Lecture 21: More Regression
API-201Z

Maya Sen

Harvard Kennedy School
http://scholar.harvard.edu/msen
Announcements

▶ Midterms graded, will circulate email as soon as they are available for pick up (along with info on distribution)

▶ Final Exercise summaries (1 paragraph) due 11/29, with Problem Set 10, will send email with poll link as soon as we can afterwards

▶ Vote for your top 3 choices by Sunday 12/2 midnight, will notify 3 “winning” teams on Monday

▶ Shiro will host 2 sections this week (Wexner 332): Today (11/27) 3-4pm, Friday (11/30) 11:45-1pm
Announcements

- Midterms graded, will circulate email as soon as they are available for pick up (along with info on distribution)
Announcements

▶ Midterms graded, will circulate email as soon as they are available for pick up (along with info on distribution)
▶ Final Exercise summaries (1 paragraph) due 11/29, with Problem Set 10, will send email with poll link as soon as we can afterwards

▶ Vote for your top 3 choices by Sunday 12/2 midnight, will notify 3 “winning” teams on Monday
▶ Shiro will host 2 sections this week (Wexner 332): Today (11/27) 3-4pm, Friday (11/30) 11:45-1pm
Announcements

- Midterms graded, will circulate email as soon as they are available for pick up (along with info on distribution)
- Final Exercise summaries (1 paragraph) due 11/29, with Problem Set 10, will send email with poll link as soon as we can afterwards
- Vote for your top 3 choices by Sunday 12/2 midnight, will notify 3 “winning” teams on Monday
Announcements

- Midterms graded, will circulate email as soon as they are available for pick up (along with info on distribution)
- Final Exercise summaries (1 paragraph) due 11/29, with Problem Set 10, will send email with poll link as soon as we can afterwards
- Vote for your top 3 choices by Sunday 12/2 midnight, will notify 3 “winning” teams on Monday
- Shiro will host 2 sections this week (Wexner 332): Today (11/27) 3-4pm, Friday (11/30) 11:45-1pm
Simple Linear Regression

Regression with an outcome and one explanatory variable

Example from last time:

What is relationship between state-level unemployment rates in U.S. in 1995 and in 2000?

For SRS of 30 states, data was collected on:

- Unemployment rate in 1995
- Unemployment rate in 2000
Simple Linear Regression

▶ Regression with an outcome and one explanatory variable

Example from last time:
What is relationship between state-level unemployment rates in U.S. in 1995 and in 2000?
For SRS of 30 states, data was collected on:
- Unemployment rate in 1995
- Unemployment rate in 2000
Simple Linear Regression

- Regression with an outcome and one explanatory variable
- Example from last time:
Simple Linear Regression

- Regression with an outcome and one explanatory variable
- Example from last time:
- What is relationship between state-level unemployment rates in U.S. in 1995 and in 2000?
Simple Linear Regression

- Regression with an outcome and one explanatory variable
- Example from last time:
  - What is relationship between state-level unemployment rates in U.S. in 1995 and in 2000?
  - For SRS of 30 states, data was collected on:
Simple Linear Regression

- Regression with an outcome and one explanatory variable
- Example from last time:
  - What is relationship between state-level unemployment rates in U.S. in 1995 and in 2000?
  - For SRS of 30 states, data was collected on:
    - Unemployment rate in 1995
Simple Linear Regression

- Regression with an outcome and one explanatory variable
- Example from last time:
  - What is relationship between state-level unemployment rates in U.S. in 1995 and in 2000?
- For SRS of 30 states, data was collected on:
  - Unemployment rate in 1995
  - Unemployment rate in 2000
State Unemployment Example
State Unemployment Example

State-Level Unemployment Rates in 1995 vs 2000

Unemployment Rate 1995
Simple Linear Regression

We assume a "true" relationship in the population between $x$ and $y$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

However: We can never observe $\beta_0$ and $\beta_1$ → these are population parameters!

Best thing we can do is estimate them using our data

Thus, we have an estimated linear relationship:

$$y_i = b_0 + b_1 x_i + e_i$$

Or using "hat" notation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$

Residuals ($e_i$) represent estimates of the random errors, $\epsilon_1$
Simple Linear Regression

- We assume a “true” relationship in the population between \( x \) and \( y \):

\[
y_i = \beta_0 + \beta_1 x_i + \epsilon_i
\]

- However: We can never observe \( \beta_0 \) and \( \beta_1 \) → these are population parameters!

- Best thing we can do is estimate them using our data

- Thus, we have an estimated linear relationship:

\[
y_i = b_0 + b_1 x_i + e_i
\]

- Or using “hat” notation

\[
\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i
\]

- Residuals (\( e_i \)) represent estimates of the random errors, \( \epsilon_1 \)
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- However: We can never observe $\beta_0$ and $\beta_1 \rightarrow$ these are population parameters!
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:
  \[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

- However: We can never observe $\beta_0$ and $\beta_1 \rightarrow$ these are population parameters!
- Best thing we can do is estimate them using our data

Residuals ($\epsilon_i$) represent estimates of the random errors.
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- However: We can never observe $\beta_0$ and $\beta_1 \rightarrow$ these are population parameters!
- Best thing we can do is estimate them using our data
- Thus, we have an estimated linear relationship:
Simple Linear Regression

⚠️ We assume a “true” relationship in the population between $x$ and $y$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

⚠️ However: We can never observe $\beta_0$ and $\beta_1$ $\rightarrow$ these are population parameters!

⚠️ Best thing we can do is estimate them using our data

⚠️ Thus, we have an estimated linear relationship:

$$y_i = b_0 + b_1 x_i + e_i$$
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

- However: We can never observe $\beta_0$ and $\beta_1 \rightarrow$ these are population parameters!
- Best thing we can do is estimate them using our data
- Thus, we have an estimated linear relationship:

\[ y_i = b_0 + b_1 x_i + e_i \]

- Or using “hat” notation
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- However: We can never observe $\beta_0$ and $\beta_1$ → these are population parameters!

- Best thing we can do is estimate them using our data

- Thus, we have an estimated linear relationship:

$$y_i = b_0 + b_1 x_i + e_i$$

- Or using “hat” notation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$$
Simple Linear Regression

- We assume a “true” relationship in the population between $x$ and $y$:

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

- However: We can never observe $\beta_0$ and $\beta_1 \rightarrow$ these are population parameters!

- Best thing we can do is estimate them using our data

- Thus, we have an estimated linear relationship:

\[ y_i = b_0 + b_1 x_i + e_i \]

- Or using “hat” notation

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i \]

- Residuals ($e_i$) represent estimates of the random errors, $\epsilon_1$
How to find the best estimated line?

Going back to our data:
How to find the best estimated line?

Going back to our data:
How to find the best estimated line?

Going back to our data:

We’ll take the line that minimizes the sum of squared residuals
How to find the best estimated line?

- Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Or:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Gives Ordinary Least Squares Estimators

- For slope $b_1$:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

- For intercept $b_0$:

$$b_0 = \bar{y} - b_1 \bar{x}$$
How to find the best estimated line?

▶ Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

▶ Or:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

▶ Gives Ordinary Least Squares Estimators

▶ For slope $b_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

▶ For intercept $b_0 = \bar{y} - b_1 \bar{x}$
How to find the best estimated line?

- Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Or:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Gives Ordinary Least Squares Estimators
How to find the best estimated line?

- Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:
  \[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

- Or:
  \[ \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \]

- Gives Ordinary Least Squares Estimators

- For slope
How to find the best estimated line?

- Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:
  $$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Or:
  $$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Gives Ordinary Least Squares Estimators
- For slope
  $$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$
How to find the best estimated line?

► Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

► Or:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

► Gives **Ordinary Least Squares Estimators**

► For slope:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

► For intercept:


How to find the best estimated line?

- Choose values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:
  \[
  \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
  \]

- Or:
  \[
  \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2
  \]

- Gives Ordinary Least Squares Estimators

- For slope
  \[
  b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
  \]

- For intercept:
  \[
  b_0 = \bar{y} - b_1 \bar{x}
  \]
State Unemployment Example

In Stata:

```
. regress yr2000 yr1995
-----------------------------------------------------
yr2000 | Coef. Std. Err. t P>|t|
-----------------------------------------------------
yr1995 | .5398317 .0818083 6.60 0.000
_cons | 1.077917 .4571589 2.36 0.026
-----------------------------------------------------

▶ Intercept Coefficient Estimate (b₀ or ˆβ₀): 1.077917
▶ Slope Coefficient Estimate (b₁ or ˆβ₁): 0.5398317
▶ Gives estimated regression line of: ˆy = 1.08 + 0.54x
▶ Can use this for prediction (but probably not outside of the support of the data)
```
State Unemployment Example

In Stata:

. regress yr2000 yr1995

-----------------------------------------------------
|       | Coef.    | Std. Err. | t     | P>|t| |
-----------------------------------------------------
| yr2000 | 0.53983  | 0.081808  | 6.60  | 0.000 |
| yr1995 |          |           |       |      |
| _cons  | 1.07792  | 0.457158  | 2.36  | 0.026 |
|        |          |           |       |      |
-----------------------------------------------------

Intercept Coefficient Estimate (b₀ or ˆβ₀): 1.077917
Slope Coefficient Estimate (b₁ or ˆβ₁): 0.5398317
Gives estimated regression line of: ˆy = 1.08 + 0.54x
Can use this for prediction (but probably not outside of the support of the data)
State Unemployment Example

In Stata:

```
. regress yr2000 yr1995
```

| yr2000 | Coef.     | Std. Err. | t     | P>|t| |
|--------|-----------|-----------|-------|-----|---|
| yr1995 | .5398317  | .0818083  | 6.60  | 0.000 |
| _cons  | 1.077917  | .4571589  | 2.36  | 0.026 |

- Intercept Coefficient Estimate ($b_0$ or $\hat{\beta}_0$): 1.077917

- Slope Coefficient Estimate ($b_1$ or $\hat{\beta}_1$): 0.5398317

Gives estimated regression line of $\hat{y} = 1.07 + 0.54x$

Can use this for prediction (but probably not outside of the support of the data)
State Unemployment Example

In Stata:

```
.regress yr2000 yr1995
```

|          | Coef.   | Std. Err. | t     | P>|t| |
|----------|---------|-----------|-------|-----|
| yr2000   | 1.077917| 0.4571589 | 2.36  | 0.026|
| yr1995   | 0.5398317| 0.0818083 | 6.60  | 0.000|

▶ Intercept Coefficient Estimate ($b_0$ or $\hat{\beta}_0$): 1.077917
▶ Slope Coefficient Estimate ($b_1$ or $\hat{\beta}_1$): 0.5398317
State Unemployment Example

In Stata:

```
. regress yr2000 yr1995
```

```
|          | Coef.  | Std. Err. | t     | P>|t| |
|----------|--------|-----------|-------|------|
| yr2000   | .5398317 | .0818083  | 6.60  | 0.000 |
| yr1995   | 1.077917 | .4571589  | 2.36  | 0.026 |
```

- Intercept Coefficient Estimate ($b_0$ or $\hat{\beta}_0$): 1.077917
- Slope Coefficient Estimate ($b_1$ or $\hat{\beta}_1$): 0.5398317
- Gives estimated regression line of: $\hat{y} = 1.08 + 0.54x$
State Unemployment Example

In Stata:

```
.regress yr2000 yr1995
```

| yr2000 | Coef.     | Std. Err. |   t   | P>|t| |
|--------|-----------|-----------|-------|------|
| yr1995 | 0.5398317 | 0.0818083 | 6.60  | 0.000|
| _cons  | 1.077917  | 0.4571589 | 2.36  | 0.026|

- Intercept Coefficient Estimate \((b_0 \text{ or } \hat{\beta}_0)\): 1.077917
- Slope Coefficient Estimate \((b_1 \text{ or } \hat{\beta}_1)\): 0.5398317
- Gives estimated regression line of: \(\hat{y} = 1.08 + 0.54x\)
- Can use this for prediction (but probably not outside of the support of the data)
Using Regression for Hypothesis Tests

\[
\begin{align*}
\hat{b}_1 &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\
\hat{b}_0 &= \bar{y} - \hat{b}_1 \bar{x}
\end{align*}
\]

are sums and means of random variables

Estimates would differ with different samples

Means that CLT kicks in

\[b_1 \text{ and } b_0 \text{ are normally distributed, with sufficiently large samples}\]
Using Regression for Hypothesis Tests

- OLS estimators for slope

\[
\begin{align*}
\hat{b}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\
\hat{b}_0 &= \bar{y} - \hat{b}_1 \bar{x}
\end{align*}
\]

- Means that CLT kicks in
- \(\hat{b}_1\) and \(\hat{b}_0\) are normally distributed, with sufficiently large samples

- Estimates would differ with different samples
Using Regression for Hypothesis Tests

- OLS estimators for slope

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept

\[ b_0 = \bar{y} - b_1 \bar{x} \]

- are sums and means of random variables

- Estimates would differ with different samples

- Means that CLT kicks in

- \( b_1 \) and \( b_0 \) are normally distributed, with sufficiently large samples
Using Regression for Hypothesis Tests

- OLS estimators for slope

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept

\[ b_0 = \bar{y} - b_1 \bar{x} \]
Using Regression for Hypothesis Tests

- OLS estimators for slope

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept

\[ b_0 = \bar{y} - b_1 \bar{x} \]
Using Regression for Hypothesis Tests

- OLS estimators for slope
  \[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept
  \[ b_0 = \bar{y} - b_1 \bar{x} \]

- are sums and means of random variables

Estimates would differ with different samples

Means that CLT kicks in

\[ b_1 \text{ and } b_0 \text{ are normally distributed, with sufficiently large samples} \]
Using Regression for Hypothesis Tests

- OLS estimators for slope

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept

\[ b_0 = \bar{y} - b_1 \bar{x} \]

- are sums and means of random variables

- Estimates would differ with different samples
Using Regression for Hypothesis Tests

- OLS estimators for slope
  \[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept
  \[ b_0 = \bar{y} - b_1 \bar{x} \]

- are sums and means of random variables
- Estimates would differ with different samples
- Means that CLT kicks in
Using Regression for Hypothesis Tests

- OLS estimators for slope
  
  \[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]

- and intercept
  
  \[ b_0 = \bar{y} - b_1 \bar{x} \]

- are sums and means of random variables
- Estimates would differ with different samples
- Means that CLT kicks in
- \( b_1 \) and \( b_0 \) are normally distributed, with sufficiently large samples
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
  - Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
  - Null and alternative hypotheses (nearly always two tailed)
    - $H_0$: $\beta_1 = 0$
    - $H_a$: $\beta_1 \neq 0$

```
regress yr2000 yr1995
```

|        | Coef. | Std. Err. | t    | P>|t| |
|--------|-------|-----------|------|-----|
| yr1995 | .5398317 | .0818083 | 6.60 | 0.000 |
| _cons  | 1.077917 | .4571589 | 2.36 | 0.026 |

- For $\beta_1$, hypothesis test yields $p$-value of 0.000
- For $\beta_0$ is testing null hypothesis that intercept equal to zero → mean of $y$ is zero when mean of $x$ is zero
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests

```
regress yr2000 yr1995

|        | Coef. | Std. Err. | t    | P>|t| |
|--------|-------|-----------|------|-----|
| yr1995 | .5398 | .0818     | 6.60 | 0.00|
| _cons  | 1.08  | .46       | 2.36 | 0.03|
```

- For $\beta_1$, hypothesis test yields p-value of 0.000
- For $\beta_0$ is testing null hypothesis that intercept equal to zero
  → mean of y is zero when mean of x is zero
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables

```
regress yr2000 yr1995

|            | Coef. | Std. Err. | t     | P>|t| |
|------------|-------|-----------|-------|-----|
| yr1995     | 0.5398317 | 0.0818083  | 6.60  | 0.000 |
| _cons      | 1.077917  | 0.4571589  | 2.36  | 0.026 |
```

- For $\beta_1$, hypothesis test yields $p$-value of 0.000
- For $\beta_0$ is testing null hypothesis that intercept equal to zero → mean of $y$ is zero when mean of $x$ is zero
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
- Null and alternative hypotheses (nearly always two tailed)
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
- Null and alternative hypotheses (nearly always two tailed)
  - $H_0$: $\beta_1 = 0$
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
- Null and alternative hypotheses (nearly always two tailed)
  - $H_0$: $\beta_1 = 0$
  - $H_a$: $\beta_1 \neq 0$
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
- Null and alternative hypotheses (nearly always two tailed)
  - $H_0$: $\beta_1 = 0$
  - $H_a$: $\beta_1 \neq 0$

```
. regress yr2000 yr1995
```

```
| yr2000 | Coef.  | Std. Err. | t     | P>|t| |
|---------|--------|-----------|-------|------|
| yr1995 | .5398317 | .0818083 | 6.60  | 0.000|
| _cons  | 1.077917 | .4571589 | 2.36  | 0.026|
```

- For $\beta_1$, hypothesis test yields $p$-value of 0.000
- For $\beta_0$ is testing null hypothesis that intercept equal to zero
  - $\text{mean of } y \text{ is zero when mean of } x \text{ is zero}$
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
- Null and alternative hypotheses (nearly always two tailed)
  - $H_0$: $\beta_1 = 0$
  - $H_a$: $\beta_1 \neq 0$

```
. regress yr2000 yr1995
```

```
| yr2000 | Coef.    | Std. Err. | t     | P>|t| |
|--------|----------|-----------|-------|-------|
| yr1995 | .5398317 | .0818083  | 6.60  | 0.000 |
| _cons  | 1.077917 | .4571589  | 2.36  | 0.026 |
```

- For $\beta_1$, hypothesis test yields $p$-value of 0.000
Using Regression for Hypothesis Tests

- Can use this fact to conduct hypothesis tests
- Specifically: If our slope $\beta_1$ is zero, then no linear relationship between the two variables
- Null and alternative hypotheses (nearly always two tailed)
  - $H_0$: $\beta_1 = 0$
  - $H_a$: $\beta_1 \neq 0$

```
.regress yr2000 yr1995
```

|        | Coef.  | Std. Err. | t       | P>|t| |
|--------|--------|-----------|---------|------|
| yr2000 | 1.077917 | .4571589  | 2.36    | 0.026|
| yr1995 | .5398317 | .0818083  | 6.60    | 0.000|

- For $\beta_1$, hypothesis test yields $p$-value of 0.000
- For $\beta_0$ is testing null hypothesis that intercept equal to zero → mean of $y$ is zero when mean of $x$ is zero
Using Regression for Confidence Intervals of Slope

Just as we can conduct hypothesis tests, we can also construct confidence intervals for the true slope $\beta_1$.

The formula for this confidence interval is:

$$ b_1 \pm t_{n-2} \left( \frac{\alpha}{2} \right) \times SE[b_1] $$

In our example:

$$ 0.5398 \pm t_{28, \frac{\alpha}{2}} \times 0.0818 \rightarrow (0.372, 0.707) $$

Interpretation: In repeated sampling, we expect 95 out of 100 confidence intervals to contain the true slope.
Using Regression for Confidence Intervals of Slope

- Just as we can conduct hypothesis tests, can also construct confidence intervals for true slope $\beta_1$

- Follows the same formula as before:

$$b_1 \pm t_{n-2} \left(\frac{\alpha}{2}\right) \times SE[b_1]$$

- In our example:

$$0.5398 \pm t_{28, \frac{\alpha}{2}} \times 0.0818 \rightarrow (0.372, 0.707)$$

- Interpretation: In repeated sampling, expect 95 out of 100 confidence intervals to contain true slope $\beta_1$. 
Using Regression for Confidence Intervals of Slope

- Just as we can conduct hypothesis tests, can also construct confidence intervals for true slope $\beta_1$
- Follows the same formula as before:

$$b_1 \pm t_{n-2} \left( \frac{\alpha}{2} \right) \times SE[b_1]$$

- In our example:
  $$0.5398 \pm t_{28, \alpha/2} \times 0.0818 \rightarrow (0.372, 0.707)$$
- Interpretation: In repeated sampling, expect 95 out of 100 confidence intervals to contain true slope
Using Regression for Confidence Intervals of Slope

- Just as we can conduct hypothesis tests, can also construct confidence intervals for true slope $\beta_1$
- Follows the same formula as before:

$$ b_1 \pm t_{n-2}(\alpha/2) \times SE[b_1] $$

In our example:

$$ 0.5398 \pm t_{28}(0.025) \times 0.0818 \rightarrow (0.372, 0.707) $$

Interpretation: In repeated sampling, expect 95 out of 100 confidence intervals to contain true slope
Using Regression for Confidence Intervals of Slope

- Just as we can conduct hypothesis tests, can also construct confidence intervals for true slope $\beta_1$
- Follows the same formula as before:

$$b_1 \pm t_{n-2}(\alpha/2) \times SE[b_1]$$

- In our example:
Using Regression for Confidence Intervals of Slope

- Just as we can conduct hypothesis tests, can also construct confidence intervals for true slope $\beta_1$
- Follows the same formula as before:

$$b_1 \pm t_{n-2}(\alpha/2) \times SE[b_1]$$

- In our example:

$$0.5398 \pm t_{28, \alpha/2} \times 0.0818 \rightarrow (0.372, 0.707)$$
Using Regression for Confidence Intervals of Slope

- Just as we can conduct hypothesis tests, can also construct confidence intervals for true slope $\beta_1$
- Follows the same formula as before:

$$ b_1 \pm t_{n-2}(\alpha/2) \times SE[b_1] $$

- In our example:

$$ 0.5398 \pm t_{28, \alpha/2} \times 0.0818 $$

$$ \rightarrow (0.372, 0.707) $$

- Interpretation: In repeated sampling, expect 95 out of 100 confidence intervals to contain true slope
State Unemployment Example

R and Stata will also report 95% CIs:

```
regress yr2000 yr1995
Source | SS df MS Number of obs = 30
-------------+------------------------------ F( 1, 28) = 43.54
Model | 13.3338426 1 13.3338426 Prob > F = 0.0000
Residual | 8.57415592 28 .306219854 R-squared = 0.6086
-------------+------------------------------ Adj R-squared = 0.5947
Total | 21.9079986 29 .755448226 Root MSE = .55337
-------------+------------------------------
yr2000 | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
yr1995 | .5398317 .0818083 6.60 0.000 .372255 .7074084
_cons | 1.077917 .4571589 2.36 0.026 .1414697 2.014365
-------------+----------------------------------------------------------------
```
**State Unemployment Example**

R and Stata will also report 95% CIs:

```
. regress yr2000 yr1995
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13.3338426</td>
<td>1</td>
<td>13.3338426</td>
<td>F( 1, 28) = 43.54</td>
</tr>
<tr>
<td>Residual</td>
<td>8.57415592</td>
<td>28</td>
<td>.306219854</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>21.9079986</td>
<td>29</td>
<td>.755448226</td>
<td>R-squared = 0.6086</td>
</tr>
</tbody>
</table>

|            | Coef.    | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|------------|----------|-----------|-------|---------|---------------------|
| yr2000     | .5398317 | .0818083  | 6.60  | 0.000   | .372255 .7074084    |
| yr1995     | .077917  | .4571589  | 2.36  | 0.026   | .1414697 2.014365   |
| _cons      | 1.077917 | .4571589  | 2.36  | 0.026   | .1414697 2.014365   |
Model Fit of a Simple Linear Regression

Model fit is a measure of how "well" the line fits the data.

In linear regression, $R^2$, the most commonly used measure,

is defined as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

→ $\sum (y_i - \hat{y}_i)^2$ is variation not explained by the model (residual sum of squares, or $SS_{res}$).

→ $\sum (y_i - \bar{y})^2$ is total variation (total sum of squares, or $SS_{tot}$).

$R^2$: Proportion of variance in $y$ explained by the linear model (conditional on $x$).
Model Fit of a Simple Linear Regression

- Model fit is a measure of how “well” the line fits the data.

\[ R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \]

- The numerator, \( \sum (y_i - \hat{y}_i)^2 \), is the variation not explained by the model (residual sum of squares, or \( SS_{res} \)).

- The denominator, \( \sum (y_i - \bar{y})^2 \), is the total variation (total sum of squares, or \( SS_{tot} \)).

- \( R^2 \) represents the proportion of variance in \( y \) explained by the linear model (conditional on \( x \)).
Model Fit of a Simple Linear Regression

- Model fit is a measure of how “well” the line fits the data
- In linear regression, $R^2$ most commonly used measure

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- $\sum (y_i - \hat{y}_i)^2$ is the variation not explained by the model (residual sum of squares, or $SS_{res}$)
- $\sum (y_i - \bar{y})^2$ is the total variation (total sum of squares, or $SS_{tot}$)

$R^2$: Proportion of variance in $y$ explained by the linear model (conditional on $x$)
Model Fit of a Simple Linear Regression

- Model fit is a measure of how “well” the line fits the data
- In linear regression, $R^2$ most commonly used measure
- $R^2$ defined as
  \[
  1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}
  \]
Model Fit of a Simple Linear Regression

- Model fit is a measure of how “well” the line fits the data.
- In linear regression, $R^2$ most commonly used measure.
- $R^2$ defined as:

$$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- $\sum (y_i - \hat{y}_i)^2$ is variation not explained by the model (residual sum of squares, or $SS_{res}$)
Model Fit of a Simple Linear Regression

- Model fit is a measure of how “well” the line fits the data.
- In linear regression, $R^2$ is the most commonly used measure.
- $R^2$ is defined as:

$$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- $\sum (y_i - \hat{y}_i)^2$ is variation not explained by the model (residual sum of squares, or $SS_{res}$).
- $\sum (y_i - \bar{y})^2$ is total variation (total sum of squares, or $SS_{tot}$).

$R^2$: Proportion of variance in $y$ explained by the linear model (conditional on $x$).
Model Fit of a Simple Linear Regression

- Model fit is a measure of how “well” the line fits the data
- In linear regression, $R^2$ most commonly used measure
- $R^2$ defined as

$$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

- $\sum (y_i - \hat{y}_i)^2$ is variation not explained by the model (residual sum of squares, or $SS_{res}$)
- $\sum (y_i - \bar{y})^2$ is total variation (total sum of squares, or $SS_{tot}$)
- $R^2$: Proportion of variance in $y$ explained by the linear model (conditional on $x$)
Model Fit of a Simple Linear Regression

Can get sum of squares from Stata or R output:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>----</td>
<td>--------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Model</td>
<td>13.33</td>
<td>1</td>
<td>13.33</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>8.57</td>
<td>28</td>
<td>0.31</td>
<td>R-squared = 0.6086</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.5947</td>
</tr>
<tr>
<td>Total</td>
<td>21.90</td>
<td>29</td>
<td>0.76</td>
<td>Root MSE = 0.5534</td>
</tr>
</tbody>
</table>

So:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{8.57}{21.90} = 0.6086$$

Note: With one explanatory variable (one x), correlation coefficient r is square root of R^2, so

$$r = \sqrt{R^2}$$

Here

$$\sqrt{0.6086} = 0.780$$

Substantive interpretation: High R^2 → model explaining a lot of variation in outcome
Model Fit of a Simple Linear Regression

- Can get sum of squares from Stata or R output:

```
Source | SS   df  MS      Number of obs = 30
-------+---------------------------------
Model  | 13.33 1 13.33  Prob > F = 0.0000
Residual | 8.57 28 .31  R-squared = 0.6086
-------+---------------------------------
Total  | 21.90 29 .76  Root MSE = 0.55
-------+---------------------------------```

- So:

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{8.57}{21.90} = 0.6086
\]

- Note: With one explanatory variable (one x), correlation coefficient \( r \) is square root of \( R^2 \), so

\[
r = \sqrt{R^2}
\]

- Here \( \sqrt{0.6086} = 0.780 \)

- Substantive interpretation: High \( R^2 \) → model explaining a lot of variation in outcome
Model Fit of a Simple Linear Regression

- Can get sum of squares from Stata or R output:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13.33</td>
<td>1</td>
<td>13.33</td>
</tr>
<tr>
<td>Residual</td>
<td>8.57</td>
<td>28</td>
<td>0.31</td>
</tr>
<tr>
<td>Total</td>
<td>21.90</td>
<td>29</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Number of obs = 30
F( 1, 28) = 43.54
Prob > F = 0.0000
R-squared = 0.6086
Adj R-squared = 0.5947
Root MSE = 0.5537

So:

R^2 = 1 - \frac{SS_{res}}{SS_{tot}}

= 1 - \frac{8.57}{21.90}

= 0.6086

Note: With one explanatory variable (one x), correlation coefficient \( r \) is square root of \( R^2 \), so \( r = \sqrt{R^2} \)

Here \( \sqrt{0.6086} = 0.780 \)

Substantive interpretation: High \( R^2 \) → model explaining a lot of variation in outcome
Model Fit of a Simple Linear Regression

- Can get sum of squares from Stata or R output:

```
Source | SS    df  MS
-------------+------------------
Model       | 13.3338426       1  13.3338426
Residual   | 8.57415992       28  .306219854
-------------+------------------
Total       | 21.9079986       29  .755448226
```

- So:

\[ R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{8.57}{21.90} = 0.6086 \]

- Note: With one explanatory variable (one \( x \)), correlation coefficient \( r \) is square root of \( R^2 \), so \( r = \sqrt{R^2} \)

\[ \sqrt{0.6086} = 0.780 \]

- Substantive interpretation: High \( R^2 \)→ model explaining a lot of variation in outcome
Model Fit of a Simple Linear Regression

- Can get sum of squares from Stata or R output:

```
Source | SS     df   MS
---------------+--------+---------
Model         13.3338426 1 13.3338426
Residual     8.57415592 28  .306219854
---------------+--------+---------
Total        21.9079986 29  .755448226
```

- So:

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \\
= 1 - \frac{8.57}{21.90} \\
= 0.6086
\]
Model Fit of a Simple Linear Regression

- Can get sum of squares from Stata or R output:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13.3338426</td>
<td>1</td>
<td>13.3338426</td>
</tr>
<tr>
<td>Residual</td>
<td>8.57415592</td>
<td>28</td>
<td>.306219854</td>
</tr>
<tr>
<td>Total</td>
<td>21.9079986</td>
<td>29</td>
<td>.755448226</td>
</tr>
</tbody>
</table>

- So:

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{8.57}{21.90} = 0.6086
\]

- Note: With one explanatory variable (one \( x \)), correlation coefficient \( r \) is square root of \( R^2 \), so \( r = \sqrt{R^2} \)
Model Fit of a Simple Linear Regression

Can get sum of squares from Stata or R output:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13.3338426</td>
<td>1</td>
<td>13.3338426</td>
<td>F( 1, 28) = 43.54</td>
</tr>
<tr>
<td>Residual</td>
<td>8.57415592</td>
<td>28</td>
<td>.306219854</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>21.9079986</td>
<td>29</td>
<td>.755448226</td>
<td>R-squared = 0.6086</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.5947</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = .55337</td>
</tr>
</tbody>
</table>

So:

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}}
\]

\[
= 1 - \frac{8.57}{21.90}
\]

\[
= 0.6086
\]

Note: With one explanatory variable (one x), correlation coefficient \( r \) is square root of \( R^2 \), so \( r = \sqrt{R^2} \)

Here \( \sqrt{0.6086} = 0.780 \)
Model Fit of a Simple Linear Regression

▶ Can get sum of squares from Stata or R output:

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13.3338426</td>
<td>1</td>
<td>13.3338426</td>
<td>F( 1, 28) = 43.54</td>
</tr>
<tr>
<td>Residual</td>
<td>8.57415592</td>
<td>28</td>
<td>.306219854</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>21.9079986</td>
<td>29</td>
<td>.755448226</td>
<td>R-squared = 0.6086</td>
</tr>
</tbody>
</table>

▶ So:

\[ R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \]

\[ = 1 - \frac{8.57}{21.90} \]

\[ = 0.6086 \]

▶ Note: With one explanatory variable (one x), correlation coefficient \( r \) is square root of \( R^2 \), so \( r = \sqrt{R^2} \)

▶ Here \( \sqrt{0.6086} = 0.780 \)

▶ Substantive interpretation: High \( R^2 \) → model explaining a lot of variation in outcome
Some Notes About Residuals

Residuals represent estimates of the random errors, \( \epsilon \). Empirically: "Left-over" distance from each observation to the regression line after fitting (e.g., after drawing it).

\[ \text{Residual} = \text{Observed } y - \text{Predicted } y \]
Some Notes About Residuals

- **Residuals** represent estimates of the random errors, $\epsilon_i$.
Some Notes About Residuals

- **Residuals** represent estimates of the random errors, $\epsilon_i$
- Empirically: “Left-over” distance from each observation to regression line after fitting (e.g., after drawing it)
Some Notes About Residuals

- **Residuals** represent estimates of the random errors, $\epsilon_i$
- Empirically: “Left-over” distance from each observation to regression line after fitting (e.g., after drawing it)

\[
\text{Residual} = \text{Observed } y - \text{Predicted } y
\]
Some Notes About Residuals

- **Residuals** represent estimates of the random errors, $\epsilon_i$.
- **Empirically:** “Left-over” distance from each observation to regression line after fitting (e.g., after drawing it)

\[
\text{Residual} = \text{Observed } y - \text{Predicted } y
\]
Some Notes About Residuals

For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i | x_i)$ to be constant

Called homoscedasticity

Implies residual values roughly constant across values of $x$

However, if $\text{Var}(\epsilon_i | x_i) = f(x_i)$

Then have heteroscedasticity

Residuals appear to increase/decrease with $x$

Suggests:

Might have non-linear relationship or other problems

May have mismeasurement in data (larger $x$ more likely to have been measured with error)

OLS might not be best estimator (and could have biased SEs)

Will discuss how to address in API 202
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant

- Called homoscedasticity

- Implies residual values roughly constant across values of $x$

- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$

- Then have heteroscedasticity

- Residuals appear to increase/decrease with $x$

- Suggests:

  - Might have non-linear relationship or other problems
  - May have mismeasurement in data (larger $x$ more likely to have been measured with error)
  - OLS might not be best estimator (and could have biased SEs)

- Will discuss how to address in API 202
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called homoscedasticity
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called **homoscedasticity**
  - Implies residual values roughly constant across values of $x$

- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have heteroscedasticity
  - Residuals appear to increase/decrease with $x$
  - Suggests:
    - Might have non-linear relationship or other problems
    - May have mismeasurement in data (larger $x$ more likely to have been measured with error)
  - OLS might not be best estimator (and could have biased SEs)

- Will discuss how to address in API 202
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called homoscedasticity
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called **homoscedasticity**
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have **heteroscedasticity**
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called homoscedasticity
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have heteroscedasticity
  - Residuals appear to increase/decrease with $x$
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called homoscedasticity
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have heteroscedasticity
  - Residuals appear to increase/decrease with $x$
  - Suggests:
  - Might have non-linear relationship or other problems
  - May have mismeasurement in data (larger $x$ more likely to have been measured with error)
  - OLS might not be best estimator (and could have biased SEs)
  - Will discuss how to address in API 202
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called homoscedasticity
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have heteroscedasticity
  - Residuals appear to increase/decrease with $x$
  - Suggests:
    - Might have non-linear relationship or other problems
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called **heteroscedasticity**
  - Implies residual values roughly constant across values of $x$

- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have **heteroscedasticity**
  - Residuals appear to increase/decrease with $x$
  - Suggests:
    - Might have non-linear relationship or other problems
    - May have mismeasurement in data (larger $x$ more likely to have been measured with error)
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called *homo*scedasticity
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have *hetero*scedasticity
  - Residuals appear to increase/decrease with $x$
  - Suggests:
    - Might have non-linear relationship or other problems
    - May have mismeasurement in data (larger $x$ more likely to have been measured with error)
    - OLS might not be best estimator (and could have biased SEs)
Some Notes About Residuals

- For OLS to be a good (BLUE) estimator, want $\text{Var}(\epsilon_i|x_i)$ to be constant
  - Called homoscedasticity
  - Implies residual values roughly constant across values of $x$
- However, if $\text{Var}(\epsilon_i|x_i) = f(x)$
  - Then have heteroscedasticity
  - Residuals appear to increase/decreas with $x$
  - Suggests:
    - Might have non-linear relationship or other problems
    - May have mismeasurement in data (larger $x$ more likely to have been measured with error)
    - OLS might not be best estimator (and could have biased SEs)
- Will discuss how to address in API 202
Some Notes About Residuals

Can see from visualizing data:

- Left hand side: Residuals get bigger with larger $x$ values $\rightarrow$ Suggests heteroscedasticity
- Right hand side: Fairly constant size of residuals $\rightarrow$ Suggests homoscedasticity
Some Notes About Residuals

Can see from visualizing data:

- **Heteroskedastic**
  - Residuals get bigger with larger $x$ values
  - Suggests heteroscedasticity

- **Homoskedastic**
  - Fairly constant size of residuals
  - Suggests homoscedasticity
Some Notes About Residuals

Can see from visualizing data:

- Left hand side: Residuals get bigger with larger x values → Suggests heteroscedasticity
Some Notes About Residuals

Can see from visualizing data:

- **Left hand side:** Residuals get bigger with larger $x$ values $\rightarrow$ Suggests **heteroscedasticity**
- **Right hand side:** Fairly constant size of residuals $\rightarrow$ Suggests **homoscedasticity**
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$.

- **Leverage point**: Observation that has an unusual $x$ value (far from the mean).

- An observation is influential if it substantially changes the regression line. That is, it is an outlier and has high leverage.

- Outlier, leverage, and influential observations raise interesting questions to examine further. Use substantive knowledge to investigate, make decisions.
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$
- **Leverage point**: Observation that has an unusual $x$ value (far from the mean)
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$
- **Leverage point**: Observation that has an unusual $x$ value (far from the mean)
- An observation is **influential** if it substantially changes the regression line
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$
- **Leverage point**: Observation that has an unusual $x$ value (far from the mean)
- An observation is **influential** if it substantially changes the regression line. That is, it is an outlier *and* has high leverage

Use substantive knowledge to investigate, make decisions.
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$

- **Leverage point**: Observation that has an unusual $x$ value (far from the mean)

- An observation is **influential** if it substantially changes the regression line → That is, it is an outlier and has high leverage

- Outlier, leverage, and influential observations raise interesting questions to examine more
Outliers and Leverage Points

- **Outlier**: Observation that has an unusual $y$ value, conditional on $x$
- **Leverage point**: Observation that has an unusual $x$ value (far from the mean)
- An observation is **influential** if it substantially changes the regression line → That is, it is an outlier and has high leverage
- Outlier, leverage, and influential observations raise interesting questions to examine more → Use substantive knowledge to investigate, make decision
Outliers and Leverage Points
Outliers and Leverage Points

State-Level Unemployment Rates in 1995 vs 2000

Unemployment Rate 1995

Fitted values
yr2000
Outliers and Leverage Points

State-Level Unemployment Rates in 1995 vs 2000

Unemployment Rate 1995

- Fitted values
- yr2000
Extending this to Multiple Regression

Simple Linear Regression appropriate when we have one outcome \((y)\) and one explanatory variable \((x)\).

Is state unemployment in year 2000 explained by state unemployment in year 1995?

But what if we have more variables that we think could explain variation in \(y\)?

Multiple Linear Regression

Extends simple linear regression to include more covariates, or additional explanatory variables.

Is state unemployment in year 2000 explained by state unemployment in year 1995, state gas prices, change in state % college educated?
Extending this to Multiple Regression

- **Simple Linear Regression** appropriate when we have one outcome \((y)\) and one explanatory variable \((x)\)

- Is state unemployment in year 2000 explained by state unemployment in year 1995?

- But what if we have more variables that we think could explain variation in \(y\)?

- **Multiple Linear Regression**

  Extends simple linear regression to include more covariates, or additional explanatory variables

  - Is state unemployment in year 2000 explained by state unemployment in year 1995, state gas prices, change in state % college educated?
Extending this to Multiple Regression

- **Simple Linear Regression** appropriate when we have one outcome \((y)\) and one explanatory variable \((x)\)
  - Is state unemployment in year 2000 explained by state unemployment in year 1995?

- But what if we have more variables that we think could explain variation in \(y\)?
- **Multiple Linear Regression**
  - Extends simple linear regression to include more covariates, or additional explanatory variables
  - Is state unemployment in year 2000 explained by state unemployment in year 1995, state gas prices, change in state % college educated?
Extending this to Multiple Regression

- **Simple Linear Regression** appropriate when we have one outcome ($y$) and one explanatory variable ($x$)
  - Is state unemployment in year 2000 explained by state unemployment in year 1995?

- But what if we have more variables that we think could explain variation in $y$?
Extending this to Multiple Regression

- **Simple Linear Regression** appropriate when we have one outcome \((y)\) and one explanatory variable \((x)\)
  - Is state unemployment in year 2000 explained by state unemployment in year 1995?
- But what if we have more variables that we think could explain variation in \(y\)?
- **Multiple Linear Regression**
Extending this to Multiple Regression

- **Simple Linear Regression** appropriate when we have one outcome \((y)\) and one explanatory variable \((x)\)
  - Is state unemployment in year 2000 explained by state unemployment in year 1995?
- But what if we have more variables that we think could explain variation in \(y\)?
- **Multiple Linear Regression**
  - Extends simple linear regression to include more **covariates**, or additional explanatory variables
Extending this to Multiple Regression

- **Simple Linear Regression** appropriate when we have one outcome \((y)\) and one explanatory variable \((x)\)
  - Is state unemployment in year 2000 explained by state unemployment in year 1995?
- But what if we have more variables that we think could explain variation in \(y\)?
- **Multiple Linear Regression**
  - Extends simple linear regression to include more *covariates*, or additional explanatory variables
  - Is state unemployment in year 2000 explained by state unemployment in year 1995, state gas prices, change in state % college educated?
Multiple Linear Regression

The model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i \]

Equivalently using the expected value notation:

\[ E(y_i | x_1, \ldots, x_k) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} \]

where \( x_1, x_2, \) and \( x_3 \) are covariates and \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \) etc, are coefficient estimates.

Note: We still are working with Outcome is continuous (will explore extensions in API 202)

Independent variables can be continuous, binary, or categorical.
Multiple Linear Regression

- Model then takes form:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i
\]

- Or, equivalently using the expected value notation:

\[
E(y_i | x_1, \ldots, x_k) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}
\]

- where \( x_{i1}, x_{i2}, \text{ and } x_{i3} \) are covariates
- and \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \text{ etc, } \) are coefficient estimates

- Note: We still are working with

- Outcome is continuous (will explore extensions in API 202)
- Independent variables can be continuous, binary, or categorical
Multiple Linear Regression

- Model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i \]
Multiple Linear Regression

- Model then takes form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \epsilon_i$$

- Or, equivalently using the expected value notation:

$$E(y_i | x_1, \ldots, x_k) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \epsilon_i$$

- Where $x_{i1}, x_{i2}$ and $x_{i3}$ are covariates
- And $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots$ are coefficient estimates

- Note: We still are working with
- Outcome is continuous (will explore extensions in API 202)
- Independent variables can be continuous, binary, or categorical
Multiple Linear Regression

- Model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i \]

- Or, equivalently using the expected value notation:

\[ E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \]
Multiple Linear Regression

- Model then takes form:

  \[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i \]

- Or, equivalently using the expected value notation:

  \[ E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \]

- where \( x_1, x_2 \) and \( x_3 \) are covariates
Multiple Linear Regression

- Model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i \]

- Or, equivalently using the expected value notation:

\[ E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \]

- where \( x_1, x_2 \) and \( x_3 \) are covariates
- and \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \) etc, are coefficient estimates
Multiple Linear Regression

- Model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \epsilon_i \]

- Or, equivalently using the expected value notation:

\[ E(y|x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k \]

- where \( x_1, x_2 \) and \( x_3 \) are covariates
- and \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \), etc, are coefficient estimates
- Note: We still are working with
Multiple Linear Regression

- Model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \epsilon_i \]

- Or, equivalently using the expected value notation:

\[ E(y|x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k \]

- where \( x_1, x_2 \) and \( x_3 \) are covariates
- and \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \), etc, are coefficient estimates
- Note: We still are working with
  - Outcome is continuous (will explore extensions in API 202)
Multiple Linear Regression

- Model then takes form:

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \epsilon_i \]

- Or, equivalently using the expected value notation:

\[ E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k \]

- where \( x_1, x_2 \) and \( x_3 \) are covariates
- and \( \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \) etc, are coefficient estimates
- Note: We still are working with
  - Outcome is continuous (will explore extensions in API 202)
  - Independent variables can be continuous, binary, or categorical
Multiple Linear Regression

Used extensively across all kinds of areas:

▶ Law: Multiple regression takes into account different defendant characteristics to estimate size of gap in sentencing between black and white defendants

▶ Education: Multiple regression takes into account different student characteristics to estimate size of math test score gap between boys and girls

▶ Medicine: Multiple regression takes into account patient characteristics such as age, gender, race to determine effect of new drugs and treatments

→ You'll visit some of the assumptions needed to make causal inferences in 202
Multiple Linear Regression

Used extensively across all kinds of areas:

▶ Law: Multiple regression takes into account different defendant characteristics to estimate size of gap in sentencing between black and white defendants

▶ Education: Multiple regression takes into account different student characteristics to estimate size of math test score gap between boys and girls

▶ Medicine: Multiple regression takes into account patient characteristics such as age, gender, race to determine effect of new drugs and treatments

→ You’ll visit some of the assumptions needed to make causal inferences in 202
Multiple Linear Regression

Used extensively across all kinds of areas:

▶ Law: Multiple regression takes into account different defendant characteristics to estimate size of gap in sentencing between black and white defendants
Multiple Linear Regression

Used extensively across all kinds of areas:

- **Law:** Multiple regression takes into account different defendant characteristics to estimate size of gap in sentencing between black and white defendants
- **Education:** Multiple regression takes into account different student characteristics to estimate size of math test score gap between boys and girls

→ You'll visit some of the assumptions needed to make causal inferences in 202
Multiple Linear Regression

Used extensively across all kinds of areas:

- **Law**: Multiple regression takes into account different defendant characteristics to estimate size of gap in sentencing between black and white defendants.
- **Education**: Multiple regression takes into account different student characteristics to estimate size of math test score gap between boys and girls.
- **Medicine**: Multiple regression takes into account patient characteristics such as age, gender, race to determine effect of new drugs and treatments.

You'll visit some of the assumptions needed to make causal inferences in 20
Multiple Linear Regression

Used extensively across all kinds of areas:

- **Law**: Multiple regression takes into account different defendant characteristics to estimate size of gap in sentencing between black and white defendants
- **Education**: Multiple regression takes into account different student characteristics to estimate size of math test score gap between boys and girls
- **Medicine**: Multiple regression takes into account patient characteristics such as age, gender, race to determine effect of new drugs and treatments

→ You’ll visit some of the assumptions needed to make causal inferences in 202
Multiple Linear Regression

Key advantages of multiple regression over other approaches:

1. Can examine relationship between several independent variables and one outcome simultaneously
2. If one variable is of particular interest, allows us to account for other factors to isolate that relationship more effectively
3. Straightforward to interpret
Multiple Linear Regression

Key advantages of multiple regression over other approaches:
Multiple Linear Regression

Key advantages of multiple regression over other approaches:

1. Can examine relationship between several independent variables and one outcome simultaneously
Multiple Linear Regression

Key advantages of multiple regression over other approaches:

1. Can examine relationship between several independent variables and one outcome simultaneously

2. If one variable is of particular interest, allows us to account for other factors to isolate that relationship more effectively
Key advantages of multiple regression over other approaches:

1. Can examine relationship between several independent variables and one outcome simultaneously
2. If one variable is of particular interest, allows us to account for other factors to isolate that relationship more effectively
3. Straightforward to interpret
Multiple Linear Regression

Let's work with an example

What variables help predict a country's average life expectancy (our outcome)

- Birthrate
- Wealth
- Region of the world one lives in

Use multiple regression to assess relationships
Multiple Linear Regression

- Let’s work with an example

- What variables help predict a country’s average life expectancy (our outcome)
  - Birthrate
  - Wealth
  - Region of the world one lives in

- Use multiple regression to assess relationships
Multiple Linear Regression

- Let’s work with an example
- What variables help predict a country’s average life expectancy (our outcome)
Multiple Linear Regression

- Let’s work with an example
- What variables help predict a country’s average life expectancy (our outcome)
  - Birthrate
Multiple Linear Regression

- Let’s work with an example
- What variables help predict a country’s average life expectancy (our outcome)
  - Birthrate
  - Wealth
Let’s work with an example

What variables help predict a country’s average life expectancy (our outcome)

- Birthrate
- Wealth
- Region of the world one lives in
Multiple Linear Regression

- Let’s work with an example
- What variables help predict a country’s average life expectancy (our outcome)
  - Birthrate
  - Wealth
  - Region of the world one lives in
- Use multiple regression to assess relationships
Multiple Linear Regression

<table>
<thead>
<tr>
<th>Country</th>
<th>Life Expectancy</th>
<th>Birthrate</th>
<th>Region</th>
<th>GDP per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>77.04227</td>
<td>12.74800</td>
<td>3</td>
<td>4148.8517</td>
</tr>
<tr>
<td>Algeria</td>
<td>73.07990</td>
<td>19.81100</td>
<td>1</td>
<td>5403.9992</td>
</tr>
<tr>
<td>Angola</td>
<td>51.05932</td>
<td>40.83800</td>
<td>1</td>
<td>5484.8350</td>
</tr>
<tr>
<td>Argentina</td>
<td>75.79766</td>
<td>17.00600</td>
<td>6</td>
<td>11557.5715</td>
</tr>
<tr>
<td>Armenia</td>
<td>73.91602</td>
<td>15.20300</td>
<td>2</td>
<td>3337.8637</td>
</tr>
<tr>
<td>Australia</td>
<td>81.84634</td>
<td>13.29069</td>
<td>4</td>
<td>67035.5712</td>
</tr>
<tr>
<td>Austria</td>
<td>81.03171</td>
<td>9.30000</td>
<td>3</td>
<td>47226.1957</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>70.65285</td>
<td>19.20000</td>
<td>2</td>
<td>7227.5007</td>
</tr>
</tbody>
</table>
## Multiple Linear Regression

<table>
<thead>
<tr>
<th>Country</th>
<th>LifeExpectancy</th>
<th>Birthrate</th>
<th>Region</th>
<th>GDP per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>77.04227</td>
<td>12.74800</td>
<td>3</td>
<td>4148.8517</td>
</tr>
<tr>
<td>Algeria</td>
<td>73.07990</td>
<td>19.81100</td>
<td>1</td>
<td>5403.9992</td>
</tr>
<tr>
<td>Angola</td>
<td>51.05932</td>
<td>40.83800</td>
<td>1</td>
<td>5484.8350</td>
</tr>
<tr>
<td>Argentina</td>
<td>75.79766</td>
<td>17.00600</td>
<td>6</td>
<td>11557.5715</td>
</tr>
<tr>
<td>Armenia</td>
<td>73.91602</td>
<td>15.20300</td>
<td>2</td>
<td>3337.8637</td>
</tr>
<tr>
<td>Australia</td>
<td>81.84634</td>
<td>13.29069</td>
<td>4</td>
<td>67035.5712</td>
</tr>
<tr>
<td>Austria</td>
<td>81.03171</td>
<td>9.30000</td>
<td>3</td>
<td>47226.1957</td>
</tr>
<tr>
<td>Azerbaijan</td>
<td>70.65285</td>
<td>19.20000</td>
<td>2</td>
<td>7227.5007</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Multiple Linear Regression

Outcome (response) variable $y$: Life expectancy at birth

Three explanatory (predictor) variables:
- Birthrate per 1000 people ($x_1$)
- GDP per capita ($x_2$)
- 6 continents, or regions of the world ($x_3$ to $x_7$)

Note: We’ll include region with an indicator or dummy variable.
Multiple Linear Regression

- Outcome (response) variable $y$: Life expectancy at birth
Multiple Linear Regression

- Outcome (response) variable $y$: Life expectancy at birth
  - Number of years average newborn infant would live if prevailing patterns of mortality at time of birth were to stay same throughout his/her life
Multiple Linear Regression

- **Outcome (response) variable $y$:** Life expectancy at birth
  - Number of years average newborn infant would live if prevailing patterns of mortality at time of birth were to stay same throughout his/her life

- **Three explanatory (predictor) variables:**

  - Birthrate per 1000 people ($x_1$)
  - GDP per capita ($x_2$)
  - 6 continents, or regions of the world ($x_3$ to $x_7$)

Note: We'll include region with an indicator or dummy variable
Multiple Linear Regression

- **Outcome (response) variable** $y$: Life expectancy at birth
  - Number of years average newborn infant would live if prevailing patterns of mortality at time of birth were to stay the same throughout his/her life

- **Three explanatory (predictor) variables:**
  - Birthrate per 1000 people ($x_1$)
Multiple Linear Regression

- **Outcome (response) variable** $y$: Life expectancy at birth
  - Number of years average newborn infant would live if prevailing patterns of mortality at time of birth were to stay same throughout his/her life

- **Three explanatory (predictor) variables:**
  - Birthrate per 1000 people ($x_1$)
  - GDP per capita ($x_2$)
Multiple Linear Regression

- **Outcome (response) variable** $y$: Life expectancy at birth
  - Number of years average newborn infant would live if prevailing patterns of mortality at time of birth were to stay same throughout his/her life

- **Three explanatory (predictor) variables:**
  - Birthrate per 1000 people ($x_1$)
  - GDP per capita ($x_2$)
  - 6 continents, or regions of the world ($x_3$ to $x_7$)

Note: We'll include region with an indicator or dummy variable
Multiple Linear Regression

- Outcome (response) variable $y$: Life expectancy at birth
  - Number of years average newborn infant would live if prevailing patterns of mortality at time of birth were to stay same throughout his/her life
- Three explanatory (predictor) variables:
  - Birthrate per 1000 people ($x_1$)
  - GDP per capita ($x_2$)
  - 6 continents, or regions of the world ($x_3$ to $x_7$)
- Note: We’ll include region with an indicator or dummy variable
Multiple Linear Regression

And then to full the model

|       | Coef.     | Std. Err. | t    | P>|t| |
|-------|-----------|-----------|------|-----|
| birthrate | -.5748516 | .1303377  | -4.41| 0.000 |
| GDPpercapita | .0000745 | .0000328  | 2.27 | 0.029 |
| Region  |           |           |      |     |
| 2      | 7.497148  | 2.515289  | 2.98 | 0.005 |
| 3      | 7.21654   | 3.316651  | 2.18 | 0.036 |
| 4      | 8.925904  | 3.085364  | 2.89 | 0.007 |
| 5      | 10.80421  | 3.020083  | 3.58 | 0.001 |
| 6      | 8.722155  | 3.139505  | 2.78 | 0.009 |
| _cons | 75.38939   | 4.59498   | 16.41| 0.000 |
And then to full the model

```
. regress LifeExpAll birthrate GDPpercapita i.Region
```

| LifeExpAll | Coef.  | Std. Err. | t     | P>|t| |
|------------|--------|-----------|-------|-----|
| birthrate  | -.5748516 | .1303377 | -4.41 | 0.000 |
| GDPpercapita | .0000745 | .0000328 | 2.27  | 0.029 |
| Region     |         |           |       |     |
| 2          | 7.497148 | 2.515289 | 2.98  | 0.005 |
| 3          | 7.21654  | 3.316651 | 2.18  | 0.036 |
| 4          | 8.925904 | 3.085364 | 2.89  | 0.007 |
| 5          | 10.80421 | 3.020083 | 3.58  | 0.001 |
| 6          | 8.722155 | 3.139505 | 2.78  | 0.009 |
| _cons      | 75.38939 | 4.59498  | 16.41 | 0.000 |
Multiple Linear Regression

This gives us the estimated regression line:

\[ \text{Life Exp} = 75.39 - 0.575 \times \text{(Birth Rate)} + 0.0000745 \times \text{(GDP)} + 7.50 \times \text{(Reg}^2) + 7.22 \times \text{(Reg}^3) + 8.93 \times \text{(Reg}^4) + 10.80 \times \text{(Reg}^5) + 8.72 \times \text{(Reg}^6) \]

Note: OLS formula for multiple regression slope and intercept estimates requires matrix algebra (since working with multiple covariates) and calculus. You do not need to know it for this class! For those of you who are comfortable with matrix algebra, the expression is:

\[ \hat{\beta} = (X'X)^{-1} X'Y \]
Multiple Linear Regression

- This gives us the estimated regression line:

\[
\text{Life Exp} = 75.39 - 0.575 \times \text{Birth Rate} + 0.0000745 \times \text{GDP} + 7.50 \times \text{Reg}_2 + 7.22 \times \text{Reg}_3 + 8.93 \times \text{Reg}_4 + 10.80 \times \text{Reg}_5 + 8.72 \times \text{Reg}_6
\]
Multiple Linear Regression

- This gives us the estimated regression line:

\[
Life \ Exp = 75.39 - 0.575 (Birth \ Rate) + 0.0000745 (GDP) \\
+ 7.50 (Reg_2) + 7.22 (Reg_3) + 8.93 (Reg_4) + 10.80 (Reg_5) + 8.72 (Reg_6)
\]
Multiple Linear Regression

- This gives us the estimated regression line:

\[
\text{Life Exp} = 75.39 - 0.575 (\text{Birth Rate}) + 0.0000745 (\text{GDP}) \\
+ 7.50 (\text{Reg}_2) + 7.22 (\text{Reg}_3) + 8.93 (\text{Reg}_4) + 10.80 (\text{Reg}_5) + 8.72 (\text{Reg}_6)
\]

- Note: OLS formula for multiple regression slope and intercept estimates requires matrix algebra (since working with multiple covariates) and calculus.
Multiple Linear Regression

- This gives us the estimated regression line:

\[
\text{Life Exp} = 75.39 - 0.575(\text{Birth Rate}) + 0.0000745(\text{GDP}) + 7.50(\text{Reg}_2) + 7.22(\text{Reg}_3) + 8.93(\text{Reg}_4) + 10.80(\text{Reg}_5) + 8.72(\text{Reg}_6)
\]

- Note: OLS formula for multiple regression slope and intercept estimates requires matrix algebra (since working with multiple covariates) and calculus.

- You do not need to know it for this class!
Multiple Linear Regression

- This gives us the estimated regression line:

\[ \text{Life Exp} = 75.39 - 0.575(Birth \ Rate) + 0.0000745(GDP) \\
+ 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6) \]

- Note: OLS formula for multiple regression slope and intercept estimates requires matrix algebra (since working with multiple covariates) and calculus.
- You do not need to know it for this class!
- For those of you who are comfortable with matrix algebra, expression is:
Multiple Linear Regression

- This gives us the estimated regression line:

\[ \text{Life Exp} = 75.39 - 0.575(\text{Birth Rate}) + 0.0000745(\text{GDP}) + 7.50(\text{Reg}_2) + 7.22(\text{Reg}_3) + 8.93(\text{Reg}_4) + 10.80(\text{Reg}_5) + 8.72(\text{Reg}_6) \]

- Note: OLS formula for multiple regression slope and intercept estimates requires matrix algebra (since working with multiple covariates) and calculus.

- You do not need to know it for this class!

- For those of you who are comfortable with matrix algebra, expression is:

\[ \hat{\beta} = (X'X)^{-1}X'Y \]
Multiple Linear Regression Interpretation

\[ \text{Life Exp} = 75.39 - 0.575 \times (\text{Birth Rate}) + 0.0000745 \times (\text{GDP}) + 7.50 \times (\text{Reg}^2) + 7.22 \times (\text{Reg}^3) + 8.93 \times (\text{Reg}^4) + 10.80 \times (\text{Reg}^5) + 8.72 \times (\text{Reg}^6) \]

▶ Need to be more cautious in interpretation
▶ For example, birth rate:
▶ An increase of 1 per 1000 in a country's birth rate is on average linked to a 0.575 decrease in a country's average life expectancy, assuming other variables in model remain the same
Multiple Linear Regression Interpretation

\[
\text{Life Exp} = 75.39 - 0.575(Birth \ Rate) + 0.0000745(GDP) \\
+ 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6)
\]

Need to be more cautious in interpretation

For example, birth rate:
An increase of 1 per 1000 in a country’s birth rate is on average linked to a 0.575 decrease in a country’s average life expectancy, assuming other variables in model remain the same.
Multiple Linear Regression Interpretation

\[
\text{Life Exp} = 75.39 - 0.575(Birth\ Rate) + 0.0000745(GDP) \\
+ 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6)
\]

- Need to be more cautious in interpretation.
Multiple Linear Regression Interpretation

\[
Life\ Exp = 75.39 - 0.575(Birth\ Rate) + 0.0000745(GDP) \\
+ 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6)
\]

- Need to be more cautious in interpretation
- For example, birth rate:
Multiple Linear Regression Interpretation

\[
\text{Life Exp} = 75.39 - 0.575(Birth\ Rate) + 0.0000745(GDP) \\
+ 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6)
\]

- Need to be more cautious in interpretation
- For example, birth rate:
- An increase of 1 per 1000 in a country’s birth rate is on average linked to a 0.575 decrease in a country’s average life expectancy, assuming other variables in model remain the same.
Multiple Linear Regression Interpretation

\[ \text{Life Exp} = 75.39 - 0.575 (\text{Birth Rate}) + 0.0000745 (\text{GDP}) + 7.50 (\text{Reg2}) + 7.22 (\text{Reg3}) + 8.93 (\text{Reg4}) + 10.80 (\text{Reg5}) + 8.72 (\text{Reg6}) \]

▶ For regions, we have a \( \beta \) coefficient for each of region (2 to 6)
▶ We set up model estimation to compare regions to Region #1 (Africa)
▶ Ex) A country being in Region 5 (N America) is linked with an average 10.80 yr life expectancy increase compared to Region 1 (Africa), other variables in model remaining constant
▶ → When using indicator or dummy variables, think about which baseline comparison group makes most sense
Multiple Linear Regression Interpretation

\[ \text{Life Exp} = 75.39 - 0.575(Birth\ Rate) + 0.0000745(GDP) + 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6) \]
Multiple Linear Regression Interpretation

\[
\text{Life Exp} = 75.39 - 0.575(Birth \ Rate) + 0.0000745(GDP) \\
+ 7.50(Reg_2) + 7.22(Reg_3) + 8.93(Reg_4) + 10.80(Reg_5) + 8.72(Reg_6)
\]

▶ For regions, we have a $\beta$ coefficient for each of region (2 to 6)
Multiple Linear Regression Interpretation

\[ \text{Life Exp} = 75.39 - 0.575(\text{Birth Rate}) + 0.0000745(\text{GDP}) \\
\quad + 7.50(\text{Reg}_2) + 7.22(\text{Reg}_3) + 8.93(\text{Reg}_4) + 10.80(\text{Reg}_5) + 8.72(\text{Reg}_6) \]

- For regions, we have a \( \beta \) coefficient for each of region (2 to 6)
- We set up model estimation to compare regions to Region #1 (Africa)
Multiple Linear Regression Interpretation

\[
\text{Life Exp} = 75.39 - 0.575(Birth \ Rate) + 0.0000745(GDP) \\
+ 7.50(\text{Reg}_2) + 7.22(\text{Reg}_3) + 8.93(\text{Reg}_4) + 10.80(\text{Reg}_5) + 8.72(\text{Reg}_6)
\]

- For regions, we have a $\beta$ coefficient for each of region (2 to 6)
- We set up model estimation to compare regions to Region #1 (Africa)
- Ex) A country being in Region 5 (N America) is linked w/ an average 10.80 yr life expectancy increase compared being in Region 1 (Africa), other variables in model remaining constant
Multiple Linear Regression Interpretation

\[
\text{Life Exp} = 75.39 - 0.575(\text{Birth Rate}) + 0.0000745(\text{GDP})
\]
\[
+ 7.50(\text{Reg}_2) + 7.22(\text{Reg}_3) + 8.93(\text{Reg}_4) + 10.80(\text{Reg}_5) + 8.72(\text{Reg}_6)
\]

- For regions, we have a \( \beta \) coefficient for each of region (2 to 6)
- We set up model estimation to compare regions to Region #1 (Africa)
- Ex) A country being in Region 5 (N America) is linked w/ an average 10.80 yr life expectancy increase compared being in Region 1 (Africa), other variables in model remaining constant
- \( \rightarrow \) When using indicator or dummy variables, think about which baseline comparison group makes most sense
Multiple Linear Regression Hypothesis Testing

Can use output to conduct hypothesis tests of each regression coefficient

Ex) Is country’s birth rate a significant predictor of life expectancy?

Null hypothesis:

\[ H_0: \beta_1 = 0 \]

Two-tailed alternative hypothesis:

\[ H_a: \beta_1 \neq 0 \]

From Stata, \( t \)-statistic = -4.41, \( p \)-value < 0.001

Strong evidence to reject null hypothesis that \( \beta_1 = 0 \)

→ Birth rate appears to be associated with life expectancy, after adjusting for region, GDP per capita
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient

  - Null hypothesis: $H_0: \beta_1 = 0$
  - Two-tailed alternative hypothesis: $H_a: \beta_1 \neq 0$

  - From Stata, $t$-statistic = -4.41, $p$-value < 0.001
  - Strong evidence to reject null hypothesis that $\beta_1 = 0$
  - Birth rate appears to be associated with life expectancy, after adjusting for region, GDP per capita
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?

Null hypothesis:

\[ H_0: \beta_1 = 0 \]

Two-tailed alternative hypothesis:

\[ H_a: \beta_1 \neq 0 \]

From Stata, \( t \)-statistic = -4.41, \( p \)-value < 0.001

Strong evidence to reject null hypothesis that \( \beta_1 = 0 \)

→ Birth rate appears to be associated with life expectancy, after adjusting for region, GDP per capita
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_1 = 0$
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_1 = 0$
- Two-tailed alternative hypothesis:
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_1 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_1 \neq 0$

From Stata, $t$-statistic = -4.41, $p$-value < 0.001

→ Birth rate appears to be associated with life expectancy, after adjusting for region, GDP per capita
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_1 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_1 \neq 0$
- From Stata, $t$-statistic = -4.41, $p$-value < 0.001
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_1 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_1 \neq 0$
- From Stata, $t$-statistic = -4.41, $p$-value < 0.001
- Strong evidence to reject null hypothesis that $\beta_1 = 0$
Multiple Linear Regression Hypothesis Testing

- Can use output to conduct hypothesis tests of each regression coefficient
- Ex) Is country’s birth rate a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_1 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_1 \neq 0$
- From Stata, $t$-statistic $= -4.41$, $p$-value $< 0.001$
- Strong evidence to reject null hypothesis that $\beta_1 = 0$
- → Birth rate appears to be associated w/ life expectancy, after adjusting for region, GDP per capita
Multiple Linear Regression Hypothesis Testing

Ex) Is country's GDP per capita a significant predictor of life expectancy?

Null hypothesis:

\[ H_0: \beta_2 = 0 \]

Two-tailed alternative hypothesis:

\[ H_a: \beta_2 \neq 0 \]

From Stata, t-statistic = 2.27, p-value = 0.029

Fairly strong evidence to reject null hypothesis that \( \beta_2 = 0 \)

GDP per capita appears to be associated with life expectancy, after adjusting for region, birth rate
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?

- Null hypothesis: $H_0: \beta_2 = 0$
- Two-tailed alternative hypothesis: $H_a: \beta_2 \neq 0$

- From Stata, $t$-statistic = 2.27, $p$-value = 0.029

- Fairly strong evidence to reject null hypothesis that $\beta_2 = 0$

- GDP per capita appears to be associated with life expectancy, after adjusting for region, birth rate
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0$: $\beta_2 = 0$
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_2 = 0$
- Two-tailed alternative hypothesis:
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_2 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_2 \neq 0$

From Stata, $t$-statistic = 2.27, $p$-value = 0.029

Fairly strong evidence to reject null hypothesis that $\beta_2 = 0$

GDP per capita appears to be associated with life expectancy, after adjusting for region, birth rate
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0$: $\beta_2 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_2 \neq 0$
- From Stata, $t$-statistic = 2.27, $p$-value = 0.029
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0: \beta_2 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_2 \neq 0$
- From Stata, $t$-statistic = 2.27, $p$-value = 0.029
- Fairly strong evidence to reject null hypothesis that $\beta_2 = 0$
Multiple Linear Regression Hypothesis Testing

- Ex) Is country’s GDP per capita a significant predictor of life expectancy?
- Null hypothesis:
  - $H_0$: $\beta_2 = 0$
- Two-tailed alternative hypothesis:
  - $H_a: \beta_2 \neq 0$
- From Stata, $t$-statistic = 2.27, $p$-value = 0.029
- Fairly strong evidence to reject null hypothesis that $\beta_2 = 0$
- GDP per capita appears to be associated w/ life expectancy, after adjusting for region, birth rate
$R^2$ and Model Fit

Stata output gives us:

- Number of obs = 43
- $F(7, 35) = 33.84$
- Prob > $F = 0.0000$
- $R^2$ = 0.8713
- Adj $R^2$ = 0.8455
- Root MSE = 3.889

Substantively: 87.13% of total variability observed between countries' life expectancies (outcome) can be explained with a model with these 3 covariates: birth rate, GDP per capita, and region.
$R^2$ and Model Fit

- $R^2$ (coefficient of determination) useful tool for assessing how well model does in capturing variability in outcome

- $R^2 = 0.8713$

- Substantively: 87.13% of total variability observed between countries life expectancies (outcome) can be explained with a model with these 3 covariates (birth rate, GDP per capita and region)

### Stata output gives us:

- Number of obs = 43
- $F(7, 35) = 33.84$
- Prob > $F = 0.0000$
- $R^2$-squared = 0.8713
- Adj $R^2$-squared = 0.8455
- Root MSE = 3.889
$R^2$ and Model Fit

- $R^2$ (coefficient of determination) useful tool for assessing how well model does in capturing variability in outcome
- Stata output gives us:

  Number of obs = 43  
  F( 7, 35) = 33.84  
  Prob > F = 0.0000  
  R-squared = 0.8713  
  Adj R-squared = 0.8455  
  Root MSE = 3.889

- $R^2 = 0.8713$
- Substantively: 87.13% of total variability observed between countries life expectancies (outcome) can be explained with a model with these 3 covariates (birth rate, GDP per capita and region)
\( R^2 \) and Model Fit

- \( R^2 \) (coefficient of determination) useful tool for assessing how well model does in capturing variability in outcome
- Stata output gives us:

\[
\begin{align*}
\text{Number of obs} &= 43 \\
F(7, 35) &= 33.84 \\
\text{Prob > F} &= 0.0000 \\
R\text{-squared} &= 0.8713 \\
\text{Adj R-squared} &= 0.8455 \\
\text{Root MSE} &= 3.889 
\end{align*}
\]

Substantively: 87.13% of total variability observed between countries life expectancies (outcome) can be explained with a model with these 3 covariates (birth rate, GDP per capita and region).
$R^2$ and Model Fit

- $R^2$ (coefficient of determination) useful tool for assessing how well model does in capturing variability in outcome
- Stata output gives us:

```
Number of obs = 43
F(  7,   35) =  33.84
Prob > F      =  0.0000
R-squared     =  0.8713
Adj R-squared =  0.8455
Root MSE      =  3.889
```

- $R^2 = 0.8713$
$R^2$ and Model Fit

- $R^2$ (coefficient of determination) useful tool for assessing how well model does in capturing variability in outcome.
- Stata output gives us:

\[
\begin{align*}
\text{Number of obs} &= 43 \\
F(7, 35) &= 33.84 \\
\text{Prob > F} &= 0.0000 \\
R\text{-squared} &= 0.8713 \\
\text{Adj R-squared} &= 0.8455 \\
\text{Root MSE} &= 3.889 \\
\end{align*}
\]

- $R^2 = 0.8713$
- Substantively: 87.13% of total variability observed between countries life expectancies (outcome) can be explained w/ a model w/ these 3 covariates (birth rate, GDP per capita and region).
$R^2$ and Model Fit

However, $R^2$ always increases with more covariates:

- A model with birthrate only: $R^2 = 0.7891$
- A model with birthrate and region: $R^2 = 0.8523$

If an additional covariate does not have major association with the outcome, $R^2$ will not increase by too much.

Adjusted $R^2$ modifies the value of $R^2$ based upon sample size and number of covariates → may be more reliable than non-adjusted $R^2$. 
$R^2$ and Model Fit

- However, $R^2$ always increases with more covariates:
$R^2$ and Model Fit

- However, $R^2$ always increases with more covariates:
  - A model with birthrate only: $R^2 = 0.7891$
$R^2$ and Model Fit

- However, $R^2$ always increases with more covariates:
  - A model with birthrate only: $R^2 = 0.7891$
  - A model with birthrate and region: $R^2 = 0.8523$
$R^2$ and Model Fit

- However, $R^2$ always increases with more covariates:
  - A model with birthrate only: $R^2 = 0.7891$
  - A model with birthrate and region: $R^2 = 0.8523$
- If an additional covariate does not have major association with outcome, $R^2$ will not increase by too much

Adjusted $R^2$ modifies the value of $R^2$ based upon sample size and number of covariates → may be more reliable than non-adjusted $R^2$
$R^2$ and Model Fit

- However, $R^2$ always increases with more covariates:
  - A model with birthrate only: $R^2 = 0.7891$
  - A model with birthrate and region: $R^2 = 0.8523$
- If an additional covariate does not have major association with outcome, $R^2$ will not increase by too much.
- Adjusted $R^2$ modifies the value of $R^2$ based upon sample size and number of covariates → may be more reliable than non-adjusted $R^2$. 
What if the covariates are related to teach other?

▶ Likely that birth date correlated with GDP per capita, and GDP per capita related to region of the world
▶ In fact, true: Correlations between covariates here reasonably high
▶ Raises possibility of Multicollinearity:
▶ Two or more explanatory variables in multiple regression have a strong linear correlation
What if the covariates are related to each other?

- Likely that birth date correlated with GDP per capita, and GDP per capita related to region of the world.
What if the covariates are related to teach other?

- Likely that birth date correlated with GDP per capita, and GDP per capita related to region of the world
- In fact, true: Correlations between covariates here reasonably high
What if the covariates are related to teach other?

- Likely that birth date correlated with GDP per capita, and GDP per capita related to region of the world
- In fact, true: Correlations between covariates here reasonably high
- Raises possibility of **Multicollinearity**:
What if the covariates are related to teach other?

- Likely that birth date correlated with GDP per capita, and GDP per capita related to region of the world
- In fact, true: Correlations between covariates here reasonably high
- Raises possibility of Multicollinearity:
  - Two or more explanatory variables in multiple regression have a strong linear correlation
Multicollinearity

With two covariates, $X$ and $Z$:

**Weak X-Z**

**Strong X-Z**
What if the covariates are related to teach other?

Why is this a problem?
What if the covariates are related to teach other?

Why is this a problem?

▶ Multicollinearity makes it difficult to sort out the roles of each of the two explanatory variables in this case (more in API 202)

Technically a problem b/c it leads to increased standard errors → increased probability of failure to reject a false null (increased probability of Type II error)
What if the covariates are related to teach other?

Why is this a problem?

- Multicollinearity makes it difficult to sort out the roles of each of the two explanatory variables in this case (more in API 202)
- Technically a problem b/c it leads to increased standard errors
What if the covariates are related to teach other?

Why is this a problem?

- Multicollinearity makes it difficult to sort out the roles of each of the two explanatory variables in this case (more in API 202)
- Technically a problem b/c it leads to increased standard errors
- If artificially larger standard errors $\rightarrow$ increased probability of failure to reject a false null (increased probability of Type II error)
Notes on Non-Linearity

Have been assuming linear relationships between the variables
However, many instances in which we do not have linear relationships
Two options (both of which you will cover further in API 202):
1. Transform of one or both variables to get linearity (e.g., a log transformation)
2. Use a regression model appropriate for nonlinear (perhaps non-continuous) outcomes (more next semester)
Notes on Non-Linearity

- Have been assuming linear relationships between the variables
Notes on Non-Linearity

- Have been assuming linear relationships between the variables
- However, many instances in which we do not have linear relationships
Notes on Non-Linearity

- Have been assuming linear relationships between the variables
- However, many instances in which we do not have linear relationships
- Two options (both of which you will cover further in API 202):
  1. Transform of one or both variables to get linearity (e.g., a log transformation)
  2. Use a regression model appropriate for nonlinear (perhaps non-continuous) outcomes (more next semester)
Notes on Non-Linearity

- Have been assuming linear relationships between the variables
- However, many instances in which we do not have linear relationships
- Two options (both of which you will cover further in API 202):
  1. Transform of one or both variables to get linearity (e.g., a log transformation)
Notes on Non-Linearity

- Have been assuming linear relationships between the variables
- However, many instances in which we do not have linear relationships
- Two options (both of which you will cover further in API 202):
  1. Transform of one or both variables to get linearity (e.g., a log transformation)
  2. Use a regression model appropriate for nonlinear (perhaps non-continuous) outcomes (more next semester)
Log Transformation Example

Solution is to take log
Log Transformation Example

Infant Mortality Rate vs GDP Per Capita

→ Solution is to take log
Log Transformation Example

→ Solution is to take log
Log Transformation Example

Can implement this in Stata by taking log of two vars and using transformed vars in regression:

```
.regress logGDPpercapita logmortrateinfant
```

```
------------------------------------------------------------
logGDPpercapita | Coef. Std. Err. t P>|t|
------------------------------------------------------------
logmortrateinfant | -0.6600 0.05086 -22.07 0.000
_cons | 11.64702 0.15544 74.93 0.000
-----------------------------------------------------------
```
Log Transformation Example

Can implement this in Stata by taking log of two vars and using transformed vars in regression:

```
.regress logGDPpercapita logmortrateinfant
```

```
|                    | Coef. | Std. Err. | t     | P>|t| |
|--------------------|-------|------------|-------|-----|
| logGDPpercapita    |       |            |       |     |
| logmortrateinfant  | -0.6600 | 0.05086    | -22.07| 0.000 |
| _cons              | 11.64702 | 0.15544    | 74.93 | 0.000 |
```
Can implement this in Stata by taking log of two vars and using transformed vars in regression:

```
. regress logGDPpercapita logmортrateinfant
```

|                       | Coef.  | Std. Err. | t      | P>|t| |
|-----------------------|--------|-----------|--------|------|
| logGDPpercapita       |        |           |        |      |
| logmортrateinfant     | -0.6600| 0.05086   | -22.07 | 0.000|
| _cons                 | 11.64702 | 0.15544  | 74.93  | 0.000|
Log Transformation Example

- Downside: interpretation of logged analysis more difficult

- When both variables logged:
  \[ y^b = x^a \]
  \[ \text{slope} \]

- For example, for 10% increase in GDP/capita:
  \[ y^b = (1.1)^a - 0.66 = 0.94 \]
  So 10% increase in GDP/capita → 6% decrease in mortality

- 100% increase in GDP/capita:
  \[ y^b = (2)^a - 0.66 = 0.63 \]
  So 100% increase in GDP/capita → 37% decrease in mortality
Log Transformation Example

- Downside → interpretation of logged analysis more difficult

\[ y^b = x^a \]

For example, for 10% increase in GDP/capita

\[ y^b = (1.1)^{−0.66} = 0.94 \]

So 10% increase in GDP/capita → 6% decrease in mortality

\[ y^b = (2)^{−0.66} = 0.63 \]

So 100% increase in GDP/capita → 37% decrease in mortality
Log Transformation Example

- Downside $\rightarrow$ interpretation of logged analysis more difficult
- When both variables logged:

$$y = x^b$$

For example, for 10% increase in GDP/capita

$$y = (1.1)^{-0.66} = 0.94$$

So 10% increase in GDP/capita $\rightarrow$ 6% decrease in mortality

$$y = (2)^{-0.66} = 0.63$$

So 100% increase in GDP/capita $\rightarrow$ 37% decrease in mortality
Log Transformation Example

- Downside → interpretation of logged analysis more difficult
- When both variables logged:

\[
\frac{y_b}{y_a} = \frac{x_b^{slope}}{x_a}
\]
Log Transformation Example

- Downside → interpretation of logged analysis more difficult
- When both variables logged:

\[
\frac{y_b}{y_a} = \frac{x_b}{x_a}^{slope}
\]

- For example, for 10% increase in GDP/capita
Log Transformation Example

▶ Downside → interpretation of logged analysis more difficult
▶ When both variables logged:

\[
\frac{y_b}{y_a} = \frac{x_b^{slope}}{x_a}
\]

▶ For example, for 10% increase in GDP/capita

\[
\frac{y_b}{y_a} = (1.1)^{-0.66} = 0.94
\]

▶ So 10% increase in GDP/capita → 6% decrease in mortality

▶ For example, 100% increase in GDP/capita

\[
\frac{y_b}{y_a} = (2)^{-0.66} = 0.63
\]

▶ So 100% increase in GDP/capita → 37% decrease in mortality
Log Transformation Example

- Downside → interpretation of logged analysis more difficult
- When both variables logged:
  \[
  \frac{y_b}{y_a} = \frac{x_b^{slope}}{x_a}
  \]

- For example, for 10% increase in GDP/capita
  \[
  \frac{y_b}{y_a} = (1.1)^{-0.66} = 0.94
  \]
  
  So 10% increase in GDP/capita → 6% decrease in mortality
Log Transformation Example

▶ Downside → interpretation of logged analysis more difficult
▶ When both variables logged:

\[
\frac{y_b}{y_a} = \frac{x_b^{slope}}{x_a}
\]

▶ For example, for 10% increase in GDP/capita

\[
\frac{y_b}{y_a} = (1.1)^{-0.66} = 0.94
\]

▶ So 10% increase in GDP/capita → 6% decrease in mortality

\[
\frac{y_b}{y_a} = (2)^{-0.66} = 0.63
\]
Log Transformation Example

- Downside $\rightarrow$ interpretation of logged analysis more difficult
- When both variables logged:
  \[
  \frac{y_b}{y_a} = \frac{x_b^{slope}}{x_a}
  \]

- For example, for 10% increase in GDP/capita
  \[
  \frac{y_b}{y_a} = (1.1)^{-0.66} = 0.94
  \]
  So 10% increase in GDP/capita $\rightarrow$ 6% decrease in mortality

- So 100% increase in GDP/capita $\rightarrow$ 37% decrease in mortality
Extensions of Linear Regression

If outcome variable not continuous:
▶ Logit or probit regression
▶ For use with dichotomous outcomes (0 or 1)
▶ Ex) Do people vote or not
▶ Ex) Is a defendant convicted or not?
▶ Ordered/multinomial regression
▶ For use with categorical outcomes
▶ Ex) Public opinion categories are on a scale from "Do not Support" to "Strongly Support"
▶ Negative Binomial, Poisson regressions
▶ For use with count outcomes
▶ Ex) Number of times Supreme Court rules against government in one session
▶ Ex) Number of children born to married women in a locality

Note: These models have different assumptions, interpretations
Extensions of Linear Regression

If outcome variable not continuous:

- Logit or probit regression
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?
- Ordered/multinomial regression
  - For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”
- Negative Binomial, Poisson regressions
  - For use with count outcomes
  - Ex) Number of times Supreme Court rules against government in one session
  - Ex) Number of children born to married women in a locality

Note: These models have different assumptions, interpretations
Extensions of Linear Regression

If outcome variable not continuous:
▶ Logit or probit regression

- For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

▶ Ordered/multinomial regression
- For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from "Do not Support" to "Strongly Support"

▶ Negative Binomial, Poisson regressions
- For use with count outcomes
  - Ex) Number of times Supreme Court rules against government in one session
  - Ex) Number of children born to married women in a locality

Note: These models have different assumptions, interpretations.
Extensions of Linear Regression

If outcome variable not continuous:

- Logit or probit regression
  - For use with dichotomous outcomes (0 or 1)

- Ordered/multinomial regression
  - For use with categorical outcomes

- Negative Binomial, Poisson regressions
  - For use with count outcomes

Note: These models have different assumptions, interpretations
Extensions of Linear Regression

If outcome variable not continuous:

▸ Logit or probit regression
  ▸ For use with dichotomous outcomes (0 or 1)
  ▸ Ex) Do people vote or not
Extensions of Linear Regression

If outcome variable not continuous:

▶ Logit or probit regression
  ▶ For use with dichotomous outcomes (0 or 1)
  ▶ Ex) Do people vote or not
  ▶ Ex) Is a defendant convicted or not?

▶ Ordered/multinomial regression
  ▶ For use with categorical outcomes
  ▶ Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”

▶ Negative Binomial, Poisson regressions
  ▶ For use with count outcomes
  ▶ Ex) Number of times Supreme Court rules against government in one session
  ▶ Ex) Number of children born to married women in a locality

Note: These models have different assumptions, interpretations
Extensions of Linear Regression

If outcome variable not continuous:

- Logit or probit regression
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

- Ordered/multinomial regression
Extensions of Linear Regression

If outcome variable not continuous:

- Logit or probit regression
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

- Ordered/multinominal regression
  - For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”
Extensions of Linear Regression

If outcome variable not continuous:

- Logit or probit regression
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

- Ordered/multinomial regression
  - For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”

- Negative Binomial, Poisson regressions
Extensions of Linear Regression

If outcome variable not continuous:

- **Logit or probit regression**
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

- **Ordered/multinomial regression**
  - For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”

- **Negative Binomial, Poisson regressions**
  - For use with count outcomes
Extensions of Linear Regression

If outcome variable not continuous:

- **Logit or probit regression**
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

- **Ordered/multinomial regression**
  - For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”

- **Negative Binomial, Poisson regressions**
  - For use with count outcomes
  - Ex) Number of times Supreme Court rules against government in one session

Note: These models have different assumptions, interpretations
Extensions of Linear Regression

If outcome variable not continuous:

- **Logit or probit regression**
  - For use with dichotomous outcomes (0 or 1)
  - Ex) Do people vote or not
  - Ex) Is a defendant convicted or not?

- **Ordered/multinomial regression**
  - For use with categorical outcomes
  - Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”

- **Negative Binomial, Poisson regressions**
  - For use with count outcomes
  - Ex) Number of times Supreme Court rules against government in one session
  - Ex) Number of children born to married women in a locality

Note: These models have different assumptions, interpretations.
Extensions of Linear Regression

If outcome variable not continuous:

» Logit or probit regression
  » For use with dichotomous outcomes (0 or 1)
  » Ex) Do people vote or not
  » Ex) Is a defendant convicted or not?

» Ordered/multinomial regression
  » For use with categorical outcomes
  » Ex) Public opinion categories are on a scale from “Do not Support” to “Strongly Support”

» Negative Binomial, Poisson regressions
  » For use with count outcomes
  » Ex) Number of times Supreme Court rules against government in one session
  » Ex) Number of children born to married women in a locality

Note: These models have different assumptions, interpretations