Ordinary Least Squares

J. Alexander Branham

Fall 2017

Introduction to Ordinary Least Squares

OLS

- Ordinary least squares regression (OLS) is probably the most widely-used model in political science

OLS

- Ordinary least squares regression (OLS) is probably the most widely-used model in political science
- At its core, it's all about drawing a line through data

OLS

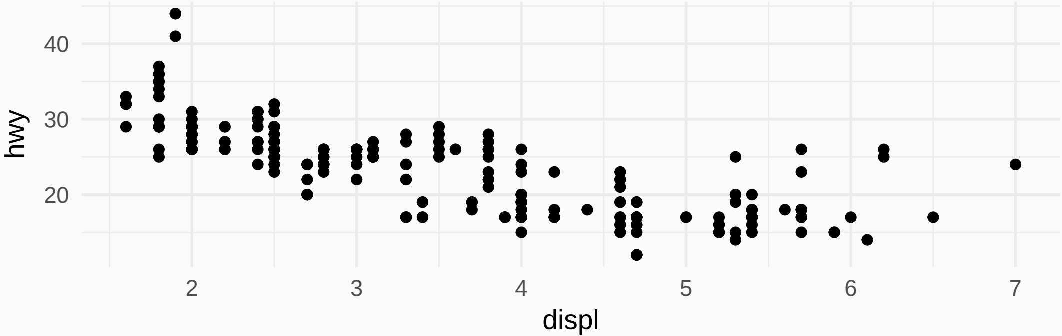
- Ordinary least squares regression (OLS) is probably the most widely-used model in political science
- At its core, it's all about drawing a line through data
- This allows us to asses the effect of x on y

OLS

- Ordinary least squares regression (OLS) is probably the most widely-used model in political science
- At its core, it's all about drawing a line through data
- This allows us to asses the effect of x on y
- Dependent variable y must be continuous

OLS

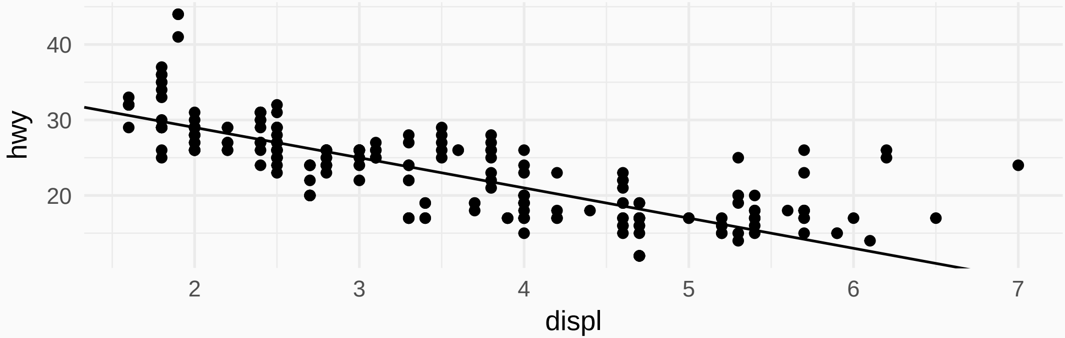
- Ordinary least squares regression (OLS) is probably the most widely-used model in political science
- At its core, it's all about drawing a line through data
- This allows us to asses the effect of x on y
- Dependent variable y must be continuous
    - OLS makes other assumptions you'll learn about in stats II



40
30
20
hwy
2
3
4
5
6
7
displ

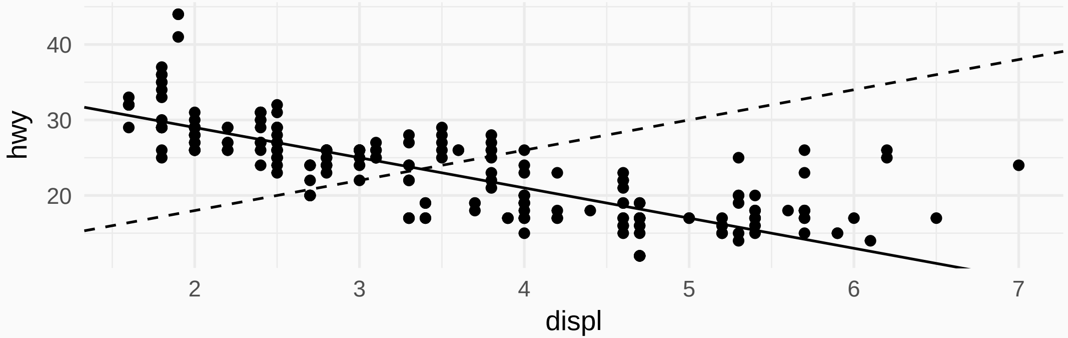
How to decide on a line?

```
p <- p +  geom_abline(slope = -4, intercept = 37)
p
```

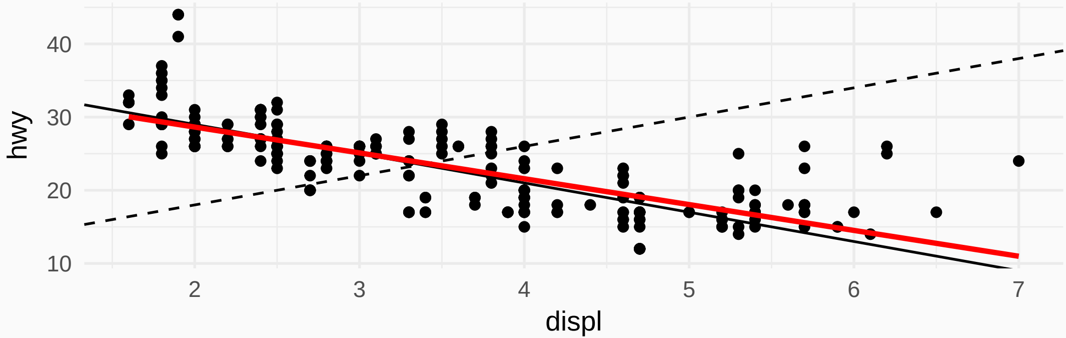


How to decide

```
p <- p +  geom_abline(slope = 4, intercept = 10, linetype = "dashed")
p
```



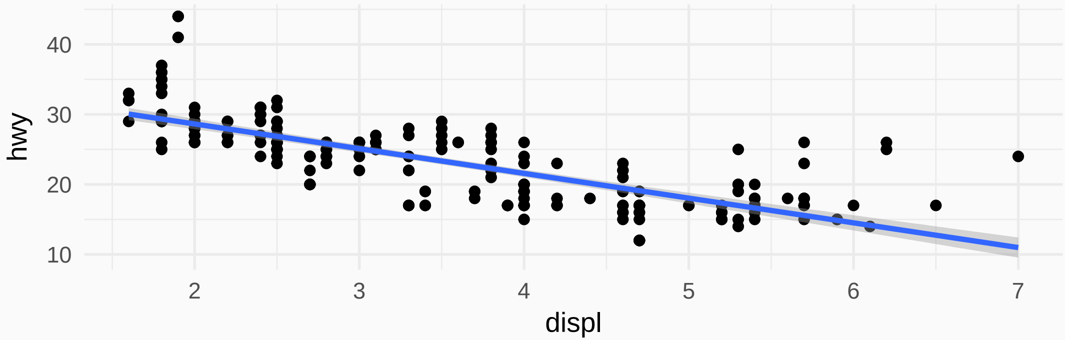
```r
p +  geom_smooth(method = "lm", se = FALSE, color = "red")
```



OLS in R

```r
lm(hwy ~ displ, data = mpg)
```

```
##
## Call:
## lm(formula = hwy ~ displ, data = mpg)
##
## Coefficients:
## (Intercept)        displ
##      35.698       -3.531
```



40
30
20
10
hwy
2
3
4
5
6
7
displ

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

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

- Intercept ($\hat{\beta}_0$) - predicted y when $x = 0$

Interpretation of coefficients

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

- Intercept ($\hat{\beta}_0$) - predicted y when $x = 0$
- Slope ($\hat{\beta}_1$) - a one unit change in x leads to a (slope) unit change in y, on average

Residuals

- Notice that our line doesn't fit our data perfectly - we always make some error with our prediction

Residuals

- Notice that our line doesn't fit our data perfectly - we always make some error with our prediction
- That's referred to as the *residual*

Residuals

- Notice that our line doesn't fit our data perfectly - we always make some error with our prediction
- That's referred to as the *residual*
- If we refer to our predicted value as $\hat{y}$, then we can calculate the residual for each observation with $e_i = y_i - \hat{y}_i$

Fitting the best line

- OLS determines the “best” line by minimizing the sum of squared residuals

Fitting the best line

- OLS determines the “best” line by minimizing the sum of squared residuals
- How to find this?

Fitting the best line

- OLS determines the "best" line by minimizing the sum of squared residuals
- How to find this?
- One option: Plug in all the values for the slope & intercept and calculate the sum of squared residuals for these infinity combinations

Fitting the best line

- OLS determines the “best” line by minimizing the sum of squared residuals
- How to find this?
- One option: Plug in all the values for the slope & intercept and calculate the sum of squared residuals for these infinity combinations
- That’s problematic...

How do we find the minimum sum of squared residuals?

Solution: Use calculus

Turns out we already know the solution - we learned it when we talked about *optimization*. We just need to *minimize* the sum of squared residuals with respect to the two coefficients:

$$\sum_{i=1}^{n} e_i^2$$

Rearrange above equation in terms of e_i:

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Rearrange above equation in terms of e_i:

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Substitute:

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

To find the minimum, we'll need to take the derivative with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. Starting with $\hat{\beta}_0$:

Derivatives

To find the minimum, we'll need to take the derivative with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. Starting with $\hat{\beta}_0$:

$$\frac{\partial}{\partial \hat{\beta}_0} \left[\sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \right]$$

To find the minimum, we'll need to take the derivative with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$. Starting with $\hat{\beta}_0$:

$$\frac{\partial}{\partial \hat{\beta}_0}\left[\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2\right]$$

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial \hat{\beta}_0}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2\right]$$

Derivatives

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial\hat{\beta}_0}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_i))^2\right]$$

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial\hat{\beta}_0}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_i))^2\right]$$

The next step is to use the **chain rule** to take the derivative of the quantity in parentheses:

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial\hat{\beta}_0}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_i))^2\right]$$

The next step is to use the **chain rule** to take the derivative of the quantity in parentheses:

$$\sum_{i=1}^{n}[-2(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)]$$

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial\hat{\beta}_0}(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))^2\right]$$

The next step is to use the **chain rule** to take the derivative of the quantity in parentheses:

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

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

Derivative for slope

Now let's take the partial derivative with respect to the slope ($\hat{\beta}_1$):

Derivative for slope

Now let's take the partial derivative with respect to the slope ($\hat{\beta}_1$):

$$\frac{\partial}{\partial \hat{\beta}_1}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial \hat{\beta}_1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

Derivative for slope

Now let's take the partial derivative with respect to the slope ($\hat{\beta}_1$):

$$\frac{\partial}{\partial \hat{\beta}_1}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

$$\sum_{i=1}^{n}\left[\frac{\partial}{\partial \hat{\beta}_1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right]$$

Using the chain rule again, we get:

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

How to find the *minimum* now that we have the partial derivatives of the sum of squared residuals?

How to find the *minimum* now that we have the partial derivatives of the sum of squared residuals?

$$\frac{\partial}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

How to find the *minimum* now that we have the partial derivatives of the sum of squared residuals?

$$\frac{\partial}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

((Solutions on board))

Multiple variables

- What if as years pass, engine size increases *and* fuel efficiency increases?

Multiple variables

- What if as years pass, engine size increases *and* fuel efficiency increases?
- Then the relationship we just observed might be *spurious*

Multiple variables

- What if as years pass, engine size increases *and* fuel efficiency increases?
- Then the relationship we just observed might be *spurious*

- What if as years pass, engine size increases *and* fuel efficiency increases?
- Then the relationship we just observed might be *spurious*

```
lm(hwy ~ displ + year, data = mpg)
```

```
##
## Call:
## lm(formula = hwy ~ displ + year, data = mpg)
##
## Coefficients:
## (Intercept)        displ         year
##   -276.1544      -3.6110       0.1558
```

Multiple variables

- How to find the effect of *one* variable (e.g. `displ`) on our y variable?

- How to find the effect of *one* variable (e.g. `displ`) on our y variable?
- Solution: partial derivatives

OLS in Matrix Form

Notation

- Let's pretend that we know the **true** model

Notation

- Let's pretend that we know the **true** model
    - Y is $nx1$ column vector

Notation

- Let's pretend that we know the **true** model
  - Y is $nx1$ column vector
  - X is $nxk(+1)$ matrix

Notation

- Let's pretend that we know the **true** model
  - Y is $nx1$ column vector
  - X is $nxk(+1)$ matrix
  - β is $kx1$ column vector

Notation

- Let's pretend that we know the **true** model
    - Y is $nx1$ column vector
    - X is $nxk(+1)$ matrix
    - β is $kx1$ column vector
    - E is $nx1$ column vector

Notation

- Let's pretend that we know the **true** model
  - Y is $nx1$ column vector
  - X is $nxk(+1)$ matrix
  - β is $kx1$ column vector
  - E is $nx1$ column vector
- Therefore, we have:

Notation

- Let's pretend that we know the **true** model
    - Y is $nx1$ column vector
    - X is $nxk(+1)$ matrix
    - β is $kx1$ column vector
    - E is $nx1$ column vector
- Therefore, we have:

Notation

- Let's pretend that we know the **true** model
    - Y is $nx1$ column vector
    - X is $nxk(+1)$ matrix
    - β is $kx1$ column vector
    - E is $nx1$ column vector
- Therefore, we have:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ 1 & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Matrix form

$$Y = X\beta + E$$

OLS minimizes the sum of squared residuals

- How to do that in matrix form?

OLS minimizes the sum of squared residuals

- How to do that in matrix form?
- First, what is sum of squared residuals?

OLS minimizes the sum of squared residuals

- How to do that in matrix form?
- First, what is sum of squared residuals?
- The residuals:

$$E = Y - X\hat{\beta}$$

OLS minimizes the sum of squared residuals

- How to do that in matrix form?
- First, what is sum of squared residuals?
- The residuals:

$$E = Y - X\hat{\beta}$$

- Sum of squared residuals:

$$E'E$$

OLS minimizes the sum of squared residuals

- How to do that in matrix form?
- First, what is sum of squared residuals?
- The residuals:

$$E = Y - X\hat{\beta}$$

- Sum of squared residuals:

$$E'E$$

- (show why on board)

OLS minimizes the sum of squared residuals

- How to do that in matrix form?
- First, what is sum of squared residuals?
- The residuals:

$$E = Y - X\hat{\beta}$$

- Sum of squared residuals:

$$E'E$$

- (show why on board)
- Alternatively,

$$\begin{aligned} E'E &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta} \end{aligned}$$

To minimize the sum of squares, we take the derivative

- Remember:

$$E'E = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}$$

To minimize the sum of squares, we take the derivative

- Remember:

$$E'E = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}$$

- The first derivative with respect to $\hat{\beta}$

$$\frac{\partial E'E}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

To minimize the sum of squares, we take the derivative

- Remember:

$$E'E = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}$$

- The first derivative with respect to $\hat{\beta}$

$$\frac{\partial E'E}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

- To check that this is a minimum, we check to make sure that the second derivative is positive

To minimize the sum of squares, we take the derivative

- Remember:

$$E'E = Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}$$

- The first derivative with respect to $\hat{\beta}$

$$\frac{\partial E'E}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

- To check that this is a minimum, we check to make sure that the second derivative is positive
- The second derivative is $2X'X$, which is positive definite so long as X is full rank

- Here ya go:

$$-2X'Y + 2X'X\hat{\beta} = 0$$

Solve for the estimator

- Here ya go:

$$-2X'Y + 2X'X\hat{\beta} = 0$$

- Move things around and divide by two:

$$X'Y = X'X\hat{\beta}$$

Solve for the estimator

- Here ya go:

$$-2X'Y + 2X'X\hat{\beta} = 0$$

- Move things around and divide by two:

$$X'Y = X'X\hat{\beta}$$

- Premultiply each side by $(X'X)^{-1}$

$$(X'X)^{-1}X'Y = (X'X)^{-1}(X'X)\hat{\beta}$$

Solve for the estimator

- Here ya go:

$$-2X'Y + 2X'X\hat{\beta} = 0$$

- Move things around and divide by two:

$$X'Y = X'X\hat{\beta}$$

- Premultiply each side by $(X'X)^{-1}$

$$(X'X)^{-1}X'Y = (X'X)^{-1}(X'X)\hat{\beta}$$

- We know that $(X'X)^{-1}(X'X) = I$

$$(X'X)^{-1}X'Y = I\hat{\beta}$$

Solve for the estimator

- Here ya go:

$$-2X'Y + 2X'X\hat{\beta} = 0$$

- Move things around and divide by two:

$$X'Y = X'X\hat{\beta}$$

- Premultiply each side by $(X'X)^{-1}$

$$(X'X)^{-1}X'Y = (X'X)^{-1}(X'X)\hat{\beta}$$

- We know that $(X'X)^{-1}(X'X) = I$

$$(X'X)^{-1}X'Y = I\hat{\beta}$$

- And I is (kinda) like multiplying by 1 so :

$$(X'X)^{-1}X'Y = \hat{\beta}$$

Interactions

- Sometimes we want to learn about the effect of X on Y *conditional* on Z.

Interactions

- Sometimes we want to learn about the effect of X on Y *conditional* on Z.
- ((example))

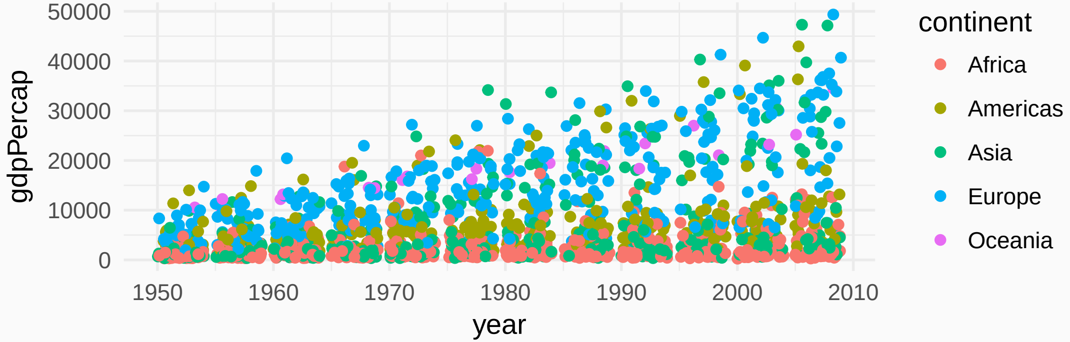
- Sometimes we want to learn about the effect of X on Y *conditional* on Z.
- ((example))
- As we go through time, GDP per capita in a country *generally* increases

Interactions

- Sometimes we want to learn about the effect of X on Y *conditional* on Z.
- ((example))
- As we go through time, GDP per capita in a country *generally* increases
- But what if this is different in different continents?

```r
library(gapminder)
library(ggplot2)
library(dplyr)
gapminder <- gapminder %>%
  filter(gdpPercap < 50000)
```

```r
ggplot(gapminder, aes(year, gdpPercap, color = continent)) +
  geom_jitter()
```

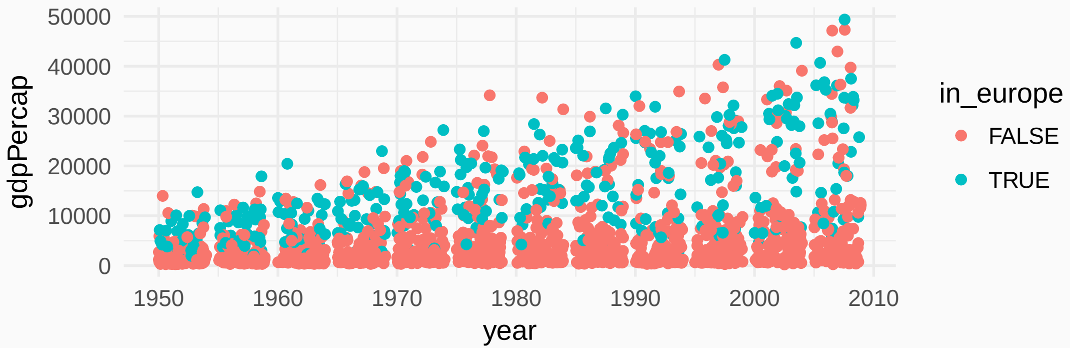


Europe?

```
gapminder$in_europe <- gapminder$continent == "Europe"
```

Europe?

```r
ggplot(gapminder, aes(year, gdpPercap, color = in_europe)) +
  geom_jitter()
```



INTERACTION TERMS

```r
summary(lm(gdpPercap ~ year * in_europe, data = gapminder))
```

```
##
## Call:
## lm(formula = gdpPercap ~ year * in_europe, data = gapminder)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -17593  -3767  -1574   1857  39731
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -190268.52   20971.87  -9.073   <2e-16 ***
```