HW 6 (Due Nov 14, 2017) **Name:**

Problem 1.

A response Y is a function of three independent variables x_1, x_2, and x_3 that are related as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon.$$

a Fit this model to the $n = 7$ data points shown in the accompanying table.

y	x_1	x_2	x_3
1	−3	5	−1
0	−2	0	1
0	−1	−3	1
1	0	−4	0
2	1	−3	−1
3	2	0	−1
3	3	5	1

b Predict Y when $x_1 = 1, x_2 = -3, x_3 = -1$. Compare with the observed response in the original data. Why are these two not equal?

c Do the data present sufficient evidence to indicate that x_3 contributes information for the prediction of Y? (Test the hypothesis H_0: $\beta_3 = 0$, using $\alpha = .05$.)

d Find a 95% confidence interval for the expected value of Y, given $x_1 = 1, x_2 = -3$, and $x_3 = -1$.

e Find a 95% prediction interval for Y, given $x_1 = 1, x_2 = -3$, and $x_3 = -1$.

Problem 2.

The data in the accompanying table come from the comparison of the growth rates for bacteria types A and B. The growth Y recorded at five equally spaced (and coded) points of time is shown in the table.

	Time				
Bacteria Type	-2	-1	0	1	2
A	8.0	9.0	9.1	10.2	10.4
B	10.0	10.3	12.2	12.6	13.9

a Fit the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

to the $n = 10$ data points. Let $x_1 = 1$ if the point refers to bacteria type B and let $x_1 = 0$ if the point refers to type A. Let $x_2 =$ coded time.

b Plot the data points and graph the two growth lines. Notice that β_3 is the difference between the slopes of the two lines and represents time–bacteria interaction.

c Predict the growth of type A at time $x_2 = 0$ and compare the answer with the graph. Repeat the process for type B.

d Do the data present sufficient evidence to indicate a difference in the rates of growth for the two types of bacteria?

e Find a 90% confidence interval for the expected growth for type B at time $x_2 = 1$.

f Find a 90% prediction interval for the growth Y of type B at time $x_2 = 1$.

Problem 3.

Utility companies, which must plan the operation and expansion of electricity generation, are vitally interested in predicting customer demand over both short and long periods of time. A short-term study was conducted to investigate the effect of each month's mean daily temperature x_1 and of cost per kilowatt-hour, x_2 on the mean daily consumption (in kWh) per household. The company officials expected the demand for electricity to rise in cold weather (due to heating), fall when the weather was moderate, and rise again when the temperature rose and there was a need for air conditioning. They expected demand to decrease as the cost per kilowatt-hour increased, reflecting greater attention to conservation. Data were available for 2 years, a period during which the cost per kilowatt-hour x_2 increased due to the increasing costs of fuel. The company officials fitted the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 + \varepsilon$$

to the data in the following table and obtained $\hat{y} = 325.606 - 11.383x_1 + .113x_1^2 - 21.699x_2 + .873x_1x_2 - .009x_1^2x_2$ with SSE $= 152.177$.

	Price per kWh (x_2)	Mean Daily Consumption (kWh) per Household					
8¢	Mean daily °F temperature (x_1)	31	34	39	42	47	56
	Mean daily consumption (y)	55	49	46	47	40	43
10¢	Mean daily °F temperature (x_1)	32	36	39	42	48	56
	Mean daily consumption (y)	50	44	42	42	38	40
8¢	Mean daily °F temperature (x_1)	62	66	68	71	75	78
	Mean daily consumption (y)	41	46	44	51	62	73
10¢	Mean daily °F temperature (x_1)	62	66	68	72	75	79
	Mean daily consumption (y)	39	44	40	44	50	55

When the model $Y = \beta_0 - \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$ was fit, the prediction equation was $\hat{y} = 130.009 - 3.302x_1 + .033x_1^2$ with SSE $= 465.134$. Test whether the terms involving $x_2 (x_2, x_1x_2, x_1^2x_2)$ contribute to a significantly better fit of the model to the data. Give bounds for the attained significance level.

Problem 4.

EP3. The brake horsepower (HORSE, Y) developed by an automobile engine is thought to be a function of the engine speed in revolutions per minute (RPM, x_1), the road octane number of the fuel (OCT, x_2), and the engine compression (COM, x_3). An experiment is run in a laboratory at twelve different times; on each run, the temperature (TEMP, x_4) is also recorded. The data from the experiment are below.

Y	x_1	x_2	x_3	x_4
225	2000	90	100	71.2
212	1800	94	95	70.3
229	2400	88	110	72.3
222	1900	91	96	69.9
219	1600	86	100	73.2
278	2500	96	110	70.0
246	3000	94	98	70.7
237	3200	90	100	70.8
233	2800	88	105	72.1
224	3400	86	97	71.8
223	1800	90	100	71.1
230	2500	89	104	70.6

As a first step, an engineer wanted to consider the **full model**

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

for $i = 1, 2, ..., 12$, where $\epsilon_i \sim$ iid $\mathcal{N}(0, \sigma^2)$, or, in matrix notation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the 12×1 response vector, $\mathbf{X}$ is the 12×5 matrix of covariates, $\boldsymbol{\beta}$ is the 5×1 vector of regression parameters, and $\boldsymbol{\epsilon}$ is a 12×1 multivariate normal random vector with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$. Here is the ANOVA table for the **full model** fit.

Analysis of Variance: FULL model

Source	DF	SS	MS	F	Pr > F
Model	4	2597.52	649.40	7.41	0.0117
Error	7	613.48	87.64		
Corrected Total	11	3211.00			

Here are the least-squares estimates (`Parm.Est`), standard errors (`Std.Err.`), t statistics, and the associated two-sided probability values for the **full model**.

Variable	DF	Parm.Est	Std.Err.	t value	Pr > |t|
Intercept	1	-402.8470	469.5873	-0.86	0.4194
RPM	1	0.0110	0.0049	2.26	0.0581
OCT	1	3.5253	1.5881	2.22	0.0619
COM	1	1.8005	0.6106	2.95	0.0214
TEMP	1	1.5127	5.0766	0.30	0.7744

Questions for you to answer:
(a) Explain what the F statistic above is used to test. What is the conclusion reached from the value of this statistic?
(b) Use the information above to determine whether or not `TEMP` adds to the model (in the presence of the other three covariates). State your hypothesis test, significance level, and conclusion (in a well-written sentence).
(c) Using only the information from the two tables immediately above, find the diagonal elements of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix. Show all of your work.

Another engineer believes that a smaller **reduced model** may be adequate for these data. Specifically, he believes the variables `OCT` and `TEMP` are not important and, thus, the reduced model $Y_i = \gamma_0 + \gamma_1 x_{i1} + \gamma_3 x_{i3} + \epsilon_i$ is adequate. Here is the ANOVA table for this **reduced model** fit.

Analysis of Variance: REDUCED model

Source	DF	SS	MS	F	Pr > F
Model	2	1519.51	759.75	4.04	0.0559
Error	9	1691.49	187.94		
Corrected Total	11	3211.00			

More questions for you to answer:
(d) Consider writing the **reduced model** in the form $\mathbf{Y} = \mathbf{X}_0\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. Give the form of $\mathbf{X}_0$ and $\boldsymbol{\gamma}$ (just write out what these are).
(e) Test whether or not the reduced model does as well as the full model in describing these data. Use $\alpha = 0.05$. Write your conclusion as a well-written sentence.
(f) Which sum of squares is the same for both models? Why is this true?
(g) Use the full model to write a 95 percent confidence interval for the mean horsepower when each covariate is equal to its mean value from the data (e.g., the mean `RPM` value is about 2408, etc.) Interpret the interval.
(h) Repeat part (g), but write a 95 percent prediction interval for a new engine's horsepower instead. Interpret the interval.

Problem 5.

Previous enrollment records at a large university indicate that of the total number of persons who apply for admission, 60% are admitted unconditionally, 5% are conditionally admitted, and the remainder are refused admission. Of 500 applicants to date for next year, 329 were admitted unconditionally, 43 were conditionally admitted, and the remainder were not admitted. Do the data indicate a departure from previous admission rates?

Use $\alpha = 0.05$.

Problem 6.

Do you hate Mondays? Researchers in Germany have provided another reason for you: They concluded that the risk of heart attack on a Monday for a working person may be as much as 50% greater than on any other day.[1] The researchers kept track of heart attacks and coronary arrests over a period of 5 years among 330,000 people who lived near Augsberg, Germany. In an attempt to verify the researcher's claim, 200 working people who had recently had heart attacks were surveyed. The day on which their heart attacks occurred appear in the following table.

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
24	36	27	26	32	26	29

Do these data present sufficient evidence to indicate that there is a difference in the percentages of heart attacks that occur on different days of the week? Test using $\alpha = .05$.

Problem 7.

The data in the following table are the frequency counts for 400 observations on the number of bacterial colonies within the field of a microscope, using samples of milk film.[2] Is there sufficient evidence to claim that the data do not fit the Poisson distribution? (Use $\alpha = .05$.)

Number of Colonies per Field	Frequency of Observation
0	56
1	104
2	80
3	62
4	42
5	27
6	9
7	9
8	5
9	3
10	2
11	0
19	1
	400

(hint: use categories: 0,1,2,3,4,5,6,7,8,9,10,11 or more).