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.
   Answer:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{10}{10} = 1,$$
$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{142}{10} = 14.2,$$
$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2} = \frac{182 - 10 \cdot 1 \cdot 14.2}{20 - 10 \cdot 1^2} = 4.$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 14.2 - 4 \cdot 1 = 10.2$$

   The fitted linear regression model is $\hat{Y} = 10.2 + 4X$.

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.
   $\hat{Y}(X = 0) = 10.2$, $\hat{Y}(X = 1) = 14.2$, $\hat{Y}(X = 2) = 18.2$, $\hat{Y}(X = 3) = 22.2$, and the residuals are in the following table.

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
$\hat{Y}_i$:	14.2	10.2	18.2	10.2	22.2	14.2	10.2	14.2	18.2	10.2
e_i:	1.8	-1.2	-1.2	1.8	-0.2	-1.2	-2.2	0.8	0.8	0.8

   Thus $\sum e_i^2 = 17.6$, and $MSE = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = 2.2$.

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.
Answer: When $X = 1$, $\hat{Y} = 14.2$. The 95% CI has lower and upper bounds

$$\hat{Y} \pm t^{n-2}_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum X_i^2 - n\bar{X}^2}} = 14.2 \pm 2.31 \cdot 1.483\sqrt{\frac{1}{10} + \frac{(1-1)^2}{20 - 10 \cdot 1^2}} = 14.2 \pm 1.083$$

Thus the 95% CI is $[13.12, 15.28]$.

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.
Answer: When $X = 3$, $\hat{Y} = 22.2$. The 95% prediction interval has lower and upper bounds

$$\hat{Y} \pm t^{n-2}_{\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{\sum X_i^2 - n\bar{X}^2}} = 22.2 \pm 2.31 \cdot 1.483\sqrt{1 + \frac{1}{10} + \frac{(3-1)^2}{20 - 10 \cdot 1^2}} = 22.2 \pm 4.20$$

Thus the 95% prediction is $[18.00, 26.40]$.

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.
Answer: The increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer is $\hat{\beta}_1$, and $\hat{\beta}_1 = 4$. Perform the hypothesis testing with H_0 is $\beta_1 = 0$,

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{SS_{xx}}} = \frac{4}{1.483/\sqrt{10}} = 8.528.$$

Compared with $t^8_{\alpha/2} = 2.31$, $8.528 > t^8_{\alpha/2}$, thus reject the null hypothesis, and β_1 is significantly not zero.

6. (2 pts) Use F-test to test if $\beta_1 = 0$. ($F^{1,8}_{0.05} = 5.32$)
Answer: $\sum(\hat{Y}_i - \bar{Y})^2 = 160$

$$F = \frac{160}{MSE} = \frac{160}{2.2} = 72.73 > F^{1,8}_{0.05}.$$

So reject the null hypothesis and the linear regression model is significant.

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.
Answer: $F = t^2 = 72.73$. To show that $F = t^2$ for the simple linear regression,

$$F = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\hat{\sigma}^2} = \frac{\sum(\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1\bar{X})^2}{\hat{\sigma}^2}$$
$$= \frac{\hat{\beta}_1^2}{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}} = t^2.$$

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.
   Answer: $t = \frac{-19.951}{3.006} = -6.636$.

2. (2 pts) Comment on the effect of THORAX.
   Answer: THORAX has significant positive effect to the longevity of the fruitflies. The bigger the body length of a fruitfly, the longer it lives.

3. (2 pts) The coefficient of Treat80 is greater than that of Treat10. Explain why this is counter-intuitive. Explain the anomaly.
   Answer: Since both Treat10 and Treat80 group use pregnant females, there is no effect of the sex activities that affecting the longevity of the fruitflies. The only important factor should be competition of food and other resources. Since Treat80 has more fruitflies than Treat10, there should be more competition on food and resources, thus Treat80 should have smaller positive effect on Treat10, but the data analysis shows otherwise. This is 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.
   Answer: Yes, the coefficients of Treat81 and Treat11 are consistent with the animal behavior. Both Treat81 and Treat11 has negative effects on longevity, and Treat81 is

much bigger than Treat11 in size, which is consistent with the common sense that the more reproduce activities, the shorter the longevity of the fruitflies.

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	5021.62	45.44	0.000
Residual Error	119	13144.7	110.5	———	———
Total	124	38252.8	———	———	———

6. (1 pt) What is the estimate of σ^2, model variance ?
Answer: $MSE = \hat{\sigma}^2 = 110.5$

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

8. (2 pts) Calculate the R^2 and R^2_{adj}?
Answer: $R^2 = 1 - \frac{SSE}{SST} = 65.65\%$, $R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 64.19\%$.

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?
Answer: It would live longer than the other by $0.2 \times 135.82 = 27.12$ days on average.

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. *Answer:* Even though the distribution of thorax across different treatment groups is the same, it does not mean the thorax is a constant within any group. So we cannot ignore thorax as a variable, unless the data analysis indicates so. Clearly thorax is a statistically significant variable from the regression, since its p-value is close to 0, we should not ignore it.

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?

Answer: Case 1: if $\sqrt{y} = \frac{x}{k_0+k_1x+k_2x^2}$. Let $z = \frac{x}{\sqrt{y}}$, then $z = k_0 + k_1x + k_2x^2$. Case 2: if $\sqrt{y} = \frac{x}{k_0+k_1x+k_2x^2}$. Let $z = -\frac{x}{\sqrt{y}}$, then $z = k_0 + k_1x + k_2x^2$. In either case, if we have data (x_i, y_i) for $i = 1, 2, \ldots, n$, we can always use the linear regression model $z_i = k_0 + k_1x_i + k_2x_i^2 + \epsilon_i$, to estimate the 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.
Answer: $\boldsymbol{H}' = (\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')' = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H}$.

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

$$\boldsymbol{H}^2 = \left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\left(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right) = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H}.$$

$(\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H}^2 = \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{H}$.

(iii) (2 pts) $\text{trace}(\boldsymbol{H}) = p$ and $\text{trace}(\boldsymbol{I} - \boldsymbol{H}) = n - p$.
Answer: Rank of $\boldsymbol{H}$ is p, because $\boldsymbol{X}$ is full-rank, i.e., $\text{rank}(\boldsymbol{X}) = p$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ has rank p.

$$\text{trace}(\boldsymbol{H}) = \text{trace}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{trace}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = \text{trace}(\boldsymbol{I}_p) = p.$$

$\text{trace}(\boldsymbol{I}_n - \boldsymbol{H}) = \text{trace}(\boldsymbol{I}_n) - p = n - p$.

(iv) (2 pts) $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$.
Answer: $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}$. $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{H}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = 0$.

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).
Answer:

1. There seem to be increasing linear trends between Y and X_2 and Y and X_4.

2. The estimated coefficients and the inference are in Table 1. The fitted linear model is $\hat{Y} = 12.2 - 0.142X_1 + 0.282X_2 + 0.619X_3 + 7.92 \times 10^{-6}X_4$.

3. From the residuals v.s. $\hat{Y}$ plot, it seems that the residuals are symmetrically distributed around 0, only it seems that there is a slightly increasing trend in the size of the residuals v.s. $\hat{Y}$. From residuals v.s. $X_1 \sim X_4$ plot, it seems that there is a curvature in the residuals v.s. X_1. The residuals v.s. $X_2 \sim X_4$ are not conclusive. From

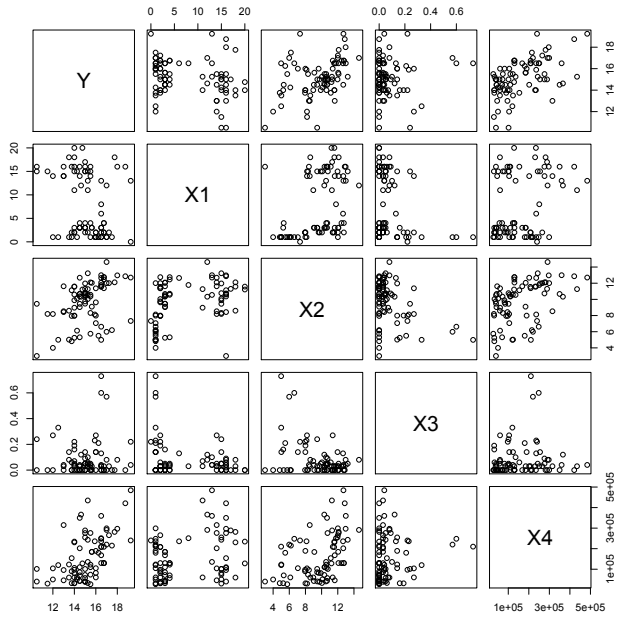

Figure 1: Scatter plot matrix of Y and other input variables.

the residuals v.s. interaction plots, It seems that the residuals has an increasing linear trend with interaction between X_1 and X_2, and decreasing linear trend with interaction between X_1 and X_3.

4. From the ANOVA table in Table 2, we can conclude that the F statistic is significant, compared with $F_{4,76}^{0.05} = 2.49$.

5. $R^2 = 0.5847$ and $R^2_{adj} = 0.5629$.

6. The following Table 3 contains point estimate, confidence interval, and prediction intervals.

7. The fitted Model 2 is $\hat{Y} = 12.37 - 0.144X_1 + 0.267X_2 + 8.178 \times 10^{-6}X_4$. The estimated coefficients and the inference are in Table 4. The partial F test is following. The p-value is 0.5704, indicating that dropping X_3 does not significantly change the model. Thus it is reasonable to drop X_3 from Model 1.

```
Analysis of Variance Table
Model 1: Y ~ X1 + X2 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     77 98.650
2     76 98.231  1   0.41975 0.3248 0.5704
```

8. Yes. There seems to be a curvature between Y and X_1, but not obvious.

Table 1: Coefficients estimate and inference for Model 1.

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	
(Intercept)	1.220e+01	5.780e-01	21.110	$< 2e-16$	***
X1	-1.420e-01	2.134e-02	-6.655	3.89e-09	***
X2	2.820e-01	6.317e-02	4.464	2.75e-05	***
X3	6.193e-01	1.087e+00	0.570	0.57	
X4	7.924e-06	1.385e-06	5.722	1.98e-07	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 2: ANOVA table of Model 1.

Source	SS	DF	MS	F-ratio
Model	138.327	4	34.582	26.75
Error	98.231	76	1.293	
Total	236.558	80		

9. The fitted Model 3 is $\hat{Y} = 12.49 - 0.404X_1 + 0.314X_2 + 8.046 \times 10^{-6}X_4 + 0.0142X_1^2$. The estimated coefficients and interference are in Table 5. From the Y v.s. $\hat{Y}$ plot, we can see that the Y and $\hat{Y}$ falls on a 45 degree diagonal lines, so it is a good fit.

10. From the following partial F test, the p-value is small enough thus we reject the null hypothesis, thus X_1^2 is a significant term.

```
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X4
Model 2: Y ~ X1 + X2 + X4 + I(X1^2)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     77 98.650
2     76 91.535  1    7.1154 5.9078 0.01743 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Table 3: Model 1: point estimate, confidence and prediction intervals.

Fit	Confidence Interval		Prediction Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
15.14850	14.76829	15.52870	12.85249	17.44450
15.54249	15.15366	15.93132	13.24504	17.83994
16.91384	16.18358	17.64410	14.53469	19.29299

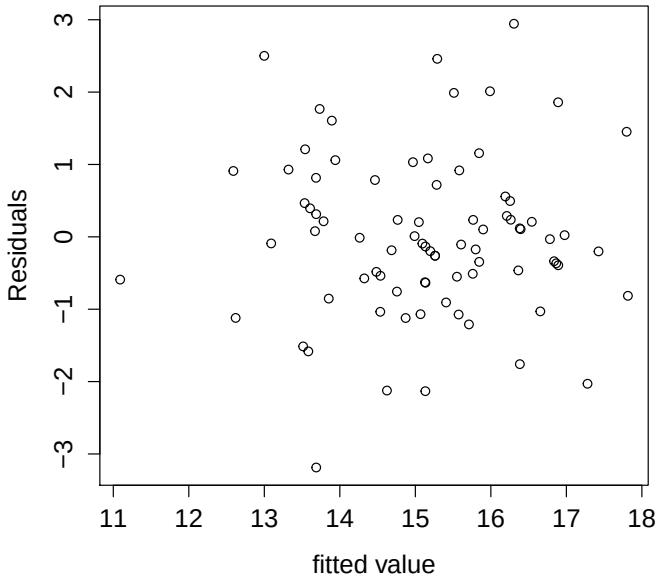

Figure 2: Model 1: Residuals v.s. fitted values $\hat{Y}$.

Table 4: Coefficients estimate and inference for Model 2.

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	
(Intercept)	1.237e+01	4.928e-01	25.100	$< 2e-16$	***
X1	-1.442e-01	2.092e-02	-6.891	1.33e-09	***
X2	2.672e-01	5.729e-02	4.663	1.29e-05	***
X4	8.178e-06	1.305e-06	6.265	1.97e-08	***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 5: Coefficients estimate and inference for Model 3.

	Estimate	Std.Error	t value	$Pr(>\|t\|)$	
(Intercept)	1.249e+01	4.805e-01	26.000	$< 2e-16$	***
X_1	-4.043e-01	1.089e-01	-3.712	0.00039	***
X_2	3.140e-01	5.880e-02	5.340	9.33e-07	***
X_4	8.046e-06	1.267e-06	6.351	1.42e-08	***
X_1^2	1.415e-02	5.821e-03	2.431	0.01743	*

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

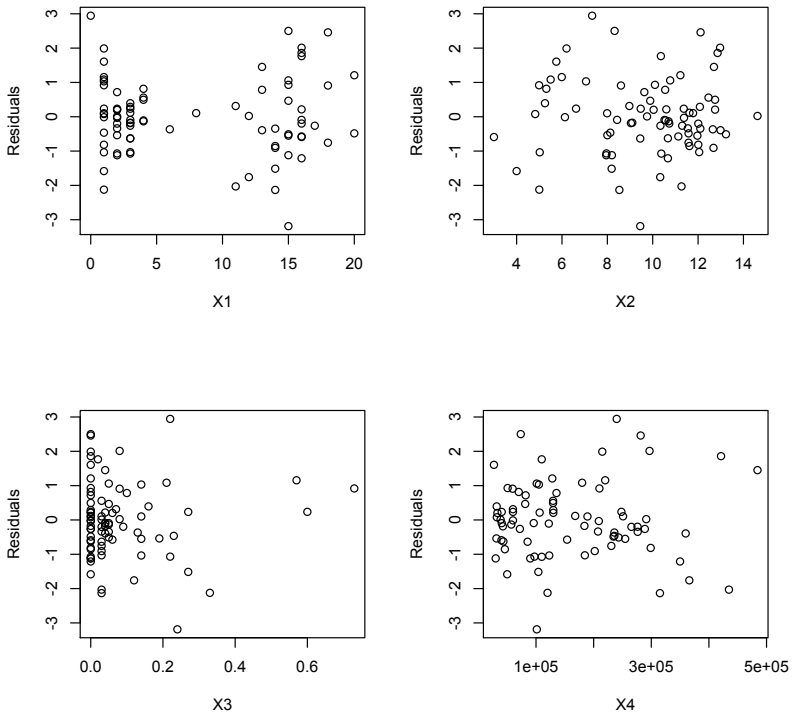

Figure 3: Model 1: Residuals v.s. predictor variables $X_1 \sim X_4$.

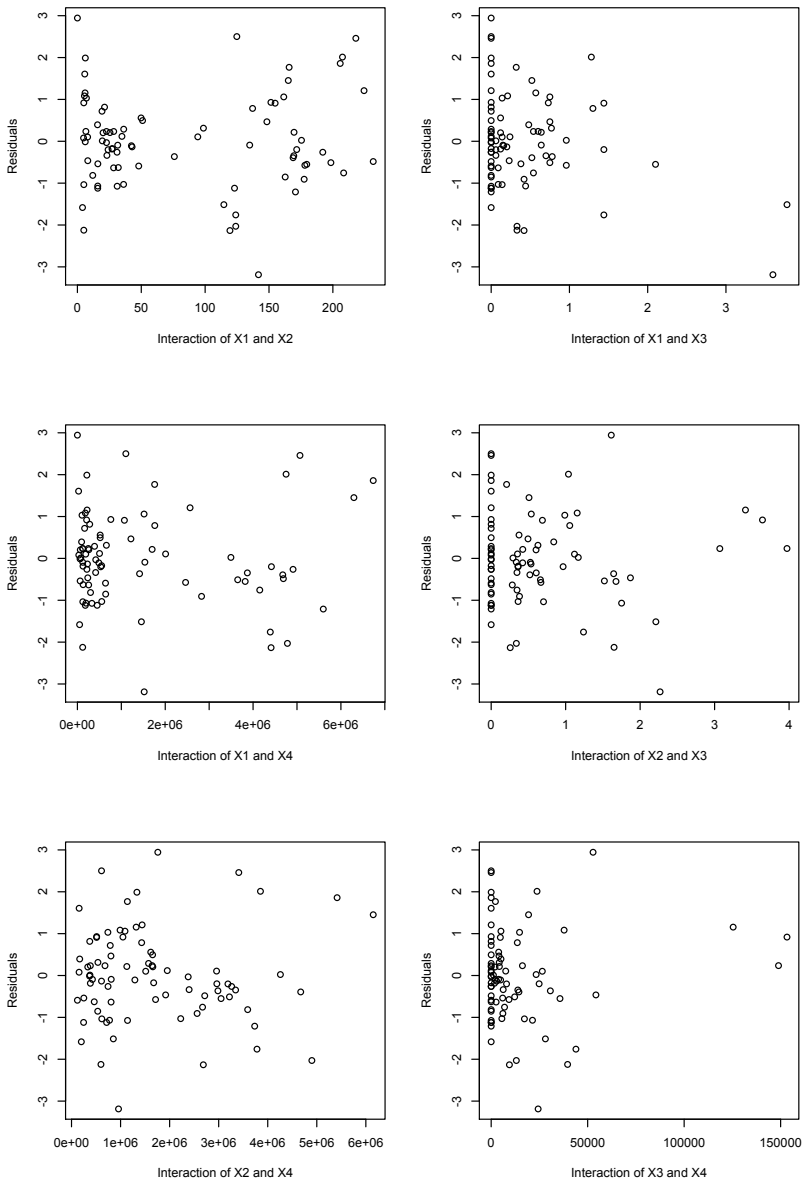

Figure 4: Model 1: Residuals v.s. interaction terms between predictor variables.

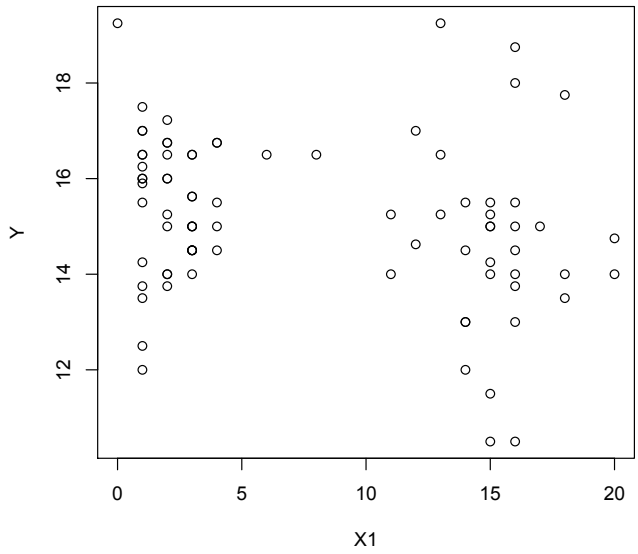

Figure 5: Scatter plot of Y v.s. X_1.

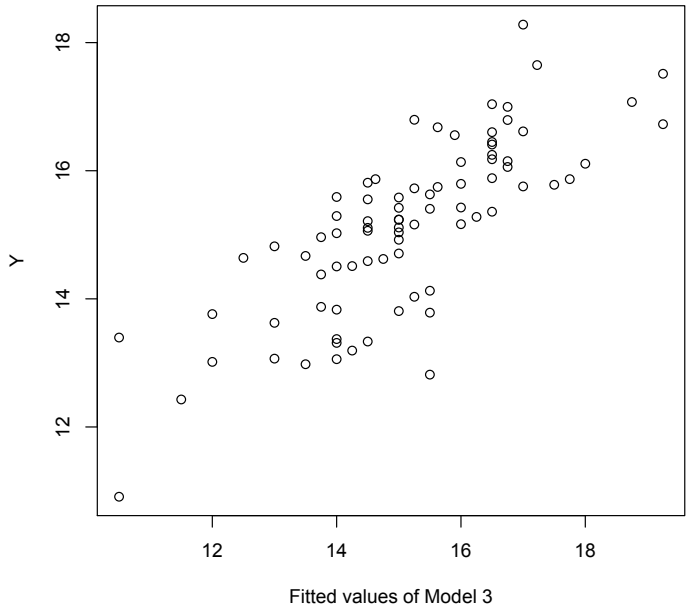

Figure 6: Y v.s. fitted values of Model 3.