Sarah Cummings CSC423 Assignment 3
page 341 #6.8, page 343 #6.9, page 343 #6.10, page 377 #7.2, page 377 #7.5, page 378 #7.10, page 381 #7.20, page 381 #7.21, page 381 #7.22

6.8 Collusive Bidding in road construction

a) Build model for low bid price (y) and use stepwise regression to find the most suitable.

```
*Fix the text error for district variable as shown in class;
data work.fix;
set perm.flag2;
data work.fix;
set perm.flag2;
district_as_number= 0 +district;
district2= (district_as_number = 2);
district3= (district_as_number = 3);
district4= (district_as_number = 4);
district5= (district_as_number = 5);
run;
proc reg data=work.fix plots=none;
model lowbid = dotest lberatio status district2 district3 district4 district5 numbids daysest rdlength pctasph pctbase pctexcav pctmobil pctstruc pcttraff/ selection=stepwise;
run;
```

Results:

Stepwise Selection: Step 4

Variable PCTTRAFF Entered: R-Square = 0.9818 and C(p) = 2.7447

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	6.91028E14	1.72757E14	2855.80	<.0001
Error	212	1.28246E13	60493413254		
Corrected Total	216	7.038526E14			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-770645	102961	3.389007E12	56.02	<.0001
DOTEST	0.89303	0.01426	2.371894E14	3920.91	<.0001
LBERATIO	822221	107900	3.512711E12	58.07	<.0001
DAYSEST	245.34735	152.12117	1.57359E11	2.60	0.1083
PCTTRAFF	-584917	377102	1.455388E11	2.41	0.1224

Bounds on condition number: 2.8123, 31.912

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DOTEST		ESTIMATED (DOT) CONTRACT AMOUNT	1	0.9763	0.9763	59.8240	8856.62	<.0001
2	LBERATIO		LOW BID/ESTIMATE(DOT) RATIO	2	0.0050	0.9813	4.0308	57.52	<.0001
3	DAYSEST		ESTIMATED WORKING DAYS	3	0.0003	0.9816	3.1250	2.92	0.0891
4	PCTTRAFF		PCT OF CONTRACT : TRAFFIC PLAN	4	0.0002	0.9818	2.7447	2.41	0.1224

NOTE I have opted to only include the section of the results that are relevant to the answer, rather than including all of the results from the code above. If this is an issue or if you prefer me to copy in all results, can you let me know in my feedback for the assignment? Thanks!

With stepwise regression, **dotest, lberatio, daysest, and pcttraff are the** important predictors.

b) Interpret betas from resulting model
B1: Holding all else constant, for every additional dollar in the Department of Transportation's estimate, the low-bid price goes up $0.89
B2: Holding all else constant, for every additional unit of low bid estimate ratio, our low-bid price goes up $822221
B3: Holding all else constant, for each additional day estimated to complete the tast, the lowest-bid price goes up $245.47
B4: Holding all else constant, for each percentage of costs allocated to traffic control, the lowest-bid price estimate goes down $584917

c) What are the dangers associated with drawing inferences in a step-wise model?
In using stepwise regression, an extremely large number of t-tests have to be conducted which leads to a high probability of error. Additionally, step-wise regression often only considers first-order and main effects terms, so we could be missing some significant higher-order or interaction variables.

6.9 Collusive bidding cont'
Are the variables in the best subset model the same as those selected by stepwise?

Code:

```
proc reg data=work.fix plots=none;
model lowbid = dotest lberatio status district2 district3 district4 district5 numbids daysest
rdlength pctasph pctbase pctexcav pctmobil pctstruc pcttraff/ selection=CP;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: LOWBID
C(p) Selection Method

Number of Observations Read	279
Number of Observations Used	217
Number of Observations with Missing Values	62

Number in Model	C(p)	R-Square	Variables in Model
5	1.7167	0.9820	DOTEST LBERATIO RDLENGTH PCTASPH PCTTRAFF

No. From our results, we see that **DOTEST, LBERATIO, RDLENGTH, PCTASPH, and PCTTRAFF are the best predictors** using the best subset with CP selection method. This is NOT the answer that is in the back of the book and I'm not sure why.

6.10 Cooling methods for gas turbines

a) Use stepwise selection to find the best predictors of heat rate.
Code:

```
proc reg data=perm.gasturbine plots=none;
model heatrate= shafts rpm cpratio inlettemp exhtemp airflow power/
selection=stepwise;
run;
```

Results:

Stepwise Selection: Step 4

Variable AIRFLOW Entered: R-Square = 0.9235 and C(p) = 3.4159

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	155055243	38763811	187.15	<.0001
Error	62	12841965	207128		
Corrected Total	66	167897208			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13618	813.30619	58070220	280.36	<.0001
RPM	0.08882	0.01344	9043367	43.66	<.0001
INLETTEMP	-9.18561	0.77040	29446014	142.16	<.0001
EXHTEMP	14.36283	2.25963	8368444	40.40	<.0001
AIRFLOW	-0.84752	0.43701	779021	3.76	0.0570

Bounds on condition number: 3.572, 50.84

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RPM		1	0.7123	0.7123	164.224	160.95	<.0001
2	INLETTEMP		2	0.1602	0.8725	39.7023	80.41	<.0001
3	EXHTEMP		3	0.0464	0.9189	5.0809	36.00	<.0001
4	AIRFLOW		4	0.0046	0.9235	3.4159	3.76	0.0570

From our results, we see that **RPM, INLETTEMP, EXHTEMP, and AIRFLOW** are the best predictors of heat rate using the stepwise selection method.

b) Use stepwise reg with backward elimination to find the best predictors of heat rate.
Code:

```
proc reg data=perm.gasturbine plots=none;
model heatrate= shafts rpm cpratio inlettemp exhtemp airflow power/ selection =
backward;
run;
```

Results:

Bounds on condition number: 47.085, 535.31

Backward Elimination: Step 3

Variable POWER Removed: R-Square = 0.9235 and C(p) = 3.4159

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	155055243	38763811	187.15	<.0001
Error	62	12841965	207128		
Corrected Total	66	167897208			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13618	813.30619	58070220	280.36	<.0001
RPM	0.08882	0.01344	9043367	43.66	<.0001
INLETTEMP	-9.18561	0.77040	29446014	142.16	<.0001
EXHTEMP	14.36283	2.25963	8368444	40.40	<.0001
AIRFLOW	-0.84752	0.43701	779021	3.76	0.0570

Bounds on condition number: 3.572, 50.84

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	CPRATIO	6	0.0000	0.9253	6.0011	0.00	0.9737
2	SHAFTS	5	0.0006	0.9247	4.4561	0.46	0.4990
3	POWER	4	0.0012	0.9235	3.4159	0.98	0.3249

As seen in our results, we see that **CPRATIO SHAFTS and POWER are the "best predictors"** using backward elimination selection method.

c) Use "all possible regressions selection"
Code:

```
proc reg data=perm.gasturbine plots=none;
model heatrate= shafts rpm cpratio inlettemp exhtemp airflow power/ selection = CP;
run;
proc reg data=perm.gasturbine plots=none;
model heatrate= shafts rpm cpratio inlettemp exhtemp airflow power/ selection =
ADJRSQ;
run;
```

Results:

Note, including parts a and b, we have now done stepwise, backward elimination, best subset by CP and best subset by R.

The REG Procedure
Model: MODEL1
Dependent Variable: HEATRATE
C(p) Selection Method

Number of Observations Read	67
Number of Observations Used	67

Number in Model	C(p)	R-Square	Variables in Model
4	3.4159	0.9235	RPM INLETTEMP EXHTEMP AIRFLOW
5	4.4561	0.9247	RPM INLETTEMP EXHTEMP AIRFLOW POWER

The REG Procedure
Model: MODEL1
Dependent Variable: HEATRATE
Adjusted R-Square Selection Method

Number of Observations Read	67
Number of Observations Used	67

Number in Model	Adjusted R-Square	R-Square	Variables in Model
4	0.9186	0.9235	RPM INLETTEMP EXHTEMP AIRFLOW
5	0.9186	0.9247	RPM INLETTEMP EXHTEMP AIRFLOW POWER

d) Compare results— which predictors are consistently selected as "best"?
RPM, INLET, EXHTEMP, AIRFLOW

Backward elimination selects different variables, but stepwise, r-sq and cp ratio all seem to agree on these terms.

e) How would we use the results to develop a model?
We would create a linear regression model with RPM (x1), INLET(x2), EXHTEMP(x3), AIRFLOW(x4)

7.2 Multicollinearity

a) The problems that result when multicollinearity is present in regression analysis: **High correlations in the independent variables increases the likelihood of rounding errors in calculations. The results might be also confusing or misleading as there is overlap between the contributions of the x variables. Additionally, multicollinearity may also affect the signs of the betas, giving an opposite sign than would be expected since it is trying to compensate for the strong correlation of the two variables.**

b) How its detected: **We can find multicollinearity by checking the correlation between our x variables. Another sign of multicollinearity is failed t tests for our coefficients when the F test does not fail, or a VIF for a beta that is greater than 10.**

c) Measures available when multicollinearity is detected: **Remove one of the x variables.**

7.5 Urban/rural ratings of counties

a) Based on the correlation matrix, is there any evidence of extreme multicollinearity? **No.** Given such small correlations among the independent variables, there is not any evidence of extreme multicollinearity.

b) Refer back to results on page 190. Based on the tests, is there any evidence of extreme multicollinearity? **No.**

7.10 FDA Investigation of meat-processing plant- Live weight v. dressed weight

a) Fit first order linear model to data:

Code:

```
proc reg data=perm.steers plots=none;
model dresswt= livewt;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: DRESSWT

Number of Observations Read	9
Number of Observations Used	9

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6739.59948	6739.59948	97.75	<.0001
Error	7	482.62274	68.94611		
Corrected Total	8	7222.22222			

Root MSE	8.30338	R-Square	0.9332
Dependent Mean	264.44444	Adj R-Sq	0.9236
Coeff Var	3.13993		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.71059	26.31520	0.22	0.8344
LIVEWT	1	0.62597	0.06331	9.89	<.0001

From the parameter section of our results, we obtain:

$E(y) = 5.71059 + 0.62597\ (x1)$

b) 95% prediction interval for a 300 pound steer:

Code:

```
*Create value to be predicted;
data work.work_to_be_predicted;
dresswt= 193;
livewt= 300;
output;
run;
* Concatenate real data and data to be predicted ;
data work.work_to_regress;
set perm.steers work.work_to_be_predicted;
run;
*find least squares and prediction;
proc reg data=work.work_to_regress plots=none;
model dresswt= livewt/cli;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: DRESSWT

Output Statistics						
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	280	268.5994	2.5912	249.7157	287.4832	11.4006
2	250	243.4896	2.6552	224.5586	262.4206	6.5104
3	310	306.2643	4.3359	285.7493	326.7792	3.7357
4	210	218.3797	3.7550	198.4832	238.2762	-8.3797
5	290	287.4318	3.2968	267.9719	306.8918	2.5682
6	280	293.7093	3.6184	273.9478	313.4708	-13.7093
7	270	274.8769	2.7712	255.8577	293.8961	-4.8769
8	240	237.2121	2.8606	218.1226	256.3016	2.7879
9	250	249.7670	2.5173	230.9365	268.5976	0.2330
10	193	193.2698	5.2787	171.6119	214.9277	-0.2698

From our results, with added 200lb steer as observation 10, 95% PI is **(171.6119 - 214.9277)**

c) I would **not recommend** the FDA use the interval above to determine whether 150 pounds is a reasonable amount of meat from a 300 pound steer. This interval shows the opposite is true, 150 is much less meat than we would expect.

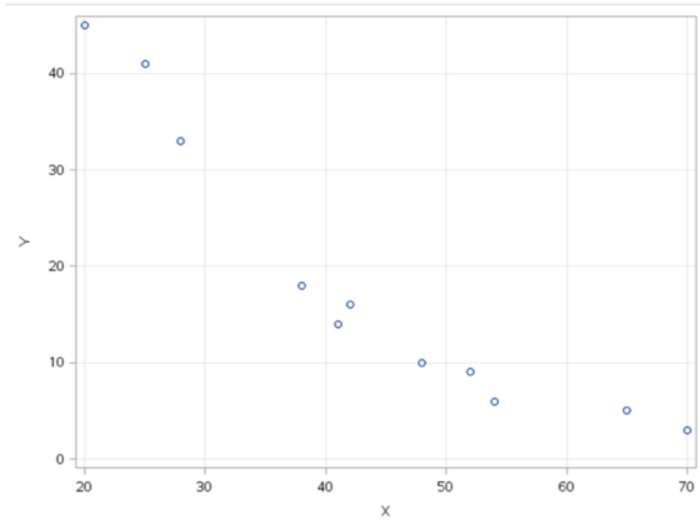
7.20- Log transformation

a) Create scatterplot if the data in the table:

Code:

```
ods graphics/reset imagemap;
proc sgplot data=perm.ex7_20;
scatter x=x y=y;
xaxis grid;
yaxis grid;
run;
ods graphics / reset;
```

Results:



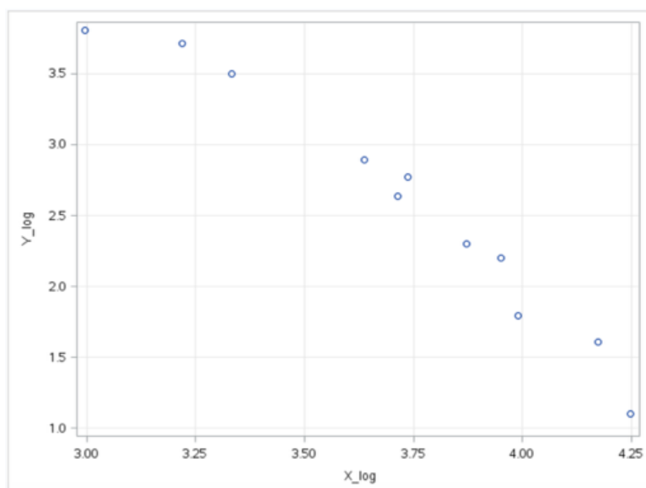
X and Y appear to have a curvilinear relationship. This graph looks like y= -log(x).

b) Plot logX v logY:

Code:

```
*Create log terms;
data work.to_be_modeled;
set perm.ex7_20;
Y_log= log(Y);
X_log= log(X);
run;
*Create scatter plot;
ods graphics/reset imagemap;
proc sgplot data=work.to_be_modeled;
scatter x=X_log y=Y_log;
xaxis grid;
yaxis grid;
run;
ods graphics / reset;
```

Results:



LogX and LogY appear to have a negative linear relationship.

c) Fit ln(y)= B0 + B1 ln(x). Is the model adequate? use a= .05

Code:

```
proc reg data=work.to_be_modeled plots=none;
model Y_log= X_log;
run;
```

From our results table, we see that the F stat p **value is <.0001, which means the overall model is adequate.**

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: Y_log

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.37999	7.37999	180.71	<.0001
Error	9	0.36755	0.04084		
Corrected Total	10	7.74754			

Root MSE	0.20209	R-Square	0.9526
Dependent Mean	2.57440	Adj R-Sq	0.9473
Coeff Var	7.84981		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.63638	0.60281	17.64	<.0001
X_log	1	-2.16985	0.16141	-13.44	<.0001

d) Use the transformed model to predict y when x=30.

Model: ln(y)= 10.636 - 2.16985(X_log)

y= e ^(10.636 - 2.16985 log(30))

y=19.745

7.21 Multicollinearity in real estate data:

a) Correlation coefficient between y and x1:

Code:

```
proc corr data=perm.hamilton;
run;
```

Results:

The CORR Procedure

3 Variables:	X1 X2 Y

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X1	15	30.00667	5.34702	450.10000	21.40000	39.00000
X2	15	69.99333	17.81450	1050	44.00000	96.60000
Y	15	120.00000	8.02167	1800	108.30000	131.30000

Pearson Correlation Coefficients, N = 15
Prob > |r| under H0: Rho=0

	X1	X2	Y
X1	1.00000	-0.89978 <.0001	0.00250 0.9930
X2	-0.89978 <.0001	1.00000	0.43407 0.1060
Y	0.00250 0.9930	0.43407 0.1060	1.00000

Correlation: 0.000250

There is not evidence of a linear relationship between sale price and appraised land value.

b) Correlation coefficient between y and x2: **0.43407**

There is no evidence of a positive linear relationship between sale price and appraised improvements.

c) Based on the results of 1 and 2, **I don't think that E(y)= B0 + B1x1+B2x2 will be useful in predicting sale price.**

d) Fit the model from c and note R-sq— does this agree with conclusion from part c?

Code:

```
 proc reg data=perm.hamilton plots=none;
model y= x1 x2;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: Y

Number of Observations Read	15
Number of Observations Used	15

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	900.72221	450.36111	39222.3	<.0001
Error	12	0.13779	0.01148		
Corrected Total	14	900.86000			

Root MSE	0.10716	R-Square	0.9998
Dependent Mean	120.00000	Adj R-Sq	0.9998
Coeff Var	0.08930		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-45.15414	0.61142	-73.85	<.0001
X1	1	3.09701	0.01227	252.31	<.0001
X2	1	1.03186	0.00368	280.08	<.0001

E(y)= -45.154+3.097x1+1.032x2
R-sq=0.9989
F stat is significant so we reject the null hypothesis. This does not fit with our conclusion from c.

e) Correlation between x1 and x2: as seen in results table from a and b, **-0.8998**

f) I would **not recommend** throwing out one of the variables. They are both too valuable to the model. Despite the fact that x and y aren't very correlated, both xs together create a nice model for y.

7.22 Socialization of doctoral students

a) Examine the correlation matrix and find the variables that are moderately or highly correlated.

Years in grad program and year GRE taken are moderately correlated: -0.602

b) If the variables in part a are left in the model, t**he resulting model might be confusing or misleading as there is overlap between the contributions of the x variables. Additionally, multicollinearity may also affect the signs of the betas, giving an opposite sign than would be expected since it is trying to compensate for the strong correlation of the two variables.**