Section G
Regression: Interpretation and Confounds
Lectures 15 and 16

Michael F. Seese

Department of Political Science
University of California San Diego

Political Science 30, Week 8

Outline

Interpreting OLS Output

Dummy Variables

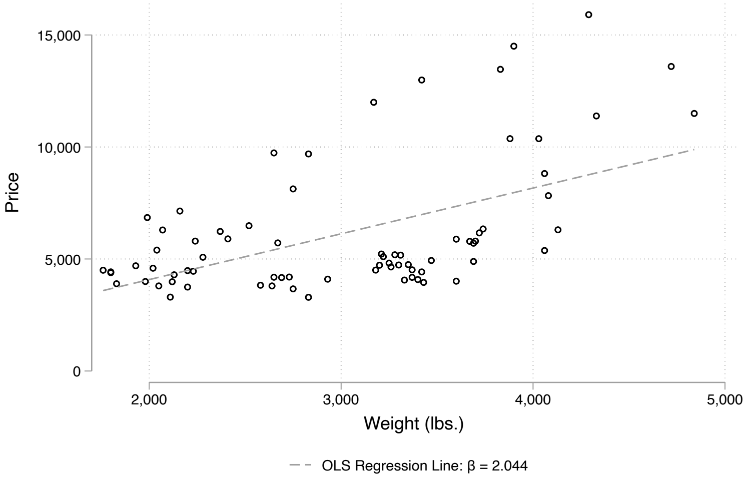
Confounds

Interpreting OLS Output

- Let's return to the "auto" dataset from Stata
- We have a theory that bigger cars are more expensive than smaller cars
- We test this theory by looking at two variables:
    1. Price
    2. Weight
- So we model price as a function of weight: $\text{Price} = \alpha + \beta(\text{Weight})$
- And we state our hypotheses:

  H_0 $\beta = 0$
  H_1 $\beta > 0$

Interpreting OLS Output



Interpreting OLS Output

```
. reg price weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     29.42
       Model |   184233937         1   184233937   Prob > F        =    0.0000
    Residual |   450831459        72  6261548.04   R-squared       =    0.2901
-------------+----------------------------------   Adj R-squared   =    0.2802
       Total |   635065396        73  8699525.97   Root MSE        =    2502.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   2.044063   .3768341     5.42   0.000     1.292857    2.795268
       _cons |  -6.707353    1174.43    -0.01   0.995     -2347.89    2334.475
------------------------------------------------------------------------------
```

Interpreting OLS Output

```
. reg price weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     29.42
       Model |  184233937         1  184233937   Prob > F        =    0.0000
    Residual |  450831459        72  6261548.04   R-squared       =    0.2901
-------------+----------------------------------   Adj R-squared   =    0.2802
       Total |  635065396        73  8699525.97   Root MSE        =    2502.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   2.044063   .3768341     5.42   0.000     1.292857    2.795268
       _cons |  -6.707353    1174.43    -0.01   0.995     -2347.89    2334.475
------------------------------------------------------------------------------
```

Interpreting OLS Output

$$\widehat{\text{Price}} = -6.707353 + (2.044063)(\text{Weight})$$

- Our y intercept, or constant, is -6.707
  - This tells us the price of a car when the weight of the car is 0
  - Is the intercept informative in this particular case?
- The slope, or coefficient, is 2.044
  - This tells us how much (in \$) the price increases for each additional lb of weight
  - A car weighs 3,000 lbs. How much more will a car that weighs 3,100 lbs cost?

$$\$2.04 \times 100 = \$204$$

Interpreting OLS Output

$$\widehat{\text{Price}} = -6.707353 + (2.044063)(\text{Weight})$$

- Our y intercept, or constant, is -6.707
  - This tells us the price of a car when the weight of the car is 0
  - Is the intercept informative in this particular case?
- The slope, or coefficient, is 2.044
  - This tells us how much (in \$) the price increases for each additional lb of weight
  - A car weighs 3,000 lbs. How much more will a car that weighs 3,100 lbs cost?

    $\$2.04 \times 100 = \$204$

Interpreting OLS Output

- Recall our hypotheses:

  H_0 $\beta = 0$

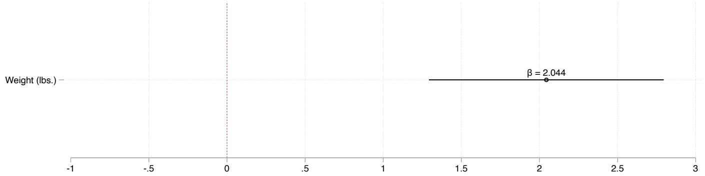
  H_1 $\beta > 0$

- Can we reject the null, based on the 95% CI Stata calculated for us?

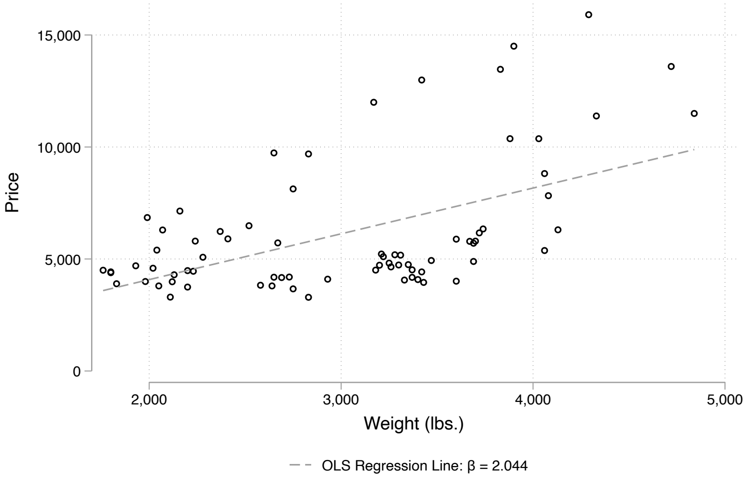
price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	2.044063	.3768341	5.42	0.000	1.292857	2.795268
_cons	-6.707353	1174.43	-0.01	0.995	-2347.89	2334.475

Interpreting OLS Output

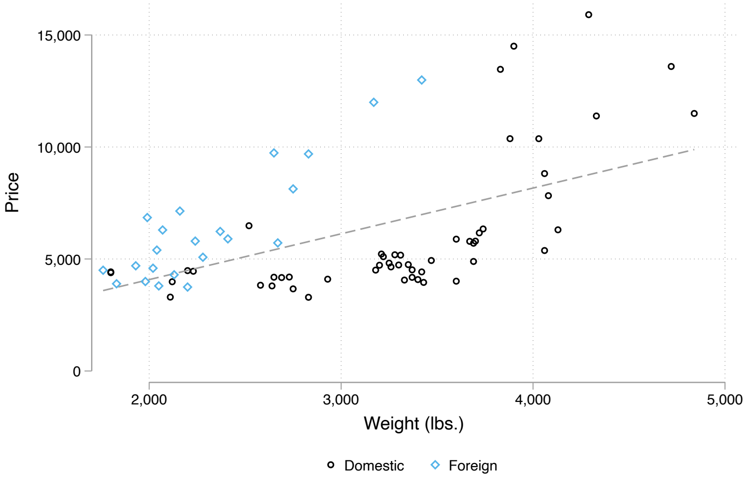
- Recall our hypotheses:
  H_0 $\beta = 0$
  H_1 $\beta > 0$
- Can we reject the null, based on the 95% CI Stata calculated for us?



Dummy Variables



Dummy Variables



Dummy Variables

- We have a theory that bigger cars are more expensive than smaller cars
- But we think that shipping costs and tariffs may also influence price
- We test this updated theory by looking at three variables:
    1. Price
    2. Weight
    3. Import Status $\longleftarrow$ Dummy Variable
- So we model price as a function of both weight *and* import status:
  $\text{Price} = \alpha + \beta_1(\text{Weight}) + \beta_2(\text{Import Status})$
- And we state our hypotheses:

  H_0 $\beta_1 = 0$
  H_1 $\beta_1 > 0$

Dummy Variables

```
. reg price weight foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     35.35
       Model |   316859273         2   158429637   Prob > F        =    0.0000
    Residual |   318206123        71  4481776.38   R-squared       =    0.4989
-------------+----------------------------------   Adj R-squared   =    0.4848
       Total |   635065396        73  8699525.97   Root MSE        =      2117

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.320737   .3958784     8.39   0.000     2.531378    4.110096
     foreign |   3637.001    668.583     5.44   0.000     2303.885    4970.118
       _cons |  -4942.844   1345.591    -3.67   0.000    -7625.876   -2259.812
------------------------------------------------------------------------------
```

Dummy Variables

$$\widehat{\text{Price}} = -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status})$$

- Our y intercept, or constant, is -4942.844
  - This tells us the price of a car when the weight of the car is 0 and when the car is built domestically (i.e., $\texttt{foreign} = 0$)
- The coefficient for Weight (β_1) is 3.320
  - This tells us how much (in \$) the price increases for each additional lb of weight, holding import status constant
- The coefficient for Import Status (β_2) is 3637.001
  - This tells us how much (in \$) the price increases when you move from a domestic car ($\texttt{foreign} = 0$) to a foreign car ($\texttt{foreign} = 1$), holding weight constant

Dummy Variables

$$\widehat{\text{Price}} = -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status})$$

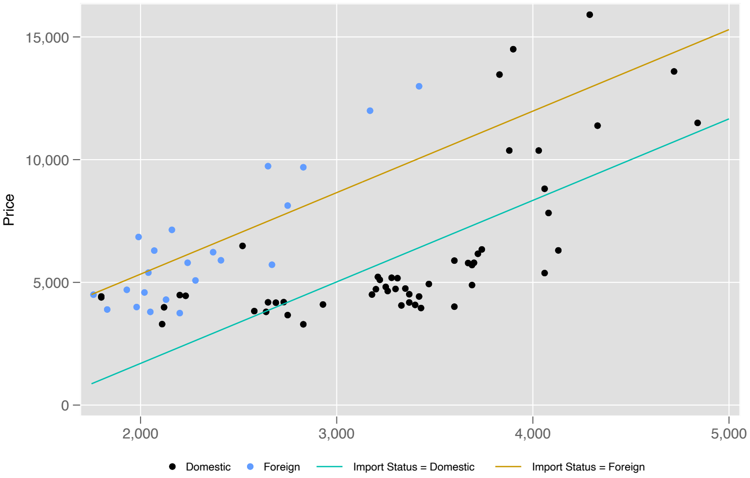
- Our y intercept, or constant, is -4942.844
  - This tells us the price of a car when the weight of the car is 0 and when the car is built domestically (i.e., $\texttt{foreign} = 0$)
- The coefficient for Weight (β_1) is 3.320
  - This tells us how much (in \$) the price increases for each additional lb of weight, holding import status constant
- The coefficient for Import Status (β_2) is 3637.001
  - This tells us how much (in \$) the price increases when you move from a domestic car ($\texttt{foreign} = 0$) to a foreign car ($\texttt{foreign} = 1$), holding weight constant

Dummy Variables

$$\widehat{\text{Price}} = -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status})$$

- Our y intercept, or constant, is -4942.844
  - This tells us the price of a car when the weight of the car is 0 and when the car is built domestically (i.e., `foreign` = 0)
- The coefficient for Weight (β_1) is 3.320
  - This tells us how much (in \$) the price increases for each additional lb of weight, holding import status constant
- The coefficient for Import Status (β_2) is 3637.001
  - This tells us how much (in \$) the price increases when you move from a domestic car (`foreign` = 0) to a foreign car (`foreign` = 1), holding weight constant

Dummy Variables



Dummy Variables

Why does the line move based on import status?

$$\begin{aligned}
\widehat{\text{Price}} &= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status}) \\
&= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(0) \\
&= -4942.844 + (3.320737)(\text{Weight})
\end{aligned}$$

But...

$$\begin{aligned}
\widehat{\text{Price}} &= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status}) \\
&= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(1) \\
&= -4942.844 + (3.320737)(\text{Weight}) + 3637.001 \\
&= (-4942.844 + 3637.001) + (3.320737)(\text{Weight}) \\
&= -1305.843 + (3.320737)(\text{Weight})
\end{aligned}$$

Dummy Variables

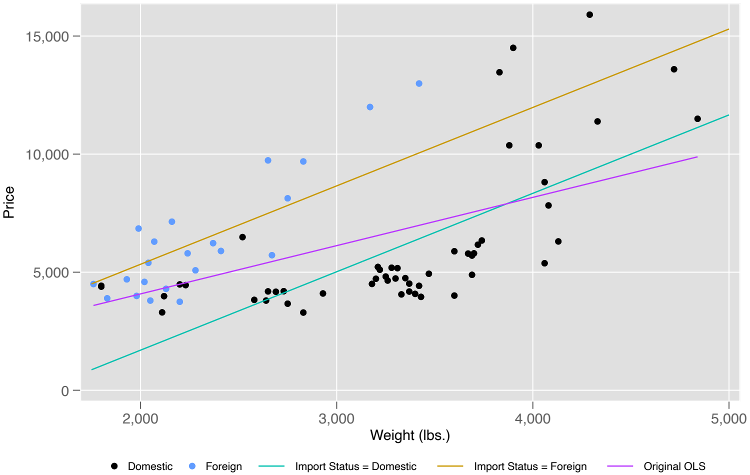
Why does the line move based on import status?

$$
\begin{aligned}
\widehat{\text{Price}} &= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status}) \\
&= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(0) \\
&= -4942.844 + (3.320737)(\text{Weight})
\end{aligned}
$$

But...

$$
\begin{aligned}
\widehat{\text{Price}} &= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(\text{Import Status}) \\
&= -4942.844 + (3.320737)(\text{Weight}) + (3637.001)(1) \\
&= -4942.844 + (3.320737)(\text{Weight}) + 3637.001 \\
&= (-4942.844 + 3637.001) + (3.320737)(\text{Weight}) \\
&= -1305.843 + (3.320737)(\text{Weight})
\end{aligned}
$$

Dummy Variables



Confounds

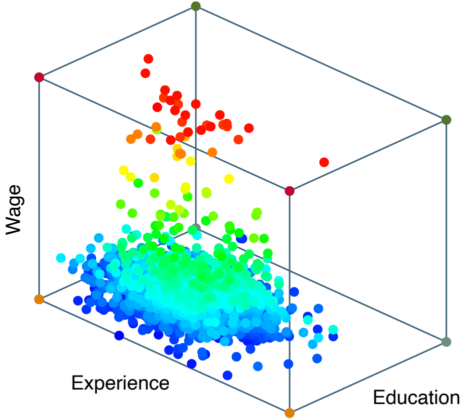
- Let's return to the labor market data we saw last week
- We have a theory that more work experience (in years) will tend to correlate with a higher hourly wage (in \$)
- But education may be a confounding factor
  - More years of education may increase wages (high skilled labor)
  - May also imply less work experience, as individuals may delay labor force entry while completing education

Confounds

So we model wage as a function of experience and education:

$$\text{Wage} = \alpha + \beta_1(\text{Experience}) + \beta_2(\text{Education})$$

Confounds



Confounds

```
. reg wage experience education

      Source |       SS           df       MS      Number of obs   =     2,244
-------------+----------------------------------   F(2, 2241)      =    194.77
       Model |     11010.6         2      5505.3   Prob > F        =    0.0000
    Residual |  63343.7305     2,241  28.2658325   R-squared       =    0.1481
-------------+----------------------------------   Adj R-squared   =    0.1473
       Total |  74354.3305     2,243  33.1495009   Root MSE        =    5.3166

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  experience |   .2616056   .0248373    10.53   0.000     .2128992     .310312
   education |   .6483343    .045426    14.27   0.000     .5592528    .7374158
       _cons |  -4.002059   .6245962    -6.41   0.000    -5.226906   -2.777211
------------------------------------------------------------------------------
```

Confounds

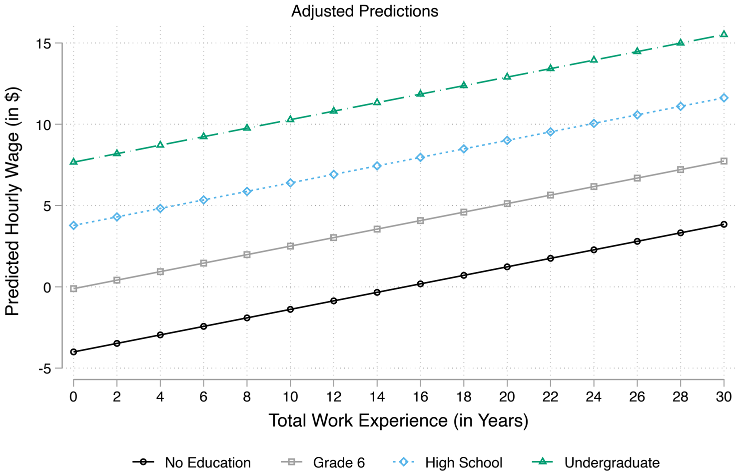
Here's our regression equation:

$$\widehat{\text{Wage}} = -4.002 + (0.261)(\text{Experience}) + (0.648)(\text{Education})$$

What does it mean to hold a variable constant?

$$\widehat{\text{Wage}} = -4.002 + (0.261)(\{0, 1, 2, \ldots, 30\}) + (0.648)(6)$$

Confounds


Adjusted Predictions
15
10
5
0
-5
Predicted Hourly Wage (in $)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
Total Work Experience (in Years)
No Education
Grade 6
High School
Undergraduate