Homework 7 Solutions

Due: Thursday 4/09/20 by 8:30am

Rubric:

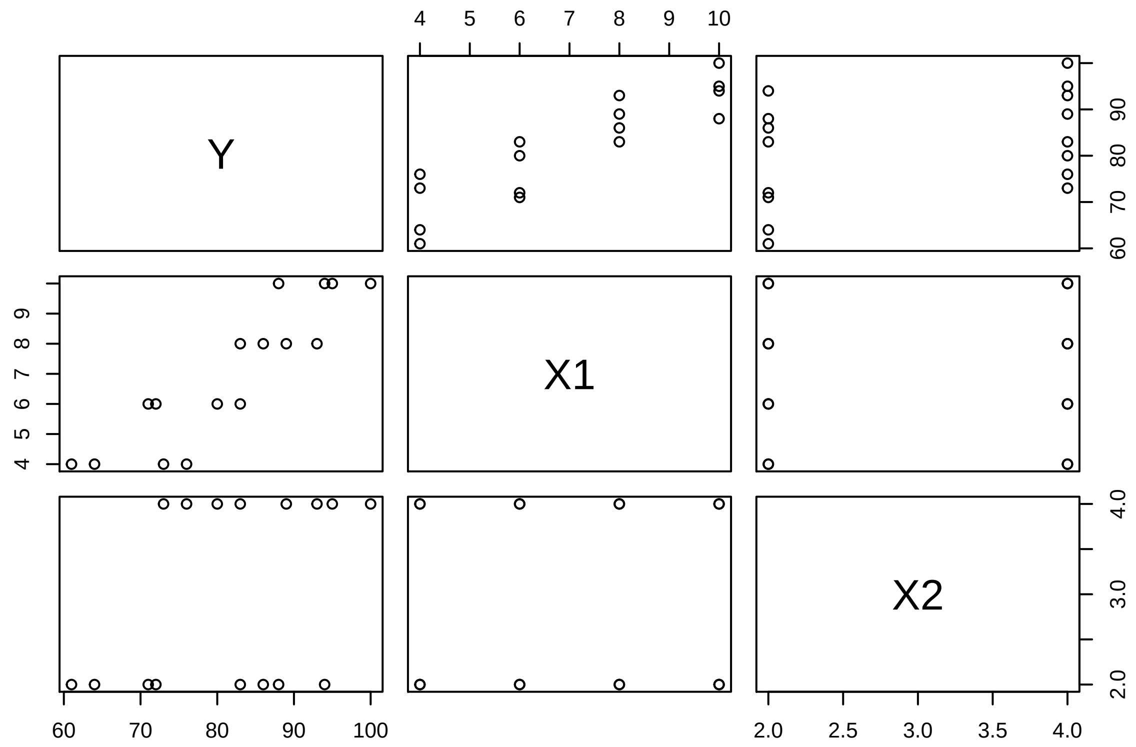
- Maximum of 2 points each for 7.-9., determined as follows:
  - 0 points for no solutions whatsoever or incomplete solutions;
  - 1 point for solutions provided for each part, but at least one incorrect solution;
  - 2 points for correct solutions to each part;
- Maximum of 3 points for 1.-6., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for mostly correct answers but at least one substantial issue;
  - 3 points for nearly/exactly correct.

1. Problem 6.5 from the `.pdf` version of the textbook, parts (a)-(d). Requires use of the `brand_preference` data that has been posted on the Homework page. We have not explictly covered the idea of making a correlation matrix in class, but you have the tools to do so. Simply make a table that contains the correlations between the response Y and the predictors X_1 and X_2 like the one shown in Figure 6.4 (b) in the `.pdf` version of the textbook.

(a)

Scatterplots of the response Y against the covariates X_1 and X_2, as well as a plot of the covariates X_1 and X_2 against each other are given below.

```
link <- url("http://maryclare.github.io/stat525/content/homework/brand_preference.RData")
load(link)
close(link)
pairs(data)
```



The correlation matrix is printed below.

Y	X_1	X_2
Y	1.00	0.89
X_1	0.89	1.00
X_2	0.39	0.00

```
cor(data)
```

The scatterplots and correlation matrix show that the response Y is approximately linear and increasing in both covariates X_1 and X_2, and that the covariates X_1 and X_2 are uncorrelated and appear to have no relationship with each other.

(b)

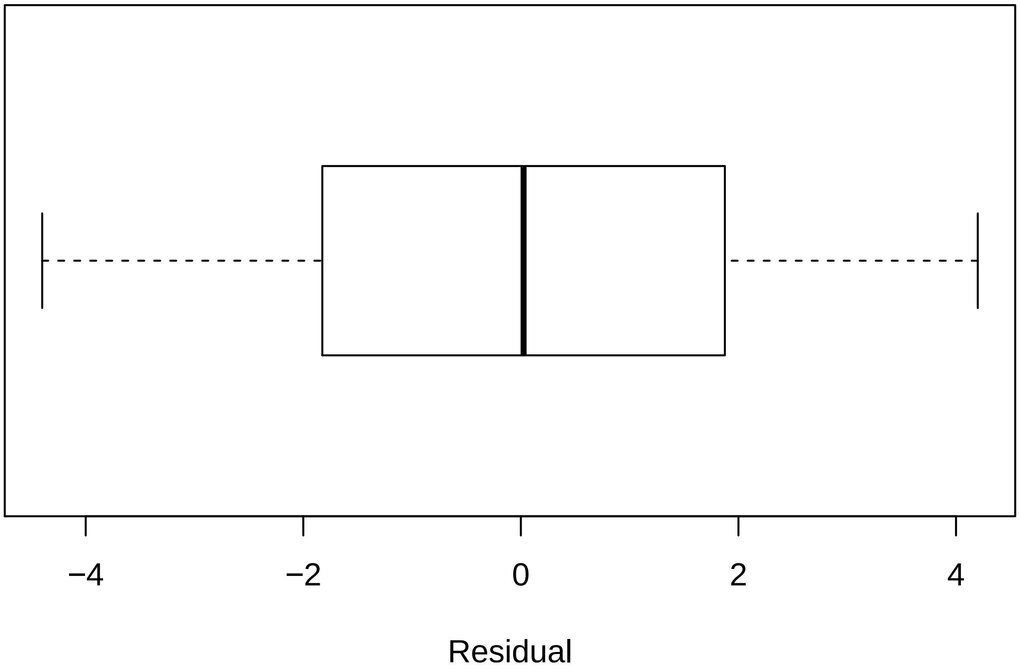
The estimated regression function is $37.650 + 4.425X_1 + 4.375X_2$. The estimated regression coefficient $b_1 = 4.425$ can be interpreted as the average change in how much a brand is liked when the moisture content is increased by one unit, holding sweetness constant.

```
linmod <- lm(Y~X1 + X2, data = data)
```

(c)

The box plot of the residuals indicates that the residuals are approximately symmetric about zero, with no outliers, as we would expect if the multiple linear regression assumptions hold.

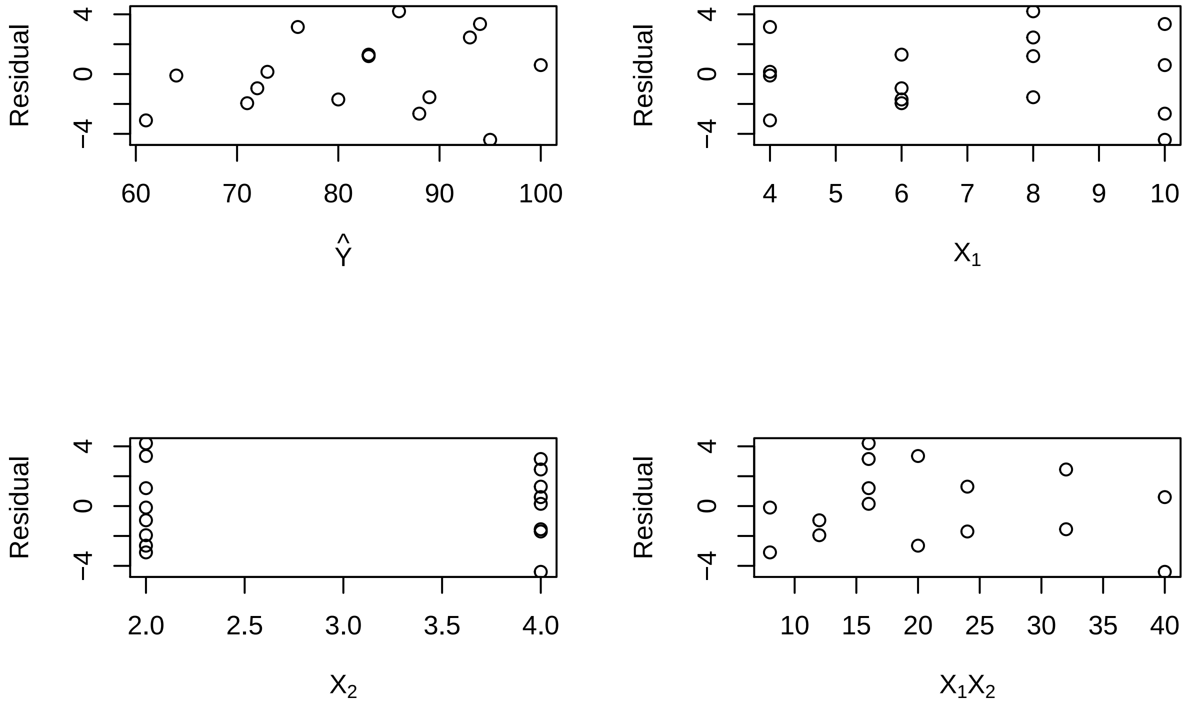
```
boxplot(linmod$residuals, horizontal = TRUE, xlab = "Residual")
```



(c)

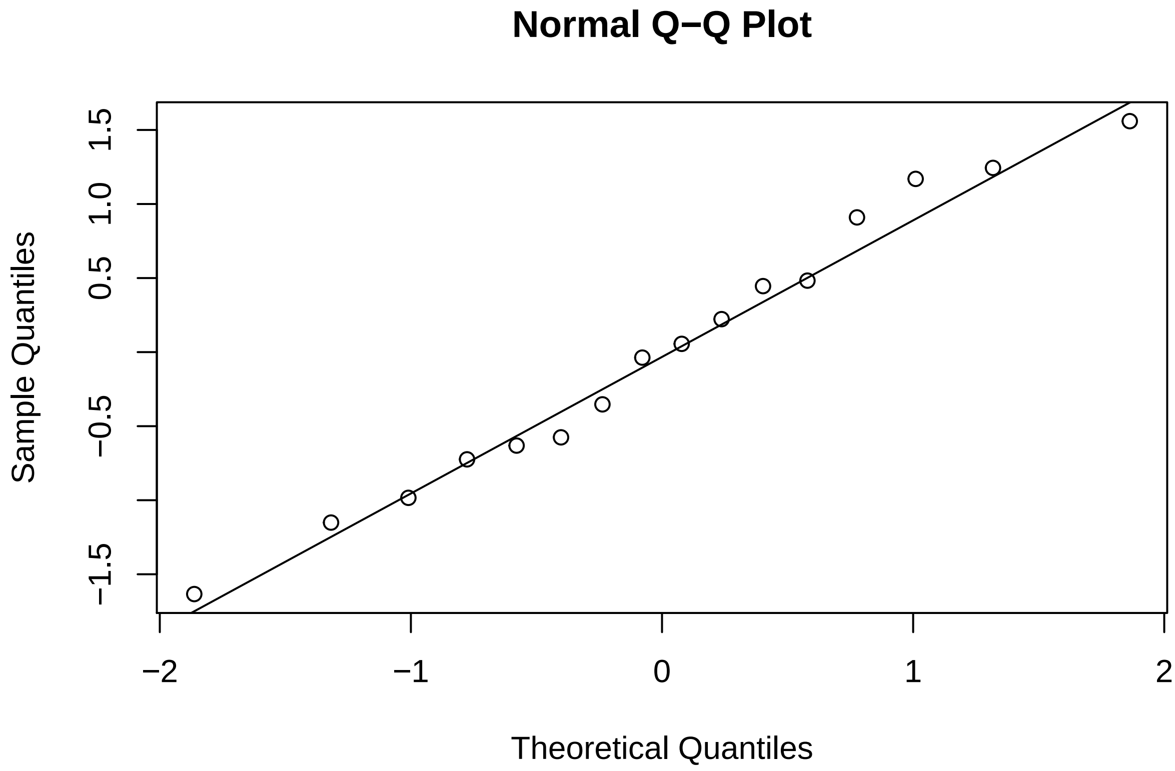
When we examine the residuals as a function of the fitted values $\hat{Y}$, we do not see any systematic trends that would suggest nonconstant variance of the response Y. When we examine the residuals as a function of each covariate X_1 and X_2 separately, we do not see any systematic trends that would suggest that the response Y is a nonlinear function of X_1 or X_2. When we examine the residuals as a function of the product of the two covariates X_1X_2, we do not see any systematic evidence that suggests an interaction term X_1X_2 should be included.

```
par(mfrow = c(2, 2))
plot(data$Y, linmod$residuals,
     xlab = expression(hat(Y)), ylab = "Residual")
plot(data$X1, linmod$residuals,
     xlab = expression(X[1]), ylab = "Residual")
plot(data$X2, linmod$residuals,
     xlab = expression(X[2]),ylab = "Residual")
plot(data$X1*data$X2, linmod$residuals,
     xlab = expression(paste(X[1], X[2], sep = "")),
     ylab = "Residual")
```



A normal probability plot of the residuals suggests that the residuals are consistent with what we would expect if the errors were independent, identically distributed, and normal with mean zero and constant variance.

```
qqnorm(linmod$residuals/summary(linmod)$sigma)
qqline(linmod$residuals/summary(linmod)$sigma)
```



2. Problem 6.6 from the `.pdf` version of the textbook, parts (a)-(b). Requires use of the `brand_preference`

data that has been posted on the Homework page.

(a)

We can perform a level-$\alpha = 0.01$ test of the null hypothesis H_0 that there is no regression relation using an F test of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$. The decision rule for a level-$\alpha = 0.01$ test based on the test statistic $F^* = MSR/MSE$ would be:

- If $F^* \leq F(0.99; 2; n-3)$, conclude H_0
- If $F^* > F(0.99; 2; n-3)$, conclude H_a.

Because $F^* = 129.08$ and $F(0.99; 2; 13) = 6.70$, we would reject H_0 and conclude H_a. This means that we conclude that one or both of the regression coefficients β_1 and β_2 are nonzero.

```
n <- nrow(data)
ssto <- sum((data$Y - mean(data$Y))^2)
sse <- sum((data$Y - linmod$fitted.values)^2)
ssr <- ssto - sse
msr <- ssr/2
mse <- sse/(n - 3)
F.star <- msr/mse
F.quantile.01 <- qf(0.99, 2, n - 3)
```

(b)

The p-value of the test in (a) is 2.66×10^{-9}.

```
p.val <- pf(F.star, 2, n - 3, lower.tail = FALSE)
```

3. Problem 6.7 from the `.pdf` version of the textbook, part (a). Requires use of the `brand_preference` data that has been posted on the Homework page.

(a)

The coefficient of multiple determination R^2 here 0.95. We can interpret this the proportionate reduction of the total variation in the response Y associated with the use of X_1 and X_2. Alternatively, we can interpret this as the proportion of variability in how much a brand is liked Y that can be explained by a linear model that is comprised of measures of moisture X_1 and sweetness X_2.

4. Problem 6.9 from the `.pdf` version of the textbook. Requires use of the `grocery_retailer` data that has been posted on the Homework page. Again, we have not explictly covered the idea of making a correlation matrix in class, but you have the tools to do so. Simply make a table that contains the correlations between the response Y and the predictors X_1, X_2, and X_3 like the one shown in Figure 6.4 (b) in the `.pdf` version of the textbook.

(a)

A stem and leaf plot of the covariate X_1 is given by: Stem | Leaf —|—- 21 | 2 22 | 8 23 | 24 | 668 25 | 027 26 | 15688999 27 | 022477 28 | 38 29 | 0023467 30 | 123367 31 | 7 32 | 22388 33 | 34 | 35 | 2 36 | 7 37 | 0 38 | 3 39 | 40 | 41 | 24 42 | 7 43 | 44 | 3 45 | 46 | 47 | 2

To be honest, a stem and leaf plot of X_1 is not very easy to construct because values of X_1 are so large. However, we can see that X_1 is concentrated in the range 200,000-320,000, and has gaps from 330,000-470,000.

A stem and leaf plot of the covariate X_2 is given by:

Stem	Leaf
46	1
48	
50	
52	

Stem	Leaf
54	
56	
58	2
60	4
62	0747
64	59
66	1256
68	2844
70	257
72	12335699
74	5
76	172279
78	35929
80	12829
82	08
84	06
86	1
88	
90	1
92	
94	
96	5

We can see that X_2 the minimum value of X_2 is an outlier. Most values of X_2 are between 5.7 and 8.5, and there are gaps from 4.6-5.6 and 8.6-9.6.

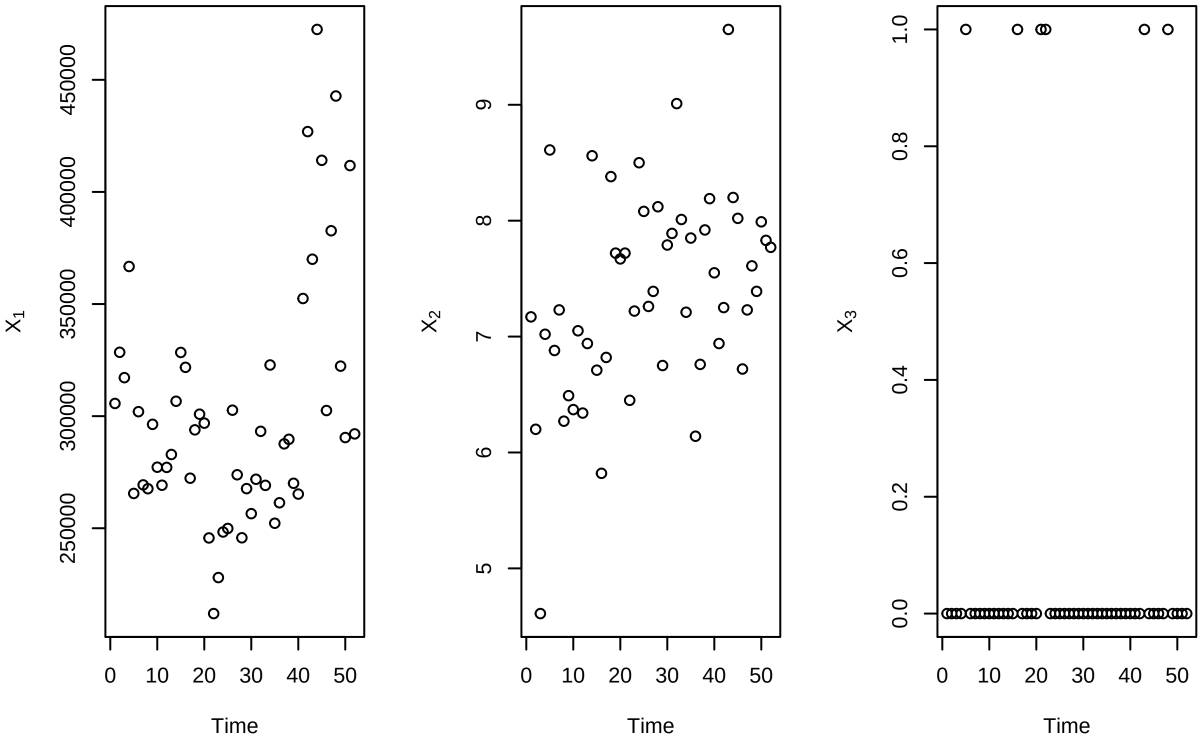
```
link <- url("http://maryclare.github.io/stat525/content/homework/grocery_retailer.RData")
load(link)
close(link)
```

```
stem(data$X1, scale = 3)
stem(data$X2, scale = 3)
```

(b)

The plots show that the values of covariates X_1 and X_2 tend to increase over time. It is more difficult to see if there is a systematic relationship between X_3 and time because X_3 is a binary variable, but there are no obvious patterns.

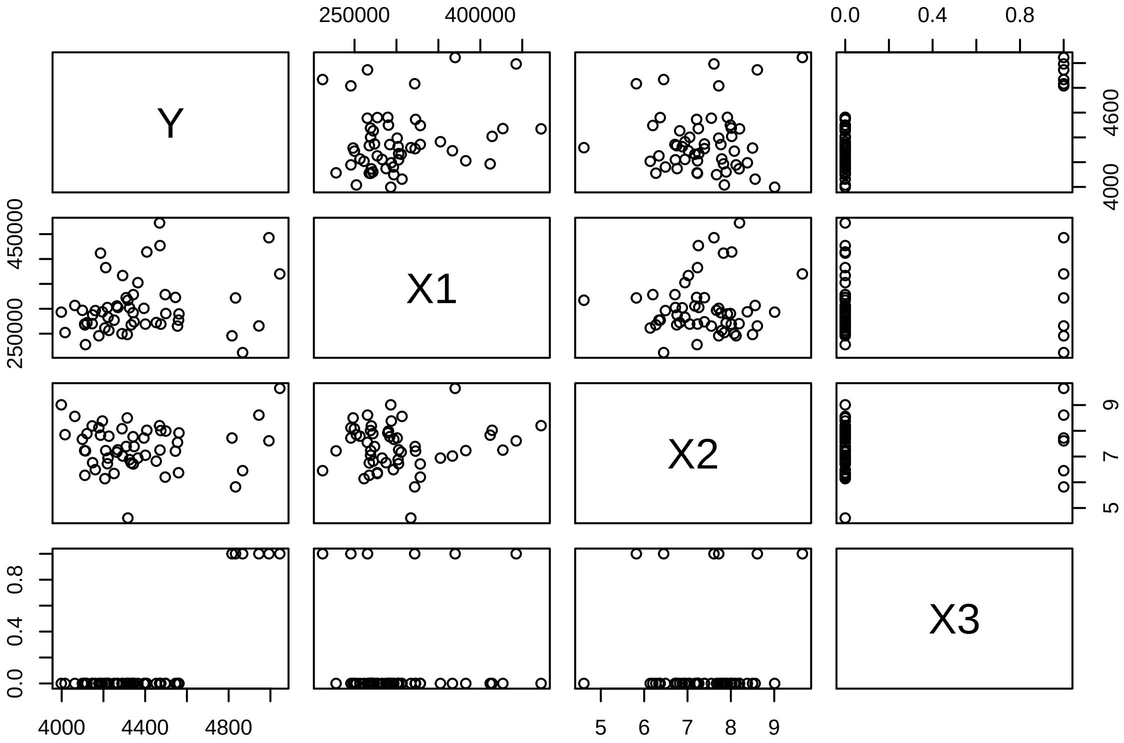
```
par(mfrow = c(1, 3))
plot(data$X1, xlab = "Time", ylab = expression(X[1]))
plot(data$X2, xlab = "Time", ylab = expression(X[2]))
plot(data$X3, xlab = "Time", ylab = expression(X[3]))
```



(c)

Scatterplots of the response Y against the covariates X_1, X_2, and X_3, as well as a plot of the covariates X_1, X_2, and X_3 against each other are given below.

```
pairs(data)
```



The correlation matrix is printed below.

Y	X_1	X_2	X_3
Y	1.00	0.21	0.06
X_1	0.21	1.00	0.08
X_2	0.06	0.08	1.00
X_3	0.81	0.05	0.11

```
cor(data)
```

The scatterplots and correlation matrix show that the response Y is approximately increasing in all three covariates X_1, X_2, and X_3. The covariates X_1 and X_2 are also positively correlated with each other, but not strangly correlated with X_3.

5. Problem 6.10 from the `.pdf` version of the textbook, parts (a)-(d). Requires use of the `grocery_retailer` data that has been posted on the Homework page.

(a)

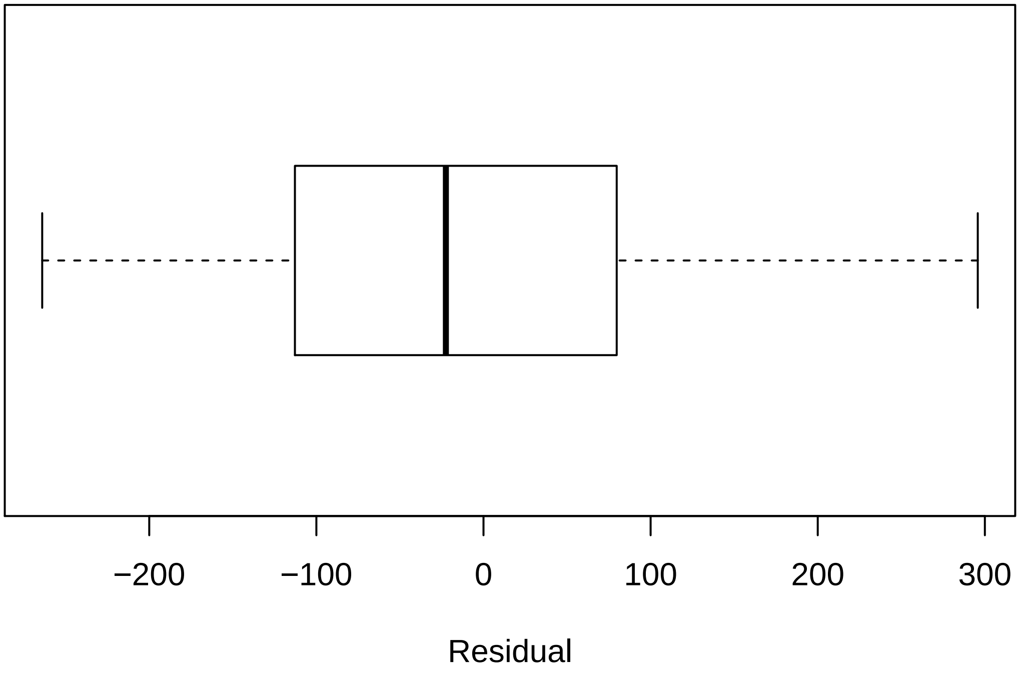
The estimated regression function is $4149.89 + 7.87 \times 10^{-4} X_1 - 13.17 X_2 + 623.55 X_3$. The estimated regression coefficient b_1 is interpreted as the average change in total labor hours when one more case is shipped, holding the indirect costs of total labor hours and whether or not a holiday occurs during the week constant. The estimated regression coefficient b_2 is interpreted as the average change in total labor hours when the indirect costs of total labor hours increase by one percent, holding the number of cases shipped and whether or not a holiday occurs during the week constant. The estimated regression coefficient b_3 is interpreted as the average change in total labor hours when a week includes a holiday, as compared to when a week does not include a holiday, holding the number of cases shipped and whether or not a holiday occurs during the week constant.

```
linmod <- lm(Y~X1 + X2 + X3, data = data)
```

(b)

A box plot of the residuals is shown below. It indicates that the residuals are approximately symmetric about 0 with no substantial outliers, as we would expect if the multiple linear regression model assumptions hold.

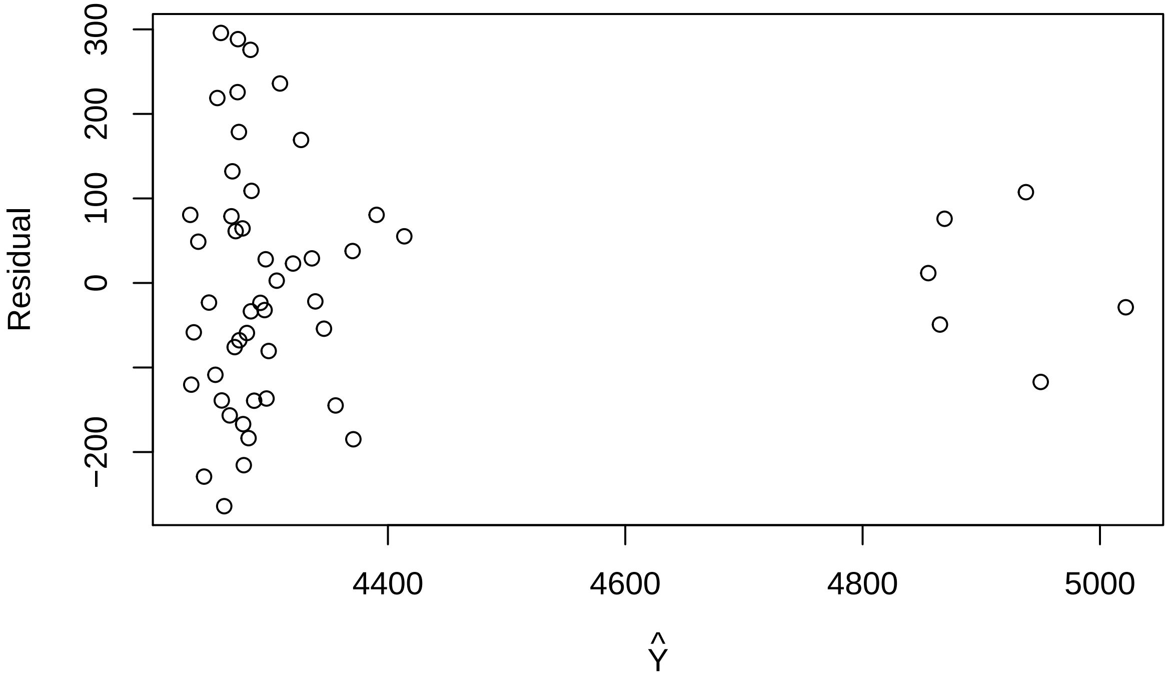
```
boxplot(linmod$residuals, horizontal = TRUE,
        xlab = "Residual")
```



(c)

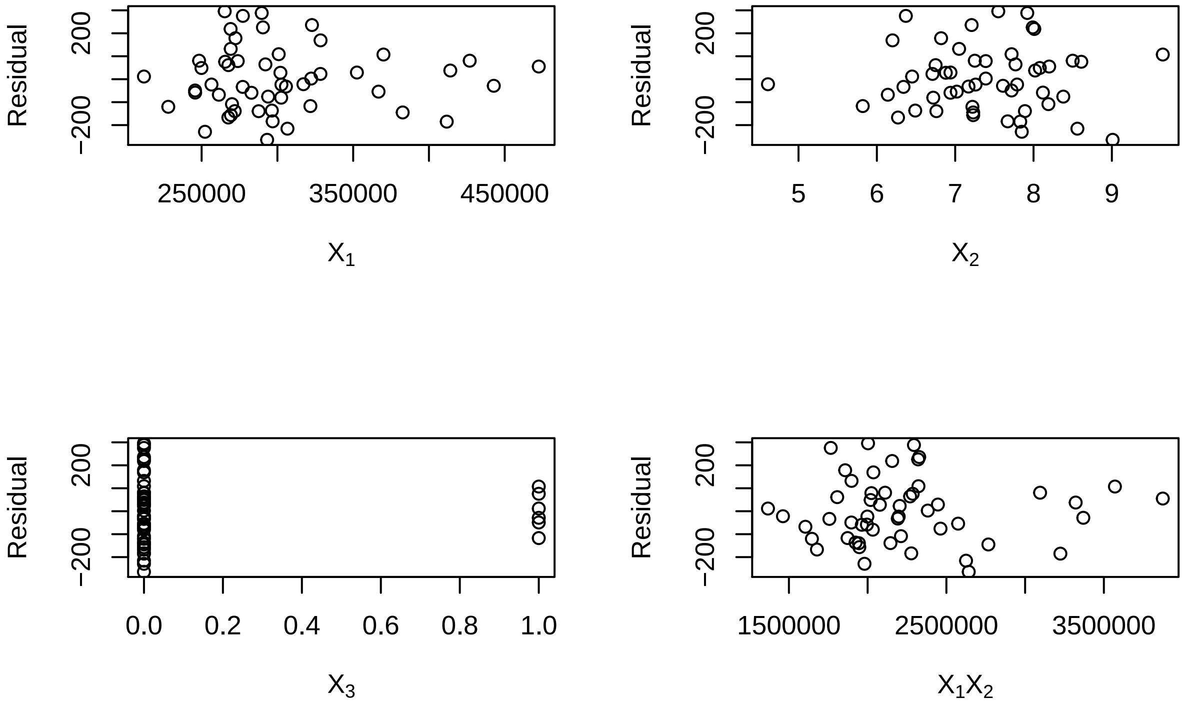
A scatterplot of the residuals against the fitted values shows some evidence of nonconstant variance - the six residuals corresponding to the largest fitted values appear to be less variable than the residuals corresponding to smaller fitted values.

```
plot(linmod$fitted.values, linmod$residuals,
     xlab = expression(hat(Y)),
     ylab = "Residual")
```



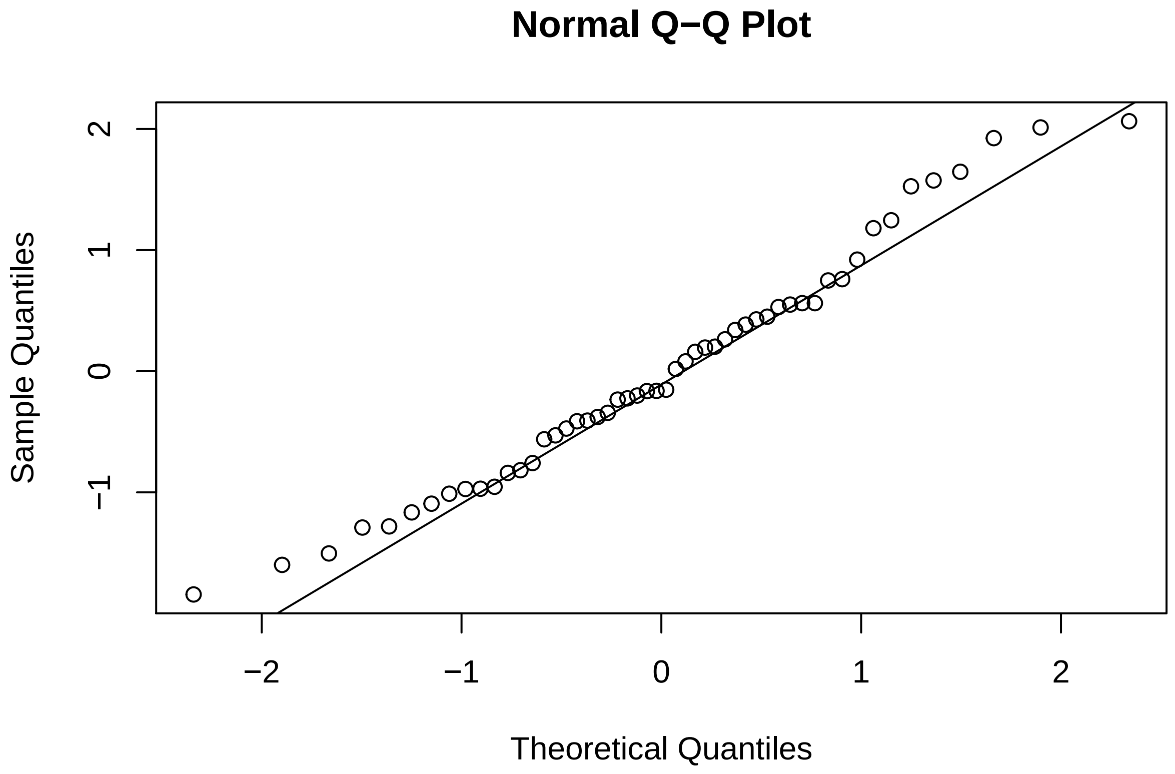
Scatterplots of the residuals against the covariates themselves and the product of covariates X_1 and X_2 do not show any systematic trends that would suggest that the covariates need to be transformed or that interactions need to be added.

```
par(mfrow = c(2, 2))
plot(data$X1, linmod$residuals,
     xlab = expression(X[1]),
     ylab = "Residual")
plot(data$X2, linmod$residuals,
     xlab = expression(X[2]),
     ylab = "Residual")
plot(data$X3, linmod$residuals,
     xlab = expression(X[3]),
     ylab = "Residual")
plot(data$X1*data$X2, linmod$residuals,
     xlab = expression(paste(X[1], X[2], sep = "")),
     ylab = "Residual")
```



The normal probability plot of the residuals indicates that most of the residuals take on values that are consistent with what we would expect if the errors were in fact mean zero, independent, and normally distributed with a constant variance.

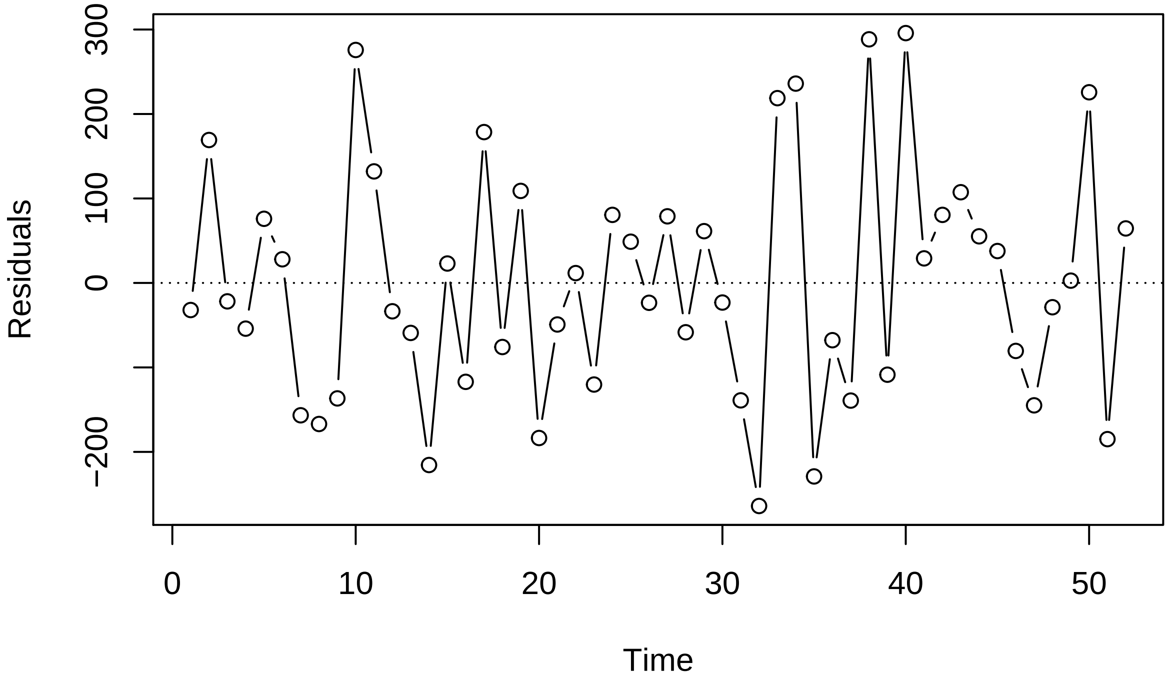
```
qqnorm(linmod$residuals/summary(linmod)$sigma)
qqline(linmod$residuals/summary(linmod)$sigma)
```



(d)

A plot of the residuals over time shows some evidence that sequential residuals are more likely to have the same sign. This suggests that the error terms may be correlated across time, which violates an assumption of the multiple linear regression model.

```
plot(linmod$residuals, xlab = "Time",
     ylab = "Residuals", type = "b")
abline(h = 0, lty = 3)
```



6. Problem 6.11 from the `.pdf` version of the textbook, part (a). Requires use of the `grocery_retailer` data that has been posted on the Homework page.

(a)

We can perform a level-$\alpha = 0.05$ test of the null hypothesis H_0 that there is no regression relation using an F test of the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ against the alternative $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$. The decision rule for a level-$\alpha = 0.05$ test based on the test statistic $F^* = MSR/MSE$ would be:

- If $F^* \leq F(0.95; 3; n-4)$, conclude H_0
- If $F^* > F(0.95; 3; n-4)$, conclude H_a.

Because $F^* = 35.34$ and $F(0.95; 3; 48) = 2.80$, we would reject H_0 and conclude H_a. This means that we conclude that at least one of the regression coefficients β_1, β_2, and β_3 are nonzero. The p-value of this test is 3.16×10^{-12}.

```
n <- nrow(data)
ssto <- sum((data$Y - mean(data$Y))^2)
sse <- sum((data$Y - linmod$fitted.values)^2)
ssr <- ssto - sse
msr <- ssr/3
mse <- sse/(n - 4)
F.star <- msr/mse
F.quantile.05 <- qf(0.95, 3, n - 4)
p.val <- pf(F.star, 3, n - 4, lower.tail = FALSE)
```

7. Problem 6.22 from the `.pdf` version of the textbook.

(a)

Yes!

(b)

It is not, but taking a log transformation allows it to be expressed in the form of a general linear regression model.

(c)

It is not linear in β_1 and β_2, but if we define $\beta_1^* = \log_{10}(\beta_1)$ then it is a general linear regression model in β_1^* and β_2.

(d)

It is not linear in β_1 and β_2, and there is no transformation that allows it to be expressed as a general linear regression model.

(e)

It is not linear in β_1 and β_2, but taking the inverse of both sides allows it to be expressed as a general linear regression model.

8. Problem 6.25 from the `.pdf` version of the textbook.

The analyst could obtain estimates of β_0, β_1, and β_3 by least squares by defining a new response $Y_i^* = Y_i - 4X_{i2}$ and then fitting the regression model $Y_i^* = \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon_i$.

9. Problem 6.27 from the `.pdf` version of the textbook.

```r
X <- cbind(c(1, 1, 1, 1, 1, 1),
           c(7, 4, 16, 3, 21, 8),
           c(33, 41, 7, 49, 5, 31))
Y <- c(42, 33, 75, 28, 91, 55)
```

(a)

We obtain $\boldsymbol{b} = \begin{pmatrix} 33.93 \\ 2.78 \\ -0.26 \end{pmatrix}$.

```r
b <- solve(t(X)%*%X)%*%t(X)%*%Y
```

(b)

We obtain $\boldsymbol{e} = \begin{pmatrix} -2.70 \\ -1.23 \\ -1.64 \\ -1.33 \\ -0.09 \\ 6.99 \end{pmatrix}$.

```r
e <- Y - X%*%b
```

(c)

We obtain $\boldsymbol{H} = \begin{pmatrix} 0.23 & 0.25 & 0.21 & 0.15 & -0.05 & 0.21 \\ 0.25 & 0.31 & 0.09 & 0.27 & -0.15 & 0.22 \\ 0.21 & 0.09 & 0.70 & -0.32 & 0.10 & 0.20 \\ 0.15 & 0.27 & -0.32 & 0.61 & 0.14 & 0.15 \\ -0.05 & -0.15 & 0.10 & 0.14 & 0.94 & 0.02 \\ 0.21 & 0.22 & 0.20 & 0.15 & 0.02 & 0.20 \end{pmatrix}$.

```
H <- X%*%solve(t(X)%*%X)%*%t(X)
```

(d)

We obtain $SSR = 62.07$.

```
ssr <- sum(e^2)
```

(e)

We obtain $\boldsymbol{s}^2\left\{\boldsymbol{b}\right\} = \begin{pmatrix} 715.47 & -34.16 & -13.59 \\ -34.16 & 1.66 & 0.64 \\ -13.59 & 0.64 & 0.26 \end{pmatrix}$.

```
n <- length(Y)
s.sq.b <- ssr*solve(t(X)%*%X)/(n - 3)
```

(f)

We obtain $\hat{Y}_h = 53.85$.

```
X.h <- c(1, 10, 30)
Y.hat.h <- X.h%*%b
```

(g)

We obtain $s^2\left\{\hat{Y}_h\right\} = 5.42$.

```
X.h <- c(1, 10, 30)
s.sq.hat.h <- t(X.h)%*%s.sq.b%*%X.h
```