

Homework 20: ECO220Y

Required Exercises: Chapter 20: 11, 13, 15, 17, 19, 23, 33

Required Problems:

(1) In doing a test of statistical significance, economists often use a simple “rule of thumb”: Is slope coefficient divided by its standard error either > 2 or < -2 . What is the sense of this rule of thumb?

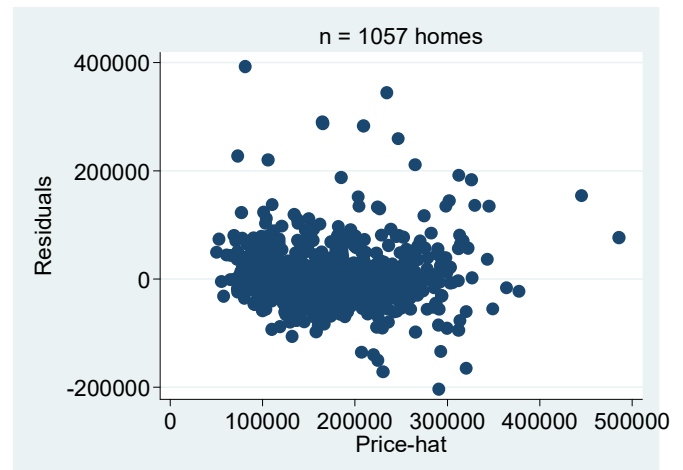
(2) Recall the housing prices example in Chapter 20 and Lecture 20. Here again are the multiple regression results.

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

Here is the graph of the residuals versus the predicted housing prices that it a great way to check for outliers, heteroscedasticity, and violations of linearity.

- (a)** What can you learn from this graph in this example?
- (b)** Do the graph and the standard deviation of the residuals reported in the STATA output match up?



(3) The housing price regression is an example of a *hedonic* regression: a regression that seeks to explain the price of something by using its features/characteristics. Next is a short excerpt from a 2016 NBER working paper “A Forward Looking Ricardian Approach: Do Land Markets Capitalize Climate Change Forecasts?”

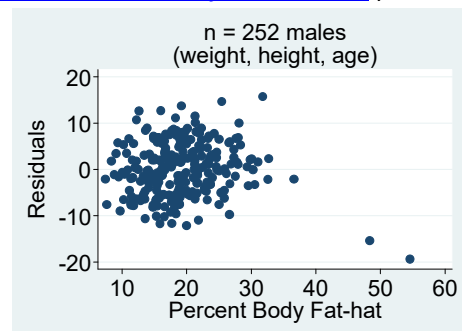
EXCERPT (p. 1): One of the greatest contributions of applied econometrics has been to provide empirical methods for estimating the economic consequences of anticipated future changes. The canonical application centers around the estimation of cross-sectional hedonic regressions using market outcome data to estimate the response of asset prices to exogenous variation in a variable of interest and that is expected to change in the future (due to change in policy, regulations, or other factors). With the estimated relationship in hand, it is straightforward to predict the costs or benefits associated with expected future changes in any variable of interest. <http://www.nber.org/papers/w22413.pdf>

For example, we could add policy variables to the housing price regression to measure local pollution levels, local school quality, etc. How would you include such policy variables into the model? Why does the excerpt say *exogenous*?

(4) In Chapter 20 and Lecture 20 we predicted male percent body fat. If you run the regression with the full sample, the results are below. (The original data are at <http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html>.)

```
. regress pct_body_fat_siri height_cm weight_kg age;
```

Source	SS	df	MS
Model	9210.64532	3	3070.21511
Residual	8368.34425	248	33.7433236
Total	17578.9896	251	70.035815



Number of obs = 252
F(3, 248) = 90.99
Prob > F = 0.0000
R-squared = 0.5240
Adj R-squared = 0.5182
Root MSE = 5.8089

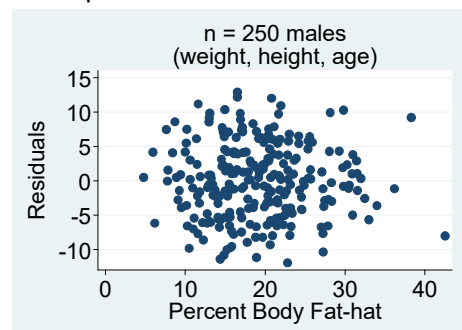
pct_body_f~i	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height_cm	-.2339897	.0420868	-5.56	0.000	-.3168828 -.1510967
weight_kg	.4368504	.0289393	15.10	0.000	.3798523 .4938485
age	.1697902	.0295603	5.74	0.000	.1115689 .2280115
_cons	17.76739	7.479351	2.38	0.018	3.036242 32.49854

(a) Identify any outliers. How would you investigate these?

(b) Here are the results without those two observations. How do the results compare with those above?

```
regress pct_body_fat_siri height_cm weight_kg age if  
(case_number~=39 & case_number~=42)
```

Source	SS	df	MS
Model	10003.7809	3	3334.59362
Residual	7125.03917	246	28.9635738
Total	17128.82	249	68.7904419



Number of obs = 250
F(3, 246) = 115.13
Prob > F = 0.0000
R-squared = 0.5840
Adj R-squared = 0.5790
Root MSE = 5.3818

pct_body_f~i	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

(5) Consider again the same percent body fat data. Here are the results if height is measured in inches (instead of cm) and weight is measured in pounds (instead of kg). In which ways are these results identical to those shown in problem (4) (b) above? In which ways are they different?

```
. regress pct_body_fat_siri height_in weight_lbs age if (case_number~=39 &
case_number~=42);
```

Source	SS	df	MS	Number of obs = 250		
Model	10003.781	3	3334.59368	F(3, 246)	=	115.13
Residual	7125.039	246	28.9635732	Prob > F	=	0.0000
				R-squared	=	0.5840
				Adj R-squared	=	0.5790
				Root MSE	=	5.3818
Total	17128.82	249	68.7904419			

pct_body_fat_i	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height_in	-1.274155	.1580123	-8.06	0.000	-1.585385	-.9629255
weight_lbs	.2536605	.0148257	17.11	0.000	.224459	.2828619
age	.1373248	.0280566	4.89	0.000	.082063	.1925865
_cons	57.27217	10.39897	5.51	0.000	36.7898	77.75454

(6) Looking at the *first graph* in Exercise 22 of Chapter 20, *approximately* what is the standard error of the residuals (s_e)? Looking at the *second graph* in Exercise 22 of Chapter 20, *approximately* what is the standard error of the residuals (s_e)?

(7) Using the following STATA output for the drug dosage example we considered in lecture, compute and interpret the missing numbers.

```
regress hrs_sleep dosage age weight;
```

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

hrs_sleep	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dosage	.5094999	.1208007	4.22	0.000		
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

(8) There is some limited (and not very convincing) evidence that sitting close to the front of the classroom improves a student's performance. Consider the research question: What is the effect of seat location on a student's performance in a course? The researcher obtains approval to conduct an experiment where students in ECO220Y are randomly assigned a seat in a classroom where they must sit for the entire course. Attendance is taken to ensure compliance in every lecture. (Note: This is a hypothetical example.) The following variables are available in the data:

- MARK_220: Student's percentage mark in ECO220Y
- ROW: Row number of student (row 1 is first row at the front of the lecture hall)
- MARK_100: Student's percentage mark in ECO100Y

Variable	Obs	Mean	Std. Dev.	Min	Max
MARK_220	250	66.42	12.35016	36.63087	97.20178
MARK_100	250	81.84	5.484407	67	98
ROW	250	13	7.225568	1	25

```
. regress MARK_220 ROW MARK_100;
```

Source	SS	df	MS	Number of obs =	250
Model	21790.3847	2	10895.1924	F(2, 247) =	166.23
Residual	16188.7148	247	65.5413556	Prob > F =	0.0000
				R-squared =	0.5737
				Adj R-squared =	0.5703
Total	37979.0996	249	152.526504	Root MSE =	8.0958

MARK_220	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ROW	-.4845315	.0710174	-6.82	0.000	-.6244085 - .3446546
MARK_100	1.56963	.0935637	16.78	0.000	1.385345 1.753914
_cons	-55.73957	7.747089	-7.19	0.000	-70.99835 -40.48079

- (a) Interpret the coefficient estimates (slopes and intercepts). Are they of the expected sign?
- (b) Do we have sufficient evidence to infer that our research hypothesis is true? (Show your work and explain.)
- (c) Given the slope of 1.57, is there regression towards the mean in terms of marks?
- (d) Considering the following simple regression with these same data, are you surprised by these results? If so, explain. If not, explain how these results are what you would expect.

```
. regress MARK_220 ROW;
```

Source	SS	df	MS	Number of obs =	250
Model	3344.68018	1	3344.68018	F(1, 248) =	23.95
Residual	34634.4194	248	139.654917	Prob > F =	0.0000
				R-squared =	0.0881
				Adj R-squared =	0.0844
Total	37979.0996	249	152.526504	Root MSE =	11.818

MARK_220	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ROW	-.5072308	.1036469	-4.89	0.000	-.7113713 - .3030903
_cons	73.014	1.540822	47.39	0.000	69.97923 76.04877

- (e) The reason that existing evidence is not very convincing is because it often relies on observational data. Describe the nature of observational data that would be available to answer the research question. Describe what would happen if a regression analysis were conducted using such data. Indicate the direction of bias on the coefficient of interest.

(9) Consider again the predicting housing prices example. Suppose we standardized all of the variables. Here are the results. Compare and contrast the results with those given in problem (2) (i.e. when the variables had not been standardized), which is reproduced again for easy comparison. Include in your answer how to interpret the coefficients when all of the variables have been standardized.

Variable	Obs	Mean	Std. Dev.	Min	Max
s_price	1057	0	1	-1.957583	5.596272
s_livingarea	1057	0	1	-1.730919	5.141484
s_bedrooms	1057	0	1	-2.950068	2.451789
s_bathrooms	1057	0	1	-1.428256	3.95517
s_fireplaces	1057	0	1	-1.134489	6.133118
s_age	1057	0	1	-.8042219	6.267465

```
. regress s_price s_livingarea s_bedrooms s_bathrooms s_fireplaces s_age;
```

Source	SS	df	MS	Number of obs =	1057
Model	638.751869	5	127.750374	F(5, 1051) =	321.79
Residual	417.248124	1051	.397001069	Prob > F =	0.0000
				R-squared =	0.6049
				Adj R-squared =	0.6030
Total	1055.99999	1056	.999999994	Root MSE =	.63008

s_price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s_livingarea	.6310485	.034444	18.32	0.000	.5634616	.6986354
s_bedrooms	-.0610493	.0263869	-2.31	0.021	-.1128263	-.0092723
s_bathrooms	.1620901	.030916	5.24	0.000	.1014259	.2227542
s_fireplaces	.0653602	.0227852	2.87	0.004	.0206506	.1100698
s_age	-.0646153	.0218536	-2.96	0.003	-.107497	-.0217336
_cons	1.27e-09	.0193802	0.00	1.000	-.0380283	.0380283

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

Source	SS	df	MS	Number of obs =	1057
Model	3.8028e+12	5	7.6055e+11	F(5, 1051) =	321.79
Residual	2.4840e+12	1051	2.3635e+09	Prob > F =	0.0000
				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