

MTH 427 - Spring 2023

Assignment #3

Due: Monday, March 6th 2023 (11:59PM)

1. **11.43** Refer to exercises 11.5 and 11.17. Use the data and model given there to construct a 95% prediction interval for the median sale price in 1980.

Year	Median Sales Price ($\times 1000$)
1972 (1)	\$27.60
1973 (2)	\$32.50
1974 (3)	\$35.90
1975 (4)	\$39.30
1976 (5)	\$44.20
1977 (6)	\$48.80
1978 (7)	\$55.70
1979 (8)	\$62.90

The summary shows that our equation is $\hat{y} = 4841.7x + 21575$

$$S = 1746 \implies S^2 = 1746^2 = 3048516. \text{ Then } SSE = (n - 2)S^2 = 6 \times 3048516 = 18291096$$

Solution: With $df = n - 2 = 6$, the t -value is $t_{0.025, 6} = 2.447$. Note $\sum x_i = 36$, $\sum x_i^2 = 204$, $\bar{x} = 4.5$. It follows that $S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 204 - \frac{36^2}{8} = 42$

To find a prediction interval for 1980 ($x^* = 9$). Then,

$$\begin{aligned} & \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \\ &= 21575 + 4841.7 \cdot 9 \pm 2.447 \cdot 1746 \sqrt{1 + \frac{1}{8} + \frac{4.5^2}{42}} \\ &= (\$59733.97, \$70566.63) \end{aligned}$$

2. Is the plant density of a species related to the altitude at which data are collected? Let Y denote the species density and X denote the altitude. A fit of a simple linear regression model using 14 observations yielded $\hat{y} = 21.6 - 7.79x$ and $r^2 = 0.61$.

(a) What is the value of the correlation coefficient r ?

Solution: $r^2 = 0.61 \iff r = \pm 0.781$, but because $\hat{\beta}_1 < 0$ then $r = -0.781$.

(b) What proportion of the variation in densities is explained by the linear model using altitude as the independent variable?

Solution: 61% of the variation is explained using altitude as the independent variable in the model.

3. Refer to exercise 11.3. Fit the model suggested there by use of matrices.

Solution:

$$\begin{array}{c|ccccc} y & 3 & 2 & 1 & 1 & 0.5 \\ \hline x & -2 & -1 & 0 & 1 & 2 \end{array}$$

$$X = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \end{bmatrix}, \quad Y = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 1 \\ 0.5 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}, \quad (X^T X)^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 7.5 \\ -6 \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} \cdot X^T Y = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{bmatrix} \begin{bmatrix} 7.5 \\ -6 \end{bmatrix} = \begin{bmatrix} 1.5 \\ -0.6 \end{bmatrix}$$

$$\therefore \hat{y} = -0.6x + 1.5$$

4. Exercise 1

This exercise relates to the **Auto** dataset, which can be found in Canvas. (This is part 2 of Additional exercise Homework #2).

Note: Screenshots of code/outputs at end of problem

- (a) Use the appropriate function in R to fit a quadratic model ($Y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \varepsilon$) with **mpg** as the response variable, and where **horsepower** and **horsepower²** are the predictors.

Solution:

```
1 library(readr)
2 df = read.csv("0:/Arr Matey/Auto.csv", header=T, na.strings="?")
3 df = na.omit(df)
4
5 mpg = df$mpg;
6 hp = df$horsepower
7
8 # 2.2 (a)
9 quadratic_model = lm(mpg~hp+I(hp^2))
10 summary(quadratic_model)
```

- (b) Write out the estimated model in equation form.

Solution:

$$\hat{y}(h) = 56.9000997 + -0.4661896h + 0.0012305h^2$$

- (c) Compute the covariance matrix for the linear regression coefficients estimated.

Solution: `vcov(quadratic_model)`

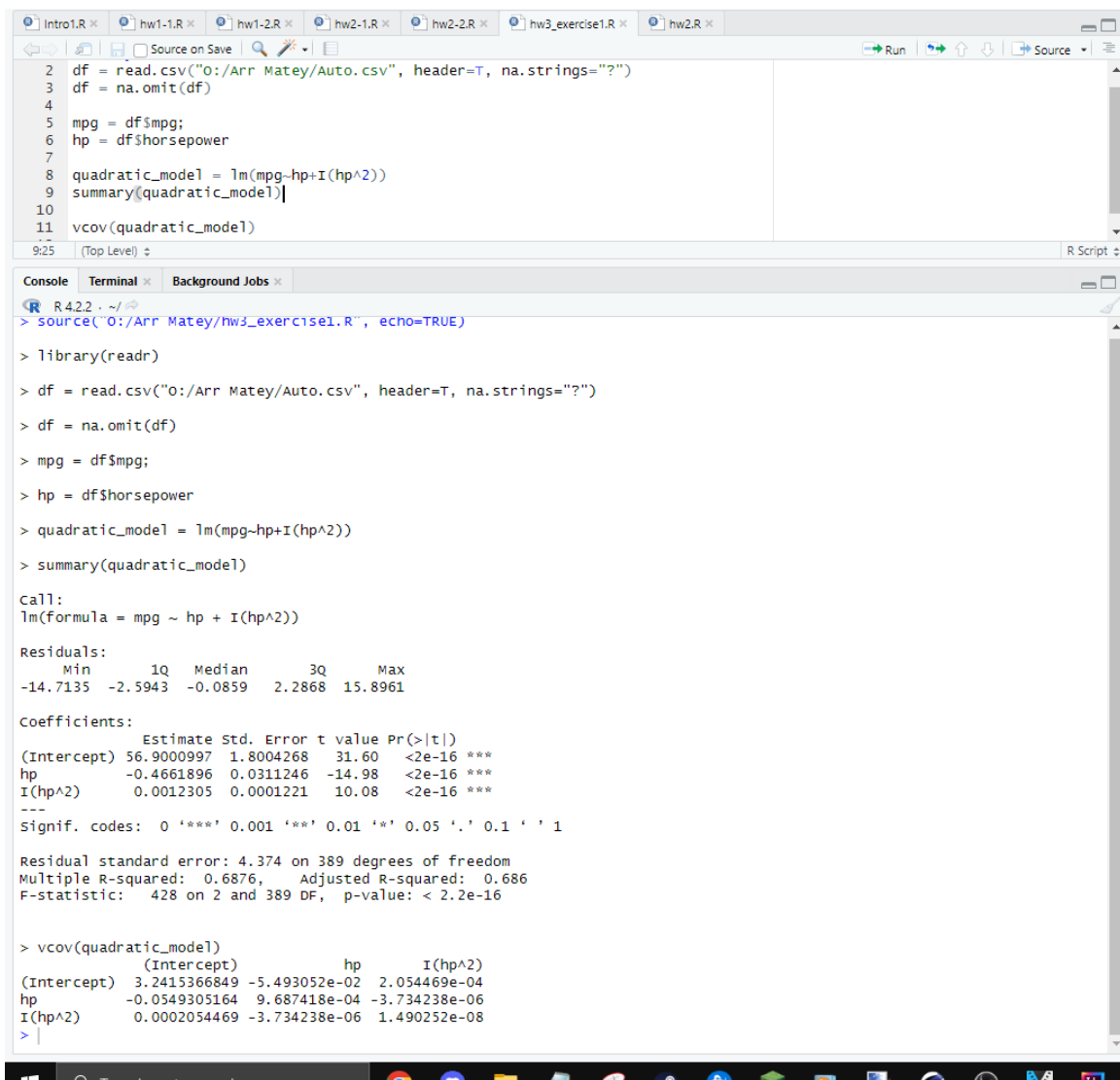
$$\begin{bmatrix} 3.2415366849 & -0.05493052 & 0.0002054469 \\ -0.0549305164 & 0.0009687418 & -3.734238 \cdot 10^{-6} \\ 0.0002054469 & -3.734238 \cdot 10^{-6} & 1.490252 \cdot 10^{-8} \end{bmatrix}$$

- (d) Do the data present sufficient evidence to indicate curvature in the response function? That is to test the hypotheses $H_0 : \hat{\beta}_2 = 0$ vs $H_a : \hat{\beta}_2 \neq 0$. (Hint: you may use the p -value from your summary in part (a))

Solution: Because $p = -2 \cdot 10^{-16} < 0.05 = \alpha$ we reject H_0 and conclude that there is evidence of curvature in the response function.

- (e) Based on the R^2 or Adjusted R^2 , compare the fits of the quadratic model in part (a) with the simple linear regression model (from Additional exercise of Homework #2) where mpg is the response variable and horsepower is the only predictor variable.

Solution: In homework 2 the degree 1 polynomial estimate had an R-squared of 0.6059. In the new model the $R^2 = 0.6876$ and adjusted $R^2 = 0.686$. The new second order term helps to explain approximately 13% more of the variance $\left(\frac{0.686}{0.6049}\right)$.



```

1 df = read.csv("O:/Arr Matey/Auto.csv", header=T, na.strings="?")
2 df = na.omit(df)
3
4
5 mpg = df$mpg;
6 hp = df$horsepower
7
8 quadratic_model = lm(mpg~hp+I(hp^2))
9 summary(quadratic_model)
10
11 vcov(quadratic_model)

```

9:25 (Top Level) R Script

Console Terminal Background Jobs

```

R 4.2.2 ~\
> source("O:/Arr Matey/hw3_exercise1.R", echo=TRUE)
> library(readr)
> df = read.csv("O:/Arr Matey/Auto.csv", header=T, na.strings="?")
> df = na.omit(df)
> mpg = df$mpg;
> hp = df$horsepower
> quadratic_model = lm(mpg~hp+I(hp^2))
> summary(quadratic_model)
call:
lm(formula = mpg ~ hp + I(hp^2))

Residuals:
    Min       1Q   Median       3Q      Max
-14.7135  -2.5943  -0.0859   2.2868  15.8961

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  56.9000997  1.8004268   31.60  <2e-16 ***
hp          -0.4661896  0.0311246  -14.98  <2e-16 ***
I(hp^2)       0.0012305  0.0001221   10.08  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,    Adjusted R-squared:  0.686
F-statistic: 428 on 2 and 389 DF,  p-value: < 2.2e-16

> vcov(quadratic_model)
              (Intercept)              hp              I(hp^2)
(Intercept)  3.2415366849 -5.493052e-02  2.054469e-04
hp          -0.0549305164  9.687418e-04 -3.734238e-06
I(hp^2)      0.0002054469 -3.734238e-06  1.490252e-08
>

```

5. Exercise 2

This question should be answered using the **Credit** dataset in Canvas.

- (a) Fit a multiple regression model to predict **Balance** using **Income**, **Limit**, **Education**, and **Rating**.

Solution:

```

1 library(readr)
2
3 df = read.csv("0:/Arr Matey/Credit.csv", header=T, na.strings="?")
4 df = na.omit(df)
5
6 balance = df$Balance;
7 income = df$Income;
8 limit = df$Limit;
9 education = df$Education;
10 rating = df$Rating;
11
12 multi_model = lm(balance ~ income + limit + education + rating);
13 summary(multi_model)

```

13:21 (Top Level) ▾

Console Terminal Background Jobs

```

R 4.2.2 ~ /
> limit = df$Limit;

> education = df$Education;

> rating = df$Rating;

> multi_model = lm(balance ~ income + limit + education + rating);

> summary(multi_model)

Call:
lm(formula = balance ~ income + limit + education + rating)

Residuals:
    Min       1Q   Median       3Q      Max
-257.08 -112.09  -35.17   49.81  562.86

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -517.34829    49.47945  -10.456 < 2e-16 ***
income       -7.71604     0.37830  -20.396 < 2e-16 ***
limit         0.08194     0.04489   1.825  0.0687 .
education     1.91714     2.61247   0.734  0.4635
rating        2.73977     0.66869   4.097 5.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 162.4 on 395 degrees of freedom
Multiple R-squared:  0.8764,    Adjusted R-squared:  0.8752
F-statistic: 700.2 on 4 and 395 DF,  p-value: < 2.2e-16

```

(b) Write out the estimated model in equation form.

Solution: For $i := \text{income}$, $l := \text{limit}$, $e := \text{education}$, $r := \text{rating}$,

$$\hat{y}(i, l, e, r) = -517.34829 - 7.71604i + 0.08194l + 1.91714e + 2.73977r$$

(c) Provide the interpretation of each coefficient in the model.

Solution: The intercept of -517.34829 means that at $\hat{y}(\vec{0})$ (all variables zero) the expected balance is $-\$517.35$.

For every unit increase of income, the expected value of balance drops by 7.71604.

For every unit increase of limit, the expected value of balance increases by 0.08194.

For every unit increase of education, the expected value of balance increases by 1.91714.

For every unit increase of rating, the expected value of balance increases by 2.73977.

(d) Obtain 95% confidence intervals for the coefficient(s)

Solution:

```
> confint(multi_model, level=0.95);
```

	2.5 %	97.5 %
(Intercept)	-6.146243e+02	-420.0722894
income	-8.459779e+00	-6.9722989
limit	-6.308071e-03	0.1701884
education	-3.218954e+00	7.0532317
rating	1.425135e+00	4.0543978

With 95% confidence, the coefficients' intervals are

- Intercept: (-\$420.07, -\$614.62)
- Income: (-\$8.46, -\$6.97)
- Limit: (-\$0.006, \$0.17)
- Education: (-\$3.22, \$7.05)
- Rating: (\$1.43, \$4.05)

(e) Test whether all the regression coefficients are zero (there is a linear relationship between the response and the predictors), i.e whether $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = \hat{\beta}_4 = 0$.

Solution: Based on the model summary, since $p = 2.2 \cdot 10^{-16} < 0.05 = \alpha$, there is sufficient evidence that all of the predictor variables are significant (non zero).

(f) Based on the p -values in part (a), which predictor(s) seem(s) to not have an association with the response variable (**Balance**)

Solution: The predictors Limit and Education seem to not be associated with Balance because their p -values are greater than α (0.0687 and 0.4635 respectively).

(g) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the response variable.

Solution: $\hat{y}(i, r) = -534.81215 - 7.67212i + 3.94926r$

```

> smaller_model = lm(balance ~ income + rating);
> summary(smaller_model)

Call:
lm(formula = balance ~ income + rating)

Residuals:
    Min       1Q   Median       3Q      Max
-278.57 -112.69  -36.21   57.92  575.24

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -534.81215    21.60270   -24.76  <2e-16 ***
income       -7.67212     0.37846   -20.27  <2e-16 ***
rating        3.94926     0.08621    45.81  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 162.9 on 397 degrees of freedom
Multiple R-squared:  0.8751,    Adjusted R-squared:  0.8745
F-statistic: 1391 on 2 and 397 DF,  p-value: < 2.2e-16

```

- (h) Test whether the coefficients of the predictor(s) in (f) are all zero. Would you drop those predictors from the full model? Why?

Solution: $H_0 : \hat{\beta}_l = \hat{\beta}_e = 0$ versus $H_a : \hat{\beta}_l \neq 0 \vee \hat{\beta}_e \neq 0$.
(next page)

```

.R × hw2-1.R × hw2-2.R × hw3_exercise1.R × hw3_exercise2.R × hw2.R ×
Source on Save Run Source
1 library(readr)
2
3 df = read.csv("0:/Arr Matey/Credit.csv", header=T, na.strings="?")
4 df = na.omit(df)
5
6 balance = df$Balance;
7 income = df$Income;
8 limit = df$Limit;
9 education = df$Education;
10 rating = df$Rating;
11
12 multi_model = lm(balance ~ income + limit + education + rating);
13 summary(multi_model)
14
15 confint(multi_model, level=0.95);
16
17 smaller_model = lm(balance ~ income + rating);
18 summary(smaller_model)
19
20 #2 h
21 k = 4
22 g = 2
23 n = 400
24
25 ssec = (n-k-1)*(162.4^2)
26 sser = (n-(k-g)-1)*(162.9^2)
27
28
29 numerator = (sser - ssec) / (k-g)
30 denominator = ssec / (n-k-1)
31 F = numerator/denominator
32 F
33
34 pf(F, 2, n-k-1, lower.tail=FALSE)
35
34:34 (Top Level) R Script
Console Terminal × Background Jobs ×
R 4.2.2 ~/
[1] 2.224172

> pf(F, 2, n-k-1, lower.tail=FALSE)
[1] 0.1095098
>

```

Since $p = 0.1095 \not\leq \alpha = 0.05$, we fail to reject H_0 . Hence there is not enough evidence to indicate a relationship between the predictors limit and education with response balance. Therefore I would remove these the full model, especially since income and rating have over a 99% significance.