raise \(\color{#6A5ACD}{b}\) to in order to get \(\color{#314f4f}{y}\)

- **Example**: \(log_{\color{#6A5ACD}{2}}(\color{#314f4f}{8})=\color{#e64173}{3}\)

The Natural Logarithm: Review II

- Logarithms can have any base, but common to use the **natural logarithm \((ln)\)** with base \(\mathbf{e=2.71828...}\) $$\text{If } e^n=y\text{, then } \ln(y)=n$$

The Natural Logarithm: Properties

- Natural logs have a lot of useful properties:

1. $ln(\frac{1}{x})=-ln(x)$

2. $ln(ab)=ln(a)+ln(b)$

3. $ln(\frac{x}{a})=ln(x)-ln(a)$

4. $ln(x^a)=a \, ln(x)$

5. $\frac{d \, ln \, x}{d \, x} = \frac{1}{x}$

The Natural Logarithm: 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:

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

- The logged difference: $$ln(101)-ln(100) = 0.00995 \approx 1\%$$

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

Elasticity

- An elasticity between any two variables, $\epsilon_{Y,X}$ describes the responsiveness (in %) of one variable (Y) to a change in another (X)

$$\epsilon_{Y,X}=\frac{\% \Delta Y}{\% \Delta X} =\cfrac{\left(\frac{\Delta Y}{Y}\right)}{\left(\frac{\Delta X}{X}\right)}$$

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

- **Interpretation**: a 1% change in X will cause a $\epsilon_{Y,X}$% chang 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
    - A 1% change in K will lead to a β% change in Y

Math FYI: Cobb Douglas Functions and Logs

Example: Cobb-Douglas production function: $$Y=2L^{0.75}K^{0.25}$$

- Taking logs:

$$\ln Y=\ln 2+0.75 \ln L + 0.25 \ln K$$

- A 1% change in L will yield a 0.75% change in output Y
- A 1% change in K will yield a 0.25% change in output Y

Logarithms in R I

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

```
gapminder <- gapminder %>%
  mutate(loggdp = log(gdpPercap)) # log GDP per capita

gapminder %>% head() # look at it
```

country <fctr>	continent <fctr>	year <int>	lifeExp <dbl>	pop <int>	gdpPercap <dbl>	loggdp <dbl>
Afghanistan	Asia	1952	28.801	8425333	779.4453	6.658583
Afghanistan	Asia	1957	30.332	9240934	820.8530	6.710344
Afghanistan	Asia	1962	31.997	10267083	853.1007	6.748878
Afghanistan	Asia	1967	34.020	11537966	836.1971	6.728864
Afghanistan	Asia	1972	36.088	13079460	739.9811	6.606625
Afghanistan	Asia	1977	38.438	14880372	786.1134	6.667101

6 rows

Logarithms in R II

- 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`

```r
log10(100)
```

```
## [1] 2
```

```r
log2(16)
```

```
## [1] 4
```

```r
log(19683, base=3)
```

```
## [1] 9
```

Logarithms in R III

- Note when running a regression, you can pre-transform the data into logs (as I did above), or just add `log()` around a variable in the regression

```
lm(lifeExp ~ loggdp,
   data = gapminder) %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117	1.934812e-13
loggdp	8.405085	0.148762	56.500206	0.000000e+00

2 rows

```
lm(lifeExp ~ log(gdpPercap),
   data = gapminder) %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117	1.934812e-13
log(gdpPercap)	8.405085	0.148762	56.500206	0.000000e+00

2 rows

Types of Logarithmic Models

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

1. **Linear-log model:** $Y_i=\beta_0+\beta_1 \color{#e64173}{\ln X_i}$

2. **Log-linear model:** $\color{#e64173}{\ln Y_i}=\beta_0+\beta_1X_i$

3. **Log-log model:** $\color{#e64173}{\ln Y_i}=\beta_0+\beta_1 \color{#e64173}{\ln X_i}$

Linear-Log Model

Linear-Log Model

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

$$\begin{align*} Y&=\beta_0+\beta_1 \color{#e64173}{\ln X_i}\\ \beta_1&=\cfrac{\Delta Y}{\big(\frac{\Delta X}{X}\big)}\\ \end{align*}$$

- **Marginal effect of $\mathbf{X \rightarrow Y}$: a 1% change in $X \rightarrow$ a $\frac{\beta_1}{100}$ unit change in Y**

Linear-Log Model in R

```
lin_log_reg <- lm(lifeExp ~ loggdp, data = gapm
library(broom)
lin_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117	1.934812e-13
loggdp	8.405085	0.148762	56.500206	0.000000e+00

2 rows

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

Linear-Log Model in R

```r
lin_log_reg <- lm(lifeExp ~ loggdp, data = gapm
library(broom)
lin_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117	1.934812e-13
loggdp	8.405085	0.148762	56.500206	0.000000e+00

2 rows

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

- A **1% change in GDP** $\rightarrow$ a $\frac{9.41}{100}=$ **0.0941 year increase** in Life Expectancy

Linear-Log Model in R

```r
lin_log_reg <- lm(lifeExp ~ loggdp, data = gapm
library(broom)
lin_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117	1.934812e-13
loggdp	8.405085	0.148762	56.500206	0.000000e+00

2 rows

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

- 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

Linear-Log Model in R

```
lin_log_reg <- lm(lifeExp ~ loggdp, data = gapm
library(broom)
lin_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-9.100889	1.227674	-7.413117	1.934812e-13
loggdp	8.405085	0.148762	56.500206	0.000000e+00

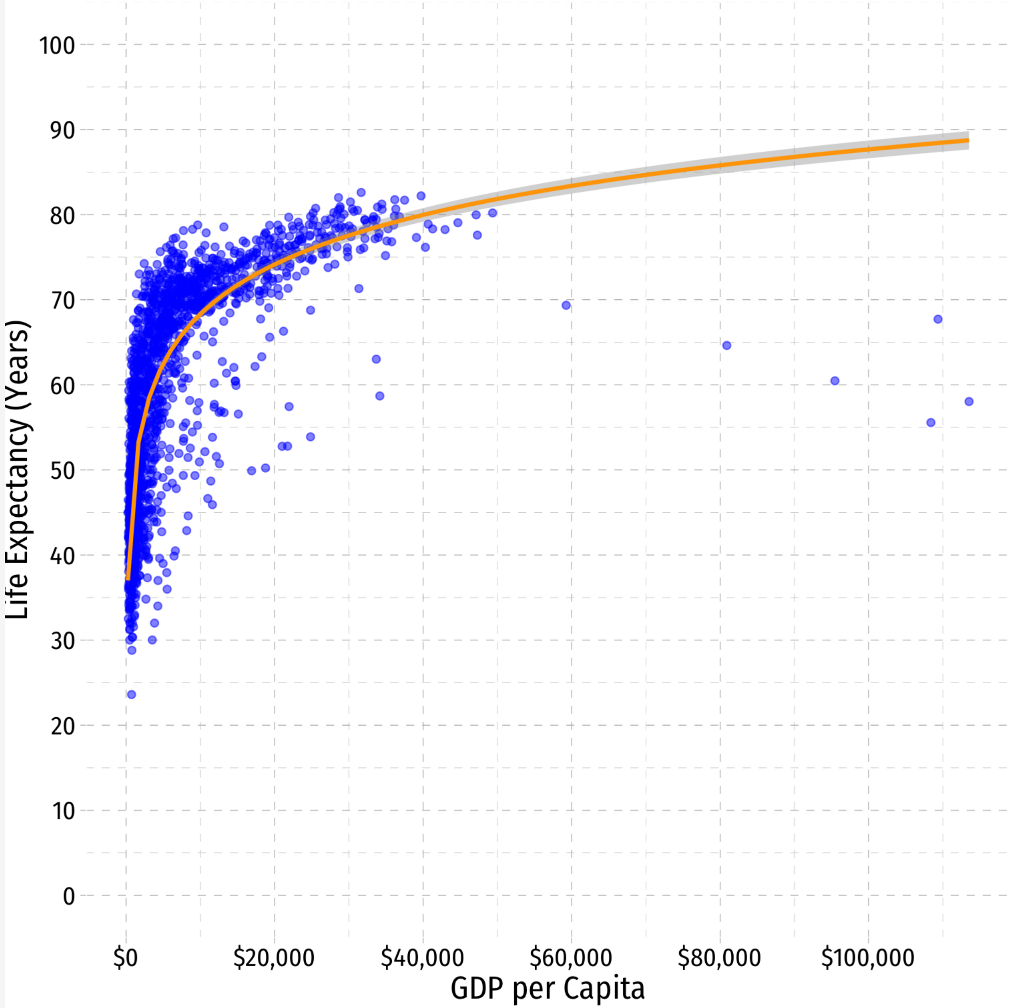
2 rows

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

- 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

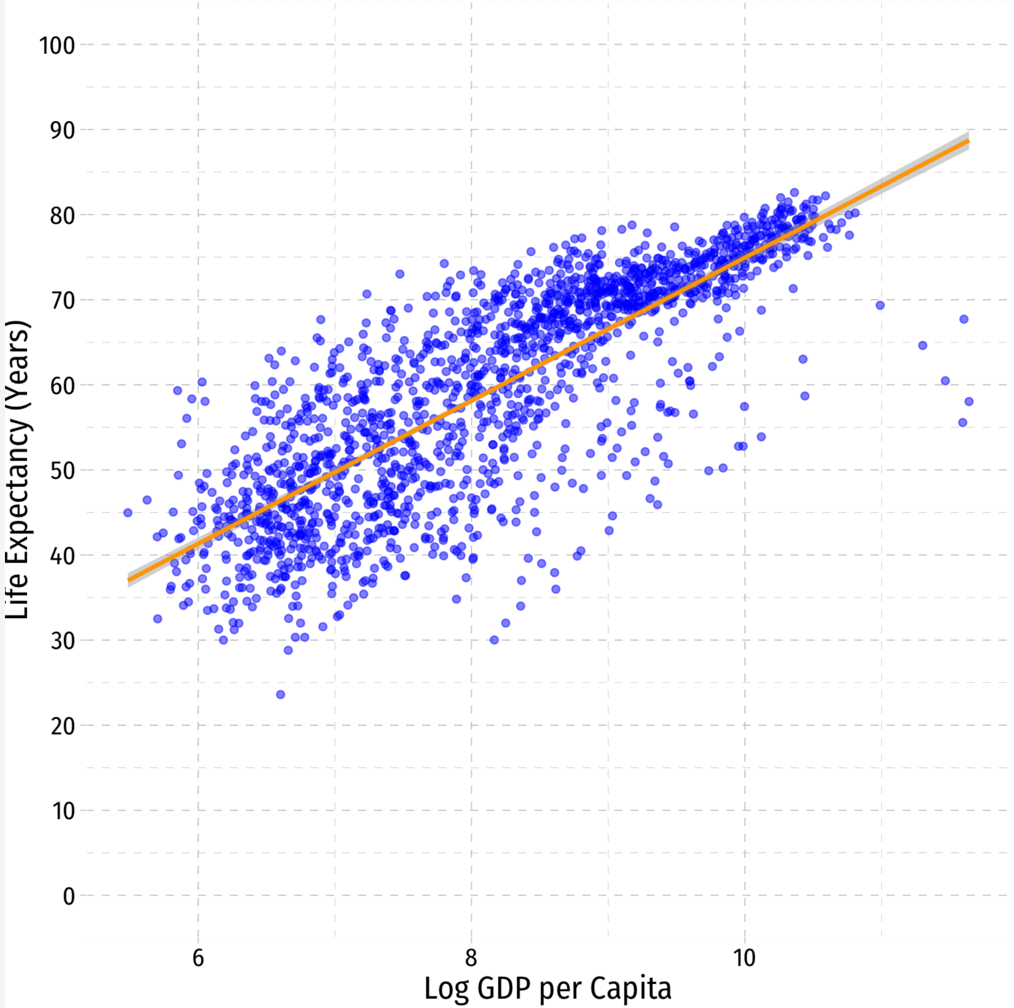
Linear-Log Model Graph I

```r
ggplot(data = gapminder)+
  aes(x = gdpPercap,
      y = lifeExp)+
  geom_point(color="blue", alpha=0.5)+
  geom_smooth(method="lm",
              formula=y~log(x),
              color="orange")+
  scale_x_continuous(labels=scales::dollar,
                     breaks=seq(0,120000,20000))+
  scale_y_continuous(breaks=seq(0,100,10),
                     limits=c(0,100))+
  labs(x = "GDP per Capita",
       y = "Life Expectancy (Years)")+
  ggthemes::theme_pander(base_family = "Fira Sans Condensed",
           base_size=16)
```



Linear-Log Model Graph II

```r
ggplot(data = gapminder)+
  aes(x = loggdp,
      y = lifeExp)+
  geom_point(color="blue", alpha=0.5)+
  geom_smooth(method="lm", color="orange")+
  scale_y_continuous(breaks=seq(0,100,10),
                     limits=c(0,100))+
  labs(x = "Log GDP per Capita",
       y = "Life Expectancy (Years)")+
  ggthemes::theme_pander(base_family = "Fira Sans Condensed",
           base_size=16)
```



Log-Linear Model

Log-Linear Model

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

$$\begin{align*} \color{#e64173}{\ln Y_i}&=\beta_0+\beta_1 X\\
\beta_1&=\cfrac{\big(\frac{\Delta Y}{Y}\big)}{\Delta X}\\ \end{align*}$$

- **Marginal effect of $\mathbf{X \rightarrow Y}$: a 1 unit change in $X \rightarrow$ a $\beta_1 \times 100$ % change in Y**

Log-Linear Model in R (Preliminaries)

- 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`

- Then log LifeExp

```
gapminder <- gapminder %>%
  mutate(gdp_t = gdpPercap/1000, # first make GDP/capita in $1000s
         loglife = log(lifeExp)) # take the log of LifeExp
gapminder %>% head() # look at it
```

country	continent	year	lifeExp	pop	gdpPercap	loggdp	gdp_t	loglife
<fctr>	<fctr>	<int>	<dbl>	<int>	<dbl>	<dbl>	<dbl>	<dbl>
Afghanistan	Asia	1952	28.801	8425333	779.4453	6.658583	0.7794453	3.360410
Afghanistan	Asia	1957	30.332	9240934	820.8530	6.710344	0.8208530	3.412203
Afghanistan	Asia	1962	31.997	10267083	853.1007	6.748878	0.8531007	3.465642
Afghanistan	Asia	1967	34.020	11537966	836.1971	6.728864	0.8361971	3.526949
Afghanistan	Asia	1972	36.088	13079460	739.9811	6.606625	0.7399811	3.585960
Afghanistan	Asia	1977	38.438	14880372	786.1134	6.667101	0.7861134	3.649047

6 rows

Log-Linear Model in R

```
log_lin_reg<-lm(loglife~gdp_t, data = gapminder
tidy(log_lin_reg)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.966639	0.0058345501	679.85339	0.000000e+00
gdp_t	0.012917	0.0004777072	27.03958	2.920378e-134

2 rows

$$\widehat{\ln\text{Life Expectancy}}_i=3.967+0.013 \, \text{GDP}_i$$

Log-Linear Model in R

```
log_lin_reg <- lm(loglife ~ gdp_t, data = gapmi
log_lin_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.966639	0.0058345501	679.85339	0.000000e+00
gdp_t	0.012917	0.0004777072	27.03958	2.920378e-134

2 rows

$$\widehat{\ln\text{Life Expectancy}}_i=3.967+0.013 \, \text{GDP}_i$$

- A **$1 (thousand) change in GDP** $\rightarrow$ a $0.013 \times 100\%=$ **1.3% increase** in Life Expectancy

Log-Linear Model in R

```
log_lin_reg <- lm(loglife ~ gdp_t, data = gapmi
log_lin_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.966639	0.0058345501	679.85339	0.000000e+00
gdp_t	0.012917	0.0004777072	27.03958	2.920378e-134

2 rows

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

- A **$1 (thousand) change in GDP** $\rightarrow$ a $0.013 \times 100\%=$ **1.3% increase** in Life Expectancy
- A **$25 (thousand) 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(loglife ~ gdp_t, data = gapmi
log_lin_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	3.966639	0.0058345501	679.85339	0.000000e+00
gdp_t	0.012917	0.0004777072	27.03958	2.920378e-134

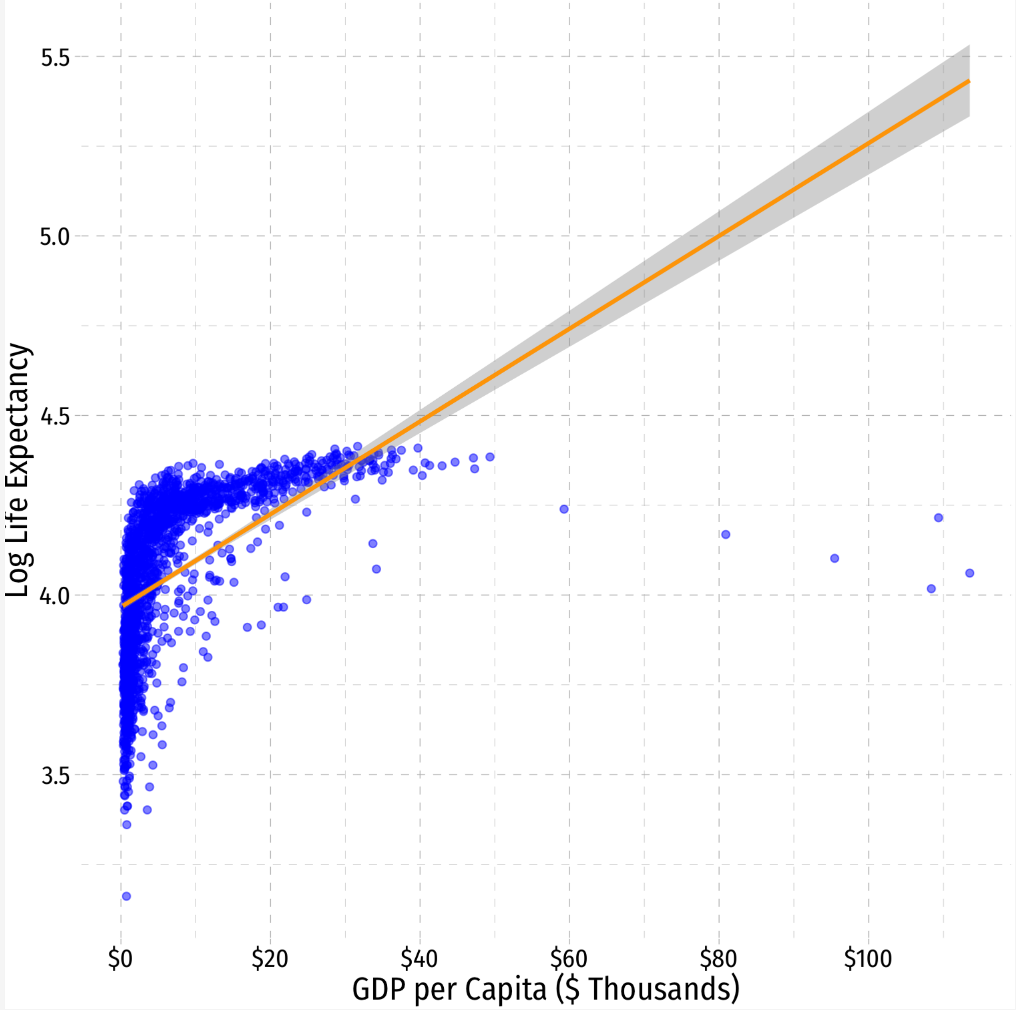
2 rows

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

- A **$1 (thousand) change in GDP** $\rightarrow$ a $0.013 \times 100\%=$ **1.3% increase** in Life Expectancy
- A **$25 (thousand) fall in GDP** $\rightarrow$ a $(-25 \times 1.3\%)=$ **32.5% decrease** in Life Expectancy
- A **$100 (thousand) rise in GDP** $\rightarrow$ a $(100 \times 1.3\%)=$ **130% increase** in Life Expectancy

Linear-Log Model Graph I

```r
ggplot(data = gapminder)+
  aes(x = gdp_t,
      y = loglife)+
  geom_point(color="blue", alpha=0.5)+
  geom_smooth(method="lm", color="orange")+
  scale_x_continuous(labels=scales::dollar,
                     breaks=seq(0,120,20))+
  labs(x = "GDP per Capita ($ Thousands)",
       y = "Log Life Expectancy")+
  ggthemes::theme_pander(base_family = "Fira Sans Condensed",
           base_size=16)
```



Log-Log Model

Log-Log Model

- **Log-log model** has both variables \((X \text{ and } Y)\) logged

$$\begin{align*} \color{#e64173}{\ln Y_i}&=\beta_0+\beta_1 \color{#e64173}{\ln X_i}\\ \beta_1&=\cfrac{\big(\frac{\Delta Y}{Y}\big)}{\big(\frac{\Delta X}{X}\big)}\\ \end{align*}$$

- **Marginal effect of \(\mathbf{X \rightarrow Y}\): a 1% change in \(X \rightarrow\) a \(\beta_1\) % change in \(Y\)**

- \(\beta_1\) is the elasticity of \(Y\) with respect to \(X\)!

Log-Log Model in R

```r
log_log_reg <- lm(loglife ~ loggdp, data = gapm
log_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	2.864177	0.02328274	123.01718	0
loggdp	0.146549	0.00282126	51.94452	0

2 rows

$$\widehat{\text{ln Life Expectancy}}_i=2.864+0.147 \, \text{ln GDP}_i$$

Log-Log Model in R

```
log_log_reg <- lm(loglife ~ loggdp, data = gapm
log_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	2.864177	0.02328274	123.01718	0
loggdp	0.146549	0.00282126	51.94452	0

2 rows

$$\widehat{\text{ln Life Expectancy}}_i=2.864+0.147 \, \text{ln GDP}_i$$

- A **1% change in GDP** $\rightarrow$ a **0.147% increase** in Life Expectancy

Log-Log Model in R

```
log_log_reg <- lm(loglife ~ loggdp, data = gapm
log_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	2.864177	0.02328274	123.01718	0
loggdp	0.146549	0.00282126	51.94452	0

2 rows

$$\widehat{\text{ln Life Expectancy}}_i=2.864+0.147 \, \text{ln GDP}_i$$

- 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

Log-Log Model in R

```
log_log_reg <- lm(loglife ~ loggdp, data = gapm
log_log_reg %>% tidy()
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	2.864177	0.02328274	123.01718	0
loggdp	0.146549	0.00282126	51.94452	0

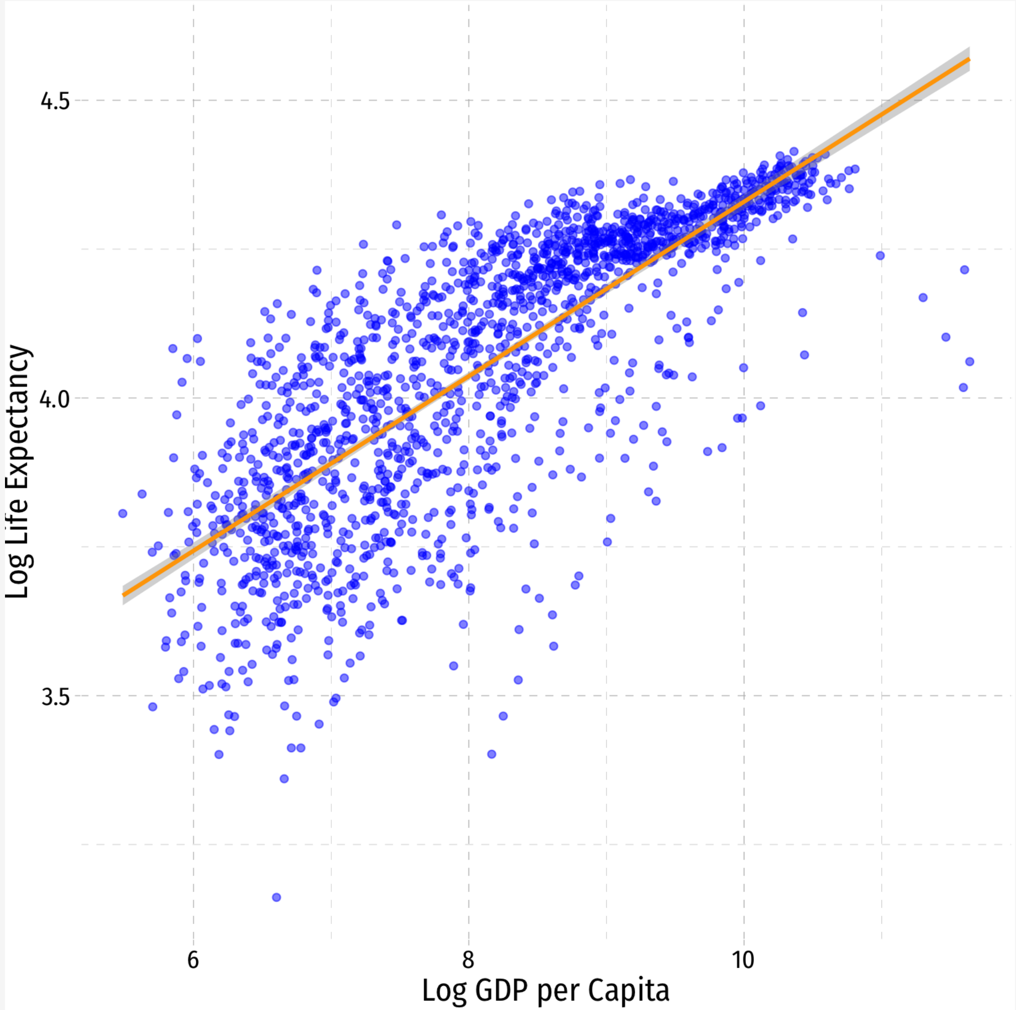
2 rows

$$\widehat{\text{ln Life Expectancy}}_i=2.864+0.147 \, \text{ln GDP}_i$$

- 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 I

```r
ggplot(data = gapminder)+
  aes(x = loggdp,
      y = loglife)+
  geom_point(color="blue", alpha=0.5)+
  geom_smooth(method="lm", color="orange")+
  labs(x = "Log GDP per Capita",
       y = "Log Life Expectancy")+
  ggthemes::theme_pander(base_family = "Fira Sans Condensed",
           base_size=16)
```



Comparing Models I

Model	Equation	Interpretation
Linear-**Log**	$Y=\beta_0+\beta_1 \color{#e64173}{\ln X}$	1% change in $X \rightarrow \frac{\hat{\beta_1}}{100}$ **unit** change in Y
Log-Linear	$\color{#e64173}{\ln Y}=\beta_0+\beta_1X$	1 **unit** change in $X \rightarrow \hat{\beta_1}\times 100$% change in Y
Log-**Log**	$\color{#e64173}{\ln Y}=\beta_0+\beta_1\color{#e64173}{\ln X}$	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

Comparing Models II

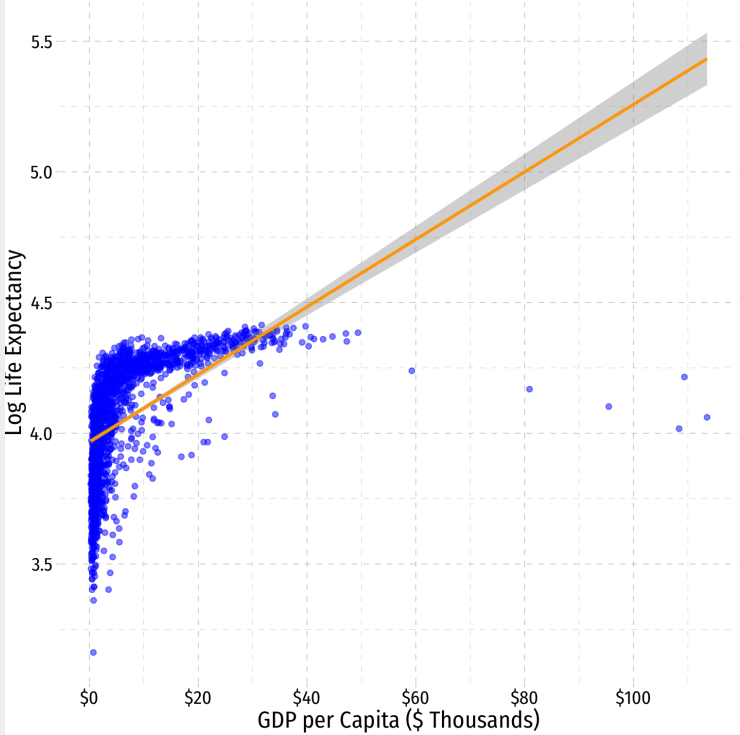
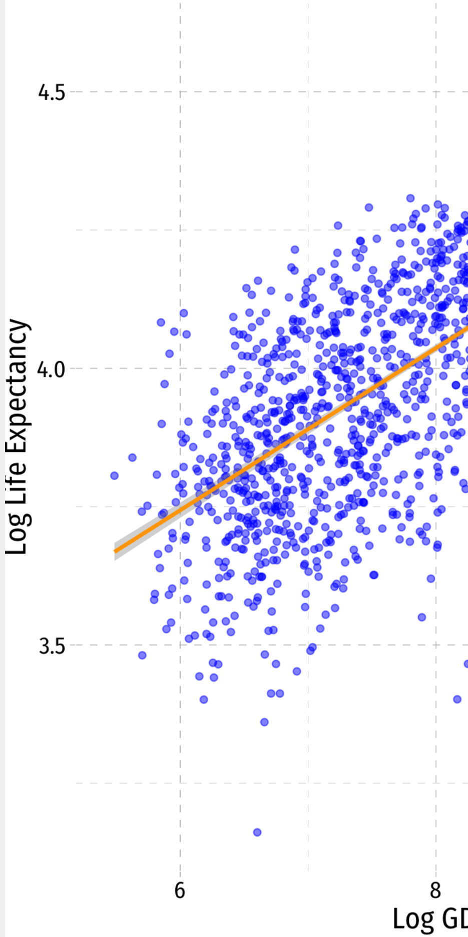
```r
library(huxtable)
huxreg("Life Exp." = lin_log_reg,
       "Log Life Exp." = log_lin_reg,
       "Log Life Exp." = log_log_reg,
       coefs = c("Constant" = "(Intercept)",
                 "GDP ($1000s)" = "gdp_t",
                 "Log GDP" = "loggdp"),
       statistics = c("N" = "nobs",
                      "R-Squared" = "r.squared",
                      "SER" = "sigma"),
       number_format = 2)
```

	Life Exp.	Log Life Exp.	Log Life Exp.
Constant	-9.10 ***	3.97 ***	2.86 ***
	(1.23)	(0.01)	(0.02)
GDP ($1000s)		0.01 ***	
		(0.00)	
Log GDP	8.41 ***		0.15 ***
	(0.15)		(0.00)
N	1704	1704	1704
R-Squared	0.65	0.30	0.61
SER	7.62	0.19	0.14

*** p < 0.001; ** p < 0.01; * p < 0.05.

- Models are very different units, how to choose?
  - Compare \(R^2\)’s
  - Compare graphs
  - Compare intution

Comparing Models III

Linear-Log	Log-Linear	Log
		

When to Log?

- In practice, the following types of variables are logged:
    - Variables that must always be positive (prices, sales, market values)
    - Very large numbers (population, GDP)
    - Variables we want to talk about as percentage changes or growth rates (money supply, population, GDP)
    - Variables that have diminishing returns (output, utility)
    - Variables that have nonlinear scatterplots

- Avoid logs for:
    - Variables that are less than one, decimals, 0, or negative
    - Categorical variables (season, gender, political party)
    - Time variables (year, week, day)

Comparing Across Units

Comparing Coefficients of Different Units 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 units?

Example: $$\begin{align*} \widehat{\text{Salary}_i}&=\beta_0+\beta_1\, \text{Batting average}_i+\beta_2\, \text{Home runs}_i\\ \widehat{\text{Salary}_i}&=-\text{2,869,439.40}+\text{12,417,629.72} \, \text{Batting average}_i+\text{129,627.36}\, \text{Home runs}_i\\ \end{align*}$$

Comparing Coefficients of Different Units II

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

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

† Also called “centering” or “scaling.”

Comparing Coefficients of Different Units: Example

Variable	Mean	Std. Dev.
Salary	\$2,024,616	\$2,764,512
Batting Average	0.267	0.031
Home Runs	12.11	10.31

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

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

- $\hat{\beta_1}$: a 1 standard deviation increase in Batting Average increases Salary by 0.14 standard deviations

$$0.14 \times \$2,764,512=\$387,032$$

- $\hat{\beta_2}$: a 1 standard deviation increase in Home Runs increases Salary by 0.48 standard deviations

$$0.48 \times \$2,764,512=\$1,326,966$$

Standardizing in R

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

```
gapminder<-gapminder %>%
  mutate(std_life = scale(lifeExp),
         std_gdp = scale(gdpPercap))

std_reg<-lm(std_life~std_gdp, data = gapminder)
tidy(std_reg)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.1e-16	0.0197	5.57e-15	1
std_gdp	0.584	0.0197	29.7	3.57e-156

Joint Hypothesis Testing

Joint Hypothesis Testing I

Example: Return again to:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Male_i+\hat{\beta_2}Northeast_i+\hat{\beta_3}M

- 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

Joint Hypothesis Testing II

- A **joint hypothesis** tests against the null hypothesis of a value for *multiple* parameters: $$\mathbf{H_0: \beta_1= \beta_2=0}$$ the hypotheses that multiple regressors are equal to zero (have no causal effect on the outcome)

- Our **alternative hypothesis** is that: $$H_1: \text{ either } \beta_1\neq0\text{ or } \beta_2\neq0\text{ or both}$$ or simply, that \(H_0\) is not true

Types of Joint Hypothesis Tests

- Three main cases of joint hypothesis tests:

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

- Testing against the claim that multiple variables don't matter
- Useful under high multicollinearity between variables
- H_a: at least one parameter $\neq$ 0

2) H_0: $\beta_1=\beta_2$

- Testing whether two variables matter the same
- Variables must be the same units
- $H_a: \beta_1 (\neq, <, \text{ or }>) \beta_2$

3) $H_0:$ ALL β's $=0$

- The "**Overall F-test"**
- Testing against claim that regression model explains *NO* variation in Y

Joint Hypothesis Tests: F-statistic

- The **F-statistic** is the test-statistic 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$\rightarrow$unrestricted model

- F has its own distribution, with *two* sets of degrees of freedom

Joint Hypothesis F-test: Example I

Example: Return again to:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Male_i+\hat{\beta_2}Northeast_i+\hat{\beta_3}M$$

- $H_0: \beta_2=\beta_3=\beta_4=0$
- H_a: H_0 is not true (at least one $\beta_i \neq 0$)

Joint Hypothesis F-test: Example II

Example: Return again to:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Male_i+\hat{\beta_2}Northeast_i+\hat{\beta_3}Midwest_i+\hat{\b

- **Unrestricted model**:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Male_i+\hat{\beta_2}Northeast_i+\hat{\beta_3}Midwest_i+\hat{\beta

- **Restricted model**:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Male_i$$

- \(F\): does going from **restricted** to **unrestricted** model statistically significantly improve \(R^2\)?

Calculating the F-statistic

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

Calculating the F-statistic

$$F_{q,(n-k-1)}=\cfrac{\left(\displaystyle\frac{(\color{#e64173}{R^2_u}-R^2_r)}{q}\right)}{\left(\displaystyle\frac{(1-\color{#e64173}{R^2_u})}{(n-k-1)}\right)}$$

- $\color{#e64173}{R^2_u}$: the R^2 from the **unrestricted model** (all variables)

Calculating the F-statistic

$$F_{q,(n-k-1)}=\cfrac{\left(\displaystyle\frac{(\color{#e64173}{R^2_u}-\color{#6A5ACD}{R^2_r})}{q}\right)}{\left(\displaystyle\frac{(1-\color{#e64173}{R^2_u})}{(n-k-1)}\right)}$$

- $\color{#e64173}{R^2_u}$: the R^2 from the **unrestricted model** (all variables)

- $\color{#6A5ACD}{R^2_r}$: the R^2 from the **restricted model** (null hypothesis)

Calculating the F-statistic

$$F_{q,(n-k-1)}=\cfrac{\left(\displaystyle\frac{(\color{#e64173}{R^2_u}-\color{#6A5ACD}{R^2_r})}{q}\right)}{\left(\displaystyle\frac{(1-\color{#e64173}{R^2_u})}{(n-k-1)}\right)}$$

- $\color{#e64173}{R^2_u}$: the R^2 from the **unrestricted model** (all variables)

- $\color{#6A5ACD}{R^2_r}$: the R^2 from the **restricted model** (null hypothesis)

- q: number of restrictions (number of $\beta's=0$) under null hypothesis

Calculating the F-statistic

$$F_{q,(n-k-1)}=\cfrac{\left(\displaystyle\frac{(\color{#e64173}{R^2_u}-\color{#6A5ACD}{R^2_r})}{q}\right)}{\left(\displaystyle\frac{(1-\color{#e64173}{R^2_u})}{(n-k-1)}\right)}$$

- $\color{#e64173}{R^2_u}$: the R^2 from the **unrestricted model** (all variables)

- $\color{#6A5ACD}{R^2_r}$: the R^2 from the **restricted model** (null hypothesis)

- q: number of restrictions (number of $\beta's=0$) under null hypothesis)

- k: number of X variables in **unrestricted model** (all variables)

Calculating the F-statistic

$$F_{q,(n-k-1)}=\cfrac{\left(\displaystyle\frac{(\color{#e64173}{R^2_u}-\color{#6A5ACD}{R^2_r})}{q}\right)}{\left(\displaystyle\frac{(1-\color{#e64173}{R^2_u})}{(n-k-1)}\right)}$$

- $\color{#e64173}{R^2_u}$: the R^2 from the **unrestricted model** (all variables)
- $\color{#6A5ACD}{R^2_r}$: the R^2 from the **restricted model** (null hypothesis)
- q: number of restrictions (number of $\beta's=0$) under null hypothesis)
- k: number of X variables in **unrestricted model** (all variables)

Calculating the F-statistic II

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

- **Key takeaway**: The bigger the difference between $(R^2_u-R^2_r)$, the greater the improvement in fit by adding variables, the larger the F!

- This formula is (believe it or not) actually a simplified version (assuming homoskedasticity)
    - I give you this formula to **build your intuition of what F is measuring**

F-test Example I

- We'll use the `wooldridge` package's `wage1` data again

```
# load in data from wooldridge package
library(wooldridge)
wages<-wooldridge::wage1

# run regressions
unrestricted_reg<-lm(wage~female+northcen+west+south, data=wages)
restricted_reg<-lm(wage~female, data=wages)
```

F-test Example II

- **Unrestricted model**:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Female_i+\hat{\beta_2}Northeast_i+\hat{\beta_3}N

- **Restricted model**:

$$\widehat{Wage_i}=\hat{\beta_0}+\hat{\beta_1}Female_i$$

- \(H_0: \beta_2 = \beta_3 = \beta_4 =0\)
- \(q = 3\) restrictions (F numerator df)
- \(n-k-1 = 526-4-1=521\) (F denominator df)

F-test Example III

- We can use the `car` package's `linearHypothesis()` command to run an \(F\)-test:
  - first argument: name of the (unrestricted) regression
  - second argument: vector of variable names (in quotes) you are testing

```
# load car package for additional regression tools
library("car")

# F-test
linearHypothesis(unrestricted_reg, c("northcen", "west", "south"))
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
524	6.33e+03				
521	6.17e+03	3	157	4.43	0.00438

Second F-test Example: Are Two Coefficients Equal?

- The second type of test is 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?

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

Second F-test Example: Data

```r
# load in data
heightwages<-read_csv("../data/heightwages.csv")

# make a "heights" variable as the sum of adolescent (height81) and adult (height85) height

heightwages <- heightwages %>%
  mutate(heights=height81+height85)

height_reg<-lm(wage96~height81+height85+male, data=heightwages)
height_restricted_reg<-lm(wage96~heights+male, data=heightwages)
```

Second F-test Example: Data

- For second argument, set two variables equal, in quotes

```
linearHypothesis(height_reg, "height81=height85") # F-test
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
6.59e+03	5.13e+06				
6.59e+03	5.13e+06	1	959	1.23	0.267

- Insufficient evidence to reject \(H_0\)!

- The effect of adolescent and adult height on wages is the same

All F-test I

```
summary(unrestricted_reg)
```

```
##
## Call:
## lm(formula = wage ~ female + northcen + west + south, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3269 -2.0105 -0.7871  1.1898 17.4146
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.5654     0.3466  21.827   <2e-16 ***
## female       -2.5652     0.3011  -8.520   <2e-16 ***
## northcen     -0.5918     0.4362  -1.357   0.1755
## west          0.4315     0.4838   0.892   0.3729
## south        -1.0262     0.4048  -2.535   0.0115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.443 on 521 degrees of freedom
## Multiple R-squared:  0.1376,    Adjusted R-squared:  0.131
## F-statistic: 20.79 on 4 and 521 DF,  p-value: 6.501e-16
```

- Last line of regression output from `summary()` is an **All F-test**
  - \(H_0:\) all \(\beta's=0\)
  - the regression explains no variation in \(Y\)
  - Calculates an `F-statistic` that, if high enough, is significant (`p-value` \(<0.05)\) enough to reject \(H_0\)

All F-test II

- Alternatively, if you use `broom` instead of `summary()`:
  - `glance()` command makes table of regression summary statistics
  - `tidy()` only shows coefficients

```
library(broom)
glance(unrestricted_reg)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.138	0.131	3.44	20.8	6.5e-16	4	-1.39e+03	2.8e+03	2.83e+03	6.17e+03	521	526

- "`statistic`" is the All F-test, "`p.value`" next to it is the p value from the F test