Q 7.2

(a) Fit a second-order polynomial that expresses weight loss as a function of the number of months since production.

```r
# import dataset
dat1 <- read.csv("data-prob-7-2.csv", header = T, sep = ",") # 10 obs. of 2 var.
# dat1$x <- scale(dat1$x, center = T, scale = F)
fit1 <- lm(y ~ x + I(x^2), data = dat1)
s1 <- summary(fit1)
s1
```

The linear regression model between weight loss and the number of months since production, its quadratic term is

\[ \hat{y} = 1.633 - 1.232x + 1.495x^2. \]

When the ratio of the number of months since production \( x \) increases 1 unit, weight loss would increase 0.263 units.

(b) Test for significance of regression.
Since F statistics is $1.859 \times 10^6$ and p-value is 0, which is less than significance level (0.05), we would reject $H_0 : \beta_1 = \beta_2 = 0$ and conclude there is a linear relationship between weight loss $y$ and any of the regressors $x, x^2$.

(c) Test the hypothesis $H_0 : \beta_2 = 0$. Comment on the need for the quadratic term in this model.

```r
anova(fit1)
```

## Analysis of Variance Table
##
## Response: y
## Df Sum Sq Mean Sq F value Pr(>F)
## x 1 42.703 42.703 3355253 < 2.2e-16 ***
## I(x^2) 1 4.607 4.607 361974 < 2.2e-16 ***
## Residuals 7 0.000 0.000
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Since F statistics is 361974 ($4.607 \approx \infty$) and p-value is 0, which is less than significance level (0.05), we would reject $H_0 : \beta_2 = 0$ and conclude that the quadratic term is significant and contributes significantly in the linear model.

(d) Are there any potential hazards in extrapolating with this model?

In general, polynomial models may turn in unanticipated and inappropriate directions, both in interpolation and in extrapolation. In this model, because of the significance of quadratic term, extrapolation may be extremely hazardous.

Q 7.8

(a) Fit a second-order model to these data using orthogonal polynomials.

```r
fit2 <- lm(y ~ poly(x, 2), data = dat1)
s2 <- summary(fit2)
s2
```

## Call:
## lm(formula = y ~ poly(x, 2), data = dat1)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.005364 -0.002727 0.001045 0.002409 0.003273
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.535000 0.001128 3133.4 <2e-16 ***
## poly(x, 2)1 6.534770 0.003568 1831.7 <2e-16 ***
## poly(x, 2)2 2.146377 0.003568 601.6 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.003568 on 7 degrees of freedom
## Multiple R-squared: 1, Adjusted R-squared: 1
## F-statistic: 1.859e+06 on 2 and 7 DF, p-value: < 2.2e-16
The linear regression model between weight loss and orthogonal polynomials of the number of months since production is

\[ \hat{y} = 3.535 + 6.355P_1(x) + 2.146P_2(x). \]

We can also fit the second-order polynomial model following the procedure in textbook. The coefficients of the orthogonal polynomials \( P_0(x_i), P_1(x_i) \) and \( P_2(x_i) \) can be obtained from Table A.5 in the textbook. Then we have \( \sum_{i=1}^{10} P_0^2(x_i) = 10, \sum_{i=1}^{10} P_1^2(x_i) = 330, \sum_{i=1}^{10} P_2^2(x_i) = 132. \)

Using the Weight Loss data, we can obtain \( \sum_{i=1}^{10} P_0(x_i)y_i = 35.35, \sum_{i=1}^{10} P_1(x_i)y_i = 118.71 \) and \( \sum_{i=1}^{10} P_2(x_i)y_i = 24.66. \) Then the model coefficients can be calculated as

\[ \hat{\alpha}_0 = \frac{\sum_{i=1}^{10} P_0(x_i)y_i}{\sum_{i=1}^{10} P_0^2(x_i)} = \frac{35.35}{10} = 3.535. \]
\[ \hat{\alpha}_1 = \frac{\sum_{i=1}^{10} P_1(x_i)y_i}{\sum_{i=1}^{10} P_1^2(x_i)} = \frac{118.71}{330} = 0.359723. \]
\[ \hat{\alpha}_2 = \frac{\sum_{i=1}^{10} P_2(x_i)y_i}{\sum_{i=1}^{10} P_2^2(x_i)} = \frac{24.66}{132} = 0.1868182. \]

The fitted model is

\[ \hat{y} = 3.535 + 0.36P_1(x) + 0.187P_2(x). \]

(b) Suppose that we wish to investigate the addition of a third-order term to this model. Comment on the necessity of this additional term. Support your conclusions with an appropriate statistical analysis.

```
anova(fit2)
```

```r
## Analysis of Variance Table

## Response: y
## Df Sum Sq Mean Sq F value Pr(>F)
## poly(x, 2)  2  47.31 23.655 1858613 < 2.2e-16 ***
## Residuals  7   0.00  0.000
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the results, we can see that the SSR in model with quadratic term is 47.31, which accounts for all of the variation in the data. Thus, the cubic term is not necessary.

Q 8.3

```
# import dataset
dat3 <- read.csv("data-prob-8-3.csv", header = T, sep = ",") # 25 obs. of 3 var.
```

(a) Develop a model that relates delivery time \( y \) to cases \( x_1 \), distance \( x_2 \), and the city in which the delivery was made. Estimate the parameters of the model.

```
s <- c(1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
b <- c(0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
a <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
fit3 <- lm(y ~ x1 + x2 + s + b + a, data = dat3)
s3 <- summary(fit3)
s3
```
## Call:
## lm(formula = y ~ x1 + x2 + s + b + a, data = dat3)
##
## Residuals:
## Min 1Q Median 3Q Max
## -4.4800 -1.5922 -0.5583  1.1045  6.1611
##
## Coefficients:
##                        Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            0.41625    2.25783   0.184   0.85569
## x1                     1.77028    0.18679  9.477  1.24e-08 ***
## x2                     0.01083    0.00379   2.862   0.00999 **
## s                      2.28510    2.41624   0.946   0.35616
## b                      3.73764    2.35702  1.586   0.12930
## a                      0.45264    2.68742 -0.168   0.86803
##
## Signif. codes:  
##     0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.986 on 19 degrees of freedom
## Multiple R-squared: 0.9707, Adjusted R-squared: 0.963
## F-statistic: 125.9 on 5 and 19 DF, p-value: 6.919e-14

We first build 3 dummy variables, $s=1$ indicates the observation is from San Diego, $b=1$ indicates the observation is from Boston, $a=1$ indicates the observation is from Austin and $s=b=a=0$ indicates the observation is from Minneapolis.

The linear regression model between delivery time and 5 regressors (cases $x_1$, distance $x_2$ and city dummy variables) is

$$ \hat{y} = 0.41625 + 1.77028x_1 + 0.01083x_2 + 2.28510s + 3.73764b - 0.45264a. $$

When the cases $x_1$ increases 1 unit, delivery time would increase 1.77028 units. When the distance $x_2$ increases 1 unit, delivery time may increase 0.01083 units. When the city is San Diego, delivery time may increase 2.28510 units compared to Minneapolis given cases and distance constant. When the city is Boston, delivery time may increase 3.73764 units similarly. When the city is Austin, delivery time may decrease 0.45264 units similarly.

(b) Is there an indication that delivery site is an important variable?

```r
anova(fit3)
```

## Analysis of Variance Table
##
## Response: y
##
##          Df Sum Sq Mean Sq  F value Pr(>F)
## x1        1 5382.4  5382.4 7.357e-16 ***
## x2        1  168.4  168.4 1.882e-02
## s         1   0.3   0.3  0.3190 0.860109
## b         1  63.7  63.7  7.1473 0.0150248 *
## a         1   0.3   0.3  0.2840 0.868026
## Residuals 19 169.5   8.9     
##
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

To determine the importance of the delivery site in this model, we use partial F test to check whether $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$. From the Anova table, we know that $SS_R(\beta_3, \beta_4, \beta_5|\beta_0, \beta_1, \beta_2) = 64.3$. Then partial
F statistics is

\[
F = \frac{SSR(\beta_3, \beta_4, \beta_5 | \beta_0, \beta_1, \beta_2)}{MSE_{full}} = \frac{64.3/3}{8.9} = 2.40824 < F_{0.95,3,19} = 3.12735.
\]

Since the F statistics is less than the critical value, we could not reject the \( H_0 \) and conclude that the delivery site may not be an important variable in this model.

(c) Analyze the residuals from this model. What conclusions can you draw regarding model adequacy?

\[
\text{fit3}\_\text{std} \leftarrow \text{rstandard(}\text{fit3}\text{)}
\]

\[
\text{qqnorm}\!(\text{fit3}\_\text{std}, \text{ylab=}'\text{Standardized Residuals}', \text{xlab=}'\text{Normal Scores}')
\]

\[
\text{qqline}\!(\text{fit3}\_\text{std}, \text{col} = 2)
\]

QQ plot for this delivery data residuals has a clear "arched" pattern which indicates of skewed data. Therefore, there may be a problem with normality.

\[
\text{plot}\!(\text{fit3}\_\text{fitted}\_\text{values}, \text{fit3}\_\text{residuals}, \text{ylab} = '\text{Residuals}', \text{xlab} = '\text{Fitted Values}', \text{main} = '\text{Residuals vs Fitted}')
\]

\[
\text{abline}\!(0, 0)
\]
Residuals versus Fitted plot for the delivery data also has a significant nonlinear pattern.

Q 8.11

```r
# import dataset
dat4 <- read.csv("data-prob-8-11.csv", header = T, sep = ",") # 25 obs. of 2 var.
```

(a) Write down the y vector and X matrix for the corresponding regression model.

We use 4 dummy variables to indicate different levels of percent, while y vector is directly from the data set. Then X is a $25 \times 5$ matrix and y is a $25 \times 1$ vector.
$$y = \begin{bmatrix} 7 \\ 7 \\ 15 \\ 11 \\ 9 \\ 12 \\ 17 \\ 12 \\ 18 \\ 18 \\ 14 \\ 18 \\ 18 \\ 19 \\ 19 \\ 19 \\ 25 \\ 22 \\ 23 \\ 7 \\ 10 \\ 11 \\ 15 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

(b) Find the least-squares estimates of the model parameters.

dat4$percent1 <- factor(dat4$percent, levels = c("35","15","20","25","30"))
fit4 <- lm(y ~ percent1, data = dat4)
s4 <- summary(fit4)
s4

# # Call:
# lm(formula = y ~ percent1, data = dat4)
# #
# # Residuals:
# # Min 1Q Median 3Q Max
# # -3.8 -2.6  0.4  1.4  5.2
# #
# # Coefficients:
# # Estimate Std. Error t value Pr(>|t|)
# #(Intercept)  10.800    1.270   8.506  4.46e-08  ***
# # percent15  -1.000    1.796  -0.557  0.58375
# # percent20   4.600    1.796   2.562  0.01859  *
# # percent25   6.800    1.796   3.787  0.00116  **
# # percent30  10.800    1.796   6.015  7.01e-06  ***
# # ---
# # Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# #
# # Residual standard error: 2.839 on 20 degrees of freedom
# # Multiple R-squared: 0.7469, Adjusted R-squared: 0.6963
The linear regression model between tensile strength and 4 regressors (percentage of cotton dummy variables) is

\[ \hat{y} = 10.8 - 1.0x_{15} + 4.6x_{20} + 6.8x_{25} + 10.8x_{30}. \]

The least-squares estimates of the model parameters are \( \hat{\beta}_0 = 10.8, \hat{\beta}_1 = -1.0, \hat{\beta}_2 = 4.6, \hat{\beta}_3 = 6.8, \hat{\beta}_4 = 10.8. \)

(c) Find a point estimate of the difference in mean strength between 15% and 25% cotton.

\[ \text{diff} = \hat{\beta}_1 - \hat{\beta}_3 = -1.0 - 6.8 = -7.8. \]

So the estimate of the difference in mean strength between 15% and 25% cotton is -7.8.

(d) Test the hypothesis that the mean tensile strength is the same for all five cotton percentages.

From the results in part (b), F statistics is 14.76 and p-value is 0, which is less than significance level (0.05). So we would reject \( H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \) and conclude that the mean tensile strength is not the same for all five cotton percentages.