Sarah Cummings
Assignment 2— CSC 423
#4.6, page 186 #4.10, page 193 #4.22, page 199 #4.28, page 207 #4.40, page 226 #4.59
page 271 #5.8, page 272 #5.10, page 281 #5.17, page 287 #5.22, page 302 #5.27, page 303
#5.30, page 320 #5.42, page 321 #5.44, page 323 #5.51 (include graphs of interaction terms)

4.6 Earnings of Mexican Street Vendors

a) Write a first order model for mean annual earnings E(y) as a function of age(x1) and hours worked(x2): **$E(y)= B_0 + B_1x_1 + B_2 x_2$**

b) Find least squares prediction line:

Code:

```
*Create regression;
proc reg data=perm.STREETVN plots=none;
model EARNINGS = AGE HOURS/ clb;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: earnings

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	5018232	2509116	8.36	0.0053
Error	12	3600196	300016		
Corrected Total	14	8618428			

Root MSE	547.73748	R-Square	0.5823
Dependent Mean	2577.13333	Adj R-Sq	0.5126
Coeff Var	21.25375		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-20.35201	652.74532	-0.03	0.9756	-1442.56189	1401.85787
age	1	13.35045	7.67168	1.74	0.1074	-3.36470	30.06559
hours	1	243.71446	63.51174	3.84	0.0024	105.33428	382.09465

From the parameter estimates of our results, we obtain: **$E(y)= -20.35 + 13.35x_1 + 243.71 x_2$**

c) Interpret betas: Our age coefficient indicates that **holding all else constant, each additional year in age corresponds to $13.35 more in annual earnings.** Our hours intercept indicates that **holding all else constant, annual earnings goes up $243.71 for each additional hour worked.**
Our intercept corresponds to annual earnings for someone with age 0 and 0 hours worked, which doesn't have a logical interpretation in this context.

d) Test for global utility of the model: **F stat p value is 0.0053** as seen in our results table, thus we can conclude this model is useful and reject the null hypothesis that **$B_1 = B_2 = 0$**.

e) Adj r-sq: As seen in our results table, adj r-sq is **0.5126.** This means the 51.26% of the variability in out data is described by our model.

f) The standard deviation of error for this model is given by root MSE in our results: **547.73**

g) As seen in the p value corresponding to the age coefficient t-test, age is not a useful predictor of earnings (**p=0.107**)

h) Given in our results table above, a 95% confidence interval for B2 is **(105.33, 382.094). This means we are 95% confident that that each additional hour worked corresponds to between $105.33 and $382.094 more in annual earnings, holding all else constant.**

4.10 Snow geese feeding trial

a) Least squares for WtChange E(y) with DigEff(x1) and ADFiber(x2)

Code:

```
proc reg data= perm.SNOWGEESE plots=none;
model WtChange = DigEff ADFiber/ clb alpha=.01;
run;
```

From the parameter of our results, we obtain

$E(y) = 12.18 + -0.026x_1 + -0.457\, x_2$

b) Interpret betas: Our Digestion Efficiency coefficient indicates that **holding all else constant, WtChange goes down 0.026 percent for each additional percentage of digestion efficiency.** Our AdFiber constant indicates that **holding all else constant, WtChange goes down 0.45783 percent for each additional percentage of acid detergent fiber.**

c) F-test for overall utility: **F=21.88, p<.0001** from results table. Thus, we can conclude this model is useful and reject the null hypothesis that **$B_1 = B_2 = 0$**.

Results

The REG Procedure
Model: MODEL1
Dependent Variable: WtChange

Number of Observations Read	42
Number of Observations Used	42

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	542.03486	271.01743	21.88	<.0001
Error	39	483.08419	12.38677		
Corrected Total	41	1025.11905			

Root MSE	3.51948	R-Square	0.5288
Dependent Mean	1.09524	Adj R-Sq	0.5046
Coeff Var	321.34427		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	99% Confidence Limits	
Intercept	1	12.18044	4.40236	2.77	0.0086	0.25923	24.10164
DigEff	1	-0.02654	0.05349	-0.50	0.6226	-0.17138	0.11830
ADFiber	1	-0.45783	0.12828	-3.57	0.0010	-0.80519	-0.11047

d) R-sq and adj r-sq: **0.5288 and 0.5046 respectively. The adj-r squared is better since it takes into account the amount of variables we have entered into the equation.** Based on our adj-r sq, **50.46% of the variability in our data is described by our model.**

e) With coefficient **t-test stat= -0.50 and p=0.622** as seen in the table, we cannot conclude that digestion efficiency is useful predictor of weight change. We fail to reject null hypothesis that B1=0.

f) 99% C1 for B2 as seen in the table **(-0.80519, -0.11047). We are 99% confident that holding all else constant, the percentage of weight change goes down between .805 and .110 percent for each additional percentage of ADFiber.**

4.22 Quasars- E(y) width in first order model redshift(x1), line flux(x2), line luminosity(x3) and AB1450 (x4). Find 95% prediction interval for fifth aberration and interpret results.
Code:

```
proc reg data= perm.QUASAR plots=none;
model RFEWIDTH = REDSHIFT LINEFLUX LUMINOSITY AB1450/ cli;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: RFEWIDTH

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	49163	12291	51.72	<.0001
Error	20	4752.76913	237.63846		
Corrected Total	24	53915			

Root MSE	15.41553	R-Square	0.9118
Dependent Mean	88.32000	Adj R-Sq	0.8942
Coeff Var	17.45417		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	21088	18553	1.14	0.2691
REDSHIFT	1	108.45084	88.73979	1.22	0.2359
LINEFLUX	1	557.90980	315.99021	1.77	0.0927
LUMINOSITY	1	-340.16553	320.76260	-1.06	0.3016
AB1450	1	85.68102	6.27334	13.66	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: RFEWIDTH

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	117	136.7572	7.1696	101.2933	172.2211	-19.7572
2	82	92.7556	3.8360	59.6187	125.8925	-10.7556
3	33	0.9725	7.8273	-35.0915	37.0365	32.0275
4	92	101.5036	9.4765	63.7572	139.2499	-9.5036
5	114	124.6313	5.2031	90.6928	158.5697	-10.6313

Our 95% prediction interval for the 5th observation is **(90.69, 158.5697)**. **We are 95% confident that an observation with the same x1-x4 values would have an RFE width between 90.69 and 158.5697 units.**

4.28 Earnings of Mexican Street Vendors

a) Least squares prediction equation for interaction model

Code:

```
*Create interaction term;
data work.to_be_modeled;
set perm.STREETVN;
AGE_HOURS= AGE*HOURS;
run;
*Create regression;
proc reg data=work.to_be_modeled plots=none;
model EARNINGS = AGE HOURS AGE_HOURS;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: earnings

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	3	5287427	1762476	5.82	0.0124
Error	11	3331000	302818		
Corrected Total	14	8618428			

Root MSE	550.28921	R-Square	0.6135
Dependent Mean	2577.13333	Adj R-Sq	0.5081
Coeff Var	21.35276		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1041.89440	1303.59326	0.80	0.4411
age	1	-13.23762	29.23395	-0.45	0.6595
hours	1	103.30564	162.01356	0.64	0.5368
AGE_HOURS	1	3.62096	3.84044	0.94	0.3660

From the parameter estimates of our results, we obtain:

$E(y)= 1041.89 -13.24x_1 +103.30\, x_2 + 3.62\, x_1x_2$

b) Estimated slope related earnings to age when hours worked is 10:

$y= 1041.89 -13.24x_1 +103.30(10) + 3.62(10)\, x_1$

Simplifying this equation, we get an estimated slope relating annual earnings to age of **22.97**. **This means that with hours worked equal to 10, the mean annual earnings is estimated to increase by 22.972 for each additional year of age.**

c) Estimated slope relating earnings to hours worked when age is 40:

$y= 1041.89 -13.24(40)+103.30\, x_2 + 3.62(40)\, x_2$

Simplifying this equation, we get an estimated slope relating annual earnings to hours worked of **248.146. This means that with age is equal to 40, the mean annual earnings is estimated to increase by 248.146 for each additional hour worked,**

d) Null hypothesis to see whether age and hours work interact: **H0: B_3 =0**

e) As seen in the results table, from the coefficient **t test stat (0.94), we obtain a p=0.366**

f) Conclusion for e: **We fail to reject the null hypothesis and cannot conclude that there is an interaction effect of age and hours on earnings.**

4.40 Failure times- E(y) failure time in a curvilinear model with solder temp (x1).

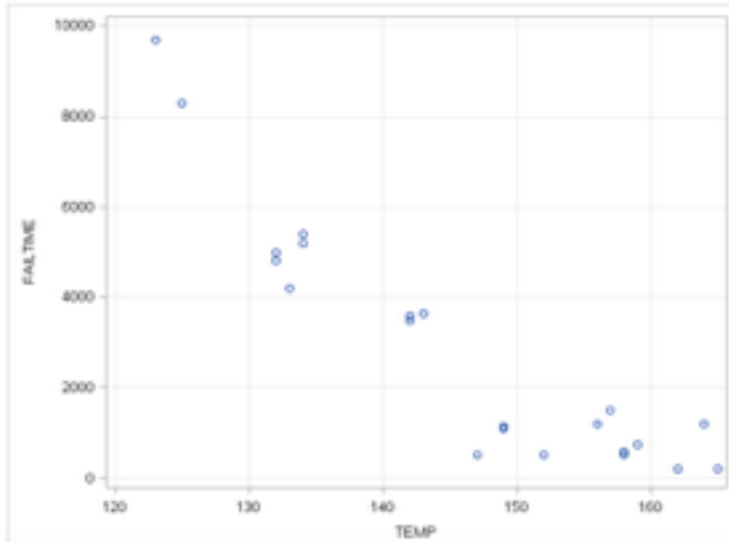
a) Scatterplot and apparent relationship:

Code:

```
*Create Scatterplot of data;
ods graphics/reset imagemap;
proc sgplot data=perm.WAFER;
scatter x=TEMP y=FAILTIME;
xaxis grid;
yaxis grid;
run;
ods graphics / reset;
```

As seen in the the results, it appears as though there is a curvilinear relationship between FailTime and Temp.

Results:



b) Fit the curvilinear model:

Code:

```
*Create squared term;
data work.to_be_modeled;
set perm.WAFER;
TEMP_SQ = TEMP**2;
run;
*Run regression;
proc reg data=work.to_be_modeled plots=none;
model FAILTIME= TEMP TEMP_SQ;
run;
```

As obtained from the parameter estimates of our results table, our regression equation is

$\mathbf{E(y)= 154232 - 1908.85x_1 + 5.93x_1^2}$

c) Test to determine if there is curvature in the relationship between failure time and temp:

As seen in our results table, the t stat for our coefficient of our squared term is 5.66 and our **p value is <.0001. Thus we reject H0: B2 =0, and conclude that the curve term is useful to our model and there is a curvilinear relationship between fail time and temp.**

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: FAILTIME

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	144830280	72415140	152.93	<.0001
Error	19	8997107	473532		
Corrected Total	21	153827386			

Root MSE	688.13657	R-Square	0.9415
Dependent Mean	2852.27273	Adj R-Sq	0.9354
Coeff Var	24.12590		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	154243	21868	7.05	<.0001
TEMP	1	-1908.85038	303.66356	-6.29	<.0001
TEMP_SQ	1	5.92895	1.04764	5.66	<.0001

4.59 RNA analysis of wheat genes- Second order model for copy number(y) with proportion of RNA(x1) and x2 (1 if MnSOD and 0 if PLD).

a) Find least squares prediction:

Code:

```
proc reg data=perm.WHEATRNA;
model y=X1 X2 X1SQ X1X2 X1SQX2;
run;
```

From the parameter estimates in the results table we obtain:

$$\mathbf{E(y)= 80.21527+156.465x_1+272.84247x_2 -24.33461x_1^2 +760.10x1x_1x_2 + 46.953x_1^2x_2}$$

b) As seen in the results table, **F=417.05 and p<.0001** thus we can conclude that our overall model is statistically useful for predicting transcript copy number. We reject the null hypothesis:

$\mathbf{B_1 = B_2 = B_3 = B_4 = B_5 = 0.}$

c) The p values for coefficients for terms including squared terms are quite large (**0.734 and 0.7898 as seen in the table). Thus we cannot confidently conclude that y is curvilinearly related to x1.**

Results:

Dependent Variable: Y

Number of Observations Read	36
Number of Observations Used	36

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6173670	1234734	417.05	<.0001
Error	30	88819	2960.62265		
Corrected Total	35	6262489			

Root MSE	54.41160	R-Square	0.9858
Dependent Mean	504.02778	Adj R-Sq	0.9835
Coeff Var	10.79536		

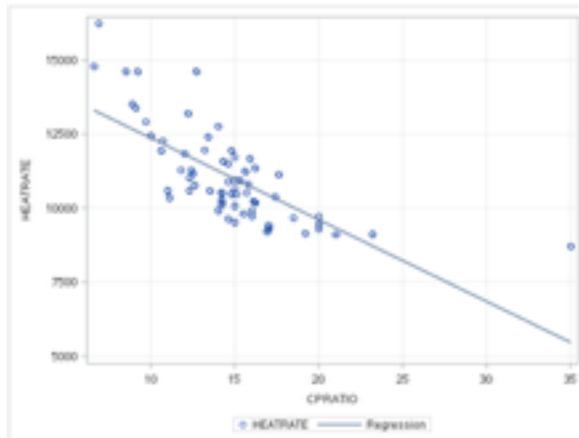
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	80.21527	30.39096	2.64	0.0130
X1	1	156.46517	128.58657	1.22	0.2332
X2	1	272.84247	42.97931	6.35	<.0001
X1SQ	1	-42.33461	123.42355	-0.34	0.7340
X1X2	1	760.10502	181.84887	4.18	0.0002
X1SQX2	1	46.95315	174.54725	0.27	0.7898

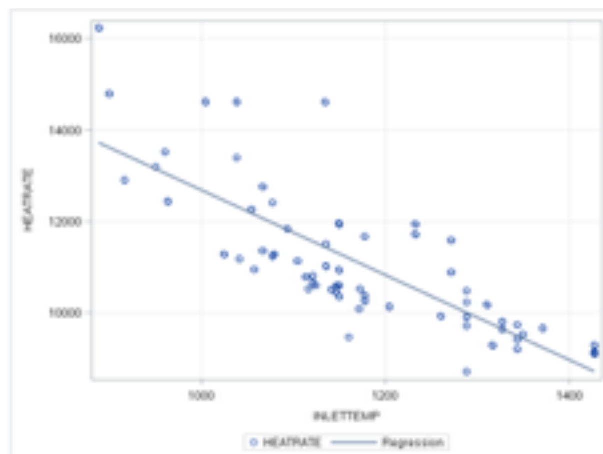
5.8 Cooling method for gas turbines- Conduct scatterplots of heat rate(y) with each of the independent variables:

Code: Results:

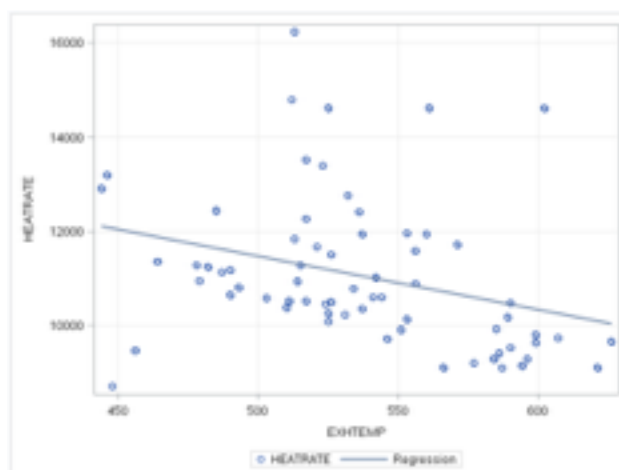
```
*Create scatterplots of heat rate with cpratio;
ods graphics/reset imagemap;
proc sgplot data=perm.GASTURBINE;
scatter x=CPRATIO y=HEATRATE;
regression x=CPRATIO y=HEATRATE;
xaxis grid;
yaxis grid;
run;
```



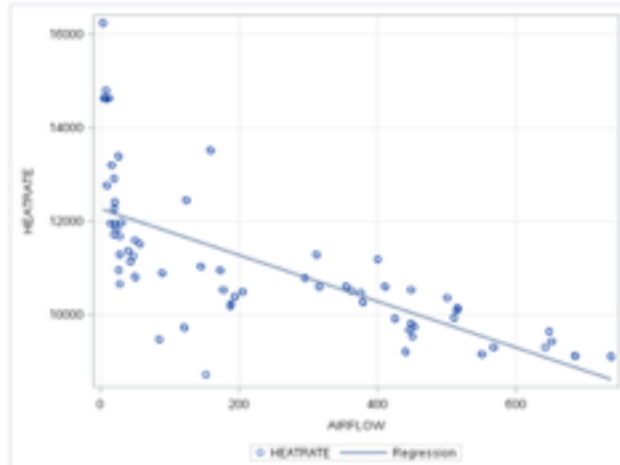

```
*Create scatterplots of heat rate with inlettemp;
ods graphics / reset;
ods graphics/reset imagemap;
proc sgplot data=perm.GASTURBINE;
scatter x=INLETTEMP y=HEATRATE;
regression x=INLETTEMP y=HEATRATE;
xaxis grid;
yaxis grid;
run;
```



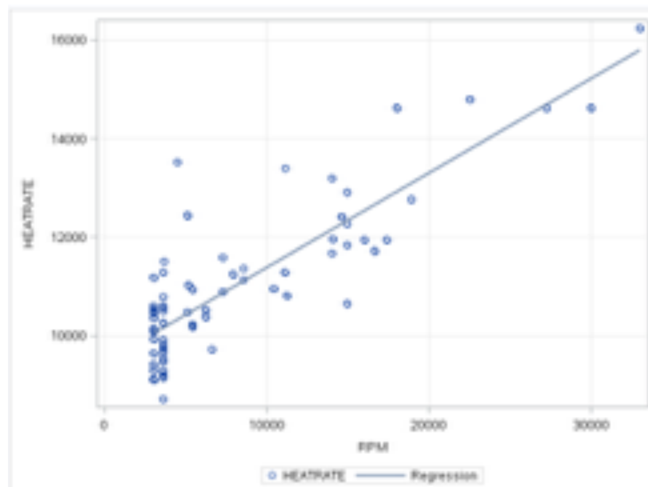

```
*Create scatterplots of heat rate with exhtemp;
ods graphics / reset;
ods graphics/reset imagemap;
proc sgplot data=perm.GASTURBINE;
scatter x=EXHTEMP y=HEATRATE;
regression x=EXHTEMP y=HEATRATE;
xaxis grid;
yaxis grid;
run;
```



```
*Create scatterplots of heat rate with airflow;
ods graphics / reset;
ods graphics/reset imagemap;
proc sgplot data=perm.GASTURBINE;
scatter x=AIRFLOW y=HEATRATE;
regression x=AIRFLOW y=HEATRATE;
xaxis grid;
yaxis grid;
run;
```




```
*Create scatterplots of heat rate with rpm;
ods graphics / reset;
ods graphics/reset imagemap;
proc sgplot data=perm.GASTURBINE;
scatter x=RPM y=HEATRATE;
regression x=RPM y=HEATRATE;
xaxis grid;
yaxis grid;
run;
ods graphics / reset;
```



Based on our scatterplots, I would hypothesize simple linear regressions with each of the independent variables.

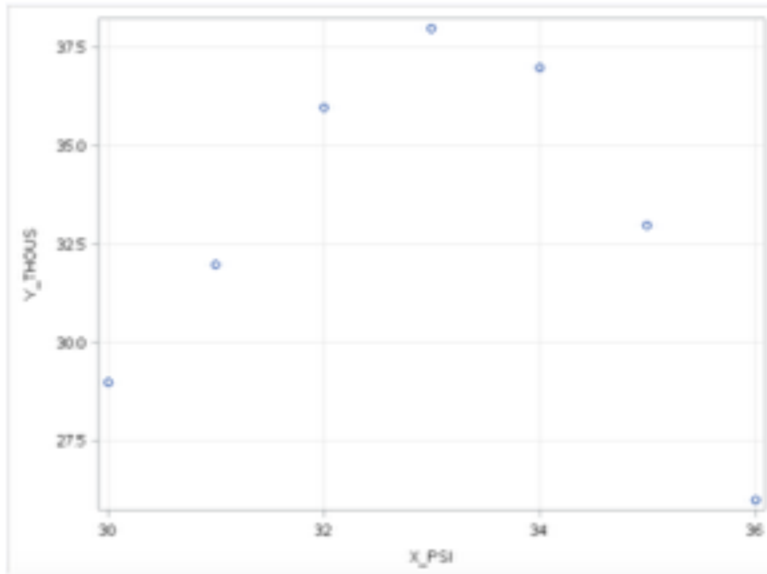
5.10- Tire wear and pressure.

a) Scatterplot of data.

Code:

```
ods graphics/reset imagemap;
proc sgplot data=perm.TIRES2;
scatter x=X_PSI y=Y_THOUS;
xaxis grid;
yaxis grid;
run;
ods graphics / reset;
```

Results:



b) If we were given the information for x=30, 31, 32 only, **I would assume psi and mileage have a positive linear relationship.**

If we were given the information for x=33, 34, 35 only, **I would assume psi and mileage have a negative linear relationship.**

Given all the data, **we can see that psi and mileage have a curvilinear relationship** that is concave down.

5.17 Quasars

a) Complete second order model for y as a function of redshift(x1), lineflux(x2), and AB1450(x3): **$E(y) = B_0 + B_1x_1 + B_2x_2 + B_3x_3 + B_4x_1^2 + B_5x_2^2 + B_6x_3^2 + B_7x_1^2 + B_8x_1x_2 + B_9x_1x_3 + B_{10}x_2x_3 + B_{11}x_1x_2x_3$**

b) Fit the model and determine if overall model is statistically significant.

Code:

```
*Create additional terms;
data work.tobemodeled;
set perm.QUASAR;
RS_Sq= REDSHIFT**2;
LF_Sq= LINEFLUX**2;
AB_Sq= AB1450**2;
RS_LF= REDSHIFT*LINEFLUX;
RS_AB= REDSHIFT*AB1450;
LF_AB= LINEFLUX*AB1450;
RS_LF_AB= REDSHIFT*LINEFLUX*AB1450;
run;
*Create regression;
proc reg data=work.tobemodeled plots=none;
model RFEWIDTH= REDSHIFT LINEFLUX AB1450
RS_Sq LF_Sq AB_Sq RS_LF RS_AB LF_AB RS_LF_AB;
run;
```

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: RFEWIDTH

Number of Observations Read	25
Number of Observations Used	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	53784	5378.35075	570.72	<.0001
Error	14	131.93254	9.42375		
Corrected Total	24	53915			

Root MSE	3.06981	R-Square	0.9976
Dependent Mean	88.32000	Adj R-Sq	0.9958
Coeff Var	3.47578		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4568.48890	12318	-0.37	0.7163
REDSHIFT	1	4511.19420	3453.33947	1.31	0.2125
LINEFLUX	1	2084.84874	1114.04499	1.87	0.0823
AB1450	1	1929.51052	524.26813	3.68	0.0025
RS_Sq	1	3.21803	4.60471	0.70	0.4961
LF_Sq	1	254.64215	26.28936	9.69	<.0001
AB_Sq	1	40.31042	2.38932	16.87	<.0001
RS_LF	1	348.18859	249.90487	1.39	0.1853
RS_AB	1	-196.49186	173.17741	-1.13	0.2756
LF_AB	1	252.90775	38.87927	6.50	<.0001
RS_LF_AB	1	-15.30300	12.46921	-1.23	0.2400

$E(y) = -4568.49 + 4211.19x_1 + 2084.84x_2 + 1929.51x_3 + 3.21x_1^2 + 254.64x_2^2 + 40.31x_3^2 + 348.19x_1x_2 - 196.49x_1x_3 + 252.91x_2x_3 - 15.30x_1x_2x_3$

$F=570.72$ and $p<.0001$ as seen in the table, thus we can conclude the overall model is statistically useful and reject the null hypothesis $B_1 = B_2 = B_3 = B_4 = B_5 = B_6 = B_7 = B_8 = B_9 = B_{10} = 0$.

c) As seen in the table from the t stat p values for the coefficients, **the LF squared and AB squared terms are statistically useful predictors of y. Both of their p values are <.0001. The RS squared term is not a statistically useful predictor of y— the p value for its t test i 0.4961.**

5.22 Failtimes- Using a quadratic model, demonstrate potential for extreme roundoff error. Then propose an alternative model.

Code:

```
*Create squared terms;
data work.tobe_modeled;
set perm.WAFER;
TEMP_SQ=TEMP**2;
run;
*Create regression;
proc reg data=work.tobe_modeled plots=none;
model FAILTIME= TEMP TEMP_SQ;
run;
```

As seen in the results, our coefficients have many sigfigs. Rounding off could drastically affect the model and decrease its efficiency.

I'm not sure which model we should use instead.

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: FAILTIME

Number of Observations Read	22
Number of Observations Used	22

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	144830280	72415140	152.93	<.0001
Error	19	8997107	473532		
Corrected Total	21	153827386			

Root MSE	688.13657	R-Square	0.9415
Dependent Mean	2852.27273	Adj R-Sq	0.9354
Coeff Var	24.12590		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	154243	21868	7.05	<.0001
TEMP	1	-1908.85038	303.66356	-6.29	<.0001
TEMP_SQ	1	5.92895	1.04764	5.66	<.0001

5.27 Quality of Bordeaux wine- wine quality(y) related to grape picking and soil type

a) Write an interaction model: $\mathbf{E(y)= B_0 + B_1 x_1 + B_2 x_2 + B_3 x_3}$
where $\mathbf{x_1}$ is 0 if automated and 1 if manual, $\mathbf{x_2}$ is 1 if gravel and 0 if not, and $\mathbf{x_3}$ is 1 if clay and 0 otherwise.

b) In this model, **$\mathbf{B_0}$ represents the predicted mean quality value for wine picked with automated grape picking from sand soil.**

c) The mean quality of grapes picked manually from clay soil is represented by $\mathbf{B_0 + B_1 + B_3}$

d) When the soil type is sand, the mean quality difference between wine picked manually and automatically is $\mathbf{B_1}$.

5.30- Modeling Faculty Salary- $E(y)= B_0+ B_1x_1$ where x=0 if lecturer, 1 if assistant prof, 2 if associate prof, and 3 if full prof.
The flaw in this model is that it assumes the mean salary for full prof is three times that of assistant prof and the mean salary for associate prof is twice that of assistant prof, etc.

Instead, add an additional term to create the following model:
$\mathbf{E(y)= B_0+ B_1x_1+ B_2x_2}$ where x1=1 if assistant prof and 0 if not, and x2 is 1 if associate prof and 0 otherwise.

5.44 Starting salaries of graduates cont.

a) Write interaction model relating salary (y), to both college and gender

$$E(y)= B_0 + B_1x_1 + B_2x_2 + +B_3x_3 + B_4x_4 + B_5x_1x_4 + B_6x_1x_4 + B_7x_2x_4 + B_8x_3x_4 + B_9x_4x_4$$

where x_1 is 1 for Bus. Admimistration college, x_2 is 1 for Engineering college, x_3 is 1 for Liberal Arts and Sciences college, and x_4 is 1 when female

b) Interpret B1: the main effect for attending Business administration college on salary.
c) Interpret B2: the main effect for attending Engineering college on salary.
d) Interpret B3: the main effect for attending Liberal Arts and Sciences college on salary.
e) Interpret B4: the main effect for being female on salary.
f) Interpret B5: the interaction effect for being a female business administration graduate on salary.
g) Explain how to test to determine whether the difference between the mean starting salaries of male and female graduates depends on college.

We would conduct t-tests on our interaction term's coefficients.

5.51 Disel engine performance

a) Test to determine whether brake power and fuel type interact

Code:

```
*Find means;
proc means data=perm.SYNFUELS noprint;
by fueltype brakepow;
by BURNRATE;
output out=work.MEANS mean=mean;
run;
* Size the graph ;
goptions reset=all border hsize=6in vsize=5in;
symbol1 interpol=join font=marker value=Z color=vibg
        width=5 height=2 line=1;  * dot, solid line ;
symbol2 interpol=join font=marker value=U color=depk
                width=5 height=2 line=2;  * square, dashed line ; * Size the graph ;
goptions reset=all border hsize=6in vsize=5in;
symbol1 interpol=join font=marker value=Z color=vibg
                width=5 height=2 line=1;  * dot, solid line ;
symbol2 interpol=join font=marker value=U color=depk
                width=5 height=2 line=2;  * square, dashed line ;

axis1 label=(angle=90 'Mean of PERFORM');

title "Line Plot of Mean (PERFORM)";
title2 "Showing Interaction Effect between FUEL and BRAND";
title3 "(Parallel lines means no interaction effect)";
proc gplot data=work.MEANS;
plot mean*fuel = brand / vaxis=axis1;
run;
title;
```

I can’t actually get this interaction graph to work— i think it’s because the example was done for SAS and I’m using SAS university addition and the graphics aren’t the same. Partial credit??