Exam 2

5/06/20

There are two questions, each of which has several parts. Neither the questions nor the parts are necessarily in order from easiest to most difficult. Make sure you have taken a look at and attempted all of the questions in the allotted time. Stop working and immediately turn in your exam when time has been called.

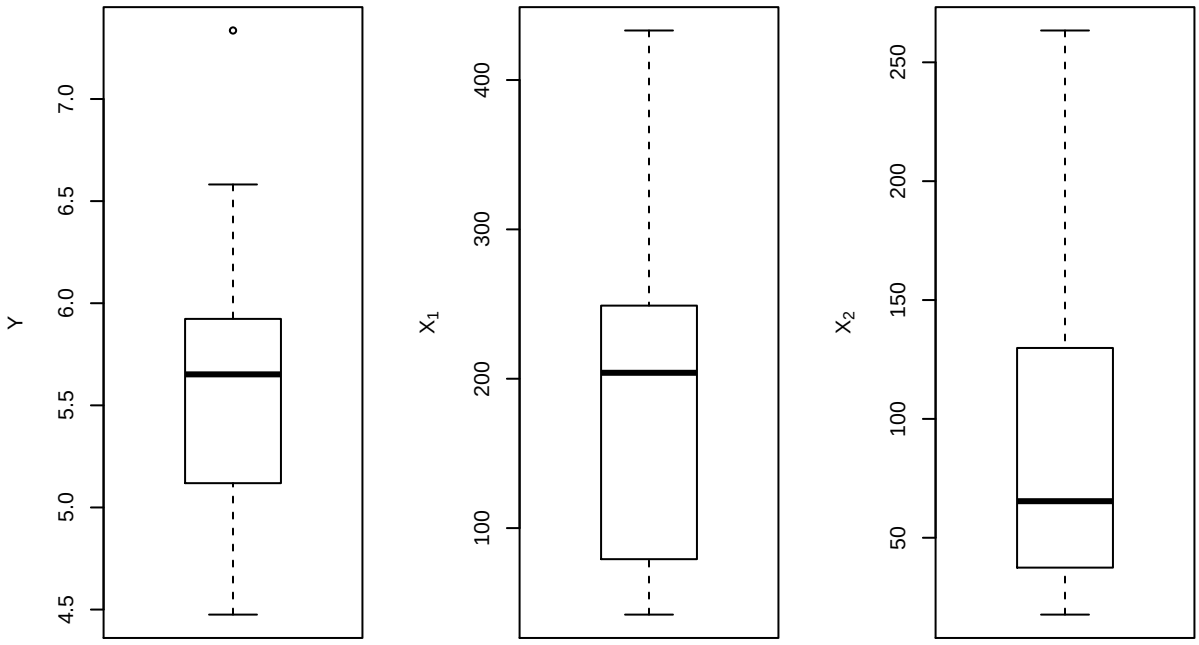
Name:	_____
Question	Points
1	_____
2	_____

Question 1

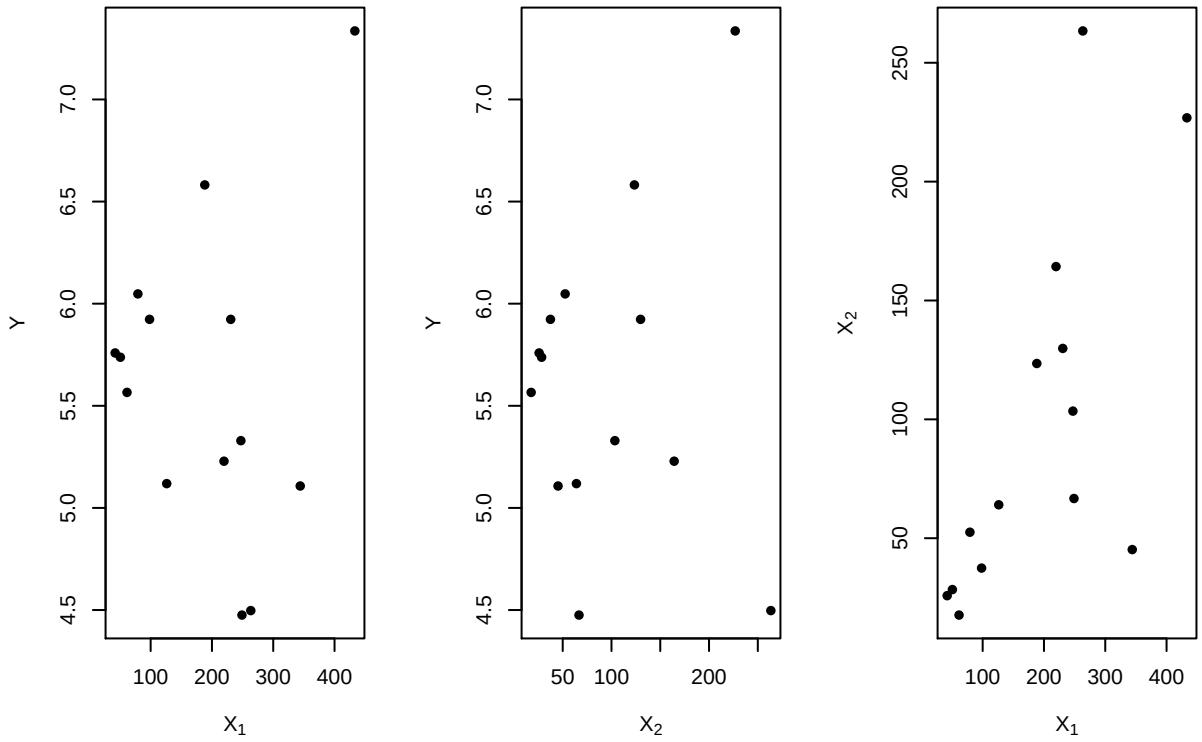
Consider the problem of modeling the genetic diversity of mice in New York City parks (Y) as a function of the total park size (X_1) and the size of the habitable park area (X_2). The data is printed below. Larger values of Y indicate a more diverse mouse population.

i	X_1	X_2	Y
1	247.23	103.47	5.33
2	98.23	37.44	5.92
3	126.24	64.06	5.12
4	433.15	226.83	7.34
5	344.05	45.23	5.11
6	79.21	52.53	6.05
7	219.66	164.26	5.23
8	188.31	123.5	6.58
9	42.09	25.84	5.76
10	230.68	129.84	5.92
11	248.96	66.71	4.48
12	263.38	263.38	4.50
13	61.44	17.68	5.57
14	50.58	28.40	5.74

(a) Boxplots of the response Y and the predictors X_1 and X_2 are printed below. What information do these plots provide? Explain in at most one sentence per plot.



(b) Scatterplots of the response Y against the predictor X_1, the response Y against the predictor X_2, and the predictor X_2 against the predictor X_1 are shown below. Based on the scatterplots, explain in at most one sentence whether or not you see evidence that suggests a multiple linear regression of the response Y on X_1 and X_2 should be performed.



(c) Write out the multiple linear regression model with normal errors for regressing the response Y on the predictors X_1 and X_2. Making sure to fully specify both the assumed regression function and any distributional assumptions.

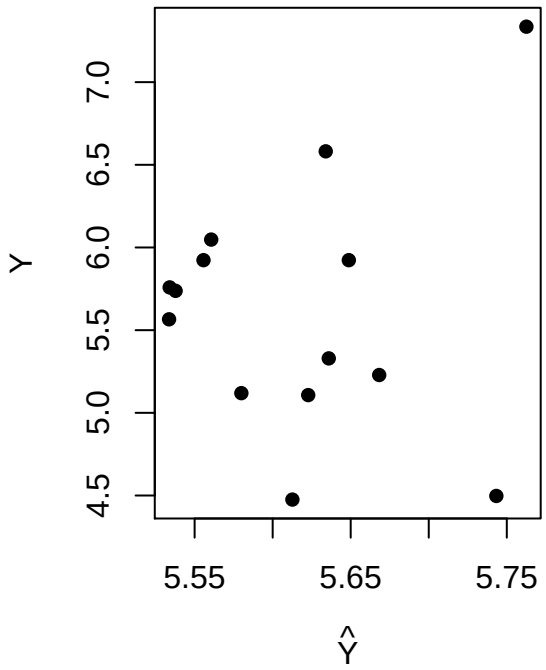
(d) The results of a multiple linear regression of the response Y on the predictors X_1 and X_2 are printed below. Provide and interpret the following, in at most one sentence each:

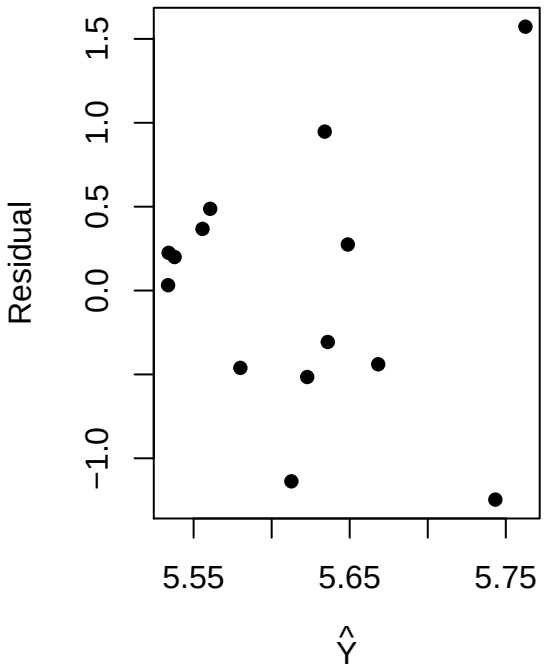
i. b_0
ii. b_1
iii. b_2
iv. R^2

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.5067077  0.4309597  12.778 6.08e-08 ***
X1          0.0002525  0.0026660   0.095    0.926
X2          0.0006458  0.0040731   0.159    0.877
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.827 on 11 degrees of freedom
Multiple R-squared:  0.009205,  Adjusted R-squared:  -0.1709
F-statistic: 0.0511 on 2 and 11 DF,  p-value: 0.9504
```

(e) The observed values Y and residuals e are both plotted against the fitted values $\hat{Y}$ below. Explain in at most one sentence why the relationship between the observed values Y and fitted values $\hat{Y}$ looks similar to the relationship between the residuals e and fitted values $\hat{Y}$, with reference to the estimated regression coefficients from (d).





(f) Referring back to the plots provided in (e), indicate whether or not you observe evidence of non-constant variance?

(g) Using only the regression results provided in (d) conduct a t test of the null hypothesis $\beta_2 = 0$ versus the alternative hypothesis $\beta_2 \neq 0$, using $\alpha = 0.05$. State the conclusion. What is the P-value of the test?

(h) Using the analysis of variance table given below, conduct an F test to determine whether or not X_2 can be dropped from the regression model given that X_1 is retained, using $\alpha = 0.05$. State the null and alternatives in terms of the regression coefficients β_1 and β_2, and the conclusion. What is the P-value of the test?

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value Pr(>F)
X1         1 0.0527 0.05269  0.0771 0.7865
X2         1 0.0172 0.01719  0.0251 0.8769
Residuals 11 7.5225 0.68387
```

(i) In at most one sentence, explain how the decision rules and conclusions in (f) and (g) relate.

(j) Now consider a simple linear regression of the response Y on X_1. The results of a simple linear regression of the response Y on the predictors X_1 are printed below. Provide b_1, interpret b_1, and explain how the interpretation of this value of b_1 relates to the value of b_1 obtained in (d).

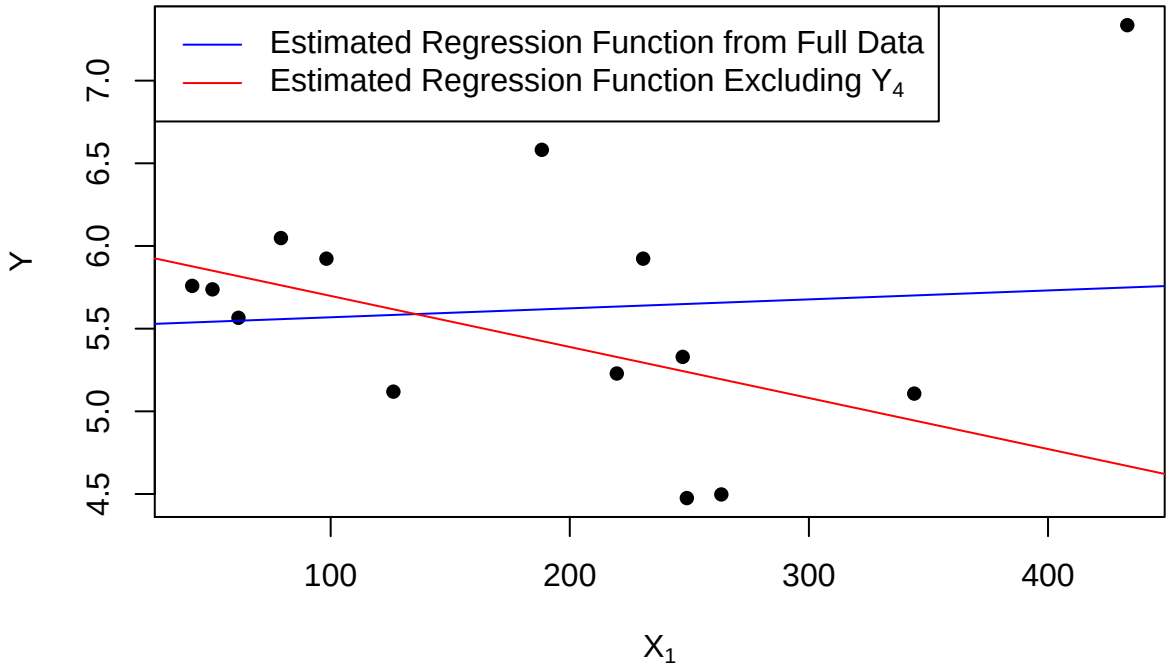
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.5146888  0.4102565   13.44 1.35e-08 ***
X1          0.0005409  0.0018679    0.29    0.777
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7927 on 12 degrees of freedom
Multiple R-squared:  0.00694,   Adjusted R-squared:  -0.07581
```

```
F-statistic: 0.08386 on 1 and 12 DF,  p-value: 0.7771
```

(k) Referring back to the regression results in (j), conduct a t test to determine whether or not there is a linear association between Y and X_1, using $\alpha = 0.05$. State the conclusion. What is the P-value of the test?

(l) The following figure shows the estimated regression functions from two simple linear regressions of Y on X_1 - one fit to the entire dataset, and another fit to a subset of the dataset that excludes the fourth observation. In at most one sentence, explain what you observe.



(m) The results of the simple linear regression of the response Y on X_1 that excludes the fourth observation are printed below. Conduct a t test to determine whether or not there is a linear association between Y and X_1, using $\alpha = 0.05$. State the conclusion and compare it to the conclusion from (k). What is the P-value of the test?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.006643   0.312784  19.204 8.26e-10 ***
X1          -0.003087   0.001615  -1.912   0.0823 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5486 on 11 degrees of freedom
Multiple R-squared:  0.2494,    Adjusted R-squared:  0.1811
F-statistic: 3.655 on 1 and 11 DF,  p-value: 0.0823
```

(n) Two 95% intervals for $\hat{Y}_4$ from the regression summarized in (m) are provided below. One is a confidence interval and one is a prediction interval. Identify the prediction interval and compare it to the observed value of Y_k.

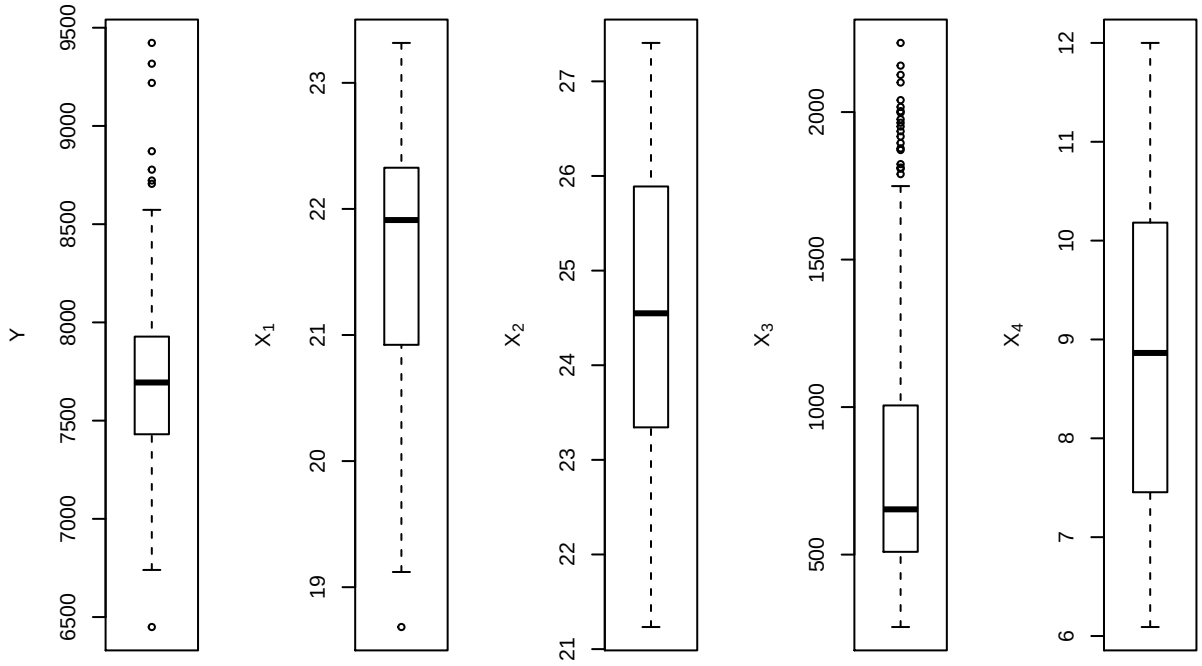
Lower	Upper
3.10	6.23
3.67	5.67

(o) Based on (a)-(n), respond to the following question in at most one sentence: "Based on the data you observed and the analysis you performed, do you expect mouse populations in large parks to be more genetically diverse than mouse populations in small parks?"

Question 2

A group of scientists collected data on 240 different bird populations in a region of Australia. Each data point represents a population at a different location, and the variables measured include a measure of the brightness of the local birds' coloring (Y), solar radiation at that location (X_1), mean annual temperature at that location (X_2), mean annual precipitation at that location (X_3), and the average growing season length (X_4).

Boxplots of the response and predictors are given below.



(a) Suppose a researcher came to you with the results of an initial multiple linear regression analysis, in which they regressed the response Y on the predictors X_1, X_2, X_3, and X_4 and remarked, "It is clear that mean temperature is most important predictor variable, because b_2 is the largest estimated regression coefficient!" Would you agree or disagree? Explain in at most one sentence.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6515.9891  1594.3872   4.087    6e-05 ***
X1             2.8069    68.2325   0.041   0.9672
X2            68.8709    36.1098   1.907   0.0577 .
X3            -0.1057     0.1339  -0.789   0.4309
X4           -53.2841    40.7351  -1.308   0.1921
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 387 on 235 degrees of freedom
Multiple R-squared:  0.2172,    Adjusted R-squared:  0.2039
F-statistic:  16.3 on 4 and 235 DF,  p-value: 8.423e-12
```

(b) Suppose the researcher then asked you to assess whether or not there is a regression relation between Y and the predictors X_1, X_2, X_3, and X_4, using $\alpha = 0.05$. State the null and alternatives in terms of the regression coefficients as well as the conclusion. What is the P-value of the test?

(c) Suppose the researcher told you that they were expecting a simpler model with fewer predictors, and asked you to test whether or not the predictors X_1, X_2 and X_3 can be eliminated from the model

given that X_2 is retained, using $\alpha = 0.05$. State the null and alternatives in terms of the regression coefficients. State the conclusion based on the analysis of variance table printed below. What is the P-value of the test?

```
Analysis of Variance Table

Model 1: Y ~ X2
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    238 38613398
2    235 35194196  3   3419202 7.6103 7.061e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(d) Suppose the researcher told you that they were expecting a simpler model with fewer predictors, and asked you to test whether or not the predictors X_1 and X_3 can be eliminated from the model given that X_2 and X_4 are retained, using $\alpha = 0.05$. State the null and alternatives in terms of the regression coefficients. State the conclusion based on the analysis of variance table printed below. What is the P-value of the test?

```
Analysis of Variance Table

Model 1: Y ~ X2 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    237 35338897
2    235 35194196  2    144701 0.4831 0.6175
```

(e) Reviewing the results of (c) and (d), what predictors would you tell the researcher to include?

(f) Suppose that the researcher was still hoping for a simpler model, and came back to you with a table of R^2 and adjusted R_a^2 values for multiple linear regression models using possible combinations of the four predictors. Which predictors would you tell the researcher to include?

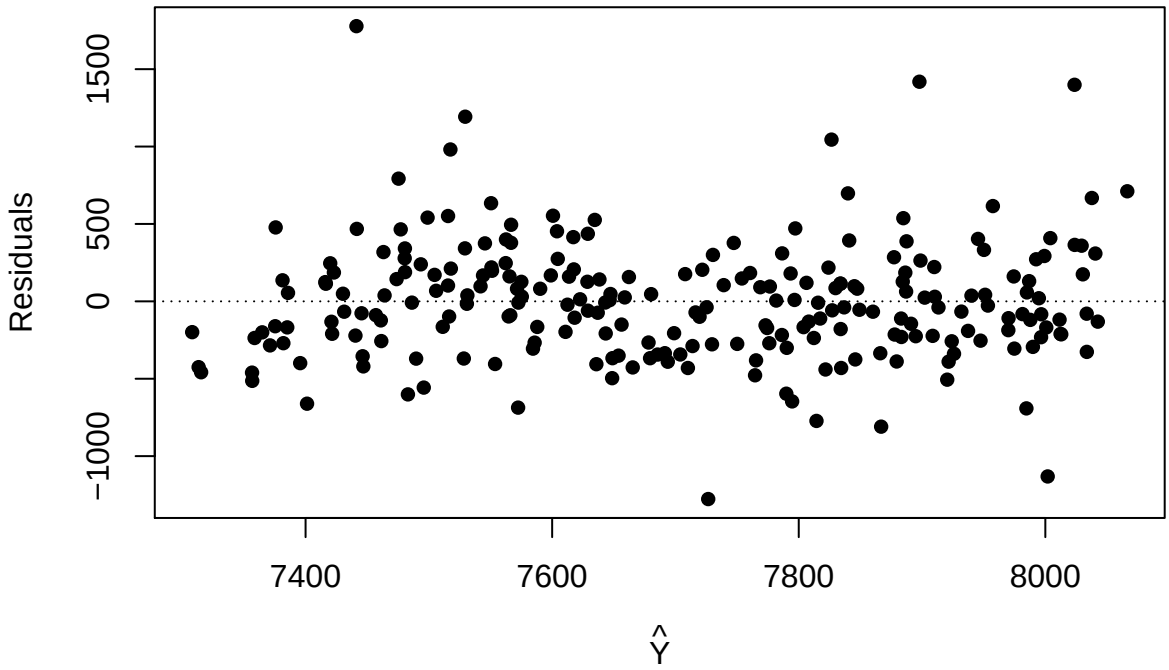
Predictors	R^2	R_a^2
X_1	0.1653	0.1618
X_2	0.1412	0.1376
X_3	0.0598	0.0559
X_4	0.1975	0.1941
X_1, X_2	0.1999	0.1931
X_1, X_3	0.1693	0.1623
X_1, X_4	0.1985	0.1917
X_2, X_3	0.2111	0.2044
X_2, X_4	0.214	0.2074
X_3, X_4	0.2016	0.1948
X_1, X_2, X_3	0.2115	0.2015
X_1, X_2, X_4	0.2152	0.2052
X_1, X_3, X_4	0.2051	0.195
X_2, X_3, X_4	0.2172	0.2073
X_1, X_2, X_3, X_4	0.2172	0.2039

(g) After several rounds of conversations with you and your peers, the researcher concludes that a multiple linear regression model with predictors X_2 and X_4 is appropriate. Now the researcher would like you to provide two clear, concise sentences that explain how the estimates b_1 and b_2 should be interpreted in the context of the original problem. Provide the two sentences using the regression results below.

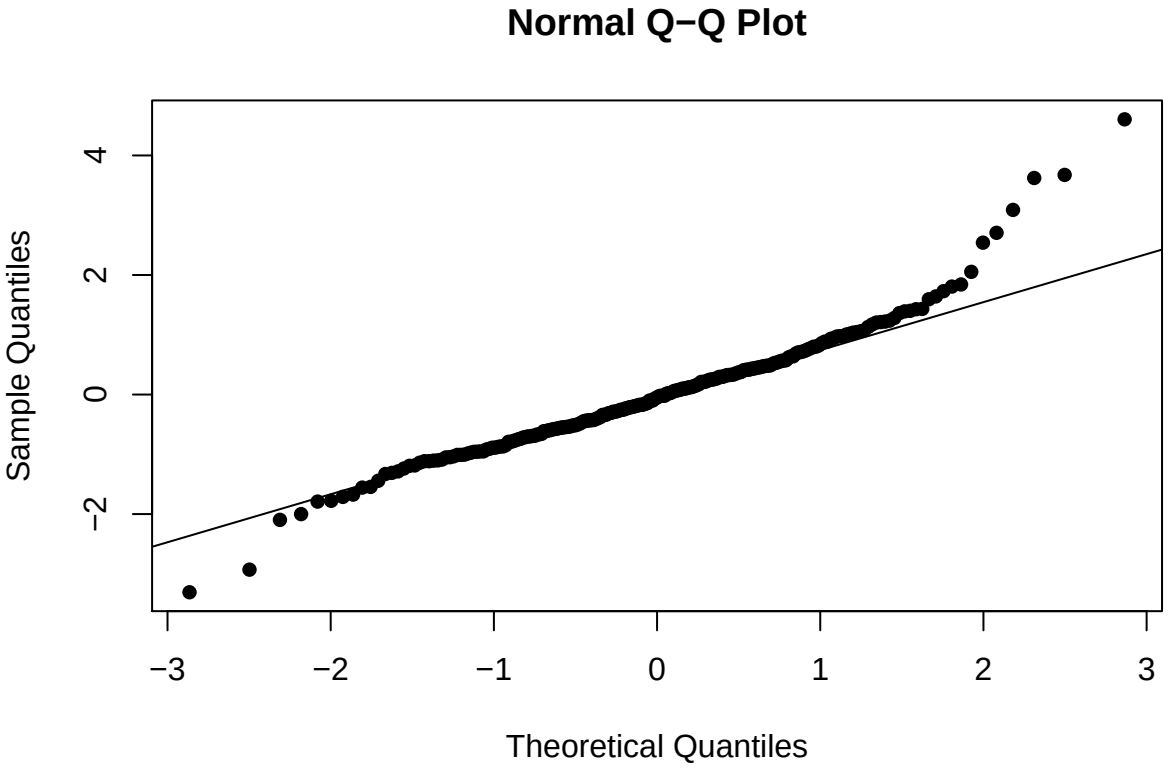
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7380.56     614.06  12.019  < 2e-16 ***
X2             45.10      20.20   2.233   0.0265 *
X4            -87.85      18.75  -4.686  4.7e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 386.1 on 237 degrees of freedom
Multiple R-squared:  0.214, Adjusted R-squared:  0.2074
F-statistic: 32.27 on 2 and 237 DF,  p-value: 4.042e-13
```

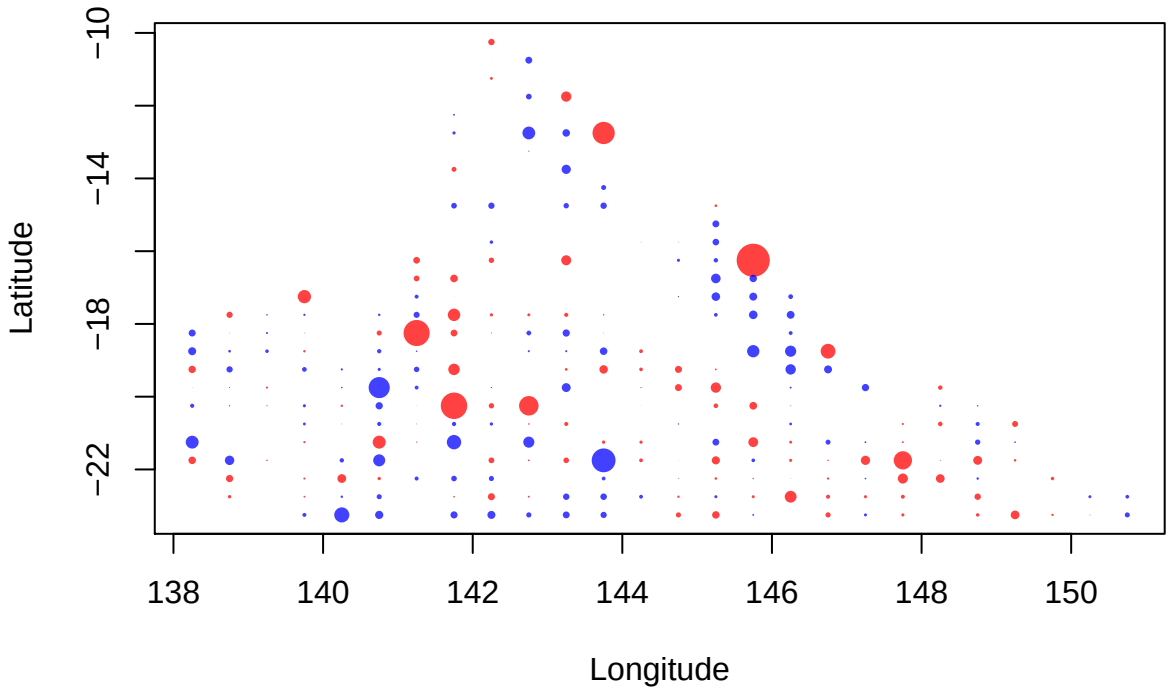
(h) To wrap up the project, the researcher would like to examine the residuals to determine whether or not there is evidence for any departures from the multiple linear regression model assumptions. First, the researcher provides a plot of the residuals from the regression model described in (g) and plots them against the corresponding fitted values. Which assumption does this plot relate to, and do you observe strong evidence of any departures from this assumption?



(i) Second, the researcher provides a normal probability plot of the studentized residuals from the regression model described in (g). Recall that the studentized residuals are obtained by dividing each residual by s, the square root of the unbiased estimate of the noise variance. Which assumption does this plot relate to, and do you observe strong evidence of any departures from this assumption?



(j) Third, the researcher reminds you that each data point corresponds to a different location, and notes that the latitude and longitude coordinates of each location were collected. The researcher creates an interesting plot that allows you to visualize how the residuals for neighboring data points compare. In the figure printed below, the data points are plotted by location and the size and color of each point indicates the magnitude and sign of the corresponding residual. Larger dots correspond to residuals that are larger in absolute value, and the color the sign of the residual (blue dots correspond to negative residuals, while red dots correspond to positive residuals). Which assumption does this plot relate to, and do you observe strong evidence of any departures from this assumption?



(k) Content, the researcher thanks you for your help. On their way out, the researcher says "Thank you for helping me show that higher temperatures cause birds to become more brightly colored!" You race after the researcher to say that that isn't what you meant. How do you explain what aspect of their last remark is incorrect?