Interpretation (Again), Overall Statistical Significance, and Fit of a Multiple Regression Model

Lecture 21

Reading: Sections 20.4 – 20.6

1

Interpretation: Nice Review (p. 704)

- "The multiple regression model looks so simple and straightforward. It *looks* like each coefficient tells us the effect of its associated predictor, x_j, on the response variable, y. But that's not true."
- "If we fail to reject the null hypothesis for a multiple regression coefficient, it does *not* mean that the corresponding predictor variable has no linear relationship to y. It means that the corresponding predictor contributes nothing to modeling *after allowing for all the other predictors*."

2

"What Explains the Flow of Foreign Fighters to ISIS"
NBER Working Paper, April 2016
http://www.nber.org/digest/jun16/w22190.html

ABSTRACT: This paper provides the first systematic analysis of the link between economic, political, and social conditions and the global phenomenon of ISIS foreign fighters. We find that poor economic conditions do not drive participation in ISIS. In contrast, the number of ISIS foreign fighters is positively correlated with a country's GDP per capita and Human Development Index (HDI). In fact, many foreign fighters originate from countries with high levels of economic development, low income inequality, and highly developed political institutions. Other factors that explain the number of ISIS foreign fighters are the size of a country's Muslim population and its ethnic homogeneity. Although we cannot directly determine why people join ISIS, our results suggest that the flow of foreign fighters to ISIS is driven not by economic or political conditions but rather by ideology and the difficulty of assimilation into homogeneous Western countries.

Observational or experimental data? y variable? x variables?

3

Table 6: Summary Statistics

	Mean	25th Perc.	Med.	75th Perc.	S.d.	Min.	Max.	Obs.
# ISIS fighters	164.3	0	0	57	594.8	0	6,000	173
Population_{2014}	36.7	1.8	7.1	23.6	139.8	0.1	1,364	193
% Muslim	24.2%	0.0%	2.7%	36.7%	36.4%	0.0%	99.9%	192
$\text{GDP per capita}_{2010}$	$14,404	$1,419	$5,056	$15,901	$22,633	$214	$145,221	193
Unemployment	8.6%	4.7%	7.6%	10.5%	5.7%	0.4%	32.0%	164
Distance to Syria (in km)	5,961	2,737	4,753	9,444	4,082	84	16,651	193
Political Rights	3.33	1	3	5	2.12	1	7	184
Ethnic Fractionalization	0.44	0.20	0.43	0.67	0.26	0	0.93	179

Notes: This table provides summary statistics for the main variables used in the paper. See main body of the manuscript for a detailed description of these sources.

4

EXCERPT, p. 6: We also include in our analysis indices for ethnic, linguistic, and religious fractionalization. These indices were built in Alesina et al. (2003) and have been updated every year since by the Quality of Government Institute at the University of Gothenburg. The indices calculate the probability that two randomly selected individuals from a given country will not share the same ethnicity, language, and religion. The indices show a great deal of variation among the countries in our sample. Korea, Japan, and Portugal are examples of countries with very low ethnic and linguistic fractionalization. Muslim countries tend to have low levels of religious fractionalization (for example, Algeria, Morocco, and Turkey are all below 0.01), whereas Australia, the United States, and South Africa are the three countries with the highest levels of religious fractionalization (their levels are 0.821, 0.824, and 0.86, respectively).

5

Tables 8 and 9: The Determinants of the Number of ISIS Foreign Fighters

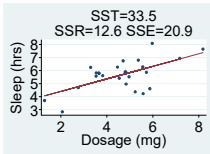
Dependent variable:	*log(1 + # ISIS foreign fighters)*		*log(# ISIS foreign fighters)*	
Explanatory variables:	(1)	(2)	(3)	(4)
$\text{Log(population)}_{2014}$	0.126	0.129	-0.281	-0.412***
	(0.113)	(0.109)	(0.176)	(0.190)
$\text{Log(Muslim population)}_{2010}$	0.417***	0.456***	0.718***	0.811***
	(0.066)	(0.065)	(0.099)	(0.118)
$\text{Log(GDP per capita)}_{2010}$	0.719***	0.663***	0.525***	0.359*
	(0.086)	(0.108)	(0.123)	(0.208)
Unemployment	0.065***	0.078***	0.064	0.066*
	(0.027)	(0.025)	(0.043)	(0.036)
Log(Distance to Syria)	-0.458*	-0.287	-0.228	-0.089
	(0.235)	(0.232)	(0.203)	(0.230)
Political Rights		0.163*		-0.030
		(0.086)		(0.145)
Ethnic Fractionalization		-2.409***		-2.589***
		(0.640)		(0.907)
R^2	0.581	0.640	Not reported	Not reported
Observations	143	141	61	60

Notes: The reported coefficients are from OLS regressions. Robust standard errors are in parentheses. *, **, *** denote statistical significance at 10%, 5% , and 1% levels, respectively.

Unit of observation? Cross-sectional, time series or panel data? 6

Recall: Analysis of Variance (Lec. 5)

- Analysis of Variance (ANOVA): How total variability of the y variable is related to the x variables versus the error term
  - Total sum of squares: $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - Regression sum of squares: $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - Sum of squared errors: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - $SST = SSR + SSE$



Meaning of $\sqrt{\frac{33.5}{25-1}} = 1.18$?

7

Recall ANOVA Table in STATA

```
. regress ln_elec_mmbtu ln_sq_feet cool_deg_days ln_num_res;

      Source |       SS           df       MS      Number of obs   =    14,044
-------------+----------------------------------   F(3, 14040)     =   2086.95
       Model |  1133.03414         3  377.678047   Prob > F        =    0.0000
    Residual |  2540.83253    14,040  .180970978   R-squared       =    0.3084
-------------+----------------------------------   Adj R-squared   =    0.3083
       Total |  3673.86668    14,043  .261615515   Root MSE        =    .42541

------------------------------------------------------------------------------
ln_elec_mmbtu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ln_sq_feet |   .4942341   .0090101    54.85   0.000     .4765731    .5118952
cool_deg_days |   .0369439    .001031    35.83   0.000     .0349231    .0389647
   ln_num_res |   .2534081   .0069756    36.33   0.000     .2397349    .2670813
         _cons |  -1.046515    .066923   -15.64   0.000    -1.177692   -.9153367
------------------------------------------------------------------------------
```

In a simple regression with just ln_sq_feet as x variable ($k = 1$), how would the SST differ? SSE? SSR?

Recall California energy data from Slides 2 and 16 of Lecture 19.

8

R^2: A Measure of a Model's Fit

- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- The R^2 measures what fraction of the total variation in y variable that's explained by variation in x variables
- $(1 - R^2)$: fraction of unexplained variation in y (explained by error)
- To interpret, lay it out in plain English, in context
  - e.g. 72.5 percent of variation in percent body fat among 250 males is explained by variation in their height, abdominal circumference, age, and weight
  - In interpreting the R^2, mention the units?

9

Overall Test of Statistical Significance

- Overall, is the model statistically significant?
  - Are coefficients jointly statistically significant?
    - Can we reject no association at all?
  - $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
  - H_1: Not all the slope parameters are zero
  - The test statistic for this hypothesis test is the F test statistic: often call this test "the F-test"
  - This is <u>not</u> the same as doing k hypothesis tests, one for each coefficient

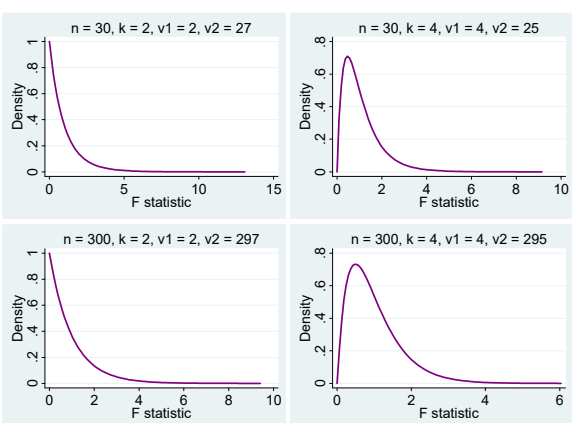
10

F statistic and its Distribution

- $F = \frac{(SST-SSE)/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$

  also $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$
  - Numerator degrees of freedom: $\nu_1 = k$
  - Denominator degrees of freedom: $\nu_2 = n - k - 1$
  - Can F be negative?
  - Want a big or small F for statistical significance?
- F distribution
  - Distribution tells how F statistic would vary given sampling error if y were entirely unrelated to the x variables
  - Continuous
  - Positively skewed
  - No density below zero
  - 2 parameters: ν_1 and ν_2

11



12

```
      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
------------------------------------------------------------------------------
```

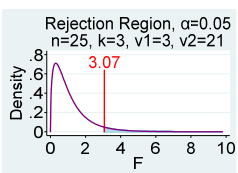
$$F = \frac{(SST - SSE)/k}{SSE/(n-k-1)} = \frac{(33.5296 - 16.0009)/3}{16.0009/(25-3-1)} = 7.67$$

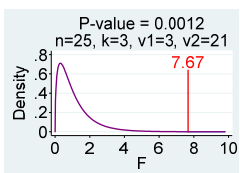
$$F = \frac{MSR}{MSE} = \frac{5.8429}{0.7619} = 7.67$$

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.5228/3}{(1-0.5228)/(25-3-1)} = 7.67$$

13

Recall F = 7.67





14

Dependent variable:	*log(1 + # ISIS foreign fighters)*
Explanatory variables:	(2)
Log(population)$_{2014}$	0.129 (0.109)
Log(Muslim population)$_{2010}$	0.456*** (0.065)
Log(GDP per capita)$_{2010}$	0.663*** (0.108)
Unemployment	0.078*** (0.025)
Log(Distance to Syria)	-0.287 (0.232)
Political Rights	0.163* (0.086)
Ethnic Fractionalization	-2.409*** (0.640)
R^2	0.640
Observations	141

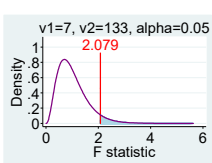
Is Specification (2) statistically significant?

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

H_1: Not all betas are zero

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{0.640/7}{(1-0.640)/133} = 33.8$$

Is an F test statistic of 33.8 big enough to reject H_0?



15

California Energy Regression Again

```
. regress ln_elec_mmbtu ln_sq_feet cool_deg_days ln_num_res;

      Source |       SS           df       MS      Number of obs   =    14,044
-------------+----------------------------------   F(3, 14040)     =   2086.95
       Model |  1133.03414         3  377.678047   Prob > F        =    0.0000
    Residual |  2540.83253    14,040  .180970978   R-squared       =    0.3084
-------------+----------------------------------   Adj R-squared   =    0.3083
       Total |  3673.86668    14,043  .261615515   Root MSE        =    .42541

------------------------------------------------------------------------------
ln_elec_mmbtu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
   ln_sq_feet |  .4942341   .0090101    54.85   0.000     .4765731    .5118952
cool_deg_days |  .0369439    .001031    35.83   0.000     .0349231    .0389647
   ln_num_res |  .2534081   .0069756    36.33   0.000     .2397349    .2670813
        _cons | -1.046515    .066923   -15.64   0.000    -1.177692   -.9153367
------------------------------------------------------------------------------
```

Can we reject $H_0: \beta_1 = \beta_2 = \beta_3 = 0$? In other words, is the regression statistically significant overall? Where to look above?

Does a highly statistically significant regression mean a good fit?

16

Recall Regression from p. 695

```
regress pct_body_fat height_cm age weight_kg if (case_number~=39 &
case_number~=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
         age |   .1373248   .0280566     4.89   0.000      .082063    .1925866
   weight_kg |    .559226   .0326851    17.11   0.000     .4948477    .6236043
       _cons |   57.27217   10.39897     5.51   0.000      36.7898    77.75454
------------------------------------------------------------------------------
```

17

Statistically Significant Correlations?

```
correlate pct_body_fat height_cm abdomen_cm age weight_kg if
(case_number~=39 & case_number~=42)

(obs=250)
             | pct_body_fat  height_cm  abdomen_cm      age  weight_kg
-------------+------------------------------------------------------
pct_body_fat |       1.0000
   height_cm |      -0.0294     1.0000
  abdomen_cm |       0.8237     0.1867      1.0000
         age |       0.2951    -0.2459      0.2428   1.0000
   weight_kg |       0.6173     0.5129      0.8737  -0.0161     1.0000
```

We can test for statistically significant correlations – the topic of Section 18.5 – by simply using an F test where k equals 1.

Statistically significant correlation btwn pct_body_fat & height?

Find $F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(-0.0294)^2/1}{(1-(-0.0294)^2)/(250-1-1)} = 0.21.$

The critical value for $\alpha = 0.10$ is about 2.7. Conclusion?

18

F test for Testing for Statistically Significant Correlations

- F test checks if a regression is stat. sig. *overall*
  - For *simple* regression, asks if 2 variables related
  - Convenient: *F* test stat. only requires R^2, n, and k
    - Simple regression: R^2 is correlation squared
    - t test (p. 617, textbook) same conclusion: redundant
  - F test reminds us that for *simple* regression these are *same*: (1) is the slope coefficient stat. sig.?, (2) is the regression model stat. sig. overall?, and (3) is the correlation between x and y stat. sig.?

19

Simple Regression: t and F Tests are Same

```
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
------------------------------------------------------------------------------
```

How can you see that the tests are the same in this output?

Fun fact: In *simple* regression the F test statistic is the t test statistic squared. (Note that $3.73^2 = 13.9$.)

20

Multiple Regression: use F, not t to test overall statistical significance

- With k separate t tests, k chances to make a Type 1 error
  - Type I error in test of "statistical significance"?
  - Too many chances for "statistical significance"
    - If have 100 x's that are <u>independent</u> of y, how many coefficients do we <u>expect</u> to be statistically significant if α = 0.05?
- With one F test we can fully control Type 1 error by picking the significance level

21

Silly Regression with Our Class List

```
. regress last_three let_fname let_lname chars_utorid;

      Source |       SS           df       MS      Number of obs   =       478
-------------+----------------------------------   F(3, 474)       =      0.55
       Model |  132234.146         3  44078.0488   Prob > F        =    0.6514
    Residual |  38299913.4       474  80801.5051   R-squared       =    0.0034
-------------+----------------------------------   Adj R-squared   =   -0.0029
       Total |  38432147.5       477  80570.5399   Root MSE        =    284.26

------------------------------------------------------------------------------
  last_three |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   let_fname |  -7.037891   6.700247    -1.05   0.294    -20.20375     6.12797
   let_lname |  -3.755335   5.136781    -0.73   0.465    -13.84901    6.338343
chars_utorid |   4.898899   30.89672     0.16   0.874    -55.81258    65.61038
       _cons |   503.7653   233.1496     2.16   0.031     45.63074    961.8999
------------------------------------------------------------------------------
```

22

Ridiculous Regression

```
. regress last_three let_fname let_lname chars_utorid s_utorid - z_utorid;

      Source |       SS           df       MS      Number of obs   =       478
-------------+----------------------------------   F(11, 466)      =      1.90
       Model |  1645991.58        11  149635.598   Prob > F        =    0.0378
    Residual |    36786156       466  78940.2489   R-squared       =    0.0428
-------------+----------------------------------   Adj R-squared   =    0.0202
       Total |  38432147.5       477  80570.5399   Root MSE        =    280.96

------------------------------------------------------------------------------
  last_three |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   let_fname |  -5.683719   6.706356    -0.85   0.397    -18.86216    7.494724
   let_lname |  -2.208591   5.359994    -0.41   0.680    -12.74134     8.32416
chars_utorid |   5.220551   30.85472     0.17   0.866    -55.41107    65.85217
    s_utorid |  -11.18485    7.04092    -1.59   0.113    -25.02073    2.651037
    t_utorid |   7.960974   7.885815     1.01   0.313    -7.535186    23.45713
    u_utorid |  -4.986156    7.37811    -0.68   0.500    -19.48464     9.51233
    v_utorid |  -6.223315   11.16217    -0.56   0.577    -28.15774    15.71111
    w_utorid |  -15.86895   10.37429    -1.53   0.127    -36.25512    4.517227
    x_utorid |  -8.006023   9.271085    -0.86   0.388    -26.22433    10.21229
    y_utorid |   11.31031   7.166759     1.58   0.115    -2.772852    25.39348
    z_utorid |   28.52639   10.22676     2.79   0.005     8.430118    48.62266
       _cons |   483.9508   231.6736     2.09   0.037     28.69663    939.2051
------------------------------------------------------------------------------
```

23

But even with F, Type I Error Possible

```
      Source |       SS       df       MS              Number of obs =     335
-------------+------------------------------           F( 11,   323) =    3.35
       Model |  28671.4939    11  2606.49945           Prob > F      =  0.0002
    Residual |  251272.249   323   777.93266           R-squared     =  0.1024
-------------+------------------------------           Adj R-squared =  0.0719
       Total |  279943.743   334   838.15492           Root MSE      =  27.891

------------------------------------------------------------------------------
    last_two |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   let_fname |  -.7565112   .8612825    -0.88   0.380    -2.450943    .9379206
   let_lname |   .1212191   .6991151     0.17   0.862    -1.254175    1.496613
chars_utorid |   5.724756   2.944279     1.94   0.053    -.0676289    11.51714
    a_utorid |   .8607217   .6738678     1.28   0.202    -.4650025    2.186446
    b_utorid |  -4.081615   1.465967    -2.78   0.006    -6.965664   -1.197566
    c_utorid |  -2.332378   .8814372    -2.65   0.009    -4.066461    -.598295
    d_utorid |  -.4513824   .9687537    -0.47   0.642    -2.357246    1.454481
    e_utorid |   1.192979   .6796103     1.76   0.080    -.1440422    2.530001
    f_utorid |   2.857886   1.504574     1.90   0.058    -.1021151    5.817887
    g_utorid |   -2.21003   .6892896    -3.21   0.001    -3.566094   -.8539657
    h_utorid |  -.4078496   .7961369    -0.51   0.609    -1.974118    1.158419
       _cons |   11.84176   20.14885     0.59   0.557    -27.79778    51.48131
------------------------------------------------------------------------------
```

But, my 2013/14 class had this wild result!

24

Population R^2, R^2, Adj. R^2

- Pop. R^2 (parameter) $= 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2} = 1 - \frac{SSE/N}{SST/N}$
- R^2 (statistic) $= 1 - \frac{SSE}{SST} = 1 - \frac{SSE/n}{SST/n}$
- Adjusted R^2 (statistic) $= 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$
  - $E[SSE/(n-k-1)] = \sigma_\varepsilon^2$
  - $E[SST/(n-1)] = \sigma_y^2$
  - BUT, E[Adj. R^2] ≠ pop. R^2

Generally,
$E\left[\frac{X}{Y}\right] \neq \frac{E[X]}{E[Y]}$

25

R^2 versus Adj-R^2

- R^2
  - Pros: Always between 0 and 1 (so long as model includes a constant term)
  - Cons: Increases even with inclusion of irrelevant variables
- Adj-R^2
  - Pros: Because of deg. of freedom correction, doesn't tend to increase with inclusion of irrelevant variables
  - Cons: Can be negative (confuses the interpretation)

However, most software (e.g. STATA and Excel) automatically report both and they are usually quite similar to each other

26

Housing Prices Again, but in $1000's

```
. regress price_1000 livingarea bedrooms bathrooms fireplaces age;

      Source |       SS       df       MS              Number of obs =    1057
-------------+------------------------------           F(  5,  1051) =  321.79
       Model |  3802752.07     5  760550.414           Prob > F      =  0.0000
    Residual |  2484049.39  1051  2363.51036           R-squared     =  0.6049
-------------+------------------------------           Adj R-squared =  0.6030
       Total |  6286801.46  1056  5953.41048           Root MSE      =  48.616

------------------------------------------------------------------------------
  price_1000 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  livingarea |   .0734464   .0040089    18.32   0.000     .0655801    .0813127
    bedrooms |  -6.361311   2.749503    -2.31   0.021    -11.75645   -.9661714
   bathrooms |   19.23668    3.66908     5.24   0.000     12.03712    26.43623
  fireplaces |   9.162792   3.194233     2.87   0.004     2.894992    15.43059
         age |  -.1427395   .0482761    -2.96   0.003     -.237468   -.0480109
       _cons |    15.7127   7.311427     2.15   0.032     1.366047    30.05936
------------------------------------------------------------------------------
```

Carefully compare with Slide 14 in Lecture 20.

27