MATH 484 & 564, 2020 Fall
Take-Home Midterm Examination
Due at 12 pm (noon) CDT on Oct 14, 2020.
(Total=80 pts)

Problem 1 (22 pts) **Airfreight breakage.** A substance used in biological and medial research is shipped by air freight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, where collected on the number of times the carton as transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model $y = \beta_0 + \beta_1 x + \epsilon$ is appropriate.

i:	1	2	3	4	5	6	7	8	9	10
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

($\sum X_i = 10$, $\sum X_i^2 = 20$, $\sum Y_i = 142$, $\sum Y_i^2 = 2194$, $\sum X_i Y_i = 182$, $t^8_{\alpha/2} = 2.31$.)

1. (3 pts) Obtain the estimated regression function.

2. (7 pts) Obtain the point estimate for $X = 0, 1, 2, 3$. Compute the sum of the residuals $\sum e_i^2$ and the MES of the linear model.

3. (2 pts) Obtain a point estimate and the corresponding 95% confidence interval of the expected number of broken ampules when $X = 1$ transfer is made.

4. (2 pts) Obtain a point prediction and the corresponding 95% prediction interval of the number of broken ampules when $X = 3$ transfers are done.

5. (2 pts)Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer. Perform a hypothesis test on whether this increase is zero (two-sided test) at the 95% confidence level.

6. (2 pts) Use F-test to test if $\beta_1 = 0$. ($F^{1,8}_{0.05} = 5.32$)

7. (4 pts)Compare the F-ratio and the square of t-ratio in (e), what do you conclude? Show that the $F = t^2$, where t-ratio is for the β_1 in simple linear regression model in general.

Problem 2 (20 pts) It has been shown that increased reproduction caused reduced longevity for female fruitflies. The objective of this study is to confirm the same for male fruitflies. The flies used were an outbred stock. Sexual activity was manipulated by supplying individual males with one or eight receptive virgin females per day. The longevity of these males was compared with that of two control types. The first control consisted of two sets of individual males kept with one or eight newly inseminated (pregnant) females. Newly inseminated females will not usually remate for at least two days, and thus served as a control for any effect of competition with the male for food or space. The second control was a set of individual males kept with no females. There were 25 males in each of the five groups, which were treated identically in number of anaesthetizations (using CO2) and provision of fresh food medium. Details can be found here http://www.amstat.org/publications/jse/datasets/fruitfly.txt. The variables in the data are

longevity − lifespan in days (for male fruitflies)
thorax − thorax (body) length in mm
treat − a five level factor representing the treatment groups. The levels are labeled as follows: "00" − no females, "10" − one pregnant female, "80" − eight pregnant females, "11" − one virgin female, "81" − eight virgin females

Predictor	Coef	SE Coef	T	P
Constant	-49.98	10.61	-4.71	0.000
Treat10	2.653	2.975	0.89	0.374
Treat11	-7.017	2.973	-2.36	0.020
Treat80	3.929	2.997	1.31	0.192
Treat81	-19.951	3.006		0.000
THORAX	135.82	12.44	10.92	0.000

1. (1 pt) Calculate the t-statistic for Treat81.
2. (2 pts) Comment on the effect of THORAX.
3. (2 pts) The coefficient of Treat80 is greater than that of Treat10. Explain why this is counter-intuitive. Explain the anomaly.
4. (2 pts) Coefficient of Treat81 is smaller than that of Treat11. Is it consistent with your common sense about animal behavior? Explain statistically.
5. (5 pts) Fill up the blanks in the ANOVA table.

Analysis of Variance

Source	DF	Sum of Square	Mean Square	F-Statistic	P value
Regression	5	25108.1			0.000
Residual Error				———	———
Total	124	38252.8	———	———	———

6. (1 pt) What is the estimate of σ^2, model variance ?

7. (1 pt) What is the unit, if any, of $\hat{\sigma}^2$?

8. (2 pts) Calculate the R^2 and R^2_{adj}?

9. (1 pt) Assuming the same treatment, how much longer would you expect a fly with a thorax length 0.2mm greater than another to live?

10. (3 pts) We can assume the distributions of thorax lengths in the five groups are essentially equal. Will it be OK to do one-way ANOVA ignoring the thorax length in the analysis ? Justify your answer.

Problem 3 (10 pts) The physics of a chemical process suggests that the relationship between output (y) and input (x) should be of the form

$$y = \left(\frac{x}{k_0 + k_1 x + k_2 x^2} \right)^2 .$$

How will you use linear regression to obtain approximate estimates for the unknown parameters k_0, k_1 and k_2?

Problem 4 (8 pts) In the following problems it is assumed that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the $n \times p$ model matrix $\boldsymbol{X}$ has rank p and its first column is formed by 1's, $p = k + 1$, where k is the number of input variables. Consider the $n \times n$ matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Show that the following holds:

(i) (2 pts) $\boldsymbol{H}$ is symmetric.

(ii) (2 pts) $\boldsymbol{H}$ and $\boldsymbol{I} - \boldsymbol{H}$ are idempotent, i.e., $\boldsymbol{H}^2 = \boldsymbol{H}$ and $(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

(iii) (2 pts) trace$(\boldsymbol{H}) = p$ and trace$(\boldsymbol{I} - \boldsymbol{H}) = n - p$.

(iv) (2 pts) $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$.

Problem 5. Commercial Properties. (20 pts) A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, more attractive, and expensive for five specific geographic areas. The data contain the columns the age (X_1), operating expense and taxes (X_2), vacancy rates (X_3), total square footage (X_4) and rental rates (Y).

1. (2 pts) Obtain the scatter plot matrix of all the variables Y and $X_1 \sim X_4$. State and interpret your findings.

2. (1 pt) Fit regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$ (call this Model 1). State the estimated regression function.

3. (4 pts) Plot the residuals against $\hat{Y}$ (one plot in one picture), against 4 predictor variables (4 plots in one picture), against each two-factor interaction (6 plots in in one picture). Are the residuals look like i.i.d. normally distributed, i.e., the pattern is unsystematic random around zero?

4. (2 pts) Show the ANOVA table of the regression model. Is the F-ratio significant? ($\alpha = 0.05$).

5. (2 pts) What are the R^2 and R^2_{adj}?

6. (3 pts) Provide the point estimate of the mean values of the response at the following three new settings of $X_1 \sim X_4$. Provide the 95% confidence intervals and prediction intervals.

	1	**2**	**3**
X_1:	4.0	6.0	12.0
X_2:	10.0	11.5	12.5
X_3:	0.10	0	0.32
X_4:	80,000	120,000	340,000

7. (3 pts) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \epsilon$ (call this Model 2). State the estimated regression function. Use partial F test to compare Model 1 and Model 2. What is your conclusion.

8. (2 pts) Plot Y against X_1, do you observe any curvature?

9. (4 pts) Fit the regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_1^2 + \epsilon$ (call this Model 3). State the estimated regression function. Plot Y against the fitted $\hat{Y}$. Does Model 3 seem to be a good fit?

10. (2 pts) Use partial F test to compare Model 2 and Model 3. Can you conclude X_1^2 is a significant term?