Homework 3 Spring 2016 STAT 8210

1. Do problems 7.18, 7.19 and 7.20 in your text.

2. Use the wine data from set B11 in the Appendix..

a) Build a model using the wine quality as the response variable. Use flavor and region as predictors. Write out the specific model for each region.
b) Comment on the tests of significance of the model and of the variables in it.
c) Add interaction terms between flavor and region. Write out the specific model for each region.
d) Which model, the one with or without interaction terms, do you feel is superior? Why?

3. Use the "Wine Quality of Young Red Wines" data in table B19 to fit a linear model. Use predictors x2 through x10 to predict the quality rating y. Using so many predictors has likely led to some multicolinearity issues.

a) Run the full model first. SAS will catch a REALLY big problem for you right from the start. Three of your predictors are involved in a 100% colinearity issue. Which 3 are involved, and how are they related?

Remove one of the 3 predictors from the model to break up the problem found above. Now continue to do the rest of the questions without it.

b) Does the correlation matrix show any indication of multicolinearity? If so, which predictors are overly correlated?

Before doing the rest, please center all of your predictors and use the centered predictors in your model.

c) Calculate the VIFs. Is there an indication of multicolinearity here? If so, which predictors are involved?
d) Calculate the condition number of the matrix. How do you assess the extent of any multicolinearity problem pointed to by this value?
e) Calculate the condition indices of each eigenvalue. Comment on how many (if any) multicolinearity issues there might be in this data and how severe you think they are. Use the variance decomposition proportions to specify which variables are involved in each potential problem.

4. (no collaboration on this one, please)

For this problem, you will be using some data collected on the distance in feet that a baseball traveled in the Metrodome (a sports stadium in Minnesota). There are 32 observations. The response variable is **dist.** We believe that the response might be influenced by the following variables:

Cond, which is a categorical variable with two levels---head (meaning there was a headwind blowing for that observation) and tail (meaning there was a tailwind blowing for that observation)

Angle (the angle at which the baseball was launched)

velocity (the velocity at which the baseball was launched in feet/second).

Here are three rows of the dataset:

Obs.	Number	Dist	Cond	Angle	Velocity
	1	338.5	Head	50.2	154.1
	2	348.5	Tail	51.0	156.0
	3	329.3	Head	50.2	149.3

a.) I created an indicator variable called **Hwind** which has the value of **1 if there is a headwind** and the value of **0 if there is a tailwind**. In my model, I will use the Hwind variable in place of the Cond variable. Explain briefly (20 words or less) why I didn't need to also create another indicator variable (called, say, Twind) which has the value of 1 if there is a tailwind and 0 if there is a headwind.

b.) The first model we consider uses **Hwind (instead of cond)**, **Angle**, and **Velocity** to predict **dist**. The model does not have any interactions or quadratic effects. For the model in matrix form, use the subset of data printed above to help you write the first three rows of the design matrix **X** as well as the first three rows of vector **y**.

Refer to the following output to answer parts c)—e) of this question.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1448.32854	482.77618	8.39	0.0003
Error	30	1727.17528	57.57251		
Corrected Total	33	3175.50382			

Root MSE	7.58766	R-Square	0.4561
Dependent Mean	353.35588	Adj R-Sq	0.4017
Coeff Var	2.14731		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	29.46408	130.29629	0.23	0.8226
Hwind	1	-6.25198	2.67281	-2.34	0.0262
Angle	1	-1.01907	1.97866	-0.52	0.6103
Velocity	1	2.43782	0.54693	4.46	0.0001

c) What is the estimated regression equation?

d) Using the available output, test for significance of regression. Be sure to state exactly what the p-value for this test is, as well as the interpretation of the test result in the context of these data.

e) What is the interpretation of the estimated coefficient of Hwind? (Hint: Recall how Hwind was defined and what a one-unit increase in Hwind implies.)

f) Based on the previous output, suppose that you decide to fit a new model, this time using only the velocity variable but also allowing velocity to have a quadratic effect. Based on the following output, does it seem that the quadratic term is needed, and why or why not?

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1099.35429	549.67714	8.21	0.0014
Error	31	2076.14954	66.97257		
Corrected Total	33	3175.50382			

Root MSE	8.18368	R-Square	0.3462
Dependent Mean	353.35588	Adj R-Sq	0.3040
Coeff Var	2.31599		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1953.10036	3542.23241	0.55	0.5853
Velocity	1	-23.00857	45.68971	-0.50	0.6181
VelSquared	1	0.08182	0.14732	0.56	0.5826

g) The model in part f has an overall significance of regression but neither the linear effect nor the quadratic effect is significant based on the t-tests. What is the likely reason for this? (A one-word answer may suffice.)