	Q1	Q2	Q3	Q4	Q5
50 Points	10	14	8	8	10

Question 1

For multiple regression

$$y = X\beta + \epsilon,\ \epsilon \sim \mathrm{N}(0,\, \sigma^2)$$

$$\underset{n\times 1}{y} \qquad \underset{n\times p}{X} \qquad \underset{p\times 1}{\beta} \qquad \underset{n\times 1}{\epsilon}$$

Derive or show that

a. $\hat{\beta} = (\mathrm{X'X})^{-1}\mathrm{X'Y}$

$$y = X\beta + \epsilon$$

$$\text{Minimize: } S(\beta) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon$$

$$\begin{aligned} S(\beta) &= (y - X\beta)'(y - X\beta) \\ &= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta \\ &\text{(since } \beta'X'y \text{ is 1 x 1, } \beta'X'y = y'X\beta) \\ &= y'y - 2\beta'X'y + \beta'X'X\beta \end{aligned}$$

So,

$$\frac{\partial S}{\partial \beta}\Big|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta}$$

$$-2X'y + 2X'X\hat{\beta} = 0$$

$$2X'X\hat{\beta} = 2X'y$$

$$X'X\hat{\beta} = X'y$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

b. $\mathrm{E}[\hat{\beta}] \ = \beta$

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \mathrm{E}[(X'X)^{-1}X'y] \\ &= (X'X)^{-1}X'\mathrm{E}[y] \\ &= (X'X)^{-1}X'(X\beta + 0) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

c. $\mathrm{V}[\hat{\beta}] \ = \sigma^2(\mathrm{X'X})^{-1}$

$$
\begin{aligned}
V[\hat{\beta}] &= V[(X'X)^{-1}X'y] \\
&= (X'X)^{-1}X' \times V[y] \times ((X'X)^{-1}X')' \\
&= (X'X)^{-1}X' \times V[y] \times X((X'X)^{-1})' \\
&= (X'X)^{-1}X' \times V[y] \times X((X'X)')^{-1} \\
&= (X'X)^{-1}X' \times V[y] \times X(X'X)^{-1} \\
&= (X'X)^{-1}X' \times X(X'X)^{-1} \times V[y] \\
&= (X'X)^{-1}X'X(X'X)^{-1} \times V[y] \\
&= (X'X)^{-1}V[y] \\
&= \sigma^2(X'X)^{-1}
\end{aligned}
$$

d. $\mathrm{E}[\hat{Y}] = \mathrm{X}\beta$

$$
\begin{aligned}
\mathrm{E}[\hat{Y}] &= \mathrm{E}[\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 ...] \\
&= \mathrm{E}[X\hat{\beta}] \\
&= X \times \mathrm{E}[\hat{\beta}] \\
&= X\beta
\end{aligned}
$$

e. $\mathrm{V}[\hat{Y}] = \sigma^2\mathrm{H}$, where H is the hat matrix and $\mathrm{H} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'$

$$
\begin{aligned}
V[\hat{Y}] &= V[\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 ...] \\
&= V[X\hat{\beta}] \\
&= XV[\hat{\beta}]X' \\
&= X\sigma^2(X'X)^{-1}X' \\
&= \sigma^2 X(X'X)^{-1}X' \\
&= \sigma^2 H
\end{aligned}
$$

Question 2 (problems 3.1 and 3.3 on page 121)

a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.8084	7.9009	-0.23	0.8209
x$x2	0.0036	0.0007	5.18	0.0000
x$x7	0.1940	0.0882	2.20	0.0378
x$x8	-0.0048	0.0013	-3.77	0.0009

b. Construct the analysis-of-variance table and test for significance of regression.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x$x2	1	76.19	76.19	26.17	0.0000
x$x7	1	139.50	139.50	47.92	0.0000
x$x8	1	41.40	41.40	14.22	0.0009
Residuals	24	69.87	2.91		

To test for significance of regression, we establish H_0 and H_a:

H_0: $\beta_2 = \beta_7 = \beta_8 = 0$

H_a: $\beta_j \neq 0$ for at least one of j = 2, 7, 8

We reject H_0 if $F_{0,j} > F_{0.05=\alpha, \quad 9, \quad 18=(28-9-1)}$ for any $F_{0,j}$ ***** 28- 9- 1 or 28- 3- 1?**

$F_{0,2} = 26.17 > 2.4563$

$F_{0,7} = 47.92 > 2.4563$

$F_{0,8} = 14.22 > 2.4563$

So, reject H_0. There is evidence to conclude that there is a linear relationship for y $\sim x_2$, y $\sim x_7$, and y $\sim x_8$

c. Calculate t statistics for testing the hypotheses H_0: $\beta_2 = 0$, H_0: $\beta_7 = 0$, H_0: $\beta_8 = 0$. What conclusions can you draw about the roles the variables x_2, x_7, and x_8 play in the model?

(1) R:

i) H_0: $\beta_2 = 0$
$\beta_2 = 0.003598$, t = 5.177, $t_{\frac{0.05}{2},24} = 2.064 \longrightarrow |5.117| > 2.064 \Rightarrow$ Reject H_0

ii) H_0: $\beta_7 = 0$
$\beta_7 = 0.193960$, t = 2.198, $t_{\frac{0.05}{2},24} = 2.064 \longrightarrow |2.198| > 2.064 \Rightarrow$ Reject H_0

iii) H_0: $\beta_8 = 0$
$\beta_8 = -0.004816$, t = -3.771, $t_{\frac{0.05}{2},24} = 2.064 \longrightarrow |-3.771| > 2.064 \Rightarrow$ Reject H_0

(2) Manual:

i) H_0: $\beta_2 = 0$
$\beta_2 = 0.003598$, $t_{\frac{0.05}{2},24} = 2.064$

$$\begin{aligned} t &= \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} \\ &= \frac{0.003598}{0.000695} \\ &= 5.177 \end{aligned}$$

$|5.117| > 2.064 \Rightarrow$ Reject H_0

ii) H_0: $\beta_7 = 0$
$\beta_7 = 0.193960$, $t_{\frac{0.05}{2},24} = 2.064$

$$\begin{aligned} t &= \frac{\hat{\beta}_7 - 0}{se(\hat{\beta}_7)} \\ &= \frac{0.193960}{0.088233} \\ &= 2.198 \end{aligned}$$

$|2.198| > 2.064 \Rightarrow$ Reject H_0

iii) H_0: $\beta_8 = 0$

β_8 = -0.004816, $t_{\frac{0.05}{2},24} = 2.064$

$$t = \frac{\hat{\beta}_8 - 0}{se(\hat{\beta}_8)} = \frac{-0.004816}{0.001277} = -3.771$$

$|-3.771| > 2.064 \Rightarrow$ Reject H_0

d. Calculate R^2 and R^2_{adj} for this model.

(1) R:

$R^2 \longrightarrow$ summary(model)$r.squared yields **0.7863069**

$R^2_{adj} \longrightarrow$ summary(model)$adj.r.squared yields **0.7595953**

(2) Manual:

Knowing: $SS_T = SS_R + SS_{res}$

From anova(model) in R:

SS_T = (76.193 + 139.501 + 41.400) (SS_R) + 69.870 (SS_{res}) = 326.964

$$R^2 = 1 - \frac{SS_{res}}{SS_T} = 1 - \frac{69.870}{326.964} = 0.7863067$$

$$R^2_{\text{adj}} = \frac{1 - \frac{SS_{\text{res}}}{(n-p)}}{\frac{SS_{\text{T}}}{(n-1)}} = 1 - \frac{SS_{res}(n-1)}{SS_T(n-k-1)} = 1 - \frac{69.870(27)}{326.964(24)} = 0.7595951$$

e. Using the partial F test, determine the contribution of x_7 to the model. How is this partial F statistic related to the t test for β_7 calculated in part c above?

Knowing:

The partial F-test is the most common method of testing for a nested normal linear regression model. "Nested" model is just a fancy way of saying a reduced model in terms of variables included.

If $F_0 > F_{\alpha,r,n-p}$, we reject H_0, concluding that at least one of the parameters in β_2 is not zero, and consequently at least one of the regressors x_{k-r+1}, x_{k-r+2}, . . . , x_k in X_2 contribute significantly to the regression model. Some authors call the test in (3.35) a partial F test because it measures the contribution of the regressors in X_2 given that the other regressors in X_1 are in the model.

Partial F-Test:

H_0: $\beta_2 = 0$

$$F_0 = \frac{SS_R(\beta_1|\beta_2)}{r \times MS_{res}}$$

where $\beta_1 = \beta$ - $\{\beta_7\}$, $\beta_2 = \beta_7$

$$SS_R(\beta_2|\beta_1) = SS_R(\beta) - SS_R(\beta_1)$$
$$SS_R(\beta_2|\beta_1) = (76.193 + 139.501 + 41.400) - (76.193 + 41.400)$$
$$= 139.501$$

r $= 1$

$MS_{res} = 2.911$

$$F_0 = \frac{139.501}{1 \times 2.911}$$
$$= 47.92202$$

$F_{\alpha,r,n-p} = F_{0.05,1,(28-(3+1)=24} = 4.2597$

Reject H_0 if $F_0 > F_{0.05,1,24}$

$47.92202 > 4.2597 \longrightarrow$ reject H_0

anova(lm(y ~x7))$F yields 11.00524

qf(0.025, df1 = 1, df2 = 24, lower.tail = F) yields 5.713369

f. Find a 95% CI on β_7. (This is part a of problem 3.3, and the following one is part b of problem 3.3.)

(1) R:

(2) Manual:

A CI for β_j is $\hat{\beta}_j$ (+ or -) $t_{\frac{\alpha}{2},n-p}SE(\hat{\beta}_j)$

$\hat{\beta}_7 = 0.193960$

$t_{\frac{\alpha}{2},n-p} = t_{0.025,28-4=24} = 2.064$

$SE(\hat{\beta}_j = 0.088233$

$(0.193960 - (2.064\times0.088233), 0.193960 + (2.064\times0.088233))$

g. Find a 95% CI on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56.0$, and $x_8 = 2100$.

(1) R:

prediction(

(2) Manual:

Note: For c, d, f, and g, please show two versions of your results: (1) obtained using R code and (2) based on your manual calculation (please show detailed step for your manual calculation. You can use the partial output from the lm or ANOVA, e.g., the SS_{reg}, SS_{res}, the estimated value of β and its variance or standard deviation). If you can show how to get the t-statistics (or CI, R-square) based on part of the output obtained from R, that will be fine.

Question 3 (Exercise 3.4 on page 122)

Reconsider the National Football League data from Problem 3.1. Fit a model to this data using only x_7 and x_8 as the regressors.

a. Test for significance of the regression.

b. Calculate R^2 and R^2_{adj}. How do these quantities compare to the values computed for the model in problem 3.1, which included an additional regressor (x^2)?

c. Calculate a 95% CI on β_7. Also, find a 95% CI on the mean number of games won by a team when $x_7 = 56.0$ and $x_8 = 2100$. Compare the lengths of these CIs to the lengths of the corresponding CIs from problem 3.3 (that is, the above part f and g in question 2)

d. What conclusions can you draw from this problem about the consequences of omitting an important regressor from a model?

Question 4 (exercise 4.2 on page 165)

Consider the multiple regression model fit to the National Football League (NFL) team performance data in problem 3.1.
can use qq norm for this one

a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

   model.resid = resid(model)

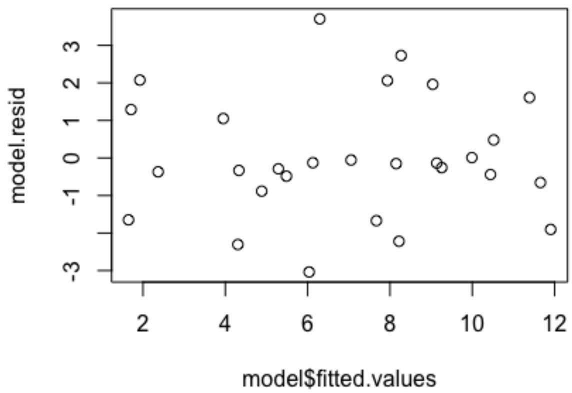
   qqnorm(model.resid, main = "Games Won vs Passing Yards / Rushing", xlab = "Passing Yards / % Rushing / Opponents Rushing", ylab = "Games Won")

   qqline(model.resid)

   I don't think so. Since the model's data follows an imagined normal distribution line fairly closely, it seems reasonable to assume normality.

b. Construct and interpret a plot of the residuals versus the predicted response.

   plot(model$fitted.values, model.resid)



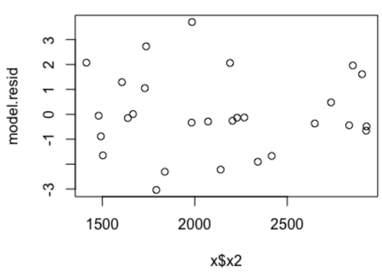
It looks like static, indicating that there is no relationship between the residuals and predicted response, supporting the assumption that the errors are independent.

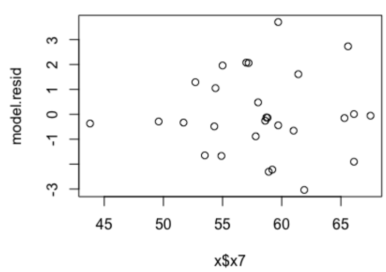
c. Construct plots of the residuals versus each of the regressor variables. Do these plots imply that the regressor is correctly specified?

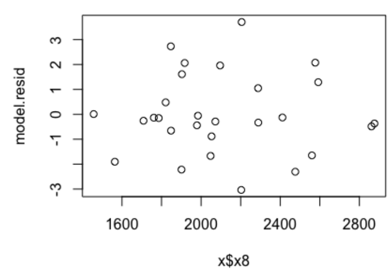
plot(x$x2, model.resid)

plot(x$x7, model.resid)

plot(x$x8, model.resid)







All 3 plots imply that the regressor is correctly specified. For x_7 specifically, it looks like the variance is a little higher on the right side, implying the variance isn't exactly constant, but it doesn't look like it changes the distribution, so it should still be good.

d. Construct the partial regression plots for this model. Compare the plots with the plots of residuals versus regressors from part c above. Discuss the type of information provided by these plots.

Question 5

Show that the hat matrix H $= X(X'X)^{-1}X'$ and I - H (where I is the identity matrix) are symmetric and idempotent. That is, please show:

a. H′ = H and HH = H (H′ means the transpose of H, HH means H * H)

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ H' &= (X(X'X)^{-1}X')' \\ &= X((X'X)^{-1})'X' \\ &= X((X'X)')^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= H \end{aligned}$$

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ HH &= (X(X'X)^{-1}X')(X(X'X)^{-1}X') \\ HH &= X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= H \end{aligned}$$

b. (I - H)′ = I - H and (I - H)(I - H) = I - H

$$\begin{aligned} (I-H)' &= (I - X(X'X)^{-1}X')' \\ &= I' - (X(X'X)^{-1}X')')' \\ &= I - (X(X'X)^{-1}X')')' \\ &= I - X(X'X)^{-1}X' \\ &= I - H \end{aligned}$$

$$\begin{aligned} (I-H)(I-H) &= (I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') \\ &= I - 2X(X'X)^{-1}X' + (X(X'X)^{-1}X')(X(X'X)^{-1}X') \\ &= I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X' \text{ by (a)} \\ &= I - X(X'X)^{-1}X' \\ &= I - H \end{aligned}$$

Hint: A = X′X is a symmetric matrix, and for a symmetric matrix, $(A')^{-1} = (A^{-1})'$. You can use this property directly in your proof of **(a)** and **(b)**. If you are interested in the proof of this property, you may check the following web page:
https://math.stackexchange.com/questions/325082/is-the-inverse-of-a-symmetric-matrix-also-symmetric