

Homework #5

A. Shawn Bandy

March 14th, 2013

1 Lab Problems

L1 STATA log:

```

-----
      name: <unnamed>
      log: /Users/shawn/src/econ485/lab5/mlb1.log
log type: text
opened on: 14 Mar 2013, 11:08:38

. use "mlb1.dta";

. /*a. Run a regression where the dependent variable is salary and the explanatory v
> ariables are teamsal
> pcinc yrsallst years hits runsyr. In other words, the total team salary, the avera
> ge income of the city
> where the team is located, the number of years on the All-Star team, the number of
> year the player has
> been in the League, the number of all-time hits (over all years), and runs per yea
> r.*/
>
> regress salary teamsal pcinc yrsallst years hits runsyr;

```

Source	SS	df	MS	Number of obs =	353
Model	3.8936e+14	6	6.4894e+13	F(6, 346) =	72.94
Residual	3.0782e+14	346	8.8966e+11	Prob > F =	0.0000
Total	6.9718e+14	352	1.9806e+12	R-squared =	0.5585
				Adj R-squared =	0.5508
				Root MSE =	9.4e+05

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
teamsal	.0143053	.006004	2.38	0.018	.0024964 .0261142
pcinc	-3.081507	17.9635	-0.17	0.864	-38.41291 32.24989
yrsallst	196930.8	40981.67	4.81	0.000	116326.2 277535.3
years	53884.84	33490.08	1.61	0.109	-11984.91 119754.6

hits		-365.6974	335.6952	-1.09	0.277	-1025.957	294.5627
runsyrr		32917.18	3553.05	9.26	0.000	25928.88	39905.47
_cons		-609625.3	374222.5	-1.63	0.104	-1345662	126411.9

```

. /*b. Look at the simple correlation between the variables in your model, using cor
> relate.*/
>
> correlate salary teamsal pcinc yrsallst years hits runsyr;
(obs=353)

```

		salary	teamsal	pcinc	yrsallst	years	hits	runsyrr
salary		1.0000						
teamsal		0.2247	1.0000					
pcinc		0.0874	0.1891	1.0000				
yrsallst		0.5856	0.1082	0.0877	1.0000			
years		0.4782	0.1774	0.0843	0.5751	1.0000		
hits		0.6271	0.1721	0.1044	0.7678	0.8824	1.0000	
runsyrr		0.7033	0.1887	0.0979	0.5677	0.5018	0.7328	1.0000

```

. /*c. Test for multicollinearity by using the vif command.*/
>
> vif;

```

Variable		VIF	1/VIF
hits		14.76	0.067746
years		6.68	0.149676
runsyrr		2.84	0.351592
yrsallst		2.84	0.352695
teamsal		1.09	0.921569
pcinc		1.04	0.958082
Mean VIF		4.88	

```

. /*d. Look at your regression results, the test for multicollinearity, and the test
> for correlation.
> Is there a problem with multicollinearity in the model? Look at the variables in
> the model. If you believe
> there is a problem with multicollinearity, explain why you might have that proble
> m.
>
> INTERPRETATION
>
> */
>
>

```

```

>
> /*e. Now, use the regression command to estimate a linear regression model where t
> he dependent variable is salary
> and the explanatory variables are teamsal pcinc allstar slugavg fldperc frstbase
> scndbase shrtstop thrdbase
> outfield. In this new model, the explanatory variables are team salary, per capit
> a income of the city where
> the team is located, percentage of years as an All-Star, career slugging average,
> career fielding percentage,
> and dummy variables for the various positions on the team, with catcher excluded.
> */
>
> regress salary teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop t
> hrdbase
> outfield;

```

Source	SS	df	MS	Number of obs =	353
Model	3.3921e+14	10	3.3921e+13	F(10, 342) =	32.41
Residual	3.5797e+14	342	1.0467e+12	Prob > F =	0.0000
				R-squared =	0.4865
				Adj R-squared =	0.4715
Total	6.9718e+14	352	1.9806e+12	Root MSE =	1.0e+06

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
teamsal	.0247805	.0064344	3.85	0.000	.0121244 .0374365
pcinc	.2530229	19.56027	0.01	0.990	-38.22055 38.7266
allstar	47920.33	3013.114	15.90	0.000	41993.77 53846.9
slugavg	2060.333	2853.358	0.72	0.471	-3552.007 7672.673
fldperc	5877.833	3111.61	1.89	0.060	-242.4695 11998.13
frstbase	320346.8	210323.1	1.52	0.129	-93342.88 734036.5
scndbase	134556.9	224728.1	0.60	0.550	-307466.3 576580.1
shrtstop	87293.44	213623.5	0.41	0.683	-332887.8 507474.7
thrdbase	382886.7	244737.8	1.56	0.119	-98494.21 864267.6
outfield	485142.8	170914.1	2.84	0.005	148967.7 821317.9
_cons	-5977180	3125933	-1.91	0.057	-1.21e+07 171294.5

```

. /*f. Test for multicollinearity again for the model in part e.*/
>
> vif;

```

Variable	VIF	1/VIF
outfield	2.33	0.428590
shrtstop	1.84	0.543536
thrdbase	1.76	0.568754
frstbase	1.66	0.602641

scndbase		1.60	0.625738
fldperc		1.27	0.788590
teamsal		1.06	0.944026
slugavg		1.06	0.945291
allstar		1.05	0.948811
pcinc		1.05	0.950678

Mean VIF		1.47	

```
. corr salary teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop thrdb
> ase outfield;
(obs=353)
```

		salary	teamsal	pcinc	allstar	slugavg	fldperc	frstbase

salary		1.0000						
teamsal		0.2247	1.0000					
pcinc		0.0874	0.1891	1.0000				
allstar		0.6638	0.1186	0.0993	1.0000			
slugavg		0.1436	0.0890	-0.0151	0.1504	1.0000		
fldperc		0.0714	-0.0414	-0.0183	0.0112	-0.0311	1.0000	
frstbase		0.0656	0.0499	0.0678	0.0469	0.0768	0.2177	1.0000
scndbase		-0.0088	0.0099	0.0005	0.0408	0.0924	-0.0084	-0.1308
shrtstop		-0.0790	0.0061	0.0262	-0.0056	-0.0997	-0.1863	-0.1535
thrdbase		0.0086	0.0117	-0.0252	0.0178	0.0029	-0.3144	-0.1248
outfield		0.1091	-0.0492	-0.0401	0.0092	0.0115	0.0126	-0.3026

		scndbase	shrtstop	thrdbase	outfield			

scndbase		1.0000						
shrtstop		-0.1374	1.0000					
thrdbase		-0.1117	-0.1311	1.0000				
outfield		-0.2709	-0.3178	-0.2585	1.0000			

```
. /*g. Is multicollinearity a problem for the model in part e. Why do you think this
> result is different from before?
>
> INTERPRETATION
>
> */
>
> /*h. Now, predict the residuals for each observation in the model from part e.*/
>
> predict res, r;

. /*i. Create a new variable resid_2 using the generate command.*/
>
```

```

> gen res2 = res^2;

. /*j. Regress resid_2 on the explanatory variables (not salary) from part f.*/
>
> regress res2 teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop thrdb
> ase outfield;

```

Source	SS	df	MS	Number of obs =	353
Model	8.0464e+25	10	8.0464e+24	F(10, 342) =	3.31
Residual	8.3174e+26	342	2.4320e+24	Prob > F =	0.0004
				R-squared =	0.0882
				Adj R-squared =	0.0615
Total	9.1221e+26	352	2.5915e+24	Root MSE =	1.6e+12

res2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
teamsal	24251.53	9808.036	2.47	0.014	4959.863 43543.2
pcinc	-5131792	2.98e+07	-0.17	0.863	-6.38e+07 5.35e+07
allstar	1.42e+10	4.59e+09	3.10	0.002	5.18e+09 2.33e+10
slugavg	6.91e+09	4.35e+09	1.59	0.113	-1.65e+09 1.55e+10
fldperc	5.68e+09	4.74e+09	1.20	0.232	-3.65e+09 1.50e+10
frstbase	3.29e+11	3.21e+11	1.03	0.306	-3.02e+11 9.59e+11
scndbase	-1.27e+09	3.43e+11	-0.00	0.997	-6.75e+11 6.73e+11
shrtstop	8.28e+10	3.26e+11	0.25	0.799	-5.58e+11 7.23e+11
thrdbase	3.41e+11	3.73e+11	0.91	0.362	-3.93e+11 1.07e+12
outfield	5.57e+11	2.61e+11	2.14	0.033	4.47e+10 1.07e+12
_cons	-5.89e+12	4.76e+12	-1.24	0.217	-1.53e+13 3.48e+12

```

. /*k. Run the regression from part e again.*/
>
> regress salary teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop th
> rdbase
> outfield;

```

Source	SS	df	MS	Number of obs =	353
Model	3.3921e+14	10	3.3921e+13	F(10, 342) =	32.41
Residual	3.5797e+14	342	1.0467e+12	Prob > F =	0.0000
				R-squared =	0.4865
				Adj R-squared =	0.4715
Total	6.9718e+14	352	1.9806e+12	Root MSE =	1.0e+06

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
teamsal	.0247805	.0064344	3.85	0.000	.0121244 .0374365
pcinc	.2530229	19.56027	0.01	0.990	-38.22055 38.7266
allstar	47920.33	3013.114	15.90	0.000	41993.77 53846.9

```

      slugavg |    2060.333    2853.358    0.72    0.471   -3552.007    7672.673
      fldperc |    5877.833    3111.61    1.89    0.060   -242.4695    11998.13
      frstbase |   320346.8    210323.1    1.52    0.129   -93342.88    734036.5
      scndbase |   134556.9    224728.1    0.60    0.550   -307466.3    576580.1
      shrtstop |    87293.44    213623.5    0.41    0.683   -332887.8    507474.7
      thrdbase |   382886.7    244737.8    1.56    0.119   -98494.21    864267.6
      outfield |   485142.8    170914.1    2.84    0.005    148967.7    821317.9
      _cons |   -5977180    3125933   -1.91    0.057   -1.21e+07    171294.5
-----

. /*l. Use the standard STATA Breusch-Pagan test for heteroskedasticity (estat hett
> est).*/
>
> estat hettest;

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
      Ho: Constant variance
      Variables: fitted values of salary

      chi2(1)      =    28.48
      Prob > chi2   =    0.0000

. /*m. Run the standard STATA Whites test for heteroskedasticity (imtest, white).*/
>
>
> imtest, white;

White's test for Ho: homoskedasticity
      against Ha: unrestricted heteroskedasticity

      chi2(50)     =    78.33
      Prob > chi2   =    0.0064

Cameron & Trivedi's decomposition of IM-test

-----
      Source |      chi2    df    p
-----+-----
      Heteroskedasticity |    78.33    50    0.0064
      Skewness |   -328.67    10    1.0000
      Kurtosis |      .      1      .
-----+-----
      Total |      .     61      .
-----

. /*n. For now, assume there is a problem with heteroskedasticity in the model from
> part e,
> estimate the model again with robust standard errors.*/
>

```

```
> regress salary teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop th
> rdbase
> outfield, r;
```

Linear regression

Number of obs = 353
 F(10, 342) = 28.36
 Prob > F = 0.0000
 R-squared = 0.4865
 Root MSE = 1.0e+06

salary	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
teamsal	.0247805	.0062627	3.96	0.000	.0124623	.0370987
pcinc	.2530229	18.49531	0.01	0.989	-36.12585	36.6319
allstar	47920.33	3449.219	13.89	0.000	41135.98	54704.69
slugavg	2060.333	3705.825	0.56	0.579	-5228.745	9349.411
fldperc	5877.833	2054.141	2.86	0.004	1837.491	9918.174
frstbase	320346.8	205440.1	1.56	0.120	-83738.5	724432.1
scndbase	134556.9	189707.4	0.71	0.479	-238583.3	507697.2
shrtstop	87293.44	176727.7	0.49	0.622	-260316.6	434903.5
thrdbase	382886.7	220391.2	1.74	0.083	-50606.24	816379.6
outfield	485142.8	156000.3	3.11	0.002	178301.9	791983.7
_cons	-5977180	2107467	-2.84	0.005	-1.01e+07	-1831951

```
. /*o. Pick a factor that you believe is missing from the regression equation in par
> t e.
> Add that variable to the regression equation for salary using STATA. Test for mult
> icollinearity
> and heteroskedasticity (using the standard STATA Breusch-Pagan test) in your model
> . Rerun your model
> to account for heteroskedasticity, if appropriate.*/
>
> regress salary teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop th
> rdbase outfield hrns;
```

Source	SS	df	MS	Number of obs = 353		
Model	3.7811e+14	11	3.4373e+13	F(11, 341) = 36.74		
Residual	3.1908e+14	341	9.3571e+11	Prob > F = 0.0000		
Total	6.9718e+14	352	1.9806e+12	R-squared = 0.5423		
				Adj R-squared = 0.5276		
				Root MSE = 9.7e+05		

salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
teamsal	.0212326	.0061086	3.48	0.001	.0092173	.0332478

pcinc		.1465059	18.49417	0.01	0.994	-36.23052	36.52353
allstar		34989.41	3484.092	10.04	0.000	28136.4	41842.43
slugavg		735.0951	2705.659	0.27	0.786	-4586.788	6056.978
fldperc		5203.643	2943.874	1.77	0.078	-586.7948	10994.08
frstbase		76981.29	202410.7	0.38	0.704	-321149.4	475112
scndbase		184091.7	212618.5	0.87	0.387	-234117.2	602300.5
shrtstop		125129.3	202065.4	0.62	0.536	-272322.3	522580.9
thrdbase		235571.4	232524.2	1.01	0.312	-221790.9	692933.6
outfield		281214.6	164665.2	1.71	0.089	-42672.72	605101.9
hruns		5902.807	915.5657	6.45	0.000	4101.939	7703.674
_cons		-5247160	2957727	-1.77	0.077	-1.11e+07	570525.9

```
-----
. vif;
```

Variable		VIF	1/VIF
outfield		2.42	0.412775
shrtstop		1.84	0.543078
thrdbase		1.78	0.563262
hruns		1.72	0.581221
frstbase		1.72	0.581682
scndbase		1.60	0.624921
allstar		1.58	0.634383
fldperc		1.27	0.787595
teamsal		1.07	0.936364
slugavg		1.06	0.939835
pcinc		1.05	0.950677

```
-----
Mean VIF | 1.56
```

```
. estat hettest;
```

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
```

```
Ho: Constant variance
```

```
Variables: fitted values of salary
```

```
chi2(1) = 83.88
```

```
Prob > chi2 = 0.0000
```

```

. /*p. Why did you decide to include the variable you chose for your model in part o
> ?
>
> INTERPRETATION
>
> */
>
> /*q. Did your estimate and statistical significance for allstar (percent of years
> as an All-Star) change from
```



```

> the model in part e to the final estimation model in part o (controlling for heter
> oskedasticity, if appropriate)?
> Explain why you might have a difference.*/
>
>
> log close;
    name: <unnamed>
    log: /Users/shawn/src/econ485/lab5/mlb1.log
    log type: text
    closed on: 14 Mar 2013, 11:08:38
-----

```

Answers:

- d. Look at your regression results, the test for multicollinearity, and the test for correlation. Is there a problem with multicollinearity in the model? Look at the variables in the model. If you believe there is a problem with multicollinearity, explain why you might have that problem.

The *VIF* for *hits* and *years* is above 5 and the correlation of those two variables and many of the rest are high so I would say that the model should be re-evaluated. It certainly makes sense that nearly every factor that relates to player salary in MLB could be described as functions with *hits* and *years* as inputs. If independent variables can be determined by other independent variables then multicollinearity is a concern.

- g. Is multicollinearity a problem for the model in part e. Why do you think this result is different from before?

Multicollinearity does not appear to be a problem for the model in part e. After reviewing the independent variables it does not seem that any are functions of any others in the model.

- p. Why did you decide to include the variable you chose for your model in part o?

The honest truth is that I know very little about baseball so *hruns* - the variable I chose - is something I recognize. I can easily imagine that there is a relationship between the number of home runs a player hits has a strong positive impact on the player's salary. One of the few times that baseball stories spill over into news sections other than sports is when there is a story involving home runs and it makes sense that a player that can garner that much attention would be highly valued by the team's management.

2 Questions

Q1 Which of the following factors can cause OLS estimators to be biased? Explain.

- a. Heteroskedasticity.
- b. Omitting an important variable.
- c. A sample correlation coefficient of 0.95 between two independent variables both included in the model.

Of these only omitted variable bias creates bias in the model. Heteroskedasticity affects the good-of-fit measurements (those that involve measurements of error, in particular). Multicollinearity can cloud the interpretation of a single coefficient in a model but does not have an impact on the interpretation of the model as a whole.

Q2 Which of the following factors will lower the variance of an OLS estimator. Explain.

- a. Increasing the number of observations.
- b. Increasing the number of explanatory variables that reduce the error variance in the model.
- c. Minimizing the multicollinearity in the model.

Q3 Use your results from L1 for this problem.

- a. Use your results from part j of L1. Calculate the Chi-square statistic for the Breusch-Pagan test. Since this model has 10 explanatory variables, this statistic has 10 degrees of freedom and the critical Chi-Square value is 18.31. Write down the null and alternative hypothesis for the Breusch-Pagan test and whether or not you reject the null (or fail to reject the null) based on your results.
- b. Is your conclusion about heteroscedasticity in part a of this problem the same as what you found using the standard STATA Breusch-Pagan test (L1, part l)? I know the values will be different, but did you come to the same conclusion regarding the errors?
- c. Based on your answer to part a of this problem and the results of the STATA Breusch- Pagan and Whites test in L1 (parts l and m), which model (L1 part e or L1 part n) is consistent with your results?
- d. Using the best model (you identified in part c) interpret the coefficient on team salary (teamsal).
- e. Are there any variables missing from this model that might be biasing the results? Which ones? Explain.