Multiple Regression: Model and Interpretation

Lecture 20

Reading: Sections 20.1 – 20.3

1

Magical Regression – Oops I Mean – Multiple Regression

- Are the salaries of female professors in ON unfairly below males?
  - y-variable? x-variable?
  - Even if lower salaries, maybe females are less experienced, in lower-paid disciplines, less productive, etc.
    - Multiple x-variables: sex, experience, discipline, …
- Multiple regression allows us to control for experience, discipline, productivity, to isolate the effect (if any) of sex
  - With observational data, we *may* be able to tackle lurking/unobserved/ omitted/confounding variables by controlling for them

2

Multiple Regression: Today and Rest of ECO220Y

- Much translates from simple to multiple regression
  - E.g. t test, CI est. of coef.
- But, two <u>*big exceptions*</u>:
  - Interpreting coefficients (today)
  - Testing *overall* statistical significance (next week, F test)
- Final few weeks: building realistic multiple regression models
  - Dummy variables for categorical information (e.g. sex, program of study, discipline of research)
    - Also, wrt panel data
  - Interaction terms

3

Multiple Regression Model

- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$
  - How many explanatory (x) variables?
  - What is the interpretation of the error (ε_i)?
- OLS estimate solves $\min_{b_0,\ldots,b_k} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$:

  $\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$
  - No simple formula for coefficients: need software
  - Residuals $e_i = y_i - \hat{y}_i$ and $s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}}$

4

Recall Six Assumptions

1) Linearity: each x linearly related to y (x variables and/or y variable can be non-linearly transformed)
2) Errors independent (common problem: autocorrelation in time series data)
3) Homoscedasticity (single variance) of errors
4) Normally distributed errors
5) Constant included (error has mean 0)
6) Each x and error unrelated; i.e. no lurking variables

5

HT and CI Estimation for Slopes

- $H_0: \beta_j = \beta_j^0$
- $H_1: \beta_j \neq \beta_j^0$ (or $>$ or $<$)
  - Use $t = \frac{b_j - \beta_j^0}{s_{b_j}}$ with $v = n - k - 1$
- Statistical significance:

  $H_0: \beta_j = 0$

  $H_1: \beta_j \neq 0$

For $H_0: \beta_2 = 1$ versus $H_1: \beta_2 > 1$
$v = 54$ and $t = \frac{2.124 - 1}{0.357} = 3.15$

Conclusion?

- Standard error of slope coef., s_{b_j} or $SE[b_j]$, obtain from software like slope coef. itself
- CI estimate of β_j:

  $b_j \pm t_{\alpha/2} s_{b_j}$

  with $v = n - k - 1$

Continuing, for 95% CI get $2.124 \pm 2.005 * 0.357$ and LCL = 1.41 and UCL = 2.84

Conclusion?

6

How to Interpret Coefficients?

- Q: If Assumptions 1 – 5 hold, the coefficient is statistically significant, and the model overall is statistically significant (Lecture 21), how to interpret multiple regression coefficient b_j?
- A: b_j measures the average change in y associated with a change in x_j *holding the other included x variables fixed* (i.e. *after controlling for the other included x variables)*

As usual, interpretations also require: being context-specific, specifying units of measurement & being clear about causality 7

Predicting Males' Percent Body Fat

	Coeff	SE(Coeff)	t-ratio	P-value
Intercept	57.272	10.399	5.51	<0.0001
Height	-0.502	0.059	-8.06	<0.0001
Weight	0.558	0.033	17.11	<0.0001
Age	0.137	0.028	4.90	<0.0001
N	250			
R^2	0.584			

Source: Our textbook, *Just Checking*, p. 695 8

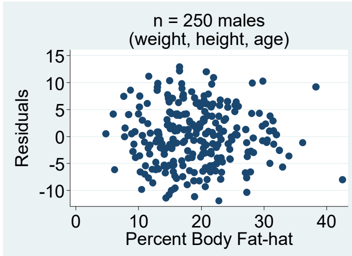
STATA Output: Percent Body Fat

```
. regress pct_body_fat height_cm weight_kg age if (case~=39 & case~=42);

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  3,   246) =  115.13
       Model |  10003.7809     3  3334.59362           Prob > F      =  0.0000
    Residual |  7125.03917   246  28.9635738           R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5790
       Total |    17128.82   249  68.7904419           Root MSE      =  5.3818

------------------------------------------------------------------------------
pct_body_fat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   height_cm |  -.5016358   .0622096    -8.06   0.000    -.6241671   -.3791045
   weight_kg |    .559226   .0326851    17.11   0.000     .4948477    .6236043
         age |   .1373248   .0280566     4.89   0.000      .082063    .1925866
       _cons |   57.27217   10.39897     5.51   0.000      36.7898    77.75454
------------------------------------------------------------------------------
```

http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html 9

Standard Deviation of Residuals

- Assumed $\varepsilon_i \sim N(0, \sigma^2)$: ε_i unknowable but we can compute e_i and its standard deviation
  - $s_e = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i - 0)^2}{n-k-1}}$
  - Roughly, what is s_e based on the graph?



No scale? Measure waist

```
. regress pct_body_fat height_cm abdomen_cm age if (case~=39 & case~=42);

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  3,   246) =  205.79
       Model |  12248.3786     3  4082.79287           Prob > F      =  0.0000
    Residual |  4880.44142   246  19.8391928           R-squared     =  0.7151
-------------+------------------------------           Adj R-squared =  0.7116
       Total |    17128.82   249  68.7904419           Root MSE      =  4.4541

------------------------------------------------------------------------------
pct_body_fat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   height_cm |  -.2192569   .0454017    -4.83   0.000    -.3086826   -.1298313
  abdomen_cm |   .6867277   .0295381    23.25   0.000     .6285478    .7449076
         age |   .0305884   .0241531     1.27   0.207    -.0169847    .0781616
       _cons |  -6.564603   8.149381    -0.81   0.421    -22.61606    9.486858
------------------------------------------------------------------------------
```

Note: Do NOT drop variables from your model simply because they are not statistically significant.

Gain Weight to Reduce Body Fat?

What does the negative (and statistically significant) coefficient on weight_kg mean?

```
. regress pct_body_fat height_cm abdomen_cm age weight_kg if (case~=39 &
case~=42);

      Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  4,   245) =  161.49
       Model |  12418.7119     4  3104.67799           Prob > F      =  0.0000
    Residual |  4710.10808   245  19.2249309           R-squared     =  0.7250
-------------+------------------------------           Adj R-squared =  0.7205
       Total |    17128.82   249  68.7904419           Root MSE      =  4.3846

------------------------------------------------------------------------------
pct_body_fat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   height_cm |  -.0884853   .0626709    -1.41   0.159    -.2119279    .0349572
  abdomen_cm |   .9133218   .0814899    11.21   0.000     .7528116    1.073832
         age |  -.0003596   .0259501    -0.01   0.989    -.0514734    .0507542
   weight_kg |  -.2221385   .0746288    -2.98   0.003    -.3691343   -.0751426
       _cons |  -31.49531   11.59772    -2.72   0.007    -54.33927   -8.651341
------------------------------------------------------------------------------
```

Guided Example, pp. 699 – 703

- Use multiple regression to predict house prices ($'s) with living area (sq. ft.), number of bedrooms, number of bathrooms, age of house (years), and the number of fireplaces
  - Check underlying conditions: see textbook
  - What is goal: describe data, forecasting, model?
  - Which kind of data are these?
  - Will Assumption #6 be violated? If yes, give a concrete example of a lurking variable?

13

Stata Output: Housing Prices

```
. regress price livingarea bedrooms bathrooms fireplaces age;

      Source |       SS       df       MS              Number of obs =    1057
-------------+------------------------------           F(  5,  1051) =  321.79
       Model |  3.8028e+12     5  7.6055e+11           Prob > F      =  0.0000
    Residual |  2.4840e+12  1051  2.3635e+09           R-squared     =  0.6049
-------------+------------------------------           Adj R-squared =  0.6030
       Total |  6.2868e+12  1056  5.9534e+09           Root MSE      =   48616

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  livingarea |    73.4464   4.008868    18.32   0.000      65.5801     81.3127
    bedrooms |  -6361.311   2749.503    -2.31   0.021    -11756.45   -966.1715
   bathrooms |   19236.68    3669.08     5.24   0.000     12037.12    26436.23
  fireplaces |   9162.791   3194.233     2.87   0.004     2894.991    15430.59
         age |  -142.7395   48.27612    -2.96   0.003     -237.468   -48.01094
       _cons |    15712.7   7311.427     2.15   0.032     1366.047    30059.36
------------------------------------------------------------------------------
```

Interpretations?

14

Without *livingarea*

```
. regress price bedrooms bathrooms fireplaces age;

      Source |       SS       df       MS              Number of obs =    1057
-------------+------------------------------           F(  4,  1052) =  241.50
       Model |  3.0094e+12     4  7.5235e+11           Prob > F      =  0.0000
    Residual |  3.2774e+12  1052  3.1154e+09           R-squared     =  0.4787
-------------+------------------------------           Adj R-squared =  0.4767
       Total |  6.2868e+12  1056  5.9534e+09           Root MSE      =   55816

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bedrooms |   18080.69   2760.189     6.55   0.000     12664.58    23496.79
   bathrooms |   53635.29   3619.122    14.82   0.000     46533.77    60736.81
  fireplaces |   27142.71   3489.901     7.78   0.000     20294.75    33990.67
         age |  -124.8008   55.41405    -2.25   0.025    -233.5355   -16.06615
       _cons |  -6557.804   8277.368    -0.79   0.428    -22799.83    9684.227
------------------------------------------------------------------------------
```

Where did *livingarea* go? Why did the s_e increase?

Bedrooms associated with ε, violating Assumption 6?

15

Recall: Experimental Drug Data

```
regress hrs_sleep dosage;

   Source |       SS       df       MS              Number of obs =      25
----------+------------------------------           F(  1,     23) =   13.89
    Model |  12.6255781     1  12.6255781           Prob > F      =  0.0011
 Residual |  20.9040126    23  .908870111           R-squared     =  0.3766
----------+------------------------------           Adj R-squared =  0.3494
    Total |  33.5295906    24  1.39706628           Root MSE      =  .95335

---------------------------------------------------------------------------
hrs_sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+----------------------------------------------------------------
   dosage |   .4816382   .1292249     3.73   0.001     .2143161    .7489602
    _cons |   3.439461   .6260549     5.49   0.000     2.144368    4.734555
---------------------------------------------------------------------------
```

16

Interpretation w/ Experimental Data

```
regress hrs_sleep dosage age weight;

   Source |       SS       df       MS              Number of obs =      25
----------+------------------------------           F(  3,     21) =    7.67
    Model |   17.528649     3  5.84288299           Prob > F      =  0.0012
 Residual |  16.0009417    21  .761949603           R-squared     =  0.5228
----------+------------------------------           Adj R-squared =  0.4546
    Total |  33.5295906    24  1.39706628           Root MSE      =   .8729

---------------------------------------------------------------------------
hrs_sleep |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------+----------------------------------------------------------------
   dosage |   .5094999   .1208007     4.22   0.000     .2582811    .7607187
      age |  -.0213827   .0131737    -1.62   0.119    -.0487789    .0060134
   weight |  -.0342918   .0164732    -2.08   0.050    -.0685497   -.0000338
    _cons |   7.005249   1.528731     4.58   0.000     3.826078    10.18442
---------------------------------------------------------------------------
```

Unlike wild swings in housing regression w/ & w/o living area, dosage coefficient is stable. Age and weight were *not* lurking/ unobserved/confounding/omitted variables: dosage coef. is NOT biased regardless of whether you control for age and weight. 17

Returns to Consumer Search: Evidence from eBay

- Research question: How much does spending time searching affect the final price that a consumer pays for a good?
  - "We assemble a dataset of search and purchase behavior from eBay to quantify the returns to consumer search on the internet." (from Abstract)
  - Will data be observational or experimental?
  - What is the x variable? y variable?
  - Do you expect a positive or negative relationship?

"Returns to Consumer Search: Evidence from eBay" *NBER Working* Paper, June 2016 http://www.nber.org/papers/w22302.pdf

18

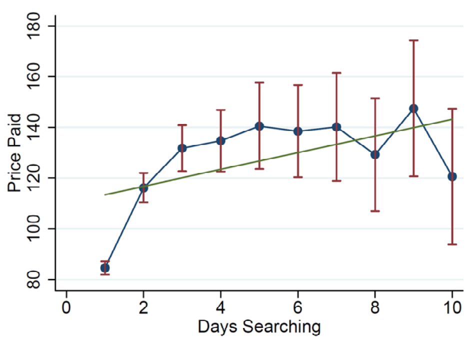
EXCERPT, p. 16: We identified all purchasers on an arbitrary date, July 27th, 2014. We then limited the sample to purchases of common and well defined goods … This allowed us to construct a distribution of prices for each of the goods in our sample.

Next we identified all search behavior of the buyer in the 6 weeks prior to the purchase. A challenge is to identify searches related to the product purchased, knowing that the queries over time may have changed due to refinements of all sorts. To do this, we first counted the number of searches that returned items which are identified as being the exact same product that was eventually purchased. …. We then identified the length of search as the time between the first search and purchase as another measure of search intensity. Finally, we counted the number of distinct days on which the user searched for the product.

So how many different variables measure the key x-variable (how much a consumer searched)?

19

EXCERPT, p. 17: Using the data we collected we explore the relationship between measures of prices paid and of search intensity, which are displayed in Figure 5. It shows the mean price paid for the different levels of the indicated search intensity (days searching, days since first search, and the number of searches).

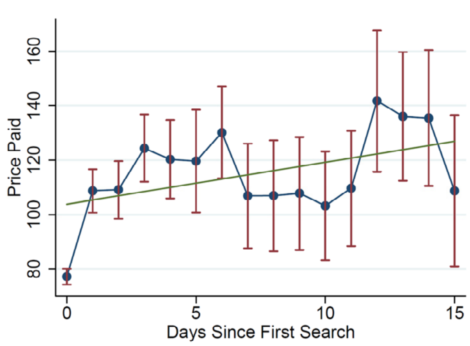


What is the green line?

Intervals around each mean (dot) show SEs or MEs (not specified). Why are they wider to the right?

20

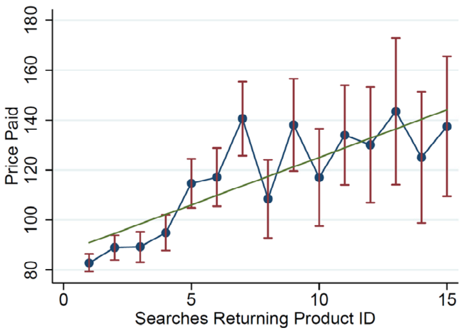
EXCERPT, p. 17: Using the data we collected we explore the relationship between measures of prices paid and of search intensity, which are displayed in Figure 5. It shows the mean price paid for the different levels of the indicated search intensity (days searching, days since first search, and the number of searches).



What is the point of this second graph? What is different from first?

21

EXCERPT, p. 17, cont'd: There is generally a positive relationship between price and search, which at first glance may be surprising. However, this does not control for the product purchased. Users presumably spend more time searching for costlier purchases because they expect to get a larger absolute value of savings from additional searches. Hence, this should not be interpreted as a causal relationship but rather one driven by selection.



What's the key lurking (unobserved, confounding, omitted) variable?

Control for Costly Purchases

- Multiple regression can control for costly purc. to help w/ endogenity of search intensity (its coefficient suffers *severe* endogeneity bias)
  - Remove variables from ε that are correlated w/ search (violating Assump. #6) by adding them as control variables (additional RHS variables)
    - "We computed the *expected product price* by taking the mean of all of the purchases of a given product in the 6 weeks prior. We treat this as the expected price one would pay for a product without search" (p. 17)

Unit of observation? Cross-sectional, time series or panel data?

Table 2: Quantifying Returns to Search

	y-variable: *Price Paid (US$)*			y-variable: *Ln(Price Paid)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Searches Returning Product ID	-0.264 (0.031)	-0.088 (0.034)	0.059 (0.054)	-0.0033 (0.0003)	-0.0012 (0.0004)	0.0004 (0.0006)
Days Since First Search		-0.317 (0.027)	-0.272 (0.030)		-0.0040 (0.0003)	-0.0035 (0.0003)
Days Searching			-0.759 (0.217)			-0.0082 (0.0023)
Product Expected Price	0.884 (0.002)	0.886 (0.002)	0.886 (0.002)	"Each additional day spent searching yields a 0.8% or 75 cents savings." p. 18		
Ln(Product Expected Price)	"Each additional search is associated with a 26 cent reduction in the price." p. 18			1.015 (0.003)	1.020 (0.003)	1.020 (0.003)
Constant	0.492 (0.469)	2.040 (0.484)	2.447 (0.498)	-0.260 (0.011)	-0.258 (0.011)	-0.254 (0.011)
Observations	14,331	14,331	14,331	14,331	14,331	14,331

Notes: Reports six separate regressions. Standard errors in parentheses.