STAT GR5205 Homework 1 [100 pts]
Due 8:40am Monday, September 24th

Problem 1 (**1.22 KNN**)

Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**model (2.1) in the notes**).

Data not displayed

Perform the following tasks:

i. Use `R` to obtain the estimated regression function.

ii. Use `R` to create a scatter plot with the line of best fit. Make the line of best fit red.

iii. Use `R` to calculate the best point estimate of σ^2.

iv. Use `R` to calculate the sample correlation coefficient and coefficient of determination.

Problem 2

Recall the sample residual is defined by $e_i = y_i - \hat{y}_i$, where y_i is the ith response value and $\hat{y}_i$ is its corresponding fitted value computed by least squares estimates $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Prove the following properties:

i.
$$\sum_{i=1}^{n} x_i e_i = 0$$

ii.
$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Problem 3

Recall that the *ith* fitted value $\hat{Y}_i$ can be expressed as a linear combination of the response values, i.e.,

$$\hat{Y}_i = \sum_{j=1} h_{ij} Y_j,$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}},$$

and

$$S_{xx} = \sum_{i=1} (x_i - \bar{x})^2.$$

Prove the following properties of the hat-values h_{ij}.

i.
$$\sum_{j=1} h_{ij}^2 = h_{ii}$$

ii.
$$\sum_{j=1} h_{ij} x_j = x_i$$

Problem 4

Consider the *regression through the origin model* given by

$$Y_i = \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{1}$$

The estimated model at observed point (x, y) is

$$\hat{y} = \hat{\beta} x,$$

where

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \tag{2}$$

Complete the following tasks

i. Show that
$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$
is an unbiased estimator of β.

ii. Compute the standard error of estimator $\hat{\beta}$.

iii. Identify the probability distribution of estimator $\hat{\beta}$.

GR5205 Homework 1

Yiqiao Yin [YY2502]

Table of Contents

PROBLEM 1

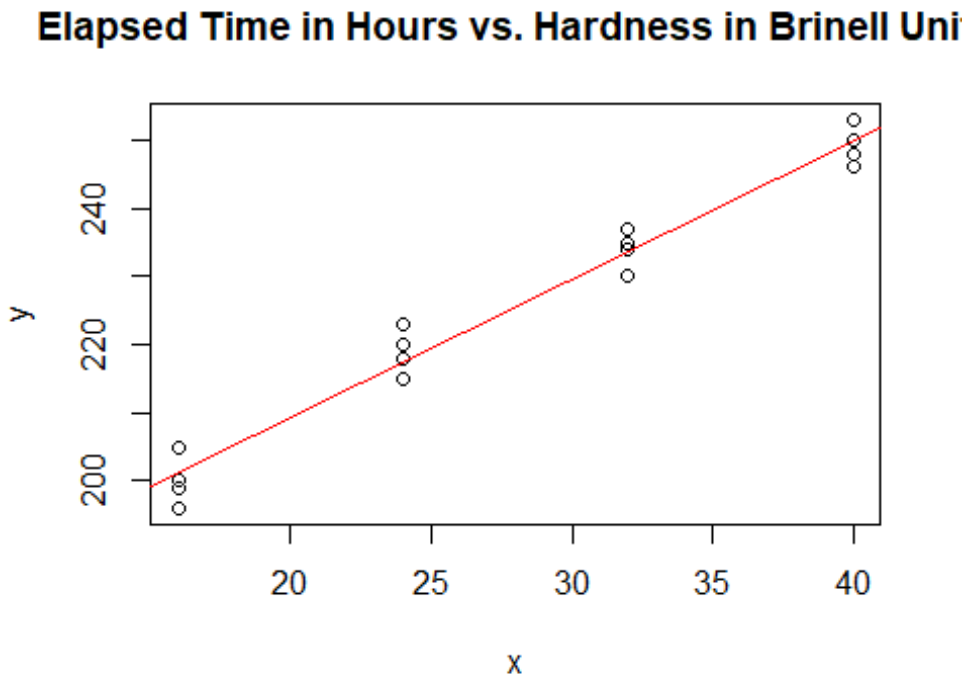
Given data, we know X is elapsed time in hours, and Y is hardness in Brinell units.

(i) Linear Regression

```
# Data set manually typed out
y <- c(199.0,205.0,196.0,200.0,218.0,220.0,215.0,223.0,237.0,234.0,235.0,230.
0,250.0,248.0,253.0,246.0)
x <- c(rep(16,4),rep(24,4),rep(32,4),rep(40,4))

# Linear Regression
# Find the line of best fit.  Define the object as model
model <- lm(y~x)
model

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     168.600        2.034
```

(ii) Scatter Plot

```
# Create a scatter plot with the line of best fit
plot(x,y,main="Elapsed Time in Hours vs. Hardness in Brinell Units")
abline(model,col="red")
```




```
# I talked with professor and he said it is okay to not have color printing.
# However, from the code above, I set the color to red according to the probl
em.
```

(iii) Best Point Estimate

```
# This is the variance of the error term
summary(model)$sigma^2
```

```
## [1] 10.45893
```

```
# Or we can do it manually
ANOVA <- anova(model); ANOVA
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 5297.5  5297.5  506.51 2.159e-12 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# 146.4 is residual sum of square
# 14 is the DF
146.4/14
```

```
## [1] 10.45714
```

```r
# The two results are siilar at 10.5
```

(iv) Sample Coef. Corr. and Coef. of Determin.

```r
# Compute
#model$coefficients # Parameters from regression model
Sum <- summary(model) # Summary of regression model
#Sum$r.squared # R-square, i.e. correlation coefficient
#sqrt(Sum$r.squared) # square root of R-square, i.e. coefficient of determina
tion

# Moreover
# we can compute the following
# Find the correlation of x & y.
cor(x,y) # This is the sample coefficient correlation
```

```
## [1] 0.9864599
```

```r
# Long way
# We can do this manually according to definition
S_xx <- sum((x-mean(x))^2)
S_yy <- sum((y-mean(y))^2)
S_xy <- sum((x-mean(x))*(y-mean(y)))
S_xy/sqrt(S_xx*S_yy)
```

```
## [1] 0.9864599
```

```r
# The two results are the same
#mean(y)-(S_xy/S_xx)*mean(x) # which should be the same of b1 and b0 from reg
ression model

# Next coefficient of determination
# is the square of r
cor(x,y)^2
```

```
## [1] 0.9731031
```

```r
Sum$r.squared
```

```
## [1] 0.9731031
```

```r
# The two results are the same
```

PROBLEM 2

Let us consider residual $e_i = y_i - \hat{y}_i$. We want to prove

1. $\sum_{i=1}^{k} X_i\, e_i = 0$
2. $\sum_{i=1}^{n} \hat{Y}_i\, e_i = 0$

Let us prove the result in the following.

1. Recall that $\hat{Y}_i = b_0 + b_1 X_i$. We consider the following

$$\begin{aligned}\sum_i X_i\, e_i &= \sum(X_i(Y_i - (b_0 + b_1 X_i))) \\ &= \sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 \\ &= b_0 \sum X_i + b b_1 \sum X_i^2 - b_0 \sum X_i - b_1 \sum X_i^2 \\ &= 0\end{aligned}$$

since $\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$ from page 17 of textbook.

2. We show

$$\begin{aligned}\sum Y_i e_i &= \sum Y_i^2 - \sum Y_i (b_0 + b_1 X_i) \\ &= \sum Y_i^2 - b_0 \sum Y_i - b_1 \sum Y_1 X_i \\ &= \sum (b_0 + b_1 X_1) e_i \\ &= b_0 \sum e_i + b_1 \sum X_i e_i\end{aligned}$$

PROBLEM 3

There are two parts of this proof.

(i) Proof: square of hat matrix is itself

We want to show $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$. Let us compute the following.

$$\begin{aligned}\sum_j h_{ij}^2 &= \sum_j \left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{s_{xx}}\right)\left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{s_{xx}}\right) \\ &= \sum_j \left(\frac{1}{n^2} + \frac{2(x_i - \overline{x})(x_j - \overline{x})}{n \cdot s_{xx}^2} + \left(\frac{(x_i - \overline{x})(x_j - \overline{x})}{s_{xx}}\right)^2\right) \\ &\qquad\qquad \underbrace{\qquad\qquad}_{=0} \\ &= \frac{1}{n} + 0 + \frac{(x_i - \overline{x}) s_{xx}}{s_{xx}^2} \\ &= h_{ii}\end{aligned}$$

Alternatively, we can approach this problem using matrix multiplication. The following is additional practice only. The above answer should suffice.

We want to consider h_{ij} to be a hat matrix. From the definition of ij index, we have

$$H := h_{ij} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1j} \\ & & & \vdots \\ h_{i1} & h_{i2} & \dots & h_{ij} \end{bmatrix}$$

and this way of writing it is sufficient to show

$$\begin{aligned} H^2 &= (X(X'X)^{-1})(X(X'X)^{-1}X') \\ &= X(X'X)^{-1}(X'X)(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= H \end{aligned}$$

and we are done.

(ii) Proof: hat matrix weighted by x is x itself

We want to show that $\sum_{j=1} h_{ij}\, x_j = x_i$. We compute the following.

$$\begin{aligned} \sum_j \left(\frac{1}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})}{s_{xx}}\right) x_i &= \sum_j \left(\frac{x_j}{n} + \frac{(x_i - \overline{x})(x_j - \overline{x})x_j}{s_{xx}}\right) \\ &= \overline{x} + \frac{(x_i - \overline{x})}{s_{xx}} \sum (x_j - \overline{x}) x_j \\ &= \overline{x} + \frac{x_i - \overline{x}}{s_{xx}} \sum_j (x_j \cdot x_j - \overline{x} x_j) \\ &= \overline{x} + \frac{x_i - \overline{x}}{s_{xx}} \sum_j (x_j^2 - n\overline{x}^2) \\ &= \overline{x} + \frac{x_i - \overline{x}}{s_{xx}} \sum_j (x_j^2 - 2\overline{x} x_j + n\overline{x}) \\ &= \overline{x} + \frac{x_i - \overline{x}}{s_{xx}} \sum_j (x_j - \overline{x})^2 \\ &= \overline{x} + \frac{x_i - \overline{x}}{s_{xx}} s_{xx} \\ &= \overline{x} + (x_i - \overline{x}) \\ &= x_i \end{aligned}$$

PROBLEM 4

This problem we want to show

1. $E[\hat{b}_1] = b_1$ and $E[\hat{b}_0] = b_0$
2. $\text{var}[\hat{b}_1] = \sigma^2/S_{xx}$
3. What is the distribution of $\hat{b}_1$?

We provide the following answer.

1. Consider the following

$$\begin{aligned} E[\hat{b}_1] &= E[\frac{s_{xy}}{s_{xx}}] \\ &= E[\frac{\sum(X_i - \overline{X})Y_i}{s_{xx}}] \\ &= E[\sum(\frac{X_i - \overline{X}}{s_{xx}}Y_i] \\ &= \frac{1}{s_{xx}}\sum(X_i - \overline{X})E[Y_i] \\ &= \frac{1}{s_{xx}}\sum(X_i - \overline{X})(b_0 + b_1X_i) \\ &= \frac{1}{s_{xx}}b_0\sum(X_i - \overline{X}) + \frac{b_1}{s_{xx}}\sum(X_i - \overline{X})X_i \\ &= 0 + b_1\frac{\sum(X_i - \overline{X})X_i}{\sum(X_i - \overline{X})X_i} \end{aligned}$$

and we can also show

$$\begin{aligned} E[\hat{b}_0] &= E[\overline{Y} - b_1\overline{X}] \\ &= \frac{1}{n}\sum E(Y_i) - E(b_1)\overline{X} \\ &= \frac{1}{n}(n[b_0 + b_1X_i] - nb_1\overline{X}) \\ &= b_0 + b_1\overline{X}_i - b_1\overline{X} \\ &= b_0 \end{aligned}$$

2. We want to show the variance is best estimate and we show the following.

$$\begin{aligned} \text{Var}[\hat{b}_1] &= \text{Var}(\sum K_iY_i) \\ &= \sum K_i^2\text{var}(Y_i) \\ &= \frac{\sigma^2}{s_{xx}} \end{aligned}$$

since $\sum K_i = \sum(\frac{X_i - \overline{X})^2}{s_{xx}} = \frac{1}{s_{xx}^2}\sum(X_i - \overline{X})^2 = \frac{1}{s_{xx}}$.

3. Since b_1 is a linear combination of normal random variable, then $\hat{b}_1 \sim N(b_1, \sigma^2/s_{xx})$.

STAT GR5205 Homework 1 KEY [100 pts]
Due 8:40am Monday, September 24th

Problem 1 (**1.22 KNN**) [25 pts]

Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**(2.1) in the notes**).

Data not displayed

Perform the following tasks:

i. Use `R` to obtain the estimated regression function. **(5 pts)**

`R` code and output

```
lm(Y~X)

Coefficients:
(Intercept)            X
    168.600        2.034
```

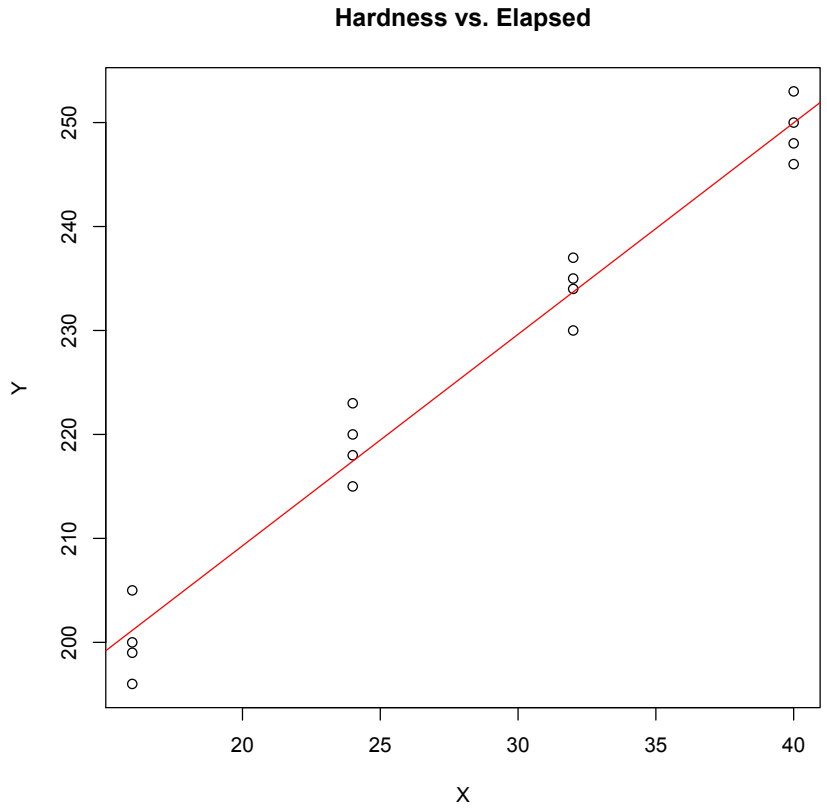
The line of best fit is:

$$\hat{y} = 168.600 + 2.034x$$

ii. Use `R` to create a scatter plot with the line of best fit. Make the line of best fit red. **(10 pts)**

`R` code

```
plot(X,Y,main="Hardness vs. Elapsed")
abline(lm(Y~X),col="Red")
```



iii. Use R to calculate the best point estimate of σ^2. **(5 pts)**

R code

```
sum((Y-fitted(lm(Y~X)))^2)/(16-2)
```

```
10.45893
```

iv. Use R to calculate the sample correlation coefficient and coefficient of determination. **(5 pts)**

R code

```
r <- cor(X,Y)
r
```

```
0.9864599
```

```
r^2
```

```
0.9731031
```

Problem 2 [25 pts]

Recall the sample residual is defined by $e_i = y_i - \hat{y}_i$, where y_i is the ith response value and $\hat{y}_i$ is its corresponding fitted value computed by least squares estimates $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Prove the following properties:

i. **(15 pts)**

$$\sum_{i=1}^{n} x_i e_i = 0$$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n} x_i e_i &= \sum_{i=1}^{n} x_i (y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i + \hat{\beta}_1 \Big(\sum_{i=1}^{n} x_i \Big)^2 - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\
&= S_{xy} - \hat{\beta}_1 S_{xx} \\
&= S_{xy} - S_{xy} \\
&= 0
\end{aligned}$$

ii. **(10 pts)**

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Proof:

$$\begin{aligned}
\sum_{i=1}^{n} \hat{y}_i e_i &= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i \\
&= \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} e_i x_i \\
&= \hat{\beta}_0 * 0 + \hat{\beta}_1 * 0 \\
&= 0
\end{aligned}$$

Problem 3 [20 pts]

Recall that the *ith* fitted value $\hat{Y}_i$ can be expressed as a linear combination of the response values, i.e.,

$$\hat{Y}_i = \sum_{j=1} h_{ij} Y_j,$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}},$$

and

$$S_{xx} = \sum_{i=1} (x_i - \bar{x})^2.$$

Prove the following properties of the hat-values h_{ij}.

i. **(10 pts)**

$$\sum_{j=1} h_{ij}^2 = h_{ii}$$

Proof:

$$\begin{aligned}
\sum_{j=1}^{n} h_{ij}^2 &= \sum_{j=1} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right)^2 \\
&= \sum_{j=1}^{n} \left(\frac{1}{n^2} + \frac{2(x_i - \bar{x})(x_j - \bar{x})}{nS_{xx}} + \frac{(x_i - \bar{x})^2 (x_j - \bar{x})^2}{S_{xx}^2} \right) \\
&= \sum_{j=1}^{n} \frac{1}{n^2} + \frac{2(x_i - \bar{x})}{nS_{xx}} \sum_{j=1}^{n} (x_j - \bar{x}) + \frac{(x_i - \bar{x})^2}{S_{xx}^2} \sum_{j=1}^{n} (x_j - \bar{x})^2 \\
&= \frac{1}{n} + 0 + \frac{(x_i - \bar{x})^2}{S_{xx}^2} S_{xx} \\
&= h_{ii}
\end{aligned}$$

ii. **(10 pts)**

$$\sum_{j=1} h_{ij}x_j = x_i$$

Proof:

$$\begin{aligned}
\sum_{j=1}^{n} h_{ij}x_j &= \sum_{j=1}^{n}\left(\frac{1}{n} + \frac{(x_i-\bar{x})(x_j-\bar{x})}{S_{xx}}\right)x_j \\
&= \frac{1}{n}\sum_{j=1}^{n} x_j + \frac{(x_i-\bar{x})}{S_{xx}}\sum_{j=1}^{n}(x_j-\bar{x})x_j \\
&= \bar{x} + \frac{(x_i-\bar{x})}{S_{xx}}S_{xx} \\
&= \bar{x} + (x_i-\bar{x}) \\
&= x_i
\end{aligned}$$

Problem 4 [30 pts]

Consider the *regression through the origin model* given by

(1) $$Y_i = \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The estimated model at observed point (x, y) is

$$\hat{y} = \hat{\beta}x,$$

where

(2) $$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Complete the following tasks

i. Show that

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}$$

is an unbiased estimator of β.

Solution **(10 pts)**

Note that under model (1), $E[Y_i] = E[\beta x_i + \epsilon_i] = \beta x_i + E[\epsilon_i] = \beta x_i$. Then we have

$$E[\hat{\beta}] = E\left[\frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2}\right] = \frac{\sum_{i=1}^{n} x_i E[Y_i]}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i \beta x_i}{\sum_{i=1}^{n} x_i^2} = \beta\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = \beta.$$

ii. Find the standard error of estimator $\hat{\beta}$. **(10 pts)**

Solution

Notice that $\hat{\beta}$ can be written as a linear combination of the response values Y_i. Namely,

$$\hat{\beta} = \sum_{i=1}^{n} c_i Y_i, \quad \text{where} \quad c_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}.$$

Also notice that $\text{Var}[Y_i] = \text{Var}[\beta x_i + \epsilon_i] = \text{Var}[\epsilon_i] = \sigma^2$. Then by independence,

$$\text{Var}[\hat{\beta}] = \text{Var}\left[\sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} c_i^2 \text{Var}[Y_i] = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \sigma^2 = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{(\sum_{i=1}^{n} x_i^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}.$$

Thus, the standard error of $\hat{\beta}$ is

$$\sigma_{\hat{\beta}} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}}$$

iii. Identify the probability distribution of estimator $\hat{\beta}$. **(10 pts)**

Solution

Since $\hat{\beta}$ is a linear combination of normal random variables,

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

STAT GR5205 Homework 2 [100 pts]

Due 8:40am Wednesday, October 10th

Problem 1 (**2.7 KNN**)

Sixteen batches of plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**(2.1) in the notes**).

Data not displayed

Use `R` to perform the following tasks:

i. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate.

ii. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = .01$.

iii. Set up the ANOVA table.

iv. Test by means of an F-test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$.

v. Does t^2_{calc} from part [ii] equal f_{calc} from part [iv]? Explain why this identity holds or does not hold.

vi. Construct 95% Bonferroni joint confidence intervals for estimating both the true intercept β_0 and the true slope β_1.

vii. Construct 95% Bonferroni joint confidence intervals for predicting the true average hardness corresponding to elapsed times 20, 28 and 36 hours.

Problem 2

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

i. Assuming $H_0 : \beta_1 = 0$ is true, use `R` to simulate the sampling distribution of the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

Assume $\beta_0 = 10$, $\sigma = 3$, $n = 30$ and run the loop 10,000 times to generate the sampling distribution. Run the following code preceding the loop so that everyone has the same seed and X data vector. Fill in the missing code to receive full credit.

```
# Set seed
set.seed(0)
# Assign sample size and create x vector
n <- 30
# Empty list for f-statistics
f.list <- NULL
x <- sample(1:100/30,n,replace=T)
# Run loop
for (i in 1:10000) {

# Fill in the body of the loop here...


}
```

ii. From the simulated sampling distribution, plot a histogram and overlay the *correct F-density* on the histogram. Adjust the bin size to *breaks=50* in the histogram. Overlay the F-density in red.

iii. Compute the $95th$ percentile of both the simulated sampling distribution and the *correct* F-distribution. Compare these values.

Problem 3

Consider splitting the response values $y_1, \ldots, y_n$ into two groups with respective sample sizes n_1 and n_2. Define the **dummy** variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{if group two} \end{cases} \tag{1}$$

Show that the least squares estimators of β_1 and β_0 are respectively

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \quad \text{and} \quad \hat{\beta}_0 = \bar{y}_2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of each group.

Problem 4

Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article *Compatibility of Outer and Feasible Interlining Fabrics in Tailored Garments* gave the accompanying data on extensibility (%) at 100 gm/cm for both high-quality (H) fabric and poor-quality (P) fabric specimens.

H	1.2	.9	.7	1.0	1.7	1.7	1.1	.9	1.7
	1.9	1.3	2.1	1.6	1.8	1.4	1.3	1.9	1.6
	.8	2.0	1.7	1.6	2.3	2.0			
P	1.6	1.5	1.1	2.1	1.5	1.3	1.0	2.6	

Use `R` to perform the following tasks.

i. Create an appropriate graphic to visualize the relationship between extensibility and quality. Do you think there is a relationship between extensibility and quality? Make sure to label the plot.

ii. Using the indicator variable

$$x = \begin{cases} 1 & \text{if high quality} \\ 0 & \text{if low quality} \end{cases}$$

run a regression analysis to test if the average fabric extensibility differs per group.

Problem 5

i. Consider the *regression through the origin model* given by

$$Y_i = \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{2}$$

Derive the maximum likelihood estimators of β and σ^2.

ii. Consider the residuals e_i related to the regression through the origin model (2). Prove that

$$\sum_{i=1}^{n} e_i x_i = 0.$$

Also, in the regression through the origin model (2), is the sum of residuals equal to zero? I.e., is the following relation true?

$$\sum_{i=1}^{n} e_i = 0.$$

Explain your answer in a few sentences or less.

iii. Consider testing the null/alternative pair

$$H_0 : \beta = \beta' \quad \text{v.s.} \quad H_A : \beta \neq \beta'.$$

Note that β' is the hypothesized value. Show that the likelihood-ratio test can be based on the rejection region $|T| > k$ with test statistic

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}x_i)^2/(n-1)}{\sum_{i=1}^{n} x_i^2}}}.$$

Note that k is some positive real number and $\hat{\beta}$ is the maximum likelihood estimator of β.

iv. Under H_0, what is the probability distribution of the above test statistic T?

Hints: To solve 5 Part iii:

(a) Compute the likelihood-ratio test statistic (λ) from Definition 2.4 on Page 45 of the class notes.

(b) When simplifying the expression, the following trick might be useful:

$$\sum_{i=1}^{n}(Y_i - \beta' x_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}x_i + \hat{\beta}x_i - \beta' x_i)^2.$$

(c) After simplifying $\lambda < c$, find a suitable transformation of λ that yields the desired test statistic and rejection rule.

GR5205 Homework 2

Yiqiao Yin [YY2502]

Table of Contents

PROBLEM 1 KNN

Sixteen batches of plastic were made, and from each other batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels. The hardness was measured after the assigned elapsed time. The results are shown below; X is elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model is appropriate

(i) Estimate Change of Beta

From the results below, we have the parameter $\hat{\beta}_1 = 2.03$ and we also have $\text{SE}(\hat{\beta}_1) = 0.09$. Using 99-percent confidence interval, we want 2.97 standard errors above and below the expectation of this parameter, which gives us the interval $[2.03 - 2.97 \times 0.09, 2.03 + 2.97 \times 0.09] = [1.76, 2.29]$

```
# Data set manually typed out
y <- c(199.0,205.0,196.0,200.0,218.0,220.0,215.0,223.0,237.0,234.0,235.0,230.
0,250.0,248.0,253.0,246.0)
x <- c(rep(16,4),rep(24,4),rep(32,4),rep(40,4))
n <- length(x)

# Linear Regression
# Find the line of best fit.  Define the object as model
model <- lm(y~x)
summary(model)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1500 -2.2188  0.1625  2.6875  5.5750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
## x             2.03438    0.09039   22.51 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.234 on 14 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9712
## F-statistic: 506.5 on 1 and 14 DF,  p-value: 2.159e-12

qt(1 - 0.01/2, n-2)

## [1] 2.976843
```

```
confint(lm(y~x), level = 1 - 0.01/2)
```

```
##                 0.25 %    99.75 %
## (Intercept) 159.763546 177.436454
## x             1.733753   2.334997
```

(ii) Standard Satisfied Or Not; Two-sided Test

We want to compute the following

$$t_{\text{calc}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\text{MSE}}{S_{xx}}}} = \frac{2.03 - 3}{0.09} = 0.33$$

which means $t^2_{\text{calc}} = 0.33^3 = 0.11$. Using alpha to be 0.01 with a two-sided test, we fail to reject null hypothesis. That means we do not have sufficient evidence to argue the plastic manufacturer's statement is false.

```
## Compute
(2.03-2)/(0.09) # t_calc
```

```
## [1] 0.3333333
```

```
((2.03-2)/(0.09))^2 # t_calc^2
```

```
## [1] 0.1111111
```

```
qt(0.998, n-2)
```

```
## [1] 3.437867
```

(iii) ANOVA

```
# ANOVA table
ANOVA <- anova(model); ANOVA
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 5297.5  5297.5  506.51 2.159e-12 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(iv) Test F-test

First, we report the results from the linear model. This gives us F-statistic of approximately 506.

```r
# Check
model.full <- lm(y~x)
anova(model.full)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 5297.5  5297.5  506.51 2.159e-12 ***
## Residuals 14  146.4    10.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
F_stat <- anova(model.full)$F[1]
F_stat
```

```
## [1] 506.5062
```

We compute that F-statistic is much larger than 1. This gives us evidence for alternative hypothesis, which states $\hat{\beta}_1 \neq 0$.

```r
model_full <- lm(y~x)
#anova(model_full)
model_reduced <- lm(y~1)
#anova(model_reduced)
anova(model_full, model_reduced)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     14  146.4
## 2     15 5443.9 -1   -5297.5 506.51 2.159e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that F-test statistic is above 500 and p-value is less than 0.01. We cannot reject null hypothesis at the 5% significance level.

(v) t-stat and F-stat

Recall from (ii) we have $t_{\text{calc}} = 22.51$ and $t_{\text{calc}}^2 = 22.51^2 = 506$. We also compute $f_{\text{calc}} = 506$. Be aware that this is using null hypothesis $H_0: \beta_1 = 0$. From (ii), we are using null hypothesis $H_0: \beta_1 = 2$, which gives us t-statistic of 0.33. The proposition, e.g. $t_{\text{calc}}^2 = F_{\text{calc}}$ will not hold under this null hypothesis.

```r
Sum <- summary(model_full)
Sum$coefficients[2,3] # t_calc
```

```
## [1] 22.50569
```

```r
Sum$coefficients[2,3]^2 # t_calc^2
```

```
## [1] 506.5062
```

```r
ANOVA <- anova(model_full, model_reduced)
ANOVA$F[2] # F_calc
```

```
## [1] 506.5062
```

(vi) Construct 95% Bonferroni C.I.

We print the results from using *confint()* function. We can also comppute the results by using *qt()* function outselves. The results are similar to the fractions and there may be rounding errors.

```r
# Linear Regression Output
model <- lm(y~x)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1500 -2.2188  0.1625  2.6875  5.5750
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
## x             2.03438    0.09039   22.51 2.16e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.234 on 14 degrees of freedom
## Multiple R-squared:  0.9731, Adjusted R-squared:  0.9712
## F-statistic: 506.5 on 1 and 14 DF,  p-value: 2.159e-12
```

```r
# Confidence Interval
confint(lm(y~x), level = 1 - 0.05/2)
```

```
##                 1.25 %    98.75 %
## (Intercept) 161.932014 175.267986
## x             1.807526   2.261224
```

```r
# Critical Value
t_critical <- qt(1 - 0.05/4, n-2); t_critical
```

```
## [1] 2.509569
```

```r
# Compute C.I. for for Intercept
168 - 2.51 * 2.67
```

```
## [1] 161.2983
```

```
168 + 2.51 * 2.67
```

```
## [1] 174.7017
```

```
# Compute C.I. for for Intercept
2.03 - 2.51 * 0.09
```

```
## [1] 1.8041
```

```
2.03 + 2.51 * 0.09
```

```
## [1] 2.2559
```

(vii) Construct 95% Bonferroni C.I.

```
x.data <- data.frame(x = c(20,28,36))
predict(model, newdata = x.data, interval = "confidence", se.fit = TRUE)
```

```
## $fit
##        fit      lwr      upr
## 1 209.2875 206.9610 211.6140
## 2 225.5625 223.8284 227.2966
## 3 241.8375 239.5110 244.1640
##
## $se.fit
##         1         2         3
## 1.0847255 0.8085067 1.0847255
##
## $df
## [1] 14
##
## $residual.scale
## [1] 3.234027
```

PROBLEM 2

Consider

$$Y_i = \beta_0 + \beta_1 x + \epsilon_i, i = 1,2,\dots,n, \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$$

(i) Hypothesis Testing

Assume $H_0: \beta_1 = 0$ is true. Use R to simulate the sampling distribution of the F-statistic

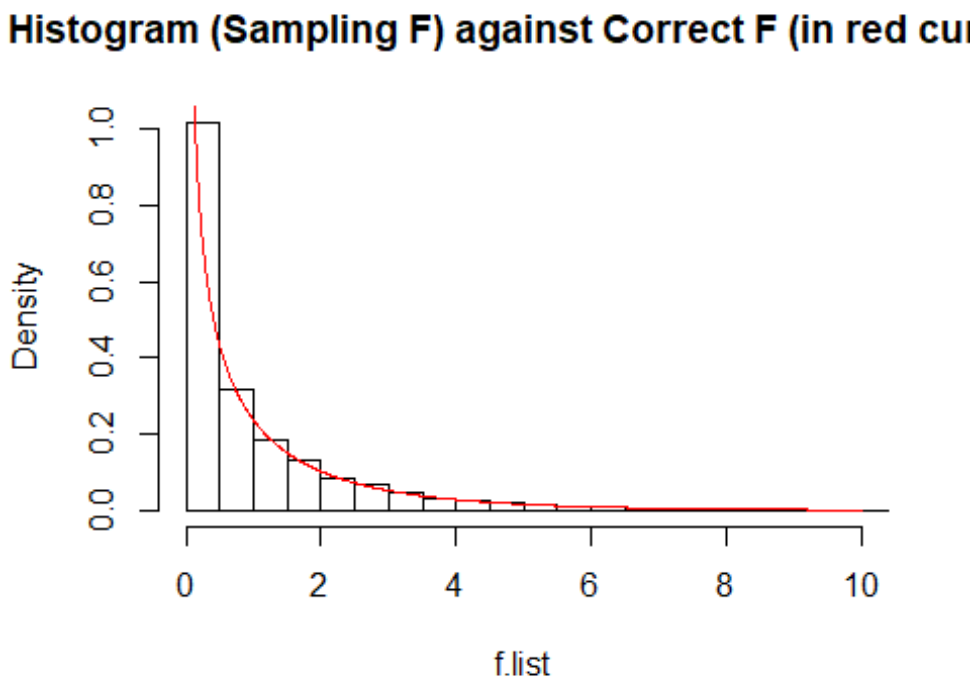
$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$$

Assume $\beta_0 = 10, \sigma = 3, n = 30$ and run the loop 10,000 times to generate the sampling distribution. Run the following code preceding the loop so that everyone has the same seed and X data vector. Fill in the missing code to receive full credits.

```r
# Set seed
set.seed(0)

# Assign sample size and create x vector
n <- 30
beta_0 = 10
sigma = 3

# Empty list for f-statistics
f.list <- NULL
x <- sample(1:100/30,n,replace=T)
y <- beta_0 + rnorm(n, mean = 0, sd = 3)
s_xx <- mean((x - mean(x))^2)
model <- summary(lm(y~x))

# Run loop
for (i in 1:10000) {
  # Fill in the body of the loop here...
  #x <- sample(1:100/30,n,replace=T)
  y <- beta_0 + rnorm(n, mean = 0, sd = 3)
  model <- lm(y~x)
  #ANOVA <- anova(model)
  f_stat <- summary(model)$fstatistic[1]
  f.list <- c(f.list, f_stat)
}
```

(ii) Plot Histogram

```r
hist(f.list, probability = TRUE, breaks=50, xlim = c(0,10), plot = TRUE,
     main = "Histogram (Sampling F) against Correct F (in red curve)")
t <- seq(0.001, 10, by = 0.01)
lines(t, df(t, 1, n-2), col = "red")
```




```
# lines(density(f.list), col="blue") # This is the sampling F-distribution
# lines(df(c(1:length(f.list)), df1=1, df2=28), col="red") # This is the corr
ect F-distribution
```

(iii) Compute 95th Percentile

The qunatile of the simulated sampling distribution and the correct F-distribution can be found by the following command. The results are similar.

```
quantile(f.list, 0.95) # for simulated sampling distribution
```

```
##      95%
## 4.248133
```

```
quantile(qf(0.95,1,n-2), 0.95) # for correct F-distribution
```

```
##      95%
## 4.195972
```

PROBLEM 3

Consider splitting the response values $y_1, \ldots, y_n$ into two groups with respective sample sizes n_1 and n_2 into two groups with respective sample sizes n_1 and n_2. Define the dummy variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{else} \end{cases}$$

Show that the least squares estimators β_1 and β_0 are respectively

$$\hat{\beta}_1 = \overline{y}_1 - \overline{y}_2 \text{ and } \hat{\beta}_0 = \overline{y}_2$$

Let us show the following. Consider simple linear model

$$y_i = \beta_0 + x\beta_1 x_i + \epsilon$$

as indicated above. We can derive the following results

$$\overline{x} = \frac{n_1}{n} = \frac{n_1}{n_1 + n_2}, \overline{x} = (\frac{n_1}{n_1 + n_2})^2, \sum_{i=1}^{n} x_i^2 = n_1, \overline{xy} = \frac{n_1}{n}(\frac{1}{n}\sum_{i=1}^{n} y_i), \sum_{i=1}^{n} x_i\, y_i$$

$$= \sum_{i=1}^{n_1} (1) y_i + \sum_{i=1}^{n_0} (0) y_i = \sum_{i=1}^{n_1} y_i$$

With these resutls in hands, we can pursue the following.

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})} \\
&= \frac{\sum x_i y_i - \sum \overline{x} y_i}{\sum x_i^2 - \sum \overline{x}^2 - 2\sum x_i \overline{x}} \\
&= \frac{\sum_{i=1}^{n_1} y_i - \frac{n_1}{n}\sum_{i=1}^{n} y_i}{n_1 + n(\frac{n_1}{n})^2 - 2(\frac{n_1}{n})n_1} \\
&= \frac{\sum_{i=1}^{n_1} y_i - n_1 \overline{y}}{n_1 - \frac{n_1^2}{n}} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n-1} y_i - \frac{n_1}{n}\frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{n_1}{n} - (\frac{n_1}{n})^2} \\
&= \frac{\frac{1}{n}[\sum_{i=1}^{n_1} - \frac{n_1}{n}(\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} y_i)]}{\frac{n_1 n_2}{n^2}} \\
&= \frac{\frac{n_1}{n}\sum_{i=1}^{n_1} y_i - \frac{n_2}{n}\sum_{i=1}^{n_2} y_i}{\frac{n_1 n_2}{n}} \\
&= \frac{1}{n_1}\sum_{i=1}^{n_1} y_i - \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \\
&= \overline{y}_1 - \overline{y}_2
\end{aligned}
$$

which gives us the first part.

Next, we proceed to show that $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 X_i = \overline{Y}_2$.

$$
\begin{aligned}
\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{n_1}{n}(\overline{y}_1 - \overline{y}_2) \\
&= \frac{1}{n}(\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_2} y_i) - \frac{n_1}{n}(\frac{1}{n_1}\sum_{i=1}^{n_1} y_i - \frac{1}{n_2}\sum_{i=1}^{n_2} y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n_1} y_i + \frac{1}{n}\sum_{i=1}^{n_2} y_i - \frac{1}{n}\sum_{i=1}^{n_1} + \frac{n_1}{nn_2}\sum_{i=1}^{n_2} y_i \\
&= \frac{1}{n}\sum_{i=1}^{n_2} y_i + \frac{n_1}{nn_2}\sum_{i=1}^{n_2} y_i \\
&= (\frac{n_1 + n_2}{n_2}\frac{1}{n})\sum_{i=1}^{n_2} y_i \\
&= \frac{1}{n_2}\sum_{i=1}^{n_2} y_i \\
&= \overline{y}_2
\end{aligned}
$$

and we are done. ▫

PROBLEM 4

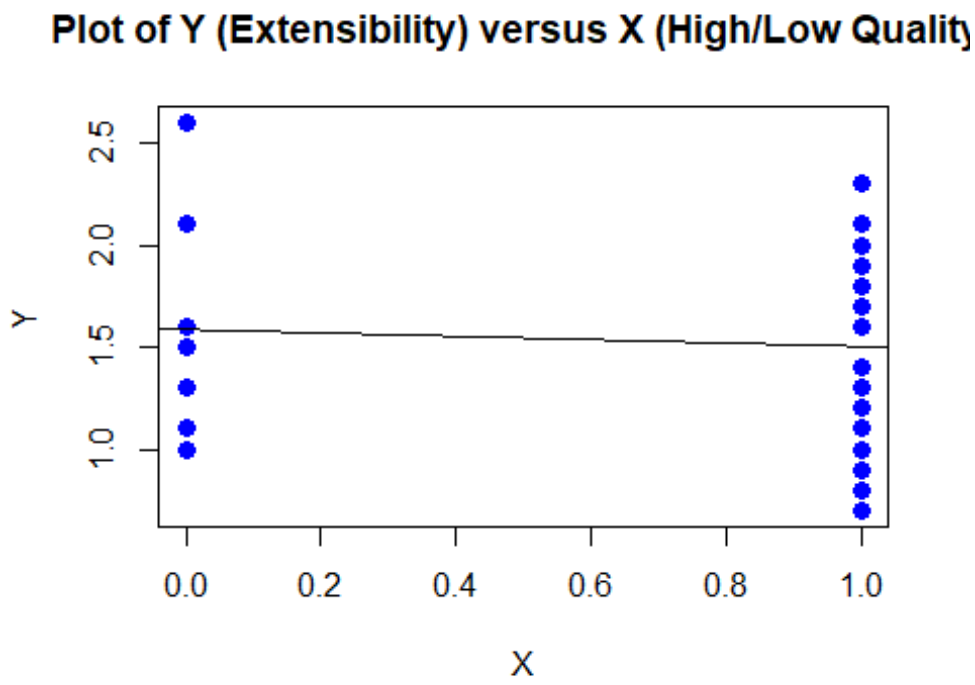
Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article gave the accompanying data an extensibility (%) at 100 gm/cm for both high-quality (H) fabric and poor-quality (P) fabric specimens

```
H <- c(1.2, 0.9, 0.7, 1.0, 1.7, 1.7, 1.1, 0.9, 1.7, 1.9, 1.3,
       2.1, 1.6, 1.8, 1.4, 1.3, 1.9, 1.6, 0.8, 2.0, 1.7, 1.6, 2.3, 2.0)
P <- c(1.6, 1.5, 1.1, 2.1, 1.5, 1.3, 1.0, 2.6)
Y <- c(H,P)
X <- c(rep(1, length(H)), rep(0, length(P)))
```

Use *R* to perform the following tasks

(i) Create Graph

```
# Visualize
plot(X, Y, main = "Plot of Y (Extensibility) versus X (High/Low Quality)",
     pch = 16, cex = 1.3, col = "blue")
abline(lm(Y ~ X))
```



(ii) Regression Analysis

This is asking the β of the explanatory variable x, that is, to test whether this parameter should be zero of not. One can consider null hypothesis $H_0: \beta = 0$ and alternative hypothesis $H_1: \beta \neq 0$.

From the results below, we observe that t-value is -0.42, which is less than 2.96 using 95% confident level. We fail to reject null hypothesis. Hence, we can conclude that we do not have sufficient evidence to say that β is not zero.

```
# Linear Model
data <- data.frame(cbind(Y, X))
summary(lm(Y~X))

##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80833 -0.33333  0.05208  0.31667  1.01250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.58750    0.16466   9.641 1.06e-10 ***
## X           -0.07917    0.19013  -0.416     0.68
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4657 on 30 degrees of freedom
## Multiple R-squared:  0.005746,   Adjusted R-squared:  -0.0274
## F-statistic: 0.1734 on 1 and 30 DF,  p-value: 0.6801

# Check:
# If quality is high, e.g. the group H;
# If quality is low, e.g. the group P
mean(data[data$X == 1, 1])

## [1] 1.508333

mean(data[data$X == 0, 1])

## [1] 1.5875

# Intuitively, this makes sense
# the two numbers are not that much different
# so using this parameter will probably not do much
# to response variable, Y
```

PROBLEM 5

There are five parts.

(i) Regression

Consider regression

$$Y_i = \beta x_i + \epsilon_i \text{ for } i = 1,2,\dots,n \text{ while } \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$$

Derive the maximum likelihood estimators of β and σ^2.

We want to check 1st order condition and also 2nd order condition. We proceed in that manner.

First we realize the density function is

$$f(Y|X) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}\frac{(Y_i - \beta x_i)^2}{\sigma^2})$$

and we can obtain likelihood function

$$l = \ln(f(Y|X) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(Y_i - \beta x_i)^2$$

Next, we want to solve for $\max_{\beta,\sigma} l(\beta, \sigma; Y, x_i)$, and we proceed the following.

Step 1. We check 1st order condition. We want to set up

$$\frac{\partial l(\cdot)}{\partial \beta} = 0, \text{and } \frac{\partial l(\cdot)}{\partial \sigma^2} = 0$$

and we solve for

$$\frac{\partial l}{\partial \beta} = \frac{1}{\sigma^2}(\sum_{i=1}^{N} x_i^T Y_i - \sum_{i=1}^{N} x_i^T x_i \beta_i)$$

and

$$\frac{\partial l}{\partial \sigma^2} = \frac{1}{2\sigma^2}(\frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - x_i\beta)^2 - N)$$

which solve for

$$\beta = (X^T X)^{-1} X^T Y, \text{and } \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - x_i\beta)^2$$

Step 2. we check the 2nd order condition. We want to compute the Hessian matrix.

$$H = \begin{bmatrix} \partial(\frac{\partial l}{\partial \beta})/\partial\beta & \partial(\frac{\partial l}{\partial \beta})/\partial\sigma \\ \partial(\frac{\partial l}{\partial \sigma})/\partial\beta & \partial(\frac{\partial l}{\partial \sigma})/\partial\sigma \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma} x_i^T x_i & -\frac{(Y_i - x_i\beta)}{(\sigma^2)^2} x_i^T \\ -\frac{(Y_i - x_i\beta)}{\sigma^4} x_i & \frac{1}{2\sigma^4} - \frac{(Y_i - x_i\beta)^2}{(\sigma^2)^2} \end{bmatrix}$$

and hence we can show $\mathbb{E}((Y_i - x_i\beta)x_i) = 0$ and thus we have a symmetric matrix

$$H = \begin{bmatrix} \frac{1}{\sigma^2}\mathbb{E}(x_i^T x_i) & 0 \\ 0 & \frac{1}{2(\sigma^2)^2} \end{bmatrix}$$

Thus, both 1st and 2nd order conditions are satisfied and we are done. Please note, we that I have answer in summation form in part (iii). They represent the same content.

(ii) Residuals

Consider the residuals e_i related to the regression through the origin model in (i). Prove that

$$\sum_{i=1}^{n} e_i x_i = 0$$

Also, in the regression in (i), is the sum of residuals equal to zero? I.e. If the following relation true?

$$\sum_{i=1}^{n} e_i = 0$$

Explain.

We consider the following.

$$\begin{aligned}\sum_{i=1}^{n} e_i\, x_i &= \sum_{i=1}^{n} (Y_i - (\beta x_i))x_i \\ &= \sum_{i=1}^{n} (Y_i x_i - \beta x_i^2) \\ &= \sum_{i=1}^{n} Y_i\, x_i - \sum_{i=1}^{n} \beta\, x_i^2 \\ &= (0 + \sum_{i=1}^{n} \beta\, x_i^2) - \sum_{i=1}^{n} \beta\, x_i^2 \\ &= 0\end{aligned}$$

Next, consider originally we have $Y = \beta_0 + \beta_1 x_i + \epsilon_i$ in matrix form. That is, we have

$$\vec{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = [\vec{1} \quad \vec{x}]$$

so the response can be projected to a plane that has the eigen-vector $\vec{1}$ and $\vec{x}$ which form a basis span$\{\vec{1}, \vec{x}\}$. In this case, we can think of the vector $\vec{1}$ and $\vec{x}$ form a 2-dimension plane and any vector y that we want to be predit to have a projection onto this 2-dimensional plane. The projection is $\hat{y}$ and the distance from y to $\hat{y}$ would be the error, $e_i = y - \hat{y}$. In this case, we have conclusions, $\sum e x_i = 0 =< \vec{0}, \vec{x} >$ and also $\sum e_i = 0 =< \vec{0}, \vec{1} >$.

With this in mind, let us use the premise in this problem. Consider $Y = \beta x_i + \epsilon_i$ in matrix form. Then we have

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

so response Y can be projected on a single vector and note that this time it is not a plane. The projection is, again, $\hat{y}$, and the distance between real value y and the projection would

be $e = Y - \hat{Y}$. Thus, we have $\sum e_i x_i = <\vec{e}, \vec{x} = 0$ but it is not gauranteed that $\sum e_i = <\vec{e}, \vec{1} > = 0$, e.g. $\sum e_i = <\vec{e}, \vec{1} > \neq 0$

(iii) Null/Alternative Pair

Consider tesitng null/alternative pair

$$H_0: \beta = \beta' \text{ v.s. } \beta \neq \beta'$$

Note that β' is the hypothesized value. Show that the likelihood-ratio test can be based on the rejection region $|T| > k$ with test statistic

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta} x_i)^2/(n-1)}{\sum_{i=1}^{n} x_i^2}}}$$

Note that k is some positive real number and $\hat{\beta}$ is the maximum likelihood estimator of β.

First, we find the likelihood function

$$\mathcal{L}(\beta) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp(-\frac{1}{2\sigma^2}\sum(Y_i - \beta x_i)^2)$$

and we need to find MLE for (1) full model, and (2) reduced model.

Full Model: we have

$$\frac{\partial l}{\partial \beta} = \partial(-\frac{1}{2\sigma^2}\sum(Y_i - \beta x_i)^2))/\partial\beta \overset{\text{set}}{=} 0$$

and we solve for

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{\beta} x_i)^2 = \frac{1}{n}\sum(Y_i - (\frac{\sum x_i Y_i}{\sum x_i^2})x_i)^2$$

Reduced Model: we have

$$\frac{\partial l}{\partial \beta} = \partial(-\frac{1}{2\sigma^2}\sum(Y_i - \beta x_i)^2))/\partial\beta \overset{\text{set}}{=} 0$$

and we solve for

$$\hat{\beta}' = \beta' \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{\beta}' x_i)^2$$

and then we can compute likelihood ratio, namely λ, to be

$$\lambda = \frac{\mathcal{L}(\beta')}{\mathcal{L}(\hat{\beta})} = \left(\frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma}^2}\sum(Y_i - \hat{\beta}\, x_i)^2\right)}{\left(\frac{1}{\sqrt{2\pi\sigma'^2}}\right)^n \exp\left(-\frac{1}{2\hat{\sigma'}^2}\sum(Y_i - \beta' x_i)^2\right)}\right)$$

and we want to show $\lambda > k$ because we have the likelihood function for full model on top. In this case, notice that the terms in exponentials will be cancelling out and the terms such as square root and π will be constant. Thus, we can simplify and obtain the following

$$\left(\sqrt{\frac{\hat{\sigma}'^2}{\hat{\sigma}^2}}\right)^n > k'$$

and it is sufficient to show

$$\left(\frac{\hat{\sigma}'^2}{\hat{\sigma}^2}\right)^2 > k'$$

From here we can plug in the MLE that we found for variance for full model and also for reduced model. Using the trick, we can obtain the following

$$\frac{\sum(Y_i - \hat{\beta}\, x_i + \hat{\beta}\, x_i + \beta' x_i)^2}{\sum(Y_i - \hat{\beta}\, x_i)} > k'''$$

and we are almost there. Opening the top of the fraction up, we realize some terms will cancel out and we obtain

$$\frac{(\hat{\beta} - \beta')^2}{\sum(Y_i - \hat{\beta}\, x_i)^2 / \sum x_i^2} > k''''$$

which allows us to set rejection region. Note that all terms on the right hand side of the inequality, e.g. k', k'', k''', k'''', are constants.

(iv) Probability Distribution

Under H_0, what is the probability distribution of the above test statistic T?

The probability distribution tends to $\chi^2(n-1)$ under null hypothesis. Let us elaborate the reasoning in the following. Recall T-statistic

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{\beta}\, x_i)^2/(n-1)}{\sum_{i=1}^n x_i^2}}}$$

as stated in premise and derived in part (ii). We can rearrange the formula to get the following

$$T = \frac{(\hat{\beta} - \beta')/\sqrt{\sigma^2/\sum x_i}}{\sqrt{\frac{\sum(Y_i - \hat{\beta}\, x_i)^2}{(n-1)\sigma^2}}}$$

and we realize the top of the fraction to be a $N(0,1)$. Then we only need to dicuss the bottom of the fraction. We claim that the bottom of the fraction $\sim \chi^2_{n-1}$. This is because we know $\sum(Y_i - \hat{\beta}\, x_i)^2 \sim \sigma^2\chi^2_{n-1}$ and we can open up the full square and we have the following.

$$\sum\sigma^2\chi_1^2 - (\frac{\sum x_i\epsilon_i}{\sum x_i})^2$$

which, in the end, gives us the form of

$$\sigma^2\chi_n^2 - \sigma^2\chi_1^2$$

and this is result tends to $\sigma^2\chi^2_{n-1}$ based on Basu's Theorem. This is last argument works. Because we know T-statistics is a function of z-statistics and some other information. In this case, z-statistics is a complete sufficient statistics and the other information we can treat them as auxilliary statistic. Hence, by Basu's Theorem, we know that they are independent. We can directly conclude that the subtraction comes down to a term that tends to chi-square distribution with degree of freedom $n - 1$.

(a) Likelihood Ratio

Compute the liklihood-ratio test statistic (λ) from Definition 2.4 from page 45 in lecture notes.

(b) Expression

Simplifying expression using the following trick

$$\sum_{i=1}^{n}(\,Y_i - \beta' x_i)^2 = \sum_{i=1}^{n}(\,Y_i - \hat{\beta}\, x_i + \hat{\beta}\, x_i - \beta' x_i)^2$$

(c) Results

After simplifying $\lambda < c$, find a suitable transformation of λ that yields the desired test statistic and rejection rule.

STAT GR5205 Homework 2 KEY [100 pts]
Due 8:40am Wednesday, October 10th

Problem 1 (2.7 KNN) [28 pts]

Sixteen batches of plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume the first-order regression model (1.1) is appropriate (**(4.1) in the notes**).

Data not displayed

Use R to perform the following tasks:

i. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate. **(4 pts)**

R code and output

```
confint(lm(Y~X),level=.99)
```

```
                0.5 %     99.5 %
(Intercept) 160.690457 176.509543
X             1.765287   2.303463
```

Interpretation

For every one hour increase in elapsed time, we expect at 99% confidence for the mean hardness to increase by 1.77 to 2.30 Brinell units.

ii. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = .01$. **(4 pts)**

R code and output

```
summary(lm(Y~X))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
X             2.03438    0.09039   22.51 2.16e-12 ***

t <- (2.03438-2)/0.09039
```

```
t
[1] 0.3803518

2*(1-pt(t,14))
[1] 0.7093927
```

Solution

Consider testing the null alternative pair $H_0 : \beta_1 = 2$ versus $H_A : \beta_1 \neq 2$. The test statistic and P-value are respectively 0.38 and 0.7093. At 1% significance, we fail to reject the null hypothesis and conclude that an increase of 2 Brinell units per hour is reasonable. Note that we could have also used the 99% confidence interval form Part (i) to perform this test. Since 2 is in the confidence interval, we fail to reject the null hypothesis.

iii. Set up the ANOVA table. **(4 pts)**

R code and output

```
anova(lm(Y~X))

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 5297.5  5297.5  506.51 2.159e-12 ***
Residuals 14  146.4    10.5
```

iv. Test by means of an F-test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$. **(4 pts)**

Solution

Consider testing the null alternative pair $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The test statistic and P-value are respectively 506.51 and $2.259 * 10^{-12}$. At 1% significance, we reject the null hypothesis and conclude that there is a linear association between the hardness of the plastic and the elapsed time.

v. Does t^2_{calc} from part [ii] equal f_{calc} from part [iv]4 Explain why this identity holds or does not hold. **(4 pts)**

Solution $t^2_{calc} \neq f_{calc}$ because the null hypothesis in Part (ii) is $H_0 : \beta_1 = 2$ and the null hypothesis in Part (iv) is $H_0 : \beta_1 = 0$.

vi. Construct 95% Bonferroni joint confidence intervals for estimating both the true intercept β_0 and the true slope β_1. **(4 pts)**

R code and output

```
confint(lm(Y~X),level=1-.05/2)

                1.25 %    98.75 %
(Intercept) 161.932014 175.267986
X             1.807526   2.261224
```

vii. Construct 95% Bonferroni joint confidence intervals for predicting the true average hardness corresponding to elapsed times 20, 28 and 36 hours. **(4 pts)**

R code and output

```
newdata <- data.frame(X=c(20,28,36))
level <- 1-.05/3
predict(lm(Y~X),newdata=newdata,interval="confidence",level=level)

       fit      lwr      upr
1 209.2875 206.3395 212.2355
2 225.5625 223.3652 227.7598
3 241.8375 238.8895 244.7855
```

Problem 2 [12 pts]

Consider the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

i. Assuming $H_0 : \beta_1 = 0$ is true, use R to simulate the sampling distribution of the F-statistic

$$F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}.$$

For this simulation, assume $\beta_0 = 10$, $\sigma = 3$ and $n = 30$. Pick an arbitrary range for x and run the loop 10,000 times to generate the sampling distribution.

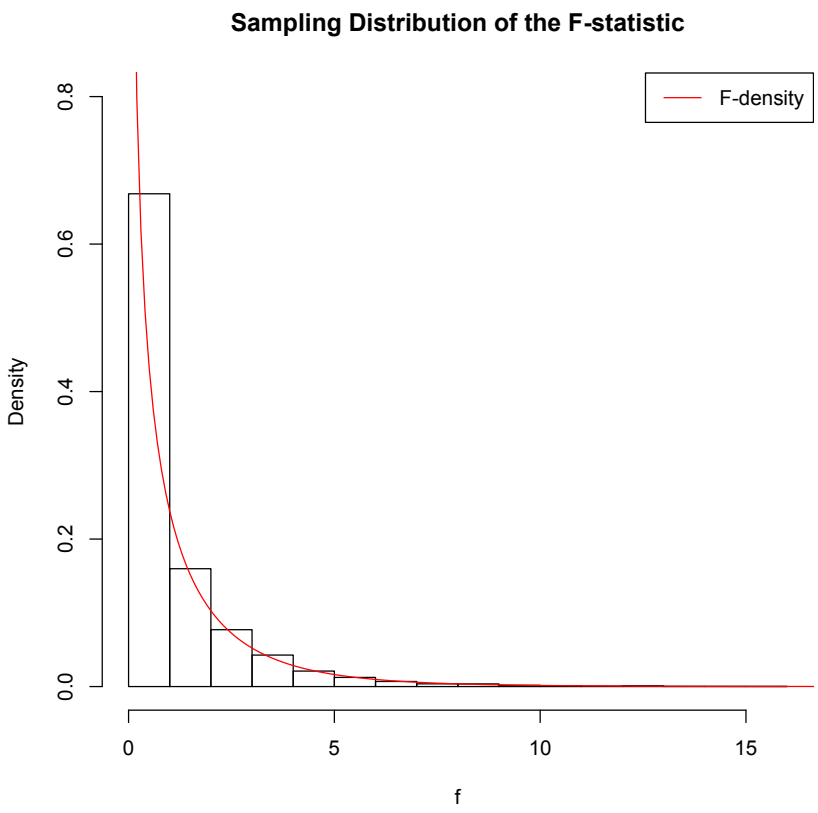
```
#Define a single X data set. We could have defined any X data set with n=30 observations.
x=1:30

#Define empty T-list
F.list <- NULL

#Start loop
for (i in 1:10000){
```

```
#Simulate n=30 random errors with mean 0 and variance 9.
error <- rnorm(30,mean=0,sd=3)

#Construct the y values from the errors.  Note: under the null H_0, the slope is zero.
y <- 10+error

#Extract the F-test statistic form the anova table
F.list[i] <- anova(lm(y~x))[[4]][1]

}

#Plot the simulated distribution
hist(F.list,freq=FALSE,ylim=c(0,0.8),main="Sampling Distribution of the F-statistic",xlab="f")
#Overlay the correct density on the histogram
f <- 0:200/10
lines(f,df(f,df1=1,df2=30-2),col=2)
legend("topright",c("F-density"),col=2,lty=1)
```

ii. From the simulated sampling distribution, plot a histogram and overlay the *correct density* on the histogram. **See the next page**.

iii. Compute the $95th$ percentile of both the simulated sampling distribution and the *correct* F-distribution. Compare these values.

```
> quantile(f.list,.95)
4.282201
> qf(.95,1,28)
[1] 4.195972
```


Sampling Distribution of the F-statistic
F-density
Density
0.0
0.2
0.4
0.6
0.8
0
5
10
15
f

Problem 3 [10 pts]

Consider splitting the response values $y_1, \ldots, y_n$ into two groups with respective sample sizes n_1 and n_2. Define the **dummy** variable

$$x_i = \begin{cases} 1 & \text{if group one} \\ 0 & \text{if group two} \end{cases} \tag{1}$$

Show that the least squares estimators of β_1 and β_0 are respectively

$$\hat{\beta}_1 = \bar{y}_1 - \bar{y}_2 \quad \text{and} \quad \hat{\beta}_0 = \bar{y}_2,$$

where $\bar{y}_1$ and $\bar{y}_2$ are the respective sample means of each group.

Solution

Note that

$$n = n_1 + n_2, \quad 1 - \frac{n_1}{n} = \frac{n_2}{n}, \quad \bar{y} = n_1\bar{y}_1 + n_2\bar{y}_2, \quad \sum x_i y_i = n_1\bar{y}_1, \quad \text{and} \quad \sum x_i = n_1$$

Also note that

$$\begin{aligned} S_{xy} &= \sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i \\ &= n_1\bar{y}_1 - \frac{1}{n}n_1(n_1\bar{y}_1 + n_2\bar{y}_2) \\ &= n_1\left(\left(1 - \frac{n_1}{n}\right)\bar{y}_1 - \frac{n_2}{n}\bar{y}_2\right) \\ &= n_1\left(\frac{n_2}{n}\bar{y}_1 - \frac{n_2}{n}\bar{y}_2\right) \\ &= \frac{n_1 n_2}{n}(\bar{y}_1 - \bar{y}_2), \end{aligned}$$

and

$$\begin{aligned} S_{xx} &= \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 \\ &= n_1 - \frac{1}{n}n_1^2 \\ &= n_1\left(1 - \frac{n_1}{n}\right) \\ &= \frac{n_1 n_2}{n}. \end{aligned}$$

Thus

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \bar{y}_1 - \bar{y}_2$$

For the intercept,

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2 - (\bar{y}_1 - \bar{y}_2)\frac{n_1}{n} \\ &= \frac{n_1}{n}\bar{y}_1 - \frac{n_1}{n}\bar{y}_1 + \left(\frac{n_1}{n} + \frac{n_2}{n}\right)\bar{y}_2 \\ &= \bar{y}_2\end{aligned}$$

Problem 4 [10 pts]

Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The article *Compatibility of Outer and Feasible Interlining Fabrics in Tailored Garments* gave the accompanying data on extensibility (%) at 100 gm/cm for both high-quality (H) fabric and poor-quality (P) fabric specimens.

H	1.2	.9	.7	1.0	1.7	1.7	1.1	.9	1.7
	1.9	1.3	2.1	1.6	1.8	1.4	1.3	1.9	1.6
	.8	2.0	1.7	1.6	2.3	2.0			
P	1.6	1.5	1.1	2.1	1.5	1.3	1.0	2.6	

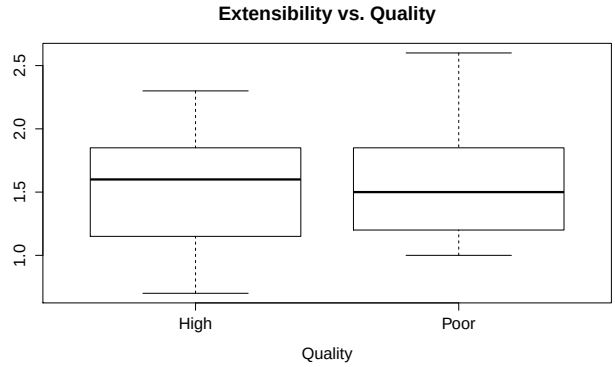
Use `R` to perform the following tasks.

i. Create an appropriate graphic to visualize the relationship between extensibility and quality. Do you think there is a relationship between extensibility and quality? Make sure to label the plot. **(5 pts)**

`R` code and output:

```
h <- c(1.2,.9,.7,1,1.7,1.7,1.1,.9,1.7,1.9,1.3,2.1,1.6,1.8,1.4,1.3,1.9,1.6,.8,2,1.7,1.6,2.3,2)
p <- c(1.6,1.5,1.1,2.1,1.5,1.3,1,2.6)
y <- c(h,p)
quality <- c(rep("High",length(h)),rep("Poor",length(p)))
boxplot(y~quality,main="Extensibility vs. Quality",xlab="Quality")
```

Based on the multiple box plot, the different fabric quality does not appear to influence fabric extensibility. Although no relationship appears to exist, a hypothesis test procedure should be preformed to formally test this claim.



ii. Using the indicator variable

$$x = \begin{cases} 1 & \text{if high quality} \\ 0 & \text{if low quality} \end{cases}$$

run a regression analysis to test if the average fabric extensibility differs per group. **(5 pts)**

R code and output:

```
y <- c(h,p)
x <- c(rep(1,length(h)),rep(0,length(p)))
summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.58750    0.16466   9.641 1.06e-10 ***
x           -0.07917    0.19013  -0.416     0.68
```

To assess if the average fabric extensibility differs per group, we test the null alternative pair $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The computed test statistic and two–tailed p–value are $t_{calc} = -0.416$ and p–value=0.68. At 5% significance, we fail to reject the null hypothesis because p–value= $0.68 > .05$. Thus, there is not sufficient evidence to show that the average fabric extensibility differs per group.

Problem 5 [40 pts]

i. Consider the *regression through the origin model* given by

$$Y_i = \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{2}$$

Derive the maximum likelihood estimators of β and σ^2. **(15 pts)**

Solution

The likelihood function is

$$\begin{aligned}
\mathcal{L}(\beta, \sigma^2; y_1, y_2, \ldots, y_n) &= f(y_1|\beta, \sigma^2) \times f(y_2|\beta, \sigma^2) \times \cdots \times f(y_n|\beta, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta x_i)^2 \right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \right).
\end{aligned}$$

The log-likelihood function is:

$$\log(\mathcal{L}(\beta, \sigma^2; y_1, y_2, \ldots, y_n)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2.$$

Taking partial derivates with respect to parameters β and σ^2 gives

$$\begin{aligned}
\frac{\partial}{\partial\beta}\log(\mathcal{L}(\beta, \sigma^2; y_1, y_2, \ldots, y_n)) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)x_i \\
\frac{\partial}{\partial\sigma^2}\log(\mathcal{L}(\beta, \sigma^2; y_1, y_2, \ldots, y_n)) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta x_i)^2.
\end{aligned}$$

Set the above partial derivatives equal to zero and solve for β and σ^2. First find the MLE of β:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0 \quad \rightarrow \quad \sum_{i=1}^{n}(y_i - \beta x_i)x_i = 0$$

$$\rightarrow \sum_{i=1}^{n} y_i x_i - \beta\sum_{i=1}^{n} x_i^2 = 0 \quad \rightarrow \quad \beta\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

$$\rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

Now find the MLE of σ^2. To accomplish this, substitute $\hat{\beta}$ into the relevant expression and solve for σ^2.

$$-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\hat{\beta}x_i)^2=0 \quad \rightarrow \quad -\frac{n\sigma^2}{2}+\frac{1}{2}\sum_{i=1}^{n}(y_i-\hat{\beta}x_i)^2=0$$

$$\rightarrow -n\sigma^2+\sum_{i=1}^{n}(y_i-\hat{\beta}x_i)^2=0 \quad \rightarrow \quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}x_i)^2$$

When using multivariate calculus to check that $L=\log(\mathcal{L}(\hat{\beta},\hat{\sigma}^2))$ is in fact a maximum, we must also validate that at least one of the second-order partial derivatives is negative,

$$\frac{\partial^2 L}{\partial^2\beta^2}L(\beta,\sigma^2)\bigg|_{\beta=\hat{\beta},\hat{\sigma}^2}<0 \quad \text{and} \quad \frac{\partial^2 L}{\partial(\sigma^2)^2}L(\beta,\sigma^2)\bigg|_{\beta=\hat{\beta},\sigma^2=\hat{\sigma}^2}<0,$$

and we must also check that the Jacobian of the second-order partial derivatives is positive,

$$\frac{\partial L}{\partial\beta^2}L(\beta,\sigma^2)\frac{\partial^2 L}{\partial(\sigma^2)^2}L(\beta,\sigma^2)-\left(\frac{\partial^2 L}{\partial\beta\partial\sigma^2}L(\beta,\sigma^2)\right)^2\bigg|_{\beta=\hat{\beta},\hat{\sigma}^2}>0.$$

These properties do check out for this application. In fact

$$\frac{\partial^2}{\partial\beta^2}\log(\mathcal{L}(\beta,\sigma^2;y_1,y_2,\ldots,y_n))=-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}x_i^2<0,$$

and the Jacobian is

$$\begin{aligned}
&-\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2\left(\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\sum_{i=1}^{n}(y_i-\beta x_i)^2\right)-\frac{1}{\sigma^8}\sum_{i=1}^{n}(y_i-\beta x_i)x_i\bigg|_{\beta=\hat{\beta},\hat{\sigma}^2}\\
=&-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}x_i^2\left(\frac{n}{2\hat{\sigma}^4}-\frac{1}{\hat{\sigma}^6}\sum_{i=1}^{n}(y_i-\beta x_i)^2\right)-\frac{1}{\hat{\sigma}^8}\sum_{i=1}^{n}(y_i-\hat{\beta}x_i)x_i\\
=&-\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}x_i^2\left(\frac{n}{2\hat{\sigma}^4}-\frac{n}{\hat{\sigma}^6}\hat{\sigma}^2\right)-\frac{1}{\hat{\sigma}^8}0=\frac{n\sum_{i=1}^{n}x_i^2}{2\hat{\sigma}^2}>0.
\end{aligned}$$

Thus the likelihood function does achieve its maximum at $\hat{\beta}$ and $\hat{\sigma}^2$.

ii. Consider the residuals e_i related to the regression through the origin model (2). Prove that

$$\sum_{i=1}^{n} e_i x_i = 0.$$

Also, in the regression through the origin model (2), is the sum of residuals equal to zero? I.e., is the following relation true?

$$\sum_{i=1}^{n} e_i = 0.$$

Explain your answer in a few sentences or less. **(10 pts)**

Solution

P1

$$\begin{aligned}
\sum_{i=1}^{n} e_i x_i &= \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) x_i \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) x_i \\
&= \sum_{i=1}^{n} y_i x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} y_i x_i - \hat{\beta} \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} y_i x_i - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} x_i^2 \\
&= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i y_i \\
&= 0
\end{aligned}$$

P2 In the regression through the origin model (2), the sum of residuals $\sum e_i$ is **not** always equal to zero. This claim can easily be validated with a counter example. Technically, the residual vector $\mathbf{e} = \begin{pmatrix} e_1 & e_2 & \cdots & e_n \end{pmatrix}^T$ is only orthogonal to $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}^T$ because the intercept wasn't used in the model. If the intercept was included, $\mathbf{e}$ would be orthogonal to the vectors $\mathbf{1} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T$ and $\mathbf{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}^T$. More formally, for the regression through the origin model, $\mathbf{e}$ is orthogonal to any vector in Span$\{\mathbf{1}, \mathbf{x}\}$. For simple linear regression, $\mathbf{e}$ is orthogonal to any vector in Span$\{\mathbf{1}, \mathbf{x}\}$.

iii. Consider testing the null/alternative pair

$$H_0 : \beta = \beta' \quad \text{v.s.} \quad H_A : \beta \neq \beta'.$$

Note that β' is the hypothesized value. Show that the likelihood-ratio test can be based on the rejection region $|T| > k$ with test statistic

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2/(n-1)}{\sum_{i=1}^n x_i^2}}}.$$

Note that k is some positive real number and $\hat{\beta}$ is the maximum likelihood estimator of β. **(15 pts)**

Solution

Before deriving the likelihood-ratio test, notice from Part 5.ii,

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \beta' x_i)^2 &= \sum_{i=1}^n (Y_i - \hat{\beta}x_i + \hat{\beta}x_i - \beta' x_i)^2 \\
&= \sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2 + 2\sum_{i=1}^n (Y_i - \hat{\beta}x_i)(\hat{\beta}x_i - \beta' x_i) + \sum_{i=1}^n (\hat{\beta}x_i - \beta' x_i)^2 \\
&= \sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2 + 2(\hat{\beta} - \beta')\sum_{i=1}^n e_i x_i + (\hat{\beta} - \beta')^2 \sum_{i=1}^n x_i^2 \\
&= \sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2 + (\hat{\beta} - \beta')^2 \sum_{i=1}^n x_i^2
\end{aligned}$$

The above expression will be used to simplify the likelihood-ratio statistic.

To derive the likelihood-ratio, first consider the *reduced* model. If $H_0 : \beta = \beta'$, then

$$Y_i = \beta' x_i + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Hence the likelihood is

$$\begin{aligned}
\mathcal{L}(\beta', \sigma^2; y_1, y_2, \ldots, y_n) &= f(y_1|\beta,' \sigma^2) \times f(y_2|\beta', \sigma^2) \times \cdots \times f(y_n|\beta', \sigma^2) \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta' x_i)^2 \right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta' x_i)^2 \right).
\end{aligned}$$

To maximize the above expression, we are only concerned with parameter σ^2. Evaluating the log-likelihood and differentiating yields

$$\frac{\partial}{\partial\sigma^2}\log(\mathcal{L}(\beta',\sigma^2;y_1,y_2,\ldots,y_n)) = -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\beta' x_i)^2.$$

Hence

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\beta' x_i)^2.$$

Also note that the second derivative test guarantees $\mathcal{L}$ is maximized at $\hat{\sigma}^2$.

The next step is to maximize $\mathcal{L}$ with respect to the full parameter space. Note that this was solved in Part 5.i. Hence the likelihood-ratio statistic is

$$\lambda = \frac{\left(\frac{1}{\sqrt{2\pi\frac{1}{n}\sum_{i=1}^{n}(y_i-\beta' x_i)^2}}\right)^n \exp(-n/2)}{\left(\frac{1}{\sqrt{2\pi\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta} x_i)^2}}\right)^n \exp(-n/2)} = \left(\frac{\sum_{i=1}^{n}(y_i-\beta' x_i)^2}{\sum_{i=1}^{n}(y_i-\hat{\beta} x_i)^2}\right)^{-n/2}.$$

Using the earlier sums of squares result, we have

$$\lambda = \left(\frac{\sum_{i=1}^{n}(Y_i-\hat{\beta}x_i)^2+(\hat{\beta}-\beta')^2\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(y_i-\hat{\beta}x_i)^2}\right)^{-n/2} = \left(1+\frac{(\hat{\beta}-\beta')^2}{\sum_{i=1}^{n}(Y_i-\hat{\beta}x_i)^2/\sum_{i=1}^{n}x_i^2}\right)^{-n/2}$$

$$= \left(1+\frac{1}{(n-1)}\left|(\hat{\beta}-\beta')\Big/\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{\beta}x_i)^2/(n-1)}{\sum_{i=1}^{n}x_i^2}}\right|^2\right)^{-n/2}$$

$$= \left(1+\frac{1}{(n-1)}|T|^2\right)^{-n/2}$$

Hence the rejection rule is defined by

$$\left(1+\frac{1}{(n-1)}|T|^2\right)^{-n/2} \leq c, \quad \text{where} \quad 0\leq c\leq 1.$$

Rearranging the above expression yields a rejection rule for T, i.e., reject H_0 if

$$|T| \geq k = \sqrt{(n-1)\Big(\exp\Big\{-2\log(c)/n\Big\}-1\Big)}, \quad \text{where} \quad k\geq 0.$$

iv. Under H_0, what is the probability distribution of the above test statistic T?

Solution

Glossing over several former details, we know that T has a students t-distribution with $n-1$ degrees of freedom, i.e.,

$$T = \frac{\hat{\beta} - \beta'}{\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{\beta}x_i)^2/(n-1)}{\sum_{i=1}^{n} x_i^2}}} \sim t(df = n-1)$$

STAT GR5205 Homework 3 [100 pts]

Due 8:40am Wednesday, October 24th

Problem 1

Consider the model

$$Y_i = \mu + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The sample mean and sample variance are defined respectively as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Use Theorem (3.6) on Page 91 to prove that the sample mean $\bar{Y}$ and sample variance S_Y^2 are independent random variables.

Problem 2

Consider the *single factor anova model* with three groups. The three groups are drug dose 1, drug dose 2 and control. Let n_1 and $\bar{y}_1$ respectively denote the number of respondents and sample mean response for drug dose 1 group. Let n_2 and $\bar{y}_2$ respectively denote the number of respondents and sample mean response for drug dose 2 group. Let n_3 and $\bar{y}_3$ respectively denote the number of respondents and sample mean response for the control group. Note that $n = n_1 + n_2 + n_3$. The *one-way anova* can be expressed using the *multiple linear regression model*

$$(1) \qquad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

$$x_{i1} = \begin{cases} 1 & \text{if drug dose 1} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if drug dose 2} \\ 0 & \text{otherwise} \end{cases}$$

i. Write down the design matrix and response vector describing model (1).

ii. Compute $(\mathbf{X}^T\mathbf{X})^{-1}$ and simplify the result. This requires inverting a 3×3 matrix.

iii. Estimate $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \beta_1 & \beta_2\end{pmatrix}^T$ using the least squares equation.

iv. Write down an expression for the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

Problem 3

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The data consists of variables age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y). For this data set, we skip residual diagnostics but in practice, that should be included in the analysis. The data set `HW3Problem3.txt` is posted on Canvas. Use `R` to perform the following tasks:

i. Regress the rental rates (Y) against all of the covariates; age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4). Write down the estimated linear model.

ii. What percentage of variation in rental rates is explained by this model?

iii. Are there any marginal relationships between the response variable and covariates? (Run t-tests on all slope parameters.)

iv. Run a F-test to see if there is an *overall relationship* between the rental rates and all of the covariates.

v. Run a F-test to simultaneously test the slopes for age (X_1) and vacancy rates (X_3).

vi. Run a F-test to see if vacancy rates (X_3) is a significant predictor after holding all other variables constant. To perform this test, use the full and reduced models. How does this test relate to the summary output from part (ii)?

vii. The researcher wishes to obtain 95% interval estimates of the mean rental rates for four typical properties specified as follows. Find the four confidence intervals using the Bonferroni procedure.

	1	2	3	4
x_1	5.0	6.0	14.0	12.0
x_2	8.25	8.50	11.50	10.25
x_3	0	0.23	0.11	0
x_4	250,000	270,000	300,000	310,000

Problem 4

i. Recall the multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i. \tag{2}$$

For each of the following regression models, indicate whether it can be expressed in the form of (2) by a suitable transformation. To receive full credit, describe the transformation if it exists.

a. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log(x_{i2}) + \beta_3 x_{i1}^2 + \epsilon_i$

b. $Y_i = \epsilon_i \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2\}$

c. $Y_i = \log(\beta_1 x_{i1}) + \beta_2 x_{i2} + \epsilon_i$

d. $Y_i = \beta_0 \exp\{\beta_1 x_{i1}\} + \epsilon_i$

e. $Y_i = [1 + \exp\{\beta_0 + \beta_1 x_{i1} + \epsilon_i\}]^{-1}$

ii. Consider the toy data set:

y	2.44	8.36	98.33	115.06	128.91	123.46	148.30	138.10	153.10	119.08
	87.66	134.88	91.71	126.81	40.41	54.94	33.03	35.74	14.99	-1.18
	2.44	8.36	28.33	45.06	48.91	43.46	118.30	108.10	233.10	199.08
	337.66	384.88								
x	0.00	0.00	1.00	1.00	2.00	2.00	3.00	3.00	4.00	4.00
	5.00	5.00	6.00	6.00	7.00	7.00	8.00	8.00	9.00	9.00
	10.00	10.00	11.00	11.00	12.00	12.00	13.00	13.00	14.00	14.00
	15.00	15.00								

The data set is provided in the file `HW3Problem4.txt` on canvas. Use multiple linear regression techniques to fit a polynomial to the above data set. To receive full credit, write down the estimated model and create a scatter plot with the estimated curve overlaid on the plot.

GR5205 Homework 3

Yiqiao Yin [YY2502]

Table of Contents

PROBLEM 1

Consider

$$Y_i = \mu + \epsilon_i \text{ for } \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

The sample mean and sample variance are defined respectively as

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

Prove sample mean and sample variance are independent random variables.

From the model, we have $Y_i = \mu + \epsilon_i$ given $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Then we know $\mu + \epsilon_i \sim N(\mu, \sigma^2)$ and thus $Y_i \sim N(0, \sigma^2)$ which is Gaussian. This implies that $Y_i - \overline{Y}$ is also Gaussian. Hence, we have $\overline{Y}$ and S_Y^2 to be independent random variables by Theorem 3.6 on Page 91.

▫

PROBLEM 2

Three groups: drug dose 1, drug dose 2, and control. Let n_1 and $\overline{y}_1$ respectively denote the number of respondents and sample mean response for drug dose 1. We define n_2 and $\overline{y}_2$, and n_3 and $\overline{y}_3$, the same way respectively. I omit the problem here for time constraints. Note that $n = n_1 + n_2 + n_3$. Consider one-way anova to be

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \text{while } i = 1,2,\dots,n \text{ and } \epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

and also

$$x_{i1} = \begin{cases} 1 & \text{if drug dose 1} \\ 0 & \text{otherwise} \end{cases}, x_{i2} = \begin{cases} 1 & \text{if drug dose 2} \\ 0 & \text{otherwise} \end{cases}$$

(i) Write design matrix and response vector

Let us write in the form of Response vector = Design Matrix. We have the following:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \begin{bmatrix} x_{11} \\ \vdots \\ x_{n1} \end{bmatrix} + \begin{bmatrix} \beta_2 \\ \vdots \\ \beta_2 \end{bmatrix} \begin{bmatrix} x_{12} \\ \vdots \\ x_{n2} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$
$$= \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

which is the matrix expression of the model above.

(ii) Computation

Compute $(X^T X)^{-1}$ and simplify result.

Given design matrix known, we have the following

$$X^T X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\overline{x} \\ n\overline{x} & \sum x_i^2 \end{bmatrix}$$

We need to find the determinant which is

$$|X^T X| = n\sum x_i^2 - (n\overline{x})^2 = n(\sum x_i^2 - n\overline{x}^2) = n s_{xx}$$

Hence, the inverse of $X^T X$, e.g. $(X^T X)^{-1}$, is $$ (X^{TX}){-1} =

=

$$

(iii) Estimation

Estimate $\vec{\beta} = (\beta_0, \beta_1, \beta_2)^T$ using least squares equation

We need to start with least squares Q and we shall derive the partial derivatives with respect to $\beta_0, \beta_1, \beta_2$, respectively.

Consider

$$Q = \sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \epsilon_i$$

and we set up

$$\begin{aligned}
\frac{\partial Q}{\partial \vec{\beta}} &= \begin{bmatrix} \partial Q/\partial \beta_0 \\ \partial Q/\partial \beta_1 \\ \partial Q/\partial \beta_2 \end{bmatrix} \\
&= \begin{bmatrix} \partial(\sum -2\beta_0(Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}))/\partial\beta_0 \\ \partial(\sum -2\beta_1(Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}))/\partial\beta_1 \\ \partial(\sum -2\beta_2(Y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}))/\partial\beta_2 \end{bmatrix} \\
&= \begin{bmatrix} -2\sum(Y_i - 2\beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) \\ -2\sum(x_{1i}Y_i - x_{1i}\beta_0 - 2\beta_1 x_{i1}^2 - \beta_2 x_{1i}x_{2i}) \\ -2\sum(x_{2i}Y_i - x_{2i}\beta_0 - x_{2i}\beta_1 x_{1i} - 2\beta_2 x_{2i}^2) \end{bmatrix} \overset{\text{set}}{=} 0
\end{aligned}$$

and we can solve for

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \overline{Y}_i - \frac{1}{n}\beta_1\sum x_{1i} - \frac{1}{n}\beta_2\sum x_{2i} \\ \frac{1}{\sum x_{1i}}(\sum Y_i - n\beta_0 - \beta_2\sum x_{2i}) \\ \frac{1}{\sum x_{2i}}(\sum Y_i - n\beta_0 - \beta_1\sum x_{1i}) \end{bmatrix}$$

(iv) Write estimated covariance matrix

Write expresstion of estimated covariance matrix $\hat{\beta}$.

The covariance matrix is

$$\Sigma_{\hat{\beta}} = \begin{bmatrix} \sigma_{\hat{\beta}_0} & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \sigma_{\hat{\beta}_1} & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \text{Cov}(\hat{\beta}_2, \hat{\beta}_0) & \text{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \sigma_{\hat{\beta}_2} \end{bmatrix}$$

PROBLEM 3

Real estate evaluates vacancy rates, square footage, rental sales, and operating expenses for commercial properties.

```
# Data
data <- read.delim("HW3Problem3.txt", sep = "")

# View
head(data); dim(data)
```

```
##   RentalRates Age OperatingExpense VacancyRates SquareFootage
## 1        13.5   1             5.02         0.14        123000
## 2        12.0  14             8.19         0.27        104079
## 3        10.5  16             3.00         0.00         39998
```

```
## 4        15.0   4            10.70         0.05         57112
## 5        14.0  11             8.97         0.07         60000
## 6        10.5  15             9.45         0.24        101385

## [1] 81  5
```

(i) Regression

This task we are asked to write down regression model for all covariates. From *R* output below, we have

$$\text{RentalRates} = 1.2 + -0.142\text{Age} + 0.282\text{OperatingExpense} + 0.619\text{VacancyRates} + (7.9 \times 10^{-6})\text{SquareFootage}$$

and, by using X_i's to represent all covariates, we have the following form

$$Y_i = 1.2 + -0.142X_1 + 0.282X_2 + 0.619X_3 + (7.9 \times 10^{-6})X_4$$

```
# Regression Model
model <- lm(data$RentalRates~., data = data)
model

##
## Call:
## lm(formula = data$RentalRates ~ ., data = data)
##
## Coefficients:
##      (Intercept)               Age  OperatingExpense      VacancyRates
##        1.220e+01        -1.420e-01         2.820e-01         6.193e-01
##    SquareFootage
##        7.924e-06
```

(ii) Percentage of variation

There is about 58% of the variation explained by this model.

```
# Regression Model
model <- lm(data$RentalRates~., data = data)
sum <- summary(model)
sum$r.squared

## [1] 0.5847496
```

(iii) t-test on all slopes

We can use *R* output from summary of the linear model using all covariates. We can generate the coefficient table and read off the t-value for each parameter. We can see that at critical value of 2.96 (e.g. 95% confidence level) we can observe VacancyRates to have a low t-value, much less than 2.96. We fail rejeject null hypothesis for VacancyRates and we can conclude that there is no sufficient evidence against null hypothesis. That is, we do not know if VacancyRates is statistically significant.

However, the rest of the parameters are all statistically significant since there t-values are above 2.96 (e.g. 95% confidence level).

```
# Regression Model
model <- lm(data$RentalRates~., data = data)
sum <- summary(model)
sum$coefficients
```

```
##                       Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)       1.220059e+01 5.779562e-01 21.1098807 1.601720e-33
## Age              -1.420336e-01 2.134261e-02 -6.6549332 3.894322e-09
## OperatingExpense  2.820165e-01 6.317235e-02  4.4642400 2.747396e-05
## VacancyRates      6.193435e-01 1.086813e+00  0.5698714 5.704457e-01
## SquareFootage     7.924302e-06 1.384775e-06  5.7224457 1.975990e-07
```

(iv) F-test to see overall relationship

We find F-calc from the summary of linear model below and then we check the p-value is less than 0.05. We reject null hypothesis: there is no overall relationship. Thus, w conclude that there is an overall relationship between rental rates and all covariates.

```
# Regression Model
model <- lm(data$RentalRates~., data = data)
sum <- summary(model)
f.calc <- sum$fstatistic[1]; f.calc

##    value
## 26.75553

1 - pf(f.calc, 4-1, nrow(data)-2)

##        value
## 4.825917e-12
```

(v) F-test to see age and vacancy rate

From *R* output below, we see that $f_{\text{calc}} = 2.6$ and we can compute p-value to be 0.11, greater than 0.05 significance level. Suppose hypothesis is that there is that $\beta_1 = \beta_3 = 0$. We fail to reject null hypothesis. Thus, we conclude that we do not have evidence to show that age and vacancy statistically influence the response variable.

```
# Regression Model
model <- lm(data$RentalRates~data$Age + data$VacancyRates, data = data)
sum <- summary(model)
f.calc <- sum$fstatistic[1]; f.calc

##    value
## 2.606831

1 - pf(f.calc, 2-1, nrow(data)-2)

##     value
## 0.1103915
```

(vi) F-test to see vacancy rates

We compute SSE(full model) and SSE(reduced model). Then we compute f-calc and the result is 35. We also check the p-value to be much less than 0.05. We fail to reject null hypothesis, and we conclude that there is a statistically significant relationship between vacancy rate and response variable.

```
# Regression Model
full.model <- lm(data$RentalRates~., data = data)
SSE.F <- sum(residuals(full.model)^2)
summary(full.model)
```

```
##
## Call:
## lm(formula = data$RentalRates ~ ., data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.220e+01  5.780e-01  21.110  < 2e-16 ***
## Age              -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## OperatingExpense  2.820e-01  6.317e-02   4.464 2.75e-05 ***
## VacancyRates      6.193e-01  1.087e+00   0.570     0.57
## SquareFootage     7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

```
reduced.model <- lm(data$RentalRates~data$VacancyRates, data = data)
SSE.R <- sum(residuals(reduced.model)^2)
summary(reduced.model)
```

```
##
## Call:
## lm(formula = data$RentalRates ~ data$VacancyRates, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.774 -1.095 -0.070  1.001  4.146
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        15.0700     0.2243  67.191   <2e-16 ***
## data$VacancyRates   0.8502     1.4347   0.593    0.555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.727 on 79 degrees of freedom
## Multiple R-squared:  0.004426,   Adjusted R-squared:  -0.008176
## F-statistic: 0.3512 on 1 and 79 DF,  p-value: 0.5551
```

```
f.calc <- ((SSE.R - SSE.F)/(79 - 76))/((SSE.F)/(76))
f.calc
```

```
## [1] 35.40403
```

```
# Check p-value
1 - pf(f.calc, 1, 76)
```

```
## [1] 7.699619e-08
```

(vii) Bonferroni procedure

We conduct Bonferroni confidence interval by using *confint()* function and we obtain the following confidence interval for all covariates.

```
# Conduct Bonferroni using confint() function
# Y ~ X1
reduced.model <- lm(data$RentalRates~data$Age, data = data)
confint(reduced.model, level = 1 - 0.05/4)

##                 0.625 %    99.375 %
## (Intercept) 14.9084395 16.38991568
## data$Age    -0.1370744  0.00729928

# Y ~ X2
reduced.model <- lm(data$RentalRates~data$OperatingExpense, data = data)
confint(reduced.model, level = 1 - 0.05/4)

##                           0.625 %   99.375 %
## (Intercept)           10.7234534 14.2170641
## data$OperatingExpense  0.1011648  0.4497414

# Y ~ X3
reduced.model <- lm(data$RentalRates~data$VacancyRates, data = data)
confint(reduced.model, level = 1 - 0.05/4)

##                      0.625 %  99.375 %
## (Intercept)       14.496712 15.643352
## data$VacancyRates -2.817118  4.517555

# Y ~ X4
reduced.model <- lm(data$RentalRates~data$SquareFootage, data = data)
confint(reduced.model, level = 1 - 0.05/4)

##                          0.625 %     99.375 %
## (Intercept)        1.304164e+01 1.452573e+01
## data$SquareFootage 4.607722e-06 1.226556e-05
```

PROBLEM 4

There is a theoretical question and a toy example.

(i) Describe transformation

Discuss whether the following can be transformed into the form of multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

(a). Consider $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log(x_{i2}) + \beta_3 x_{i1}^2 + \epsilon_i$. Write $x_{i2}{}' = \log(x_{i2})$ and we would have a polynomial regression

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}{}' + \beta_3 x_{i1}^2 + \epsilon_i$$

(b). Consider $Y_i = \epsilon_i \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2\}$. We can take log on both sides to obtain

$$\log(Y_i) = \log(\epsilon_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2$$

then we write $Y_i' = \log(Y_i)$ and $\epsilon_i' = \log(\epsilon_i)$. Thus, we obtain

$$Y_i' = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i'$$

(c). We cannot successfully transform the equation into multiple linear regression format.

(d). We cannot successfully transform the equation into multiple linear regression format.

(e). Consider $Y_i = [1 + \exp\{\beta_0 + \beta_1 x_{i1} + \epsilon_i\}]^{-1}$. We can look flip the top and bottom of the fraction on both sides to obtain

$$\frac{1}{Y_i} = 1 + \exp\{\beta_0 + \beta_1 x_{i1} + \epsilon_i\}$$

But we cannot transform further more from here.

(ii) Toy example

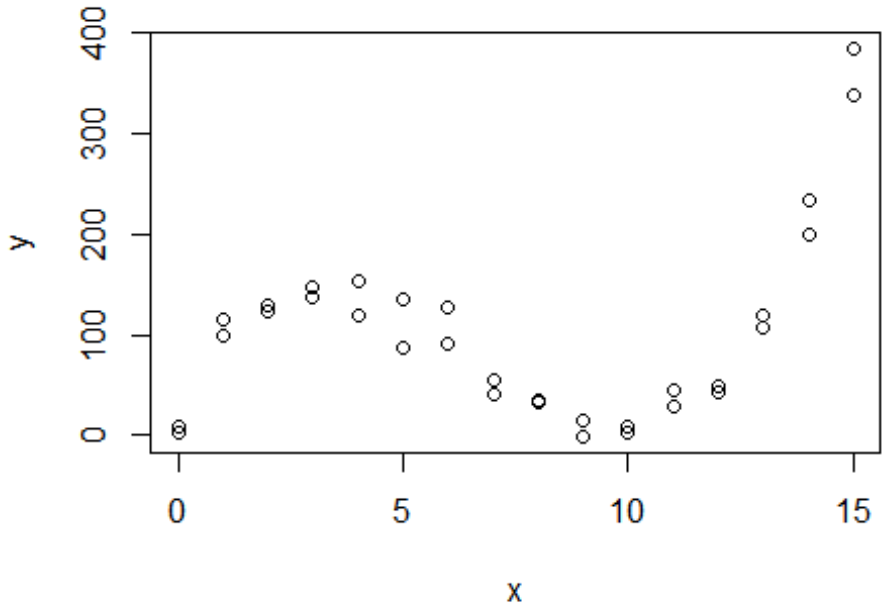
Let use the toy example from problem. We upload data and take a quick look at the data by the first six rows and a plot of x and y.

```
# Data
data <- read.delim("HW3Problem4.txt", sep = "")
y <- data$y # Response variable
x <- data$x # Explanatory variable

# View
head(data); dim(data)
```

```
##            y x
## 1   2.438432 0
## 2   8.358128 0
## 3  98.325343 1
## 4 115.060153 1
## 5 128.905388 2
## 6 123.461811 2
```

```
## [1] 32  2
```

```
plot(x, y)
```



We use the following code to run a multiply polynomial regression. We obtain

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$$

to be

$$Y_i = 16.97 + 88.17 x_1 + -18.09 x_1^2 + 0.92 x_1^3$$

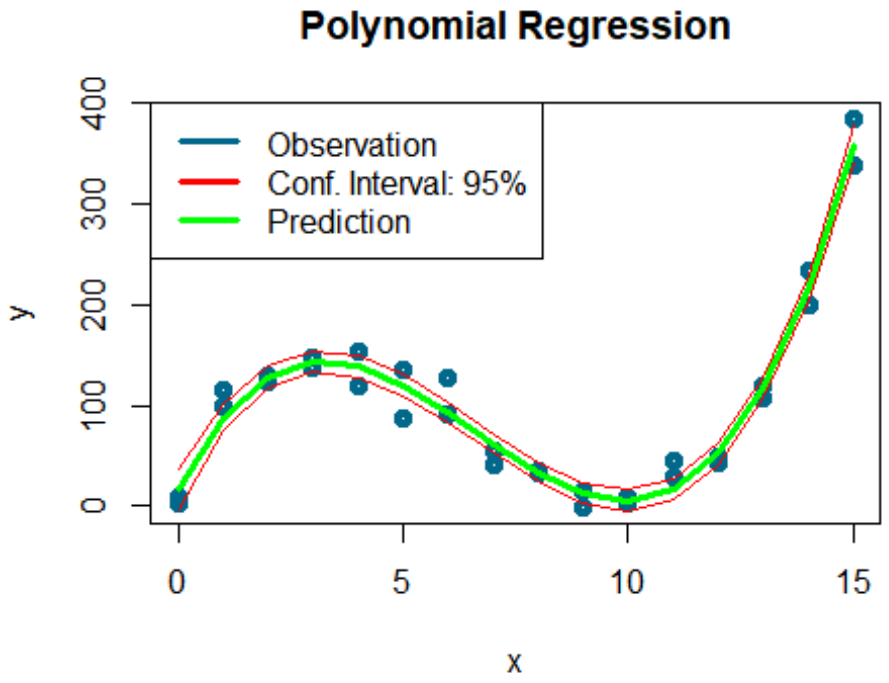
```
# Fit polynomial
model <- lm(y ~ x + I(x^2) + I(x^3))
summary(model)

##
## Call:
## lm(formula = y ~ x + I(x^2) + I(x^3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.387  -9.029  -0.623  10.671  34.244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.97286    9.49724   1.787   0.0847 .
## x            88.16596    5.66718  15.557 2.61e-15 ***
## I(x^2)      -18.08742    0.89527 -20.203  < 2e-16 ***
## I(x^3)        0.91547    0.03918  23.366  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.69 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.9698, Adjusted R-squared:  0.9665
## F-statistic: 299.2 on 3 and 28 DF,  p-value: < 2.2e-16
```

```r
confint(model, level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept)  -2.4813592  36.4270708
## x            76.5572645  99.7746482
## I(x^2)      -19.9213056 -16.2535436
## I(x^3)        0.8352178   0.9957316
```

```r
# Plot
predicted.intervals <- predict(model, data.frame(x), interval = "confidence", level =
0.95)
plot(x, y, main = "Polynomial Regression", col='deepskyblue4', lwd=4)
lines(x,predicted.intervals[,1],col='green',lwd=3)
lines(x,predicted.intervals[,2],col='red',lwd=1)
lines(x,predicted.intervals[,3],col='red',lwd=1)
legend("topleft",
       c("Observation","Conf. Interval: 95%","Prediction"),
       col=c("deepskyblue4","red","green"), lwd=3)
```



STAT GR5205 Homework 3 [100 pts]

Due 8:40am Wednesday, October 24th

Problem 1 [20 pts]

Consider the model

$$Y_i = \mu + \epsilon_i \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The sample mean and sample variance are defined respectively as

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Use Theorem (3.6) on Page 91 to prove that the sample mean $\bar{Y}$ and sample variance S_Y^2 are independent random variables.

Solution

First note that $\mathbf{Y}$ is distributed multivariate normal with mean $\boldsymbol{\mu} = \begin{pmatrix}\mu & \mu & \cdots & \mu\end{pmatrix}^T$ and variance-covariance matrix $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. The sample mean and sums of squares can be respectively written as

$$\bar{Y} = \mathbf{KY} = \begin{pmatrix}\frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n}\end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

and

$$(n-1)S_Y^2 = \mathbf{Y}^T\mathbf{AY} = \mathbf{Y}^T\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}.$$

Also notice that $\mathbf{A}$ is symmetric, i.e.,

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)^T = \mathbf{I}^T - \left(\frac{1}{n}\mathbf{J}\right)^T = \mathbf{I} - \frac{1}{n}\mathbf{J}.$$

Then since,

$$\mathbf{1}^T\mathbf{J} = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} n & n & \cdots & n \end{pmatrix} = n\mathbf{1}^T,$$

we have

$$\begin{aligned} \mathbf{K\Sigma A} &= \mathbf{K}(\sigma^2\mathbf{I})\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = \sigma^2\left(\mathbf{K} - \mathbf{K}\frac{1}{n}\mathbf{J}\right) \\ &= \sigma^2\left(\mathbf{K} - \frac{1}{n}(\mathbf{1})^T\frac{1}{n}\mathbf{J}\right) = \sigma^2\left(\mathbf{K} - \frac{1}{n^2}(\mathbf{1}^T\mathbf{J})\right) \\ &= \sigma^2\left(\mathbf{K} - \frac{1}{n^2}(n\mathbf{1}^T)\right) = \sigma^2\left(\mathbf{K} - \mathbf{K}\right) \\ &= \mathbf{0}. \end{aligned}$$

Thus by Theorem (3.6) in the notes, $\bar{Y}$ and S_Y^2 are independent.

Problem 2 [35 pts]

Consider the *single factor anova model* with three groups. The three groups are drug dose 1, drug dose 2 and control. Let n_1 and $\bar{y}_1$ respectively denote the number of respondents and sample mean response for drug dose 1 group. Let n_2 and $\bar{y}_2$ respectively denote the number of respondents and sample mean response for drug dose 2 group. Let n_3 and $\bar{y}_3$ respectively denote the number of respondents and sample mean response for the control group. Note that $n = n_1 + n_2 + n_3$. The *one-way anova* can be expressed using the *multiple linear regression model*

$$(1) \qquad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \qquad i = 1, 2, \ldots, n \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

$$x_{i1} = \begin{cases} 1 & \text{if drug dose 1} \\ 0 & \text{otherwise} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if drug dose 2} \\ 0 & \text{otherwise} \end{cases}$$

i. Write down the design matrix and response vector describing model (1). **(5 pts)**

Solution

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ \hline 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

ii. Compute $(\mathbf{X}^T\mathbf{X})^{-1}$ and simplify the result. This requires inverting a 3×3 matrix. **(10 pts)**

Solution

First note that

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} n & n_1 & n_2 \\ n_1 & n_1 & 0 \\ n_2 & 0 & n_2 \end{pmatrix}.$$

The determinate of $\mathbf{X}^T\mathbf{X}$ is

$$\begin{aligned} \det(\mathbf{X}^T\mathbf{X}) &= nn_1n_2 + 0 + 0 - 0 - n_2n_1n_2 - n_1n_1n_2 \\ &= (n_1 + n_2 + n_3)n_1n_2 - n_2n_1n_2 - n_1n_1n_2 \\ &= n_1n_2n_3 \end{aligned}$$

The inverse of $\mathbf{X}^T\mathbf{X}$ is

$$
\begin{aligned}
(\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{\det(\mathbf{X}^T\mathbf{X})}\begin{pmatrix} n_1n_2 & -n_1n_2 & -n_1n_2 \\ -n_1n_2 & n_1n_2+n_2n_3 & n_1n_2 \\ -n_1n_2 & n_1n_2 & n_1n_2+n_3n_1 \end{pmatrix} \\
&= \frac{1}{n_1n_2n_3}\begin{pmatrix} n_1n_2 & -n_1n_2 & -n_1n_2 \\ -n_1n_2 & n_1n_2+n_2n_3 & n_1n_2 \\ -n_1n_2 & n_1n_2 & n_1n_2+n_3n_1 \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{n_3} & -\frac{1}{n_3} & -\frac{1}{n_3} \\ -\frac{1}{n_3} & \frac{1}{n_3}+\frac{1}{n_1} & \frac{1}{n_3} \\ -\frac{1}{n_3} & \frac{1}{n_3} & \frac{1}{n_3}+\frac{1}{n_2} \end{pmatrix}
\end{aligned}
$$

iii. Estimate $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \beta_1 & \beta_2\end{pmatrix}^T$ using the least squares equation. **(10 pts)**

Solution

$$
\begin{aligned}
(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} &= \begin{pmatrix} \frac{1}{n_3} & -\frac{1}{n_3} & -\frac{1}{n_3} \\ -\frac{1}{n_3} & \frac{1}{n_3}+\frac{1}{n_1} & \frac{1}{n_3} \\ -\frac{1}{n_3} & \frac{1}{n_3} & \frac{1}{n_3}+\frac{1}{n_2} \end{pmatrix}\mathbf{X}^T\mathbf{Y} \\
&= \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 & \frac{1}{n_3} & \cdots & \frac{1}{n_3} \\ \frac{1}{n_1} & \cdots & \frac{1}{n_1} & 0 & \cdots & 0 & -\frac{1}{n_3} & \cdots & -\frac{1}{n_3} \\ 0 & \cdots & 0 & \frac{1}{n_2} & \cdots & \frac{1}{n_2} & -\frac{1}{n_3} & \cdots & -\frac{1}{n_3} \end{pmatrix}\mathbf{Y} \\
&= \begin{pmatrix} \bar{y}_3 \\ \bar{y}_1-\bar{y}_3 \\ \bar{y}_2-\bar{y}_3 \end{pmatrix}
\end{aligned}
$$

Note:

$$\hat{y}_i = \bar{y}_3 + (\bar{y}_1-\bar{y}_3)x_{i1} + (\bar{y}_2-\bar{y}_3)x_{i2}$$

iv. Write down an expression for the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. **(10 pts)**

Solution

$$MSE(\mathbf{X}^T\mathbf{X})^{-1} = MSE\begin{pmatrix} \frac{1}{n_3} & -\frac{1}{n_3} & -\frac{1}{n_3} \\ -\frac{1}{n_3} & \frac{1}{n_3}+\frac{1}{n_1} & \frac{1}{n_3} \\ -\frac{1}{n_3} & \frac{1}{n_3} & \frac{1}{n_3}+\frac{1}{n_2} \end{pmatrix},$$

where

$$\begin{aligned} MSE &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-3} \\ &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n_1-1)+(n_2-1)+(n_3-1)} \\ &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{(n_1-1)+(n_2-1)+(n_3-1)}, \end{aligned}$$

and s_1^2, s_3^2, s_3^2 are the respective sample variances per group.

For the grader: Give full credit if students only wrote $MSE = SSE/(n-3)$

Problem 3 [25 pts]

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The data consists of variables age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y). For this data set, we skip residual diagnostics but in practice, that should be included in the analysis. The data set `HW3Problem3.txt` is posted on Canvas. Use `R` to perform the following tasks:

i. Regress the rental rates (Y) against all of the covariates; age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4). Write down the estimated linear model. **(2 pts)**

Solution

`R code:`

```
data <- read.table("HW4Problem3.txt",header=TRUE)
model <- lm(RentalRates~Age+OperatingExpense+VacancyRates+SquareFootage,data=data
```

The estimated model is:

$$\hat{y} = 12.2 - 0.142x_1 + 0.282x_2 + 0.619x_3 + 0.000008x_4.$$

ii. What percentage of variation in rental rates is explained by this model? **(2 pts)**

Solution

R code:

```
summary(model)
```

R output:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.220e+01  5.780e-01  21.110  < 2e-16 ***
Age              -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
OperatingExpense  2.820e-01  6.317e-02   4.464 2.75e-05 ***
VacancyRates      6.193e-01  1.087e+00   0.570     0.57
SquareFootage     7.924e-06  1.385e-06   5.722 1.98e-07 ***
---

Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared:  0.5847,Adjusted R-squared:  0.5629
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

Based on the regression summary output above, 58.5% of the variation in rental rates is explained by this model.

iii. Are there any marginal relationships between the response variable and covariates? (Run t-tests on all slope parameters.) **(5 pts)**

Solution

R code:

```
summary(model)
```

R output:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.220e+01  5.780e-01  21.110  < 2e-16 ***
Age              -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
OperatingExpense  2.820e-01  6.317e-02   4.464 2.75e-05 ***
VacancyRates      6.193e-01  1.087e+00   0.570     0.57
SquareFootage     7.924e-06  1.385e-06   5.722 1.98e-07 ***
```

```
---
Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared:  0.5847,Adjusted R-squared:  0.5629
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

Marginal tests

- $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$,
  $t_{calc} = -6.655$, P–value$< 3.89 * 10^{-9} < 0.05$, Reject H_0.
- $H_0 : \beta_2 = 0$ vs. $H_A : \beta_2 \neq 0$,
  $t_{calc} = 4.464$, P–value$= 2.75 * 10^{-5} < 0.05$, Reject H_0.
- $H_0 : \beta_3 = 0$ vs. $H_A : \beta_3 \neq 0$,
  $t_{calc} = 0.570$, P–value$= 0.57 > 0.05$, FTR H_0.
- $H_0 : \beta_4 = 0$ vs. $H_A : \beta_4 \neq 0$,
  $t_{calc} = 5.722$, P–value$= 1.98 * 10^{-7} < 0.05$, Reject H_0.

At 5% significance, after holding all other variables constant, every covariate tested statistically significant except for vacancy rates (X_3). This indicates that there are marginal relationships between age (X_1), operating expenses & taxes (X_2) and total square footage (X_4) against the rental rates (Y). There is not a marginal relationship between vacancy rates (X_3) and rental rates (Y).

iv. Run a F-test to see if there is an *overall relationship* between the rental rates and all of the covariates. **(3 pts)**

Solution

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ versus H_A : *at least one* $\beta_j \neq 0$.

Based on the summary output from Part (ii), $f_{calc} = 26.76$ and P–value$=7.27 * 10^{-14}$. Since the P–value is less than any reasonable level of significance, we reject H_0 and conclude that there is an overall relationship between rental rates (Y) and the covariates.

v. Run a F-test to simultaneously test the slopes for age (X_1) and vacancy rates (X_3). **(5 pts)**

Solution

Full:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Reduced:

$$Y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$$

R code:

```
#Sample size
n=81

# model
model=lm(RentalRates~Age+OperatingExpense+VacancyRates+SquareFootage,data=data)

#reduced model
model.reduced=lm(RentalRates~OperatingExpense+SquareFootage,data=data)

# ANOVA (F-stat)
anova(model.reduced,model)
```

Hypothesis test

$H_0 : \beta_1 = \beta_3 = 0$ versus $H_A : \beta_1 \neq 0$ *or* $\beta_3 \neq 0$.

Based on the R code, $f_{calc} = 23.698$ and P–value=$1.003 * 10^{-8}$. At 5% significance, we reject the null hypothesis and conclude that age (X_1) **or** vacancy rates (X_3) is statistically related to rental rates (Y).

vi. Run a F-test to see if vacancy rates (X_3) is a significant predictor after holding all other variables constant. To perform this test, use the and reduced models. How does this test relate to the summary output from part (ii)? **(3 pts)**

Solution

Full:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Reduced:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$$

R code:

```
#  model
model=lm(RentalRates~Age+OperatingExpense+VacancyRates+SquareFootage,data=data)

# Reduced model
model.reduced=lm(RentalRates~Age+OperatingExpense+SquareFootage,data=data)

# ANOVA (F-stat)
anova(model.reduced,model)
```

Hypothesis test

$H_0 : \beta_3 = 0$ versus $H_A : \beta_3 \neq 0$.

Based on the R code, $f_{calc} = 0.3248$ and P–value=0.5704. At 5% significance, we reject the null hypothesis and conclude that vacancy rates (X_3) is not statistically related to the rental rates (Y), after controlling for all other variables in the model.

Note: From the summary output, $t^2_{calc} = (.570)^2 = 0.325 = f_{calc}$.

vii. The researcher wishes to obtain 95% interval estimates of the mean rental rates for four typical properties specified as follows. Find the four confidence intervals using the Bonferroni procedure. **(5 pts)**

	1	2	3	4
x_1	5.0	6.0	14.0	12.0
x_2	8.25	8.50	11.50	10.25
x_3	0	0.23	0.11	0
x_4	250,000	270,000	300,000	310,000

Solution

Note: $1 - \alpha/K = 1 - 0.05/4 = 0.9875$

R code

```
X.h=data.frame(Age=c(5.0,6.0,14.0,12.0),OperatingExpense=c(8.25,8.50,11.50,10.25),
VacancyRates=c(0,0.23,0.11,0),SquareFootage=c(250000,270000,300000,310000))
predict(model,newdata=X.h,interval="confidence",level=.9875)
```

R output

```
       fit      lwr      upr
1 15.79813 15.08664 16.50962
2 16.02754 15.42391 16.63116
3 15.90072 15.33232 16.46913
4 15.84339 15.18040 16.50638
```

Problem 4 [20 pts]

i. Recall the multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i. \tag{2}$$

For each of the following regression models, indicate whether it can be expressed in the form of (2) by a suitable transformation. To receive full credit, describe the transformation if it exists.

a. $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log(x_{i2}) + \beta_3 x_{i1}^2 + \epsilon_i$ **(3 pts)**

Solution to a

Yes it can expressed in the form of (2).

Let $z_{i1} = x_{i1},\ z_{i2} = \log(x_{i2})$ and $z_{i3} = x_{i1}^3$. Then the transformed model is

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \epsilon_i.$$

b. $Y_i = \epsilon_i \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2\}$ **(3 pts)**

Solution to b

Yes, but only if we assume positive support on the distribution of ϵ. Note the model can expressed in the form of (2).

Let $Y_i^* = \log(Y_i)$, $\epsilon_i^* = \log(\epsilon_i)$, $z_{i1} = x_{i1}$, and $z_{i2} = x_{i1}^2$. Then the transformed model is

$$Y_i^* = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \epsilon_i^*$$

We cannot assume normality on the errors ϵ_i because of the natural-log.

Note: If students claim that this model cannot be transformed because of the $\log \epsilon$ term, still give them full credit.

c. $Y_i = \log(\beta_1 x_{i1}) + \beta_2 x_{i2} + \epsilon_i$ **(3 pts)**

Solution to c

No it cannot be expressed in the form of (2).

Notice that $Y_i = \log(\beta_1) + \log(x_{i1}) + \beta_2 x_{i2}$ is not linear with respect to the "beta's".

d. $Y_i = \beta_0 \exp\{\beta_1 x_{i1}\} + \epsilon_i$ **(3 pts)**

Solution to d

No it cannot be expressed in the form of (2).

e. $Y_i = [1 + \exp\{\beta_0 + \beta_1 x_{i1} + \epsilon_i\}]^{-1}$ **(3 pts)**

Solution to e

Yes it can expressed in the form of (2).

Let $Y_i^* = \log\left(\frac{1}{Y_i} - 1\right)$. Then the transformed model is

$$Y_i^* = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

Note: The response Y can only take on values from $0 < Y < 1$ to satisfy the above model.

ii. Consider the toy data set:

y	2.44	8.36	98.33	115.06	128.91	123.46	148.30	138.10	153.10	119.08
	87.66	134.88	91.71	126.81	40.41	54.94	33.03	35.74	14.99	-1.18
	2.44	8.36	28.33	45.06	48.91	43.46	118.30	108.10	233.10	199.08
	337.66	384.88								
x	0.00	0.00	1.00	1.00	2.00	2.00	3.00	3.00	4.00	4.00
	5.00	5.00	6.00	6.00	7.00	7.00	8.00	8.00	9.00	9.00
	10.00	10.00	11.00	11.00	12.00	12.00	13.00	13.00	14.00	14.00
	15.00	15.00								

The data set is provided in the file `HW3Problem4.txt` on canvas. Use multiple linear regression techniques to fit a polynomial to the above data set. To receive credit, write down the estimated model and create a scatter plot with the estimated curve overlaid on the plot. **(5 pts)**

Solution

After examining the scatter plot, a cubic polynomial looks appropriate for this data set. Consider the model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The transformed model is:

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),$$

where $z_{i1} = x_i$, $z_{i2} = x_i^2$ and $z_{i3} = x_i^3$.

R code:

```
model <- lm(y x+I(x^2)+I(x^3),data=data)
```

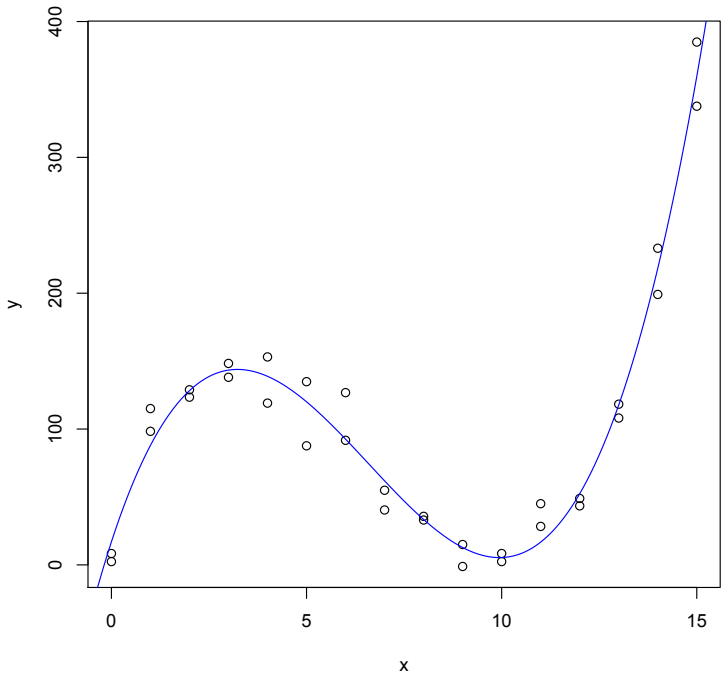
Or

```
model <- lm(y poly(x,3,raw=T),data=data)
```

The estimated cubic model is:

$$\hat{y} = 16.9729 + 88.1660x - 18.0874x^2 + 0.9155x^3.$$

Figure 1: Scatter plot with estimated cubic model.



R code to plot the curve:

```
plot(data$x,data$y,xlab="x",ylab="y")
x.list <- seq(0,16,by=.1)
y.pred <- predict(model,newdata=data.frame(x=x.list))
lines(x.list,y.pred,col=4)
```

STAT GR5205 Homework 4 [100 pts]

Due 8:40am Wednesday, November 7th

Problem 1

The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, x is the number of copiers serviced and Y is the total number of minutes spent by the service person. The data set `HW4Problem1.txt` is posted on canvas.

Use `R` to perform the following tasks:

i. Obtain the estimated regression function.

ii. Create a scatterplot of the data set with the line of best fit overlaid on the graph. Create a QQ plot of the studentized deleted residuals, histogram of the studentized deleted residuals, line plot of the studentized deleted residuals, plot the studentized deleted residuals verses predicted values $\hat{y}$, and studentized residuals verses predictor variable x. Based on the plots, discuss whether any of the regression assumptions have been violated. In your descriptions, relate your explanations to the relevant plots.

iii. Perform a Box-Cox procedure on the data set. What is the estimated value of λ? Is it necessary to perform this transformation on the response variable? Briefly explain your reasoning.

Problem 2

Sixteen of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. For this data set; x is the elapsed time in hours and Y is hardness in Brinell units. Use `R` to run a F- lack-of-fit test to see if a linear relationship is appropriate for this data set. The data set `1_22.txt` is posted on Canvas.

Problem 3

Consider the *non-constant variance* linear model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \tag{1}$$

with

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_i^2), \quad i = 1, \ldots, n.$$

Define the reciprocal of the variance σ_i^2 as the weight w_i:

$$w_i = \frac{1}{\sigma_i^2}$$

and let

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{pmatrix}.$$

We can estimate the *non-constant variance* model by minimizing the objective function

$$Q_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots \beta_{p-1} x_{i,p-1})^2 \tag{2}$$

Task: Derive the weighted least squares equation

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \tag{3}$$

Problem 4

Observations on Y are to be taken when $x = 10, 20, 30, 40$, and 50, respectively. The true regression function is $E[Y] = 20 + 10x$. The error terms are independent and normally distributed with $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = .8x$.

i. Generate a random Y observation for each x level and calculate both the ordinary and weighted least squares estimates of the regression coefficient β_1 in the simple linear regression function.

ii. Repeat part (a) 10,000 times, generating new random numbers each time.

iii. Calculate the mean and variance of the 10,000 ordinary least squares estimates of β_1 and do the same for the 10,000 weighted least squares estimates.

iv. Do both the ordinary least squares and weighted least squares estimators appear to be unbiased? Explain. Which estimator appears to be more precise here? Comment.

GR5205 Homework 4

Yiqiao Yin [YY2502]

Table of Contents

PROBLEM 1

```
# Upload data
setwd("E:/Course/CU Stats/STATS GR5205 - Linear Regression Model/3. Homework/HW4")
data <- read.delim("HW4Problem1.txt", sep = "")
n <- nrow(data)
```

(i) Linear Regression

We obtain the results of linear regression model through the following R code. From output, we conclude that we have linear regression model, $Y = -0.58 + 15.04x + e$.

```
linear.model <- lm(data$y~data$x, data = data)
summary(linear.model)
```

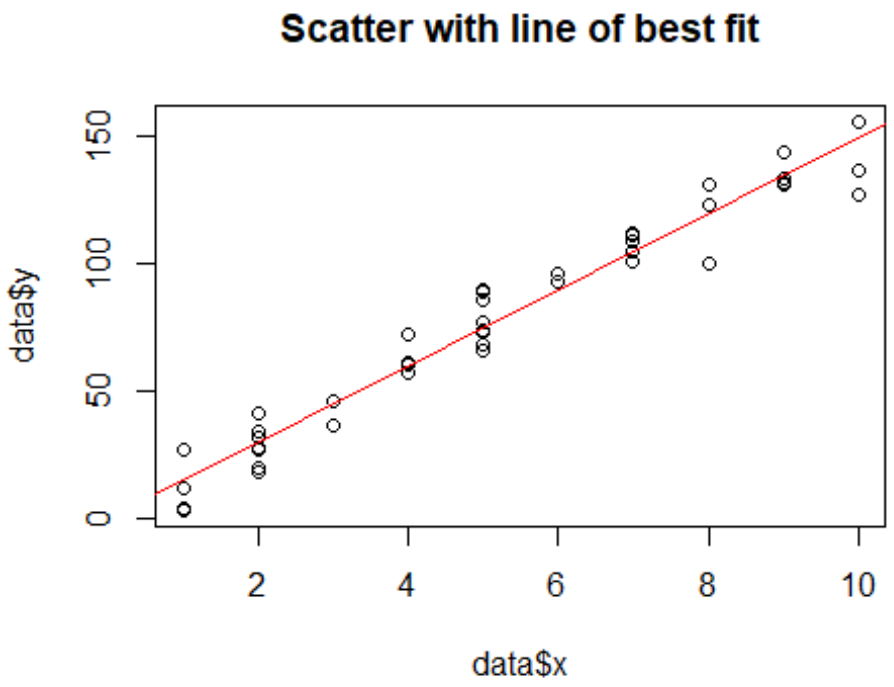
```
##
## Call:
## lm(formula = data$y ~ data$x, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## data$x       15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

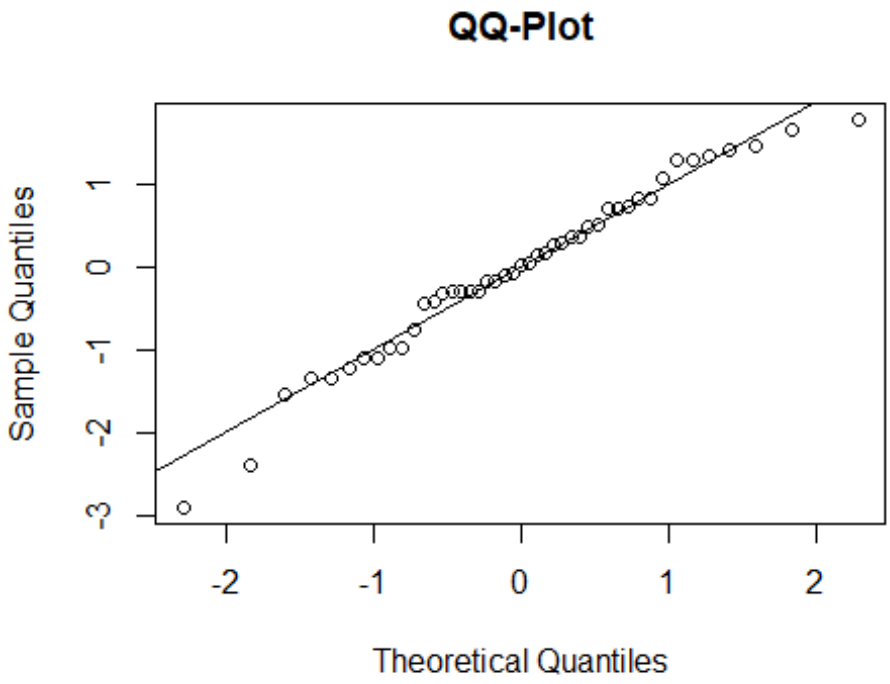
(ii) Scatter Plot

We use the following *R* code to generate the graphs requried from the problem.

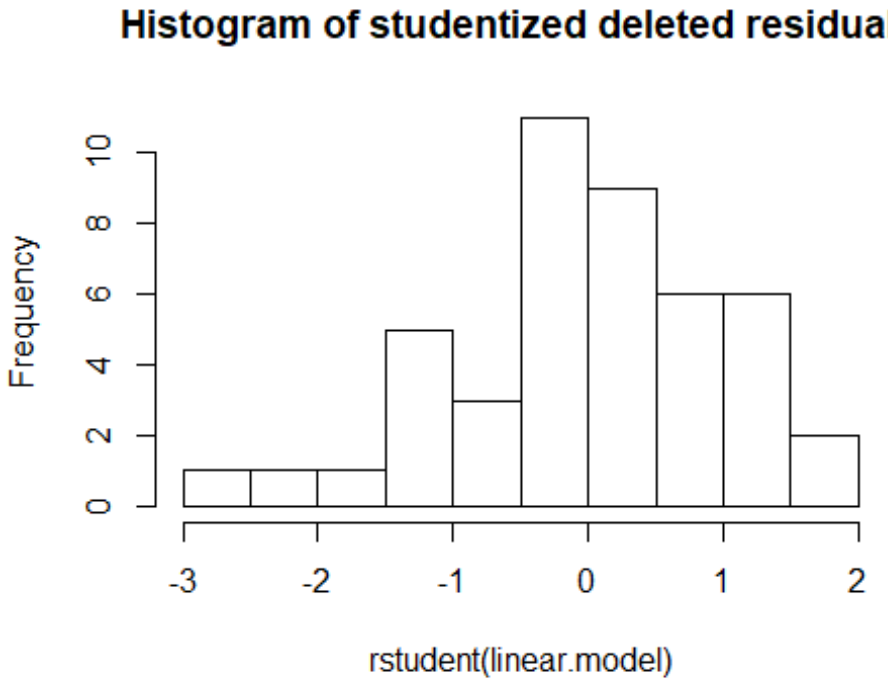
```
# Scatter plot of data with line of best fit
plot(data$x, data$y, main = "Scatter with line of best fit")
abline(linear.model, col = "red")
```



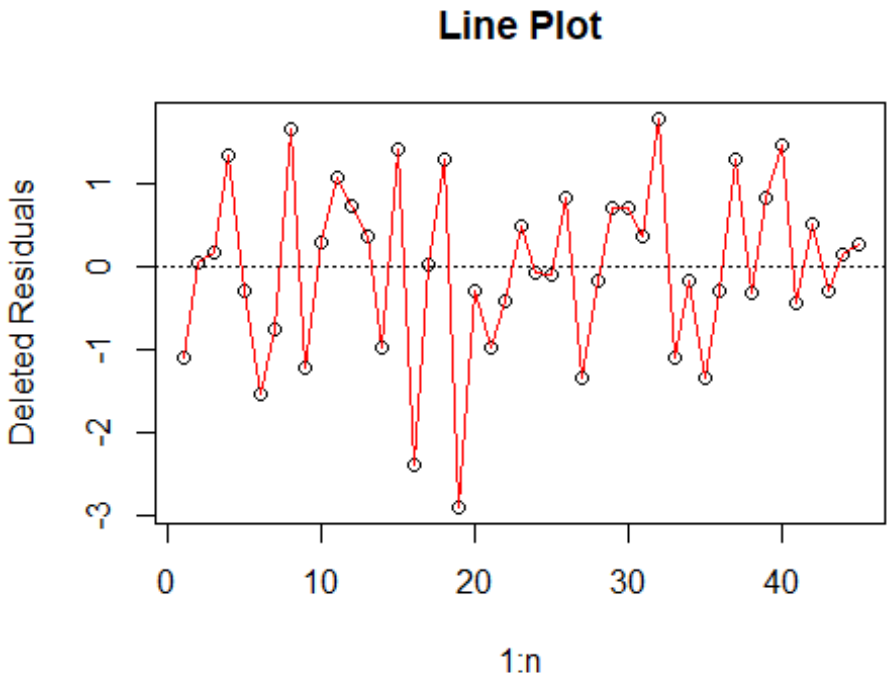

```
# QQ plot of the studentized deleted residuals
qqnorm(rstudent(linear.model), main = "QQ-Plot"); abline(0,1)
```



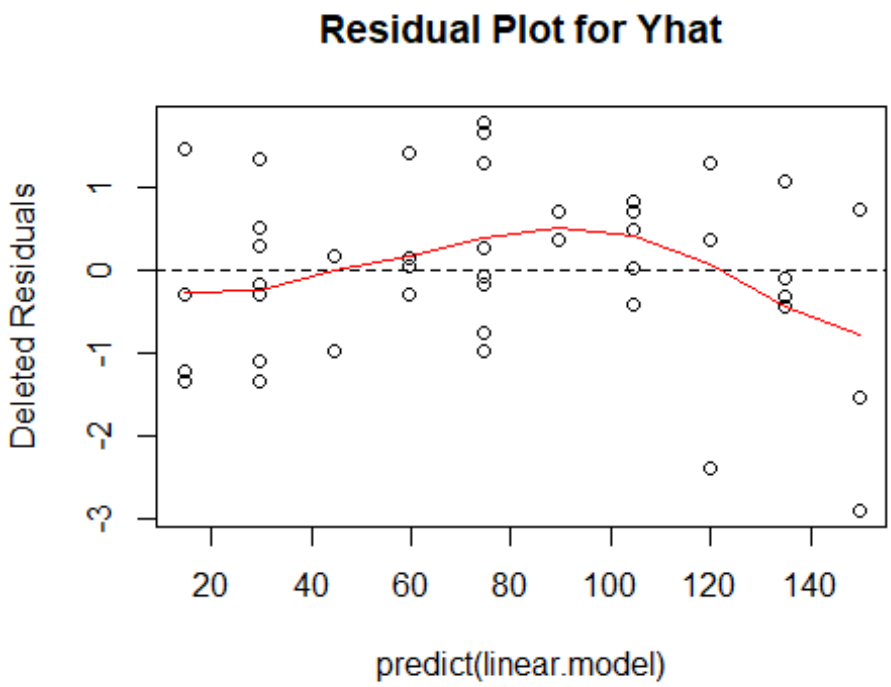

```r
# Histogram of the studentized deleted residuals
hist(rstudent(linear.model), main = "Histogram of studentized deleted residuals")
```



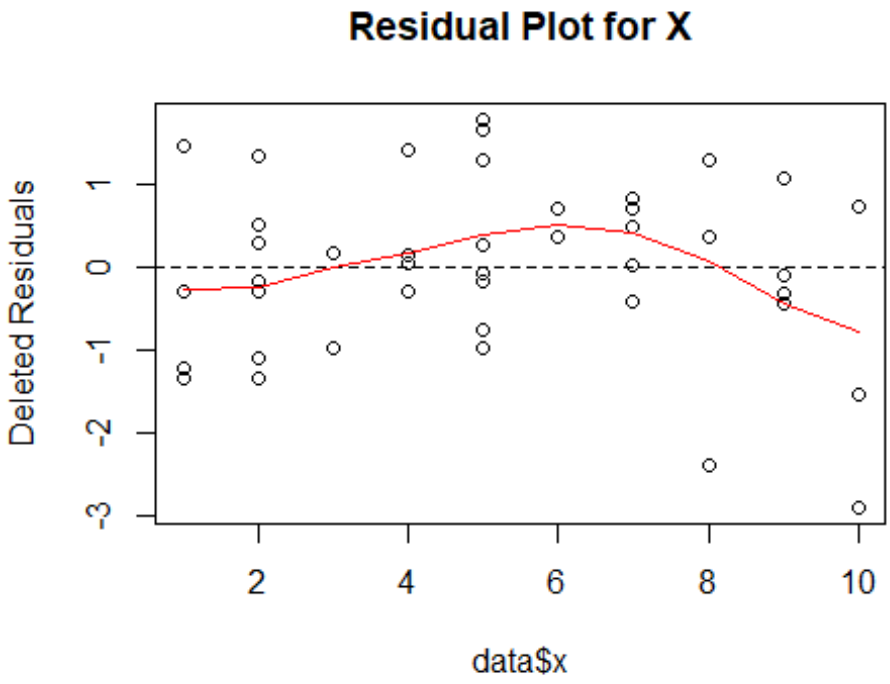
```r
# Line plot of the studentized deleted residuals
plot(1:n, rstudent(linear.model), main = "Line Plot", ylab = "Deleted Residuals"); ab
line(h=0, lty=3);
lines(1:n, rstudent(linear.model), col = 2)
```




```r
# Studentized deleted residuals vs predicted values y^#
plot(predict(linear.model), rstudent(linear.model), main = "Residual Plot for Yhat",
ylab = "Deleted Residuals");
abline(h=0, lty=2); lines(supsmu(predict(linear.model), rstudent(linear.model)), col
= 2)
```




```r
# Studentized deleted residuals vs x#
plot(data$x, rstudent(linear.model), main = "Residual Plot for X", ylab = "Deleted Re
siduals");
abline(h=0, lty=2); lines(supsmu(data$x, rstudent(linear.model)), col = 2)
```



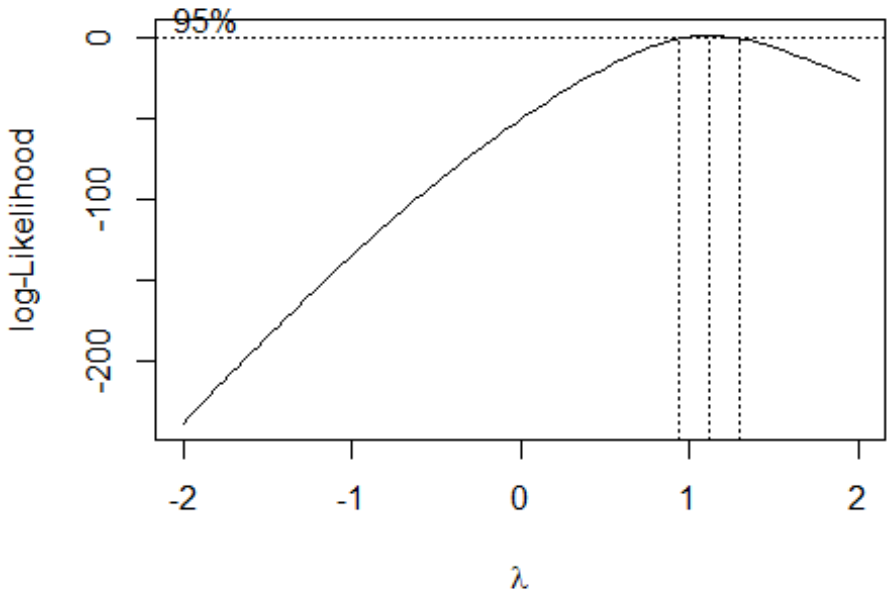
We conclude the following:

- Linearity assumption: It is not been violated. We can tell the results from the qq plot.
- Normality assumption: This has been violated since we observe from histogram of the qq plot that the plot skewed to the left.
- Constant variance assumption: It's been violated for a little. We do not observe constant variance from the plots of studentized deleted residuals vs $\hat{Y}$ and the covariance of x.
- Independence assumption: It's been violated for a little. The dependence of error is illustrated from the line plot.

(iii) Box-Cox

We can use R output below to find the estimated $\lambda = 1.11$. We can transform $Y' = 1.11Y$. However, the results will probability be similar.

```
library(MASS)
y <- data$y
x <- data$x
bac.box = boxcox(y~x)
```




```
bac.lambda = bac.box$x[which(bac.box$y==max(bac.box$y))]
bac.lambda

## [1] 1.111111
```

PROBLEM 2

Consider the following data

```
# Upload data
setwd("E:/Course/CU Stats/STATS GR5205 - Linear Regression Model/3. Homework/HW4")
data <- read.delim("1_22.txt", sep = "")
```

Let us conduct F-lack-of-fitness test to this data set. Use hypothesis:

$$H_0: E[Y] = \beta_0 + \beta_1 x \text{ versus } H_A: E[Y] \neq \beta_0 + \beta_1 x$$

```
# Define explanatory variable and response variable
x <- data$X16.0
y <- data$X199.0
n <- length(y)
c <- length(levels(as.factor(x)))

# Compute F-stat
SSE.R <- anova(lm(y~x))[[2]][2]
fac.x <- factor(x)
SSE.F <- anova(lm(y~fac.x))[[2]][2]
f.calc <- ((SSE.R - SSE.F)/(c-2))/(SSE.F/(n-2))
```

```
# Compute p-value
f.calc

## [1] 0.6838546

1 - pf(f.calc, c-2, n-2)

## [1] 0.5219307
```

Fail to reject null hypothesis. We conclude that there is not sufficient evidence to say that $E[Y]$ cannot be explained by $\beta_0 + \beta_1 x$.

PROBLEM 3

Derive the weighted least squares equation.

.

Start with the last esquares equation and we can derive the following.

$$\begin{aligned} Q_w(b) &= \sum_{i=1}^{n} w_i \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i,p-1}\right)^2 \\ &= (Y - xb)^T w (Y - xb) \\ &= Y^T wY - Y^T wxb - (xb)^T wY + (xb)^T wxb \\ &= Y^T wY - Y^T wxb - b^T x^T wT + b^T x^T wxb \end{aligned}$$

Next, we can take derivatives

$$\begin{aligned} \frac{\partial Y^T wxb}{\partial b} &= Y^T wx \\ \frac{\partial b^T x^T qwY}{\partial b} &= (x^T wY)^T = Y^T w^T x \\ \frac{\partial b^T x^T wxb}{\partial b} &= b^T (x^T wx) + b^T (x^T wx)^T = b^T (x^T wx) + b^T (x^T w^T x) \\ \frac{\partial Q_w(b)}{\partial b} &= 0 - Y^T wx - Y^T w^T x + b^T (x^T wx) + b^T (x^T w^T x) \overset{\text{Set}}{=} 0 \end{aligned}$$

and then we can solve for

$$\begin{aligned} b^T (x^T wx + x^T wx) &= Y^T wx + Y^T wx \\ (x^T wx + x^T w^T x)^T &= (Y^T wx + Y^T w^T x)^T \\ (x^T w^T x + x^T wx) b &= (x^T w^T Y + x^T wY) \end{aligned}$$

and since we have definition for weights

$$w = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix} w^T = \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{bmatrix}$$

thus we have

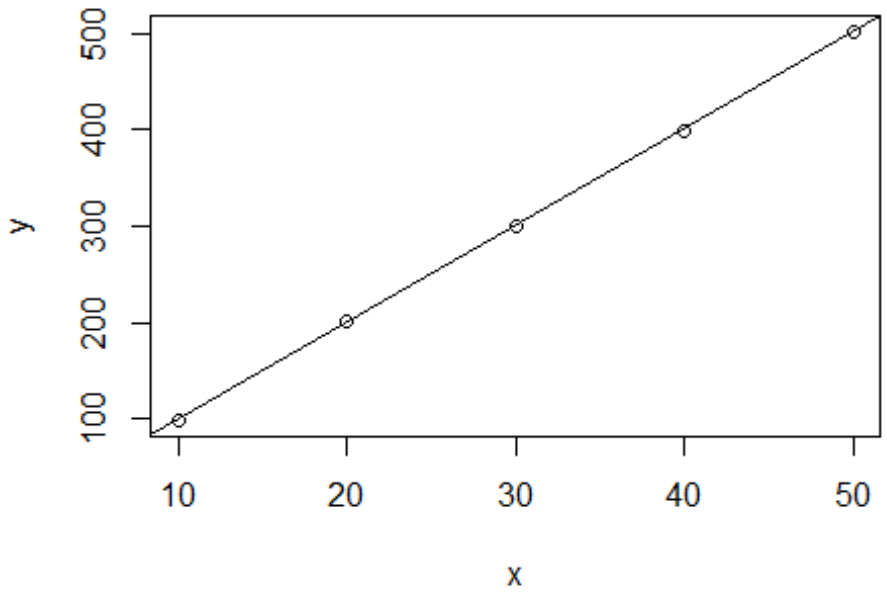
$$\begin{aligned} 2(x^T wx) b &= 2(x^T wY) \\ b &= (x^T wx)^{-1} (x^T wY) \end{aligned}$$

PROBLEM 4

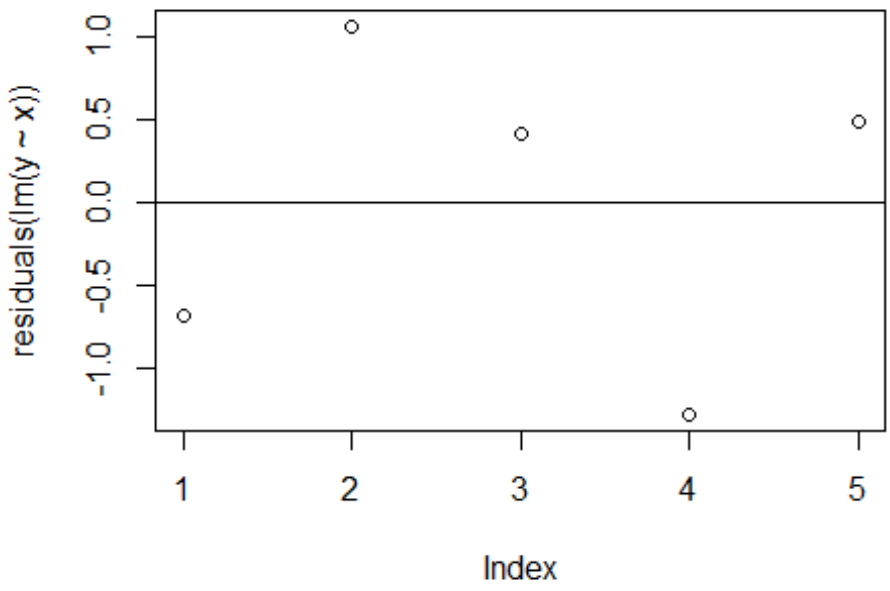
(i) Generate Response

Consider data defined in the *R* code below. We define x as well as y. Next, we use *R* to generate the linear model and we find the weights of ordinary β_1 and also weighted β_1.

```
# Create data
x <- c(10, 20, 30, 40, 50)
y <- 10*x + rnorm(length(x), mean = 0, sd = 1)

# Plot
plot(x,y); abline(lm(y~x))
```




```
plot(residuals(lm(y~x))); abline(a=0, b=0)
```




```
# Model
unweighted.model <- lm(y~x)
e <- residuals(unweighted.model)
wts <- 1/(0.8*x) # In this problem we know variance
wts <- 1/fitted(lm(abs(residuals(unweighted.model)) ~ x))^2 # In practice we have to
estimate the variance
weighted.lm <- lm(y~x, weights = wts)
ordinary.b1 <- unweighted.model$coefficients[2];
weighted.b1 <- weighted.lm$coefficients[2]
ordinary.b1; weighted.b1

##        x
## 10.03457

##        x
## 10.03426
```

(ii) Generate 10000 Times

This question I write a function defined as *MC* in order to use it in the next problem.

```
# Define function for regenerate purpose
MC <- function() {
  # Create data
  x <- c(10, 20, 30, 40, 50)
  y <- 20 + 10*x + rnorm(length(x), mean = 0, sd = sqrt(0.8*x))

  # Plot
  #plot(x,y); abline(lm(y~x))
  #plot(residuals(lm(y~x))); abline(a=0, b=0)
```

```
  # Model
  unweighted.model <- lm(y~x)
  e <- residuals(unweighted.model)
  wts <- 1/(0.8*x)
  #wts <- 1/fitted(lm(abs(residuals(unweighted.model)) ~ x))^2
  weighted.lm <- lm(y~x, weights = wts)
  ordinary.b1 <- unweighted.model$coefficients[2];
  weighted.b1 <- weighted.lm$coefficients[2]

  # Output
  return(list(
    Ordinary.b1 = ordinary.b1,
    Weighted.b1 = weighted.b1
  ))
}
```

(iii) Calculate Mean and Variance

This problem we use function *replicate* on the function defined in (ii) many times. This *replicate* function is similar to *for* loop but more efficient.

```
# Run
MC.Result <- data.frame(t(replicate(10000, unlist(MC()))))
apply(MC.Result, 2, mean, na.rm = TRUE)

## Ordinary.b1.x Weighted.b1.x
##      9.997948      9.999103

apply(MC.Result, 2, var, na.rm = TRUE)

## Ordinary.b1.x Weighted.b1.x
##    0.02339938    0.01912524
```

(iv) Are the results unbiased?

The results are unbiased. Observe that the weighted model gives us a beta coefficient that is at a similar value but with much less, e.g. half of, the standard deviation of that of the ordinary model. In other words, weighted model is somehow producing less variance when it comes to prediction which leads to the art of model selection. If we want less variance, we should pick weighted model.

STAT GR5205 Homework 4 [100 pts]

KEY

Problem 1 [30 pts]

The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, x is the number of copiers serviced and Y is the total number of minutes spent by the service person. The data set `HW4Problem1.txt` is posted on canvas.

Use `R` to perform the following tasks:

i. Obtain the estimated regression function. **(3 pts)**

`R` output:

```
Coefficients:
(Intercept)            x
    -0.5802      15.0352
```

The line is: $\hat{y} = -0.58 + 15.04x$

ii. Create a scatterplot of the data set with the line of best fit overlaid on the graph. Create a QQ plot of the sudentized residuals, histogram of the sudentized residuals, line plot of the studentized residuals, plot the sudentized residuals verses predicted values $\hat{y}$, and sudentized residuals verses predictor variable x. Based on the plots, discuss whether any of the regression assumptions have been violated. In your descriptions, relate your explanations to the relevant plots. **Note:** For this question, you can also use studentized deleted residuals. **(21 pts)**

`R` code:

```
#Plot 1
#Plot the simulated model with the LOBF

plot(x,y,main="Scatter Plot",xlab="Number of copiers serviced",ylab="Time spent")
abline(lm(y~x),col=4)

#Plot 2
#QQ plot
qqnorm(rstudent(model),main="QQ-Plot")
```

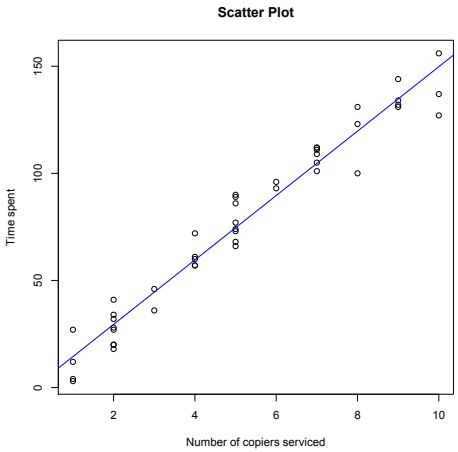
```
abline(a=0,b=1,lty=3)

#Plot 3
#Box plot
hist(rstudent(model),main="Box Plot",ylab="Studentized Deleted")

# Plot 4
#Line plot to look at temporal dependence. (Serial correlation or other patterns)
plot(1:n,rstudent(model),main="Line Plot",ylab="Studentized Deleted",xlab="i")
abline(h=0,lty=3)
lines(1:n,rstudent(model),col=2)

# Plot 5
#Studentized deleted residuals verses predected values
plot(predict(model),rstudent(model),main="Residual Plot",xlab="Y-hat",ylab="Studentized Deleted"
abline(h=0,lty=2)
lines(supsmu(predict(model),rstudent(model)),col=2)

# Plot 6
#Studentized deleted residuals verses covariate x
plot(x,rstudent(model),main="Residual Plot",xlab="Number of copiers serviced",ylab="Studentized
abline(h=0,lty=2)
lines(supsmu(x,rstudent(model)),col=2)
```

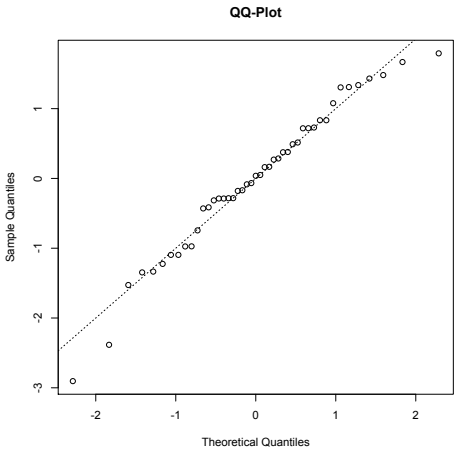
Interpretation

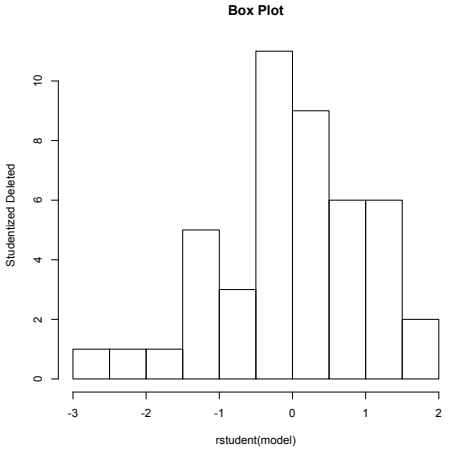
- The scatter plot shows a slight curvature to the trend. There may be an issue with linearity. To see this better, we will look at more residual plots. (**This is very subtle. Students should not get marked off if they missed this.**)
- The QQ-plot shows slight skewness indicating normality may be an issue. (**This is open for interpretation. Do not mark off if students missed this.**)
- The histogram also shows this slight left skewness. (**This is open for interpretation. Do not mark off if students missed this.**)
- The line plot looks good.
- The studentized deleted residuals plot against x shows the presence of non-linearity.
- The studentized deleted residuals plot against $\hat{y}$ also shows the presence of non-linearity.

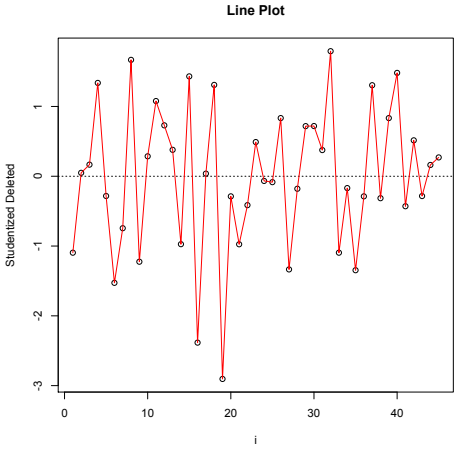
Plots:

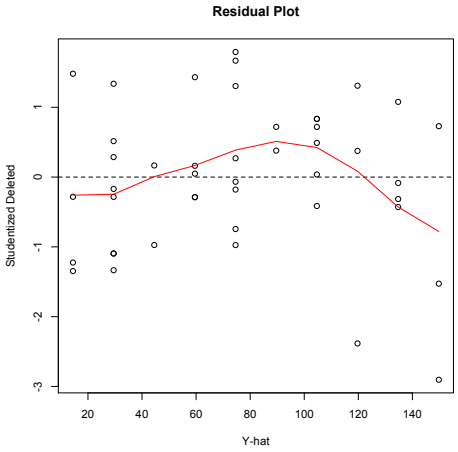
Also displayed on the next page.

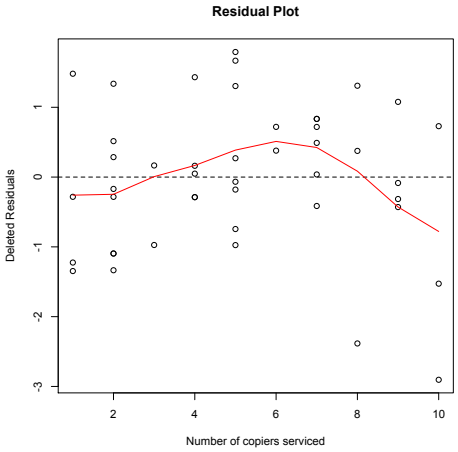

Scatter Plot
Time spent
Number of copiers serviced


QQ-Plot
Sample Quantiles
Theoretical Quantiles


Box Plot
Studentized Deleted
rstudent(model)


Line Plot
Studentized Deleted
i


Residual Plot
Studentized Deleted
Y-hat

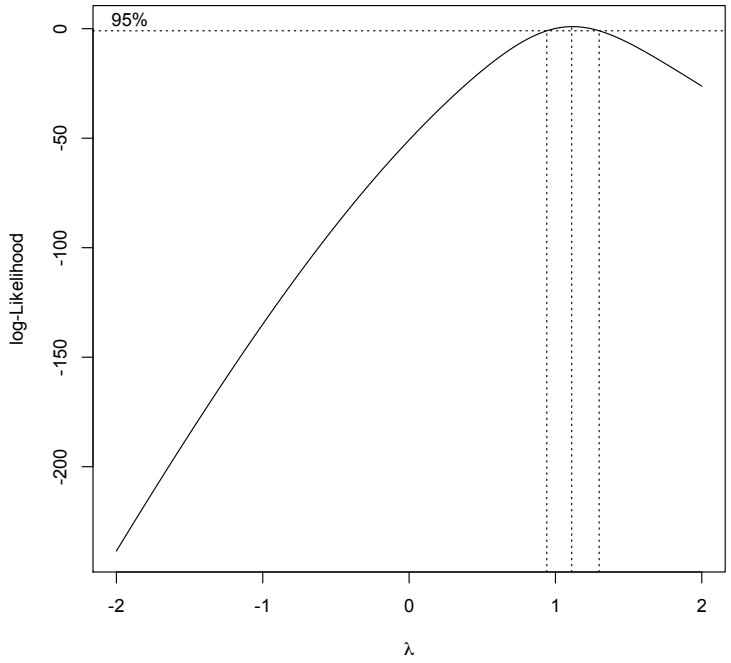

Residual Plot
Deleted Residuals
Number of copiers serviced

iii. Perform a Box-Cox procedure on the data set. What is the estimated value of λ? Is it necessary to perform this transformation on the response variable? Briefly explain your reasoning. **(6 pts)**

R code:

```
bac.box = boxcox(y~x)
bac.lambda = bac.box$x[which(bac.box$y==max(bac.box$y))]
bac.lambda
```

Figure 1

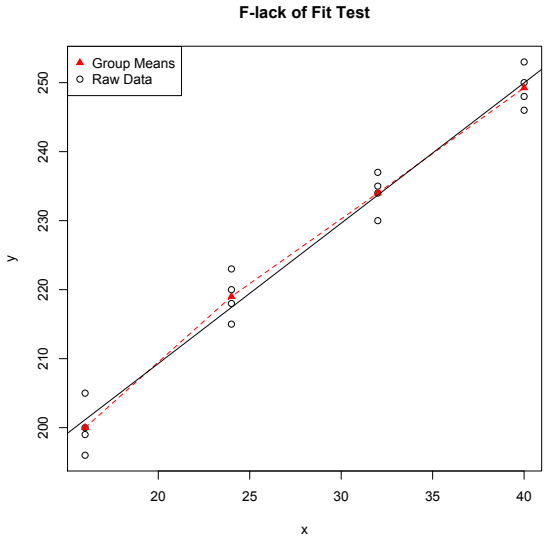


Interpretation

From the Box-Cox procedure, the estimated power transformation on the response Y is $\hat{\lambda} = 1.11$, i.e., $Y^{1.11}$. Based on the confidence interval displayed in Figure 1, the response variable Y does not need to be transformed because 1 is in the 95% confidence interval.

Problem 2 [20 pts]

Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. For this data set; x is the elapsed time in hours and Y is hardness in Brinell units. Use `R` to run a F- lack-of-fit test to see if a linear relationship is appropriate for this data set. The data set `1_22.txt` is posted on Canvas.



Note for the grader: The plot is not required.

R code:

```
n=16
c=4
sse.R=anova(lm(y~x))[[2]][2]
fac.x=factor(x)
sse.F=anova(lm(y~fac.x))[[2]][2]
f.calc=((sse.R-sse.F)/(c-2))/(sse.F/(n-c))
1-pf(f.calc,c-2,n-c)
```

Solution

Consider testing the null alternative pair $H_0 : E[Y] = \beta_0 + \beta_1 x$ versus $H_A : E[Y] \neq \beta_0 + \beta_1 x$. The F-Statistic and P–value for this test are respectively 0.8237 and 0.4622. Thus at any reasonable level of significance, we fail to reject the null hypothesis and conclude that a linear relationship is appropriate for this application.

Problem 3 [25 pts]

Consider the *non-constant variance* linear model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \tag{1}$$

with

$$\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_i^2), \quad i = 1, \ldots, n.$$

Define the reciprocal of the variance σ_i^2 as the weight w_i:

$$w_i = \frac{1}{\sigma_i^2}$$

and let

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{pmatrix}.$$

We can estimate the *non-constant variance* model by minimizing the objective function

$$Q_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots \beta_{p-1} x_{i,p-1})^2 \tag{2}$$

Task: Derive the weighted least squares equation

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y} \tag{3}$$

Solution

Consider the objective function $Q_w(\mathbf{b})$ defined by

$$\begin{aligned} Q_w(\mathbf{b}) &= Q_w(b_0, b_1, \ldots, b_{p-1}) \\ &= \sum_{i=1}^{n} w_i (y_i - (b_0 + b_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}))^2. \end{aligned}$$

We want to minimize $Q_w(b_0, b_1, \ldots, b_{p-1})$ with respect to $b_0, b_1, \ldots, b_{p-1}$, equivalently minimize the objective function with respect to the vector

$$\mathbf{b} = \begin{pmatrix} b_0 & b_1 & \ldots, b_{p-1} \end{pmatrix}^T.$$

Notice the objective function can be expressed as

$$\begin{aligned} Q_w(\mathbf{b}) &= (\mathbf{Y}-\mathbf{X}\mathbf{b})^T\mathbf{W}(\mathbf{Y}-\mathbf{X}\mathbf{b}) \\ &= \mathbf{Y}^T\mathbf{W}\mathbf{Y} - \mathbf{Y}^T\mathbf{W}\mathbf{X}\mathbf{b} - (\mathbf{X}\mathbf{b})^T\mathbf{W}\mathbf{Y} + (\mathbf{X}\mathbf{b})^T\mathbf{W}\mathbf{X}\mathbf{b} \\ &= \mathbf{Y}^T\mathbf{W}\mathbf{Y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{Y} + \mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b}. \end{aligned}$$

Taking derivatives with respect to $\mathbf{b}$ yields

$$\begin{aligned} \frac{\partial}{\partial \mathbf{b}} Q_w &= \frac{\partial}{\partial \mathbf{b}}\mathbf{Y}^T\mathbf{W}\mathbf{Y} - 2\frac{\partial}{\partial \mathbf{b}}\mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{Y} + \frac{\partial}{\partial \mathbf{b}}\mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b} \\ &= -2\mathbf{b}^T\mathbf{X}^T\mathbf{W}\mathbf{Y} + \mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^T\mathbf{b} \\ &= -2\mathbf{X}^T\mathbf{W}\mathbf{Y} + \mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{b} + \mathbf{X}^T\mathbf{W}^T\mathbf{X}\mathbf{b} \\ &= -2\mathbf{X}^T\mathbf{W}\mathbf{Y} + 2\mathbf{X}^T\mathbf{W}^T\mathbf{X}\mathbf{b} \end{aligned}$$

Equating the partial derivative to zero and solving for $\mathbf{b}$ yields the least squares solution. The work follows below:

$$0 = -2\mathbf{X}^T\mathbf{W}\mathbf{Y} + 2\mathbf{X}^T\mathbf{W}^T\mathbf{X}\mathbf{b}.$$

$$\mathbf{X}^T\mathbf{W}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{W}\mathbf{Y}$$

Notice that $\mathbf{X}^T\mathbf{W}^T\mathbf{X}$ is full rank, which yields the solution

$$(\mathbf{X}^T\mathbf{W}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^T\mathbf{X}\mathbf{b} = (\mathbf{X}^T\mathbf{W}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{W}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$

Note: You should also show that the second order derivative is positive definite. Don't mark off if students did not show this step.

Problem 4 [25 pts]

Observations on Y are to be taken when $x = 10, 20, 30, 40,$ and 50, respectively. The true regression function is $E[Y] = 20 + 10x$. The error terms are independent and normally distributed with $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = .8x$. **For the grader: Do not mark off if students did not set the seed to 0. Students will have different simulated values.**

1. Generate a random Y observation for each x level and calculate both the ordinary and weighted least squares estimates of the regression coefficient β_1 in the simple linear regression function. **(5 pts)**

   R code

```
# Set seed
set.seed(0)

# X data
x <- c(10,20,30,40,50)

# Variance list
var <- .8*x

# Random simulated data set
Y <- 20+10*x+rnorm(5,sd=sqrt(var))

# Normal LS
lm(Y~x)

# Weighted LS
lm(Y~x,weights=1/var)
```

   R output

```
> # Normal LS

Coefficients:
(Intercept)            x
       8.29        10.43

> # Weighted LS

Coefficients:
```

```
(Intercept)           x
      11.00       10.34
```

2. Repeat part (a) 10000 times, generating new random numbers each time. **(5 pts)**

   R code

```
set.seed(0)
normal.LS <- NULL
weighted.LS <- NULL
for (i in 1:1000) {
  # Random simulated data set
  Y <- 20+10*x+rnorm(5,sd=sqrt(var))

  # Normal LS
  normal.LS[i] <- lm(Y~x)$coefficients[2]

  # Weighted LS
  weighted.LS[i] <- lm(Y~x,weights=1/var)$coefficients[2]

}
```

3. Calculate the mean and variance of the 200 ordinary least squares estimates of β_1 and do the same for the 200 weighted least squares estimates. **(5 pts)**

```
> # Normal LS
> mean(normal.LS)
[1] 9.995553
> var(normal.LS)
[1] 0.0241369
>
> # Weighted  Normal LS
> mean(weighted.LS)
[1] 9.994609
> var(weighted.LS)
[1] 0.02036719
```

4. Do both the ordinary least squares and weighted least squares estimators appear to be unbiased? Explain. Which estimator appears to be more precise here? Comment. **(5 pts)**

   Solution

   Based on the simulation results, both estimators (normal and weighted) appear to

unbiased because the means are both very close to the true slope $\beta_1 = 10$. The weighted LS is more precise because its variance is smaller. This correct specification of the model induces a more precise estimate of the true slope.

STAT GR5205 Homework 5 [100 pts]

Due 8:40am Wednesday, November 28

Problem 1

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The data consists of variables age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y). The data set `HW3Problem3.txt` is posted on canvas. Use `R` to perform the following exercises. Using type I sums of squares (F-stat), test if age (X_1) is a significant predictor, after holding operating expenses & taxes (X_2), vacancy rates (X_3) and total square footage (X_4) constant.

Problem 2

Show that: $SSR(x_1, x_2, x_3, x_4) = SSR(x_1) + SSR(x_2, x_3|x_1) + SSR(x_4|x_1, x_2, x_3)$.

Problem 3

Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

Assuming that the sample correlation between x_1 and x_2 is zero, i.e.,

$$\frac{1}{(n-1)s_1 s_2} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0,$$

show that

$$SSR(x_2|x_1) = SSR(x_2).$$

Problem 4

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Let Y denote sales (in thousand cases). The assistant was instructed to study the influence variables X_1 and X_2 have on sales Y. The data set `CosmeticsSales.txt` is posted on Canvas.

Use `R` to perform the following tasks:

i. Run the simple linear regression $Y \sim X_1$. Test if expenditures for point-of-sale displays in beauty salons and department stores (X_1) statistically influences sales (Y).

ii. Run the simple linear regression $Y \sim X_2$. Test if expenditures for local media advertising (X_2) statistically influences sales (Y).

iii. Now run the the full regression $Y \sim X_1 + X_2 + X_3$. Perform *marginal* t-tests to see if X_1 statistically influences sales (Y) and if X_2 statistically influences sales (Y), after controlling for the variance of X_3. Briefly compare the results to Parts i. and ii. and comment on any discrepancies. **Note: No need for Bonferroni.**

iv. You should have noticed some discrepancies in Part iii. Explain why these discrepancies are occurring and provide graphical or exploratory evidence to complement your argument.

Problem 5

Consider the *standardized regression model*

$$Y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \epsilon_i^*.$$

The variables Y_i^*, x_{i1}^* and x_{21}^*∗ are *standardized* versions of Y_i, x_{i1} and x_{i2}, i.e.,

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right), \quad x_{i1}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{i1} - \bar{x}_1}{s_1}\right), \quad x_{i2}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{i2} - \bar{x}_1}{s_2}\right)$$

Task: Using the least squares equation, derive the estimators

$$\hat{\beta}_1^* = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2} \quad \text{and} \quad \hat{\beta}_2^* = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2}.$$

Note that the standardized regression model does relate the regular regression model. The description follows below:

Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

It is easy to show that the parameters β_1^*, β_2^* in the standardized regression model and the original parameters $\beta_0, \beta_1, \beta_2$ are related as follows:

$$\beta_1 = \frac{s_Y}{s_1}\beta_1^*, \quad \beta_2 = \frac{s_Y}{s_2}\beta_2^*, \quad \text{and} \quad \beta_0 = \bar{Y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2.$$

Reference pages 273 to 278 of the textbook for further details.

GR5205 Homework 5

Yiqiao Yin [YY2502]

Table of Contents

PROBLEM 1

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas.

Let us do the following

- Upload data;
- Define explanatory variables and response variable;
- Last we fit a linear model and we make conclusions using type I sums of squares to test if Age is a significant predictor, after holding the other variables constant.

```
# Upload data
setwd("E:/Course/CU Stats/STATS GR5205 - Linear Regression Model/3. Homework/HW5")
data <- read.delim("HW3Problem3.txt", sep = "")
n <- nrow(data)

# Introduce data
y <- data$RentalRates
x1 <- data$Age
x2 <- data$OperatingExpense
x3 <- data$VacancyRates
x4 <- data$SquareFootage
```

```r
# Linear model
lin.model <- lm(y~x1+x2+x3+x4)
anova(lin.model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 14.819  14.819 11.4649  0.001125 **
## x2         1 72.802  72.802 56.3262 9.699e-11 ***
## x3         1  8.381   8.381  6.4846  0.012904 *
## x4         1 42.325  42.325 32.7464 1.976e-07 ***
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe the F-stat and p-value for X_1, Age, and that it is statistically significant in explaining Rental Rates.

PROBLEM 2

Show that $\text{SSR}(x_1, x_2, x_3, x_4) = \text{SSR}(x_1) + \text{SSR}(x_2, x_3|x_1) + \text{SSR}(x_4|x_1, x_2, x_3)$.

Proof: Before we proceed our proof, recall the identity $\text{SSR}(X2X_3|X_1) = \text{SSR}(X_1X_2X_3) - \text{SSR}(X_1)$. Now let us proceed as the following

$$\begin{aligned} & \text{SSR}(X_1X_2X_3X_4) \\ = \; & \text{SSR}(X_1) + \text{SSR}(X_2X_3|X_1) + \text{SSR}(X_4|X_1X_2X_3) \\ = \; & \text{SSR}(X_1X_2X_3) + \text{SSR}(X_1X_2X_3X_4) - \text{SSR}(X_1) - \text{SSR}(X_2|X_1) - \text{SSR}(X_3|X_1X_2) \\ = \; & \text{SSR}(X_1X_2X_3) + \text{SSR}(X_1X_2X_3X_4) - \text{SSR}(X_1) \\ & -\text{SSR}(X_2|X_1) - (\text{SSR}(X_1X_2X_3) - \text{SSR}(X_1) - \text{SSR}(X_2|X_1)) \\ = \; & \text{SSR}(X_1X_2X_3X_4) \end{aligned}$$

and we are done.

————————————————————————————————– Q.E.D.

PROBLEM 3

Consider model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Assuming that sample correlation between x_1 and x_2 is zero, i.e.

$$\frac{1}{(n-1)s_1 s_2} \sum_{i=1}^{n} (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2) = 0$$

show that

$$\text{SRR}(x_2|x_1) = \text{SRR}(x_2)$$

Proof: First, we recall $\text{SSR}(X_2|X_1) = \text{SSR}(X_1X_2) - \text{SSR}(X_1)$. All we need is to show that $\text{SSR}(X_1X_2) = \text{SSR}(X_1) + \text{SSR}(X_2)$ because the two variables are not correlated.

Let us proceed the following. Recall that $\text{SSR}(X_1X_2) = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}) = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta} X_{i2} - \overline{Y})^2$. Then since $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$ then we have the following

$$\begin{aligned}
\text{SSR}(X_1X_2) &= \sum_{i=1}^{n}(\overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2 + \hat{\beta}_1 X_{i1} + \hat{\beta} X_{i2} - \overline{Y})^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_1 (X_{i1} - \overline{X}_1) + \hat{\beta} (X_{i2} - \overline{X}_2))^2 \\
&= \sum_{i=1}^{n}(\hat{\beta}_1^2 (X_{i1} - \overline{X}_1)^2 + \hat{\beta}_2^2 (X_{i2} - \overline{X}_2)^2 + \hat{\beta}_1 \hat{\beta}_2 (X_{i1} - \overline{X}_1)(X_{i2} - X_2)) \\
&= \hat{\beta}_1^2 \sum_{i=1}^{n}(X_{i1} - \overline{X}_1)^2 + \hat{\beta}_2^2 \sum_{i=1}^{n}(X_{i2} - \overline{X}_2)^2 \\
&= \text{SSR}(X_1) + \text{SSR}(X_2)
\end{aligned}$$

and thus, plugging the result, we have

$$\text{SSR}(X_2|X_1) = \text{SSR}(X_1X_2) - \text{SSR}(X_1) = \text{SSR}(X_1) - \text{SSR}(X_2) - \text{SSR}(X_1) = \text{SSR}(X_2)$$

and we are done.

————————————————————————————————– Q.E.D.

PROBLEM 4

(i) Regression Y ~ X1

```
# Upload data
setwd("E:/Course/CU Stats/STATS GR5205 - Linear Regression Model/3. Homework/
HW5")
data <- read.delim("CosmeticsSales.txt", sep = "")
n <- nrow(data)

# Introduce data
y <- data$Sales
x1 <- data$Expenditures
x2 <- data$LocalExpenditures
x3 <- data$ProratedShare

# Linear model
lm.X1 <- lm(y~x1)
summary(lm.X1)

##
## Call:
```

```
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0060 -0.7919  0.1584  1.2961  3.4824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1628     0.6712   4.712 2.69e-05 ***
## x1            1.6581     0.1641  10.104 8.23e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.892 on 42 degrees of freedom
## Multiple R-squared:  0.7085, Adjusted R-squared:  0.7016
## F-statistic: 102.1 on 1 and 42 DF,  p-value: 8.231e-13
```

We observe from *R* output above that the p-value is less than 0.05, which means X_1, e.g. Expenditures, statistically influences Sales, response variable.

(ii) Regression Y ~ X2

```
# Linear model
lm.X2 <- lm(y~x2)
summary(lm.X2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4287 -1.2874  0.2027  1.0759  3.6742
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8315     0.6990   4.051 0.000215 ***
## x2            1.7926     0.1769  10.135 7.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.888 on 42 degrees of freedom
## Multiple R-squared:  0.7098, Adjusted R-squared:  0.7029
## F-statistic: 102.7 on 1 and 42 DF,  p-value: 7.507e-13
```

We observe from the above *R* output that the p-value is less than 0.05, which means for X_2, e.g. local expenditures statistically influences the response variable.

(iii) Regression Y ~ X1 + X2 + X3

```
# Linear model
lm.X1.X2.X3 <- lm(y~x1+x2+x3)
summary(lm.X1.X2.X3)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4217 -0.9115  0.0703  1.1420  3.5479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0233     1.2029   0.851   0.4000
## x1            0.9657     0.7092   1.362   0.1809
## x2            0.6292     0.7783   0.808   0.4237
## x3            0.6760     0.3557   1.900   0.0646 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 40 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7223
## F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

We observe p-value for X_1 is 0.18 and p-value for X_2 is 0.18 holding others constant. We conclude that X_1 and X_2 are not statistically significant in explaining response. Comparing to (i) and (ii), the descrepancy is that they are not significant anymore.

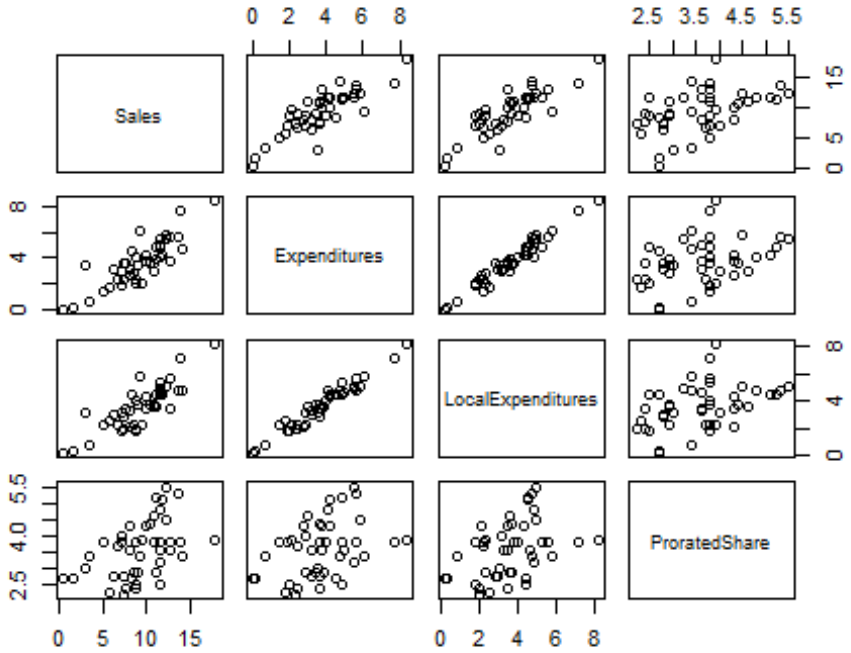
(iv) Explore

To futher explore this descripancy we observed from above, we can use correlation matrix and pairs plot to explore the data. We observe that (1) X_1 and X_2 are highly correlated and also (2) that the plot for X_1 and X_2 almost form a linear line.

```
cor(data)
```

```
##                       Sales Expenditures LocalExpenditures ProratedShare
## Sales             1.0000000    0.8417342         0.8424849     0.4740581
## Expenditures      0.8417342    1.0000000         0.9744313     0.3759509
## LocalExpenditures 0.8424849    0.9744313         1.0000000     0.4099208
## ProratedShare     0.4740581    0.3759509         0.4099208     1.0000000
```

```
pairs(data)
```



PROBLEM 5

Consider standardized regression model

$$Y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \epsilon_i^*$$

The variables Y_i^*, x_{i1}^*, and x_{i2}^* are standardized versions of Y_i, x_{i1}, and x_{i2}, i.e.

$$Y_i * = \frac{1}{n-1}\left(\frac{Y_i - \overline{Y}}{s_Y}\right), x_{i1}^* = \frac{1}{n-1}\left(\frac{x_{i1} - \overline{x}_1}{s_1}\right), x_{i2}^* = \frac{1}{n-1}\left(\frac{x_{i2} - \overline{x}_1}{s_2}\right)$$

Show that

$$\hat{\beta}_1^* = \frac{r_{Y_1} - r_{12} r_{Y_2}}{1 - r_{12}^2} \text{ and } \hat{\beta}_2^* = \frac{r_{Y_2} - r_{12} r_{Y_1}}{1 - r_{12}^2}$$

Proof: We need to compute r_{12}, r_{Y_1}, and r_{Y_2} to derive the estimators $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$. Thus,

$$r_{12} = \sum_{i=1}^{n} X_{i1}^* X_{i1}^* = \sum_{i=1}^{n} \left(\frac{X_{i1} - \overline{X}_1}{\sqrt{n-1}s_1}\right)\left(\frac{X_{i2} - \overline{X}_2}{\sqrt{n-1}s_2}\right) = \frac{1}{n-1}\frac{1}{s_1 s_2}\sum_{i=1}^{n}(X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2)$$

$$r_{Y_1} = \sum_{i=1}^{n} X_{i2}^* Y_i^* = \sum_{i=1}^{n} \left(\frac{X_{i1} - \overline{X}_1}{\sqrt{n-1}s_1}\right)\left(\frac{Y_i - \overline{Y}}{\sqrt{n-1}s_Y}\right) = \frac{1}{n-1}\frac{1}{s_1 s_Y}\sum_{i=1}^{n}(X_{i1} - \overline{X})(Y_i - \overline{Y})$$

and

$$r_{Y_2} = \sum_{i=1}^{n} X_{i2}^{*} Y_i^{*} = \frac{1}{n-1}\frac{1}{s_2 s_Y}\sum_{i=1}^{n}(X_{i2} - \overline{X})(Y_i - \overline{Y})$$

Recall least squares for the ordinary multiple regression model:

$$\mathbf{XX'b} = \mathbf{X'Y}$$

and the least squares estimators

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

can be expressed simply for the transformed variables. In this case, we have two parameters and we write

$$r_{XX} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}$$

$$r_{YX} \begin{bmatrix} r_{Y_1} \\ r_{Y_2} \end{bmatrix}$$

and

$$r_{XX}^{-1} = \frac{1}{1 - r_{12}^2}\begin{bmatrix} 1 & -r12 \\ -r_{12} & 1 \end{bmatrix}$$

Then we recall that from $\mathbf{X'X} = r_{XX}$ we can solve for $\mathbf{b} = r_{XX}^{-1} r_{YX}$ and hence

$$\mathbf{b} = \frac{1}{1 - r_{12}^2} = \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}\begin{bmatrix} r_{Y_1} \\ r_{Y_2} \end{bmatrix} = \begin{bmatrix} r_{Y_1} - r_{12} r_{Y_2} \\ r_{Y_2} - r_{12} r_{Y_1} \end{bmatrix}$$

and thus we solve for

$$b_1^{*} = \frac{r_{Y_1} - r_{12} r_{Y_2}}{1 - r_{12}^2}$$

$$b_2^{*} = \frac{r_{Y_2} - r_{12} r_{Y_1}}{1 - r_{12}^2}$$

and we are done.

———————————————————————————————– Q.E.D.

STAT GR5205 Homework 5 [100 pts]

Problem 1 [15 pts]

A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas. The data consists of variables age (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y). The data set `HW3Problem3.txt` is posted on canvas. Use `R` to perform the following exercises. Using type I sums of squares (F-stat), test if age (X_1) is a significant predictor, after holding operating expenses & taxes (X_2), vacancy rates (X_3) and total square footage (X_4) constant.

Solution

`R code`

```
model <- lm(RentalRates~OperatingExpense+VacancyRates+SquareFootage+Age,data=data)
anova(model)
```

`R output`

```
Analysis of Variance Table

Response: RentalRates
                 Df Sum Sq Mean Sq F value    Pr(>F)
OperatingExpense  1 40.503  40.503  31.337 3.291e-07 ***
VacancyRates      1 13.829  13.829  10.699  0.001613 **
SquareFootage     1 26.752  26.752  20.698 2.002e-05 ***
Age               1 57.243  57.243  44.288 3.894e-09 ***
Residuals        76 98.231   1.293
```

Problem 2 [15 pts]

Show that: $SSR(x_1, x_2, x_3, x_4) = SSR(x_1) + SSR(x_2, x_3|x_1) + SSR(x_4|x_1, x_2, x_3)$.

Solution

$$\begin{aligned} &SSR(x_1) + SSR(x_2, x_3|x_1) + SSR(x_4|x_1, x_2, x_3) \\ &\quad = SSR(x_1) + (SSR(x_1, x_2, x_3) - SSR(x_1)) + (SSR(x_1, x_2, x_3, 4) - SSR(x_1, x_2, x_3)) \\ &\quad = SSR(x_1, x_2, x_3, x_4) \end{aligned}$$

Problem 3 [20 pts]

Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

Assuming that the sample correlation between x_1 and x_2 is zero, i.e.,

$$\frac{1}{(n-1)s_1 s_2} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0,$$

show that

$$SSR(x_2|x_1) = SSR(x_2).$$

Solution

Note that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

Then

$$\begin{aligned}
SSR(x_1, x_2) &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} - \bar{y})^2 \\
&= \sum_{i=1}^{n} ((\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2) + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_1 (x_{i1} - \bar{x}_1) + \hat{\beta}_2 (x_{i2} - \bar{x}_2))^2 \\
&= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \hat{\beta}_2^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 + \hat{\beta}_1 \hat{\beta}_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i12} - \bar{x}_2) \\
&= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \hat{\beta}_2^2 \sum_{i=1}^{n} (x_{i12} - \bar{x}_2)^2 + 0 \qquad \textbf{Note: } r_{12} = 0 \\
&= SSR(x_1) + SSR(x_2).
\end{aligned}$$

Thus

$$\begin{aligned}
SSR(x_2|x_1) &= SSR(x_1, x_2) - SSR(x_1) \\
&= SSR(x_1) + SSR(x_2) - SSR(x_1) \\
&= SSR(x_2).
\end{aligned}$$

Problem 4 [20 pts]

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Let Y denote sales (in thousand cases). The assistant was instructed to study the influence variables X_1 and X_2 have on sales Y. The data set `CosmeticsSales.txt` is posted on Canvas.

Use `R` to perform the following tasks:

i. Run the simple linear regression $Y \sim X_1$. Test if expenditures for point-of-sale displays in beauty salons and department stores (X_1) statistically influences sales (Y). **(5 pts)**

Solution

```
> summary(lm(Sales~Expenditures,data=data))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.1628     0.6712   4.712 2.69e-05 ***
Expenditures   1.6581     0.1641  10.104 8.23e-13 ***

Residual standard error: 1.892 on 42 degrees of freedom
Multiple R-squared:  0.7085,Adjusted R-squared:  0.7016
F-statistic: 102.1 on 1 and 42 DF,  p-value: 8.231e-13
```

At any reasonable level of significance, expenditures for point-of-sale displays in beauty salons and department stores (X_1) does statistically influences sales (Y).

ii. Run the simple linear regression $Y \sim X_2$. Test if expenditures for local media advertising (X_2) statistically influences sales (Y). **(5 pts)**

Solution

```
> summary(lm(Sales~LocalExpenditures,data=data))

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.8315     0.6990   4.051 0.000215 ***
LocalExpenditures  1.7926     0.1769  10.135 7.51e-13 ***
```

```
Residual standard error: 1.888 on 42 degrees of freedom
Multiple R-squared:  0.7098,Adjusted R-squared:  0.7029
F-statistic: 102.7 on 1 and 42 DF,  p-value: 7.507e-13
```

At any reasonable level of significance, expenditures for local media advertising (X_2) statistically influences sales (Y).

iii. Now run the the full regression $Y \sim X_1 + X_2 + X_3$. Perform *marginal* t-tests to see if X_1 statistically influences sales (Y) and if X_2 statistically influences sales (Y), after controlling for the variance of X_3. Briefly compare the results to Parts i. and ii. and comment on any discrepancies. **Note: No need for Bonferroni. (5 pts)**

Solution

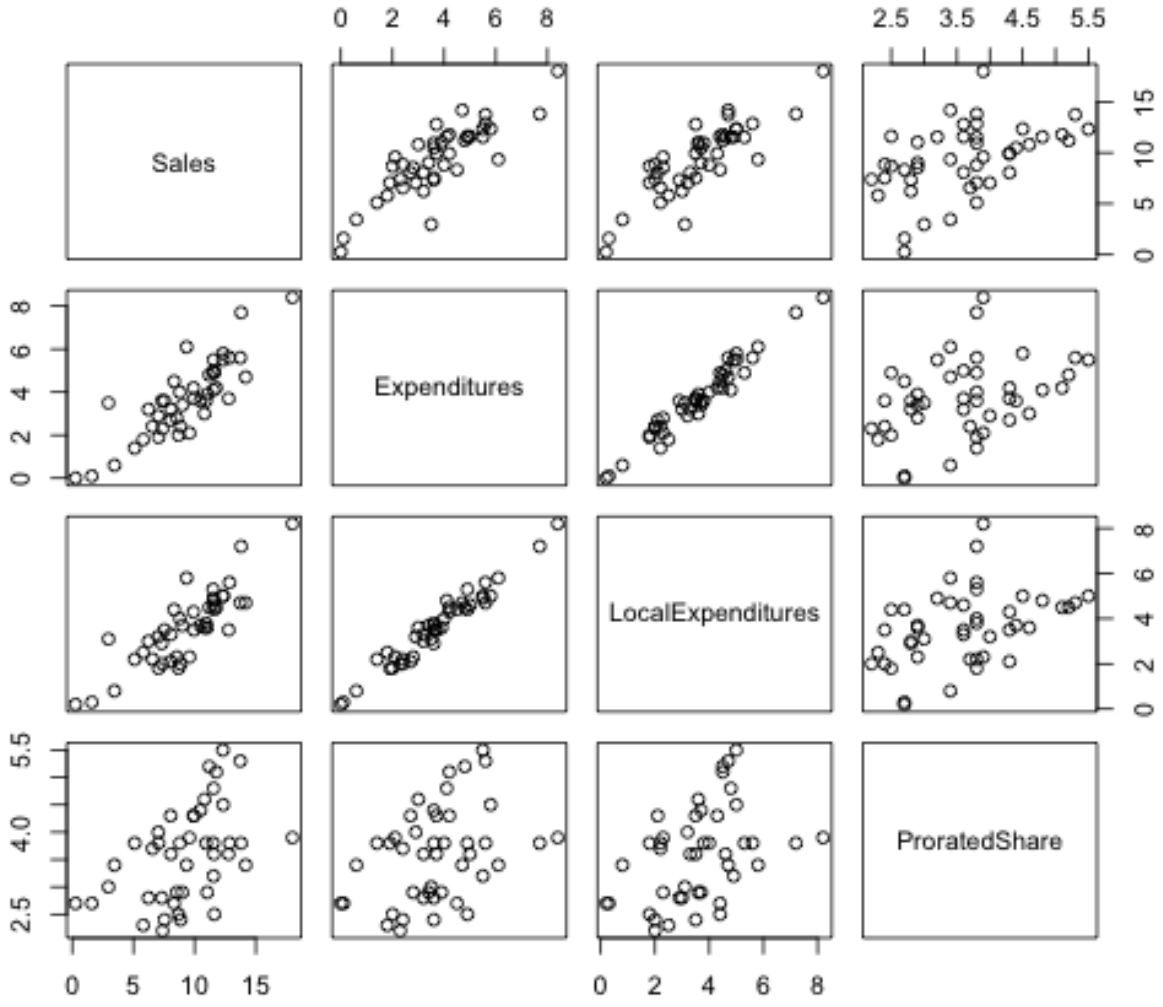
```
> summary(lm(Sales~Expenditures+LocalExpenditures+ProratedShare,data=data))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.0233     1.2029   0.851   0.4000
Expenditures        0.9657     0.7092   1.362   0.1809
LocalExpenditures   0.6292     0.7783   0.808   0.4237
ProratedShare       0.6760     0.3557   1.900   0.0646 .

Residual standard error: 1.825 on 40 degrees of freedom
Multiple R-squared:  0.7417,Adjusted R-squared:  0.7223
F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

Based on the *marginal* t-tests, both f X_1 and X_2 do not statistically influences sales (Y), after controlling for the variance of X_3. This result seems to contradict the conclusions from Parts i and ii.

iv. You should have noticed some discrepancies in Part iii. Explain why these discrepancies are occurring and provide graphical or exploratory evidence to complement your argument. **(5 pts)**

Solution

Based on the scatterplot matrix, X_1 and X_2 appear to be highly correlated. Thus multicollinearity is causing instability in the coefficients and standard errors, which in-turn is causing inconsistencies from simple linear regression and multiple regression results.


Sales
Expenditures
LocalExpenditures
ProratedShare
0 2 4 6 8
2.5 3.5 4.5 5.5
0 5 10 15
0 2 4 6 8
0 2 4 6 8
2.5 4.0 5.5
0 5 10 15
0 2 4 6 8

Problem 5 [30 pts]

Consider the *standardized regression model*

$$Y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \epsilon_i^*.$$

The variables Y_i^*, x_{i1}^* and x_{21}^*∗ are *standardized* versions of Y_i, x_{i1} and x_{i2}, i.e.,

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right), \quad x_{i1}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{i1} - \bar{x}_1}{s_1}\right), \quad x_{i2}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{i2} - \bar{x}_2}{s_2}\right)$$

Task: Using the least squares equation, derive the estimators

$$\hat{\beta}_1^* = \frac{r_{Y1} - r_{12}r_{Y2}}{1 - r_{12}^2} \quad \text{and} \quad \hat{\beta}_2^* = \frac{r_{Y2} - r_{12}r_{Y1}}{1 - r_{12}^2}.$$

Note that the standardized regression model does relate the regular regression model. The description follows below:

Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, n, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

It is easy to show that the parameters β_1^*, β_2^* in the standardized regression model and the original parameters $\beta_0, \beta_1, \beta_2$ are related as follows:

$$\beta_1 = \frac{s_Y}{s_1}\beta_1^*, \quad \beta_2 = \frac{s_Y}{s_2}\beta_2^*, \quad \text{and} \quad \beta_0 = \bar{Y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2.$$

Reference pages 273 to 278 of the textbook for further details.

Solution:

Define the design matrix of the standardized regression model by

$$\mathbf{X}^* = \begin{pmatrix} x_{11}^* & x_{12}^* \\ x_{21}^* & x_{22}^* \\ \vdots & \vdots \\ x_{n1}^* & x_{n2}^* \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n-1}}\left(\frac{x_{11}-\bar{x}_1}{s_1}\right) & \frac{1}{\sqrt{n-1}}\left(\frac{x_{12}-\bar{x}_2}{s_2}\right) \\ \frac{1}{\sqrt{n-1}}\left(\frac{x_{21}-\bar{x}_1}{s_1}\right) & \frac{1}{\sqrt{n-1}}\left(\frac{x_{22}-\bar{x}_2}{s_2}\right) \\ \vdots & \vdots \\ \frac{1}{\sqrt{n-1}}\left(\frac{x_{n1}-\bar{x}_1}{s_1}\right) & \frac{1}{\sqrt{n-1}}\left(\frac{x_{n2}-\bar{x}_2}{s_2}\right) \end{pmatrix}.$$

Define the response vector of the standardized regression model by

$$\mathbf{Y}^* = \begin{pmatrix} y_1^* & y_2^* & \cdots & y_n^* \end{pmatrix}^T = \left(\frac{1}{\sqrt{n-1}}\left(\frac{y_1-\bar{y}}{s_y}\right) \quad \frac{1}{\sqrt{n-1}}\left(\frac{y_2-\bar{y}}{s_y}\right) \quad \cdots \quad \frac{1}{\sqrt{n-1}}\left(\frac{y_n-\bar{y}}{s_y}\right)\right)^T.$$

Notice that

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i1}-\bar{x}_1}{s_1}\right)^2 = \frac{1}{s_1^2(n-1)}s_1^2(n-1) = 1$$
$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)^2 = \frac{1}{s_2^2(n-1)}s_2^2(n-1) = 1$$
$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right) = r_{12}$$
$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i1}-\bar{x}_1}{s_1}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = r_{Y1}$$
$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = r_{Y2}.$$

Then we have

$$(\mathbf{X}^*)^T(\mathbf{X}^*) = \begin{pmatrix} \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i1}-\bar{x}_1}{s_1}\right)^2 & \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right) \\ \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right) & \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)^2 \end{pmatrix} = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}.$$

$$(\mathbf{X}^*)^T(\mathbf{Y}^*) = \begin{pmatrix} \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i1}-\bar{x}_1}{s_1}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\ \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_{i2}-\bar{x}_2}{s_2}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \end{pmatrix} = \begin{pmatrix} r_{Y1} \\ r_{Y2} \end{pmatrix}.$$

Thus

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= [(\mathbf{X}^*)^T(\mathbf{X}^*)]^{-1}(\mathbf{X}^*)^T(\mathbf{Y}^*) = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}^{-1}\begin{pmatrix} r_{Y1} \\ r_{Y2} \end{pmatrix} \\ &= \frac{1}{1-r_{12}^2}\begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}\begin{pmatrix} r_{Y1} \\ r_{Y2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{r_{Y1}-r_{12}}{1-r_{12}^2} \\ \frac{r_{Y2}-r_{12}}{1-r_{12}^2} \end{pmatrix}\end{aligned}$$

STAT GR5205 Homework 6

Practice

Problem 1

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Let Y denote sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant, and was told to use *the most appropriate* model based on the covariance structure of explanatory variables X_1, X_2, X_3. The increase in expected sales should be estimated with its uncertainty, i.e., construct the estimate with its 95% confidence interval. The data set `CosmeticsSales.txt` is posted on Canvas.

Use `R` to perform the following tasks:

i. Create a scatterplot matrix and a correlation matrix of the variables. Based on the output, do you see any potential problems with using ordinary least squares to test the researcher's claim?

ii. Find the variance inflation factors for each covariate. Does there appear to be a problem with multicollinearity?

iii. Write a function in `R` that computes the unstandardized ridge regression estimates using the `CosmeticsSales` data set. The function should have two inputs: the first is a data frame that defaults to the `CosmeticsSales` data set, and the second is a scalar for the the tuning parameter c. The output of the function should be a vector of the unstandardized ridge regression estimates. Test the function with $c = 0$, which should return the regular least squares estimates.

iv. VIF values for ridge regression coefficients are defined analogously to those for ordinary least squares regression coefficients. Namely, the VIF value for $\hat{\beta}_k^R$ measures how large the variance of $\hat{\beta}_k^R$ is relative to what the variance would be if the predictor variables were uncorrelated. It can be shown that the VIF values for the ridge regression coefficients

$\hat{\beta}_k^R$ are the diagonal elements of the following $(p-1) \times (p-1)$ matrix:

$$(\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{XX}(\mathbf{r}_{XX} + c\mathbf{I})^{-1}.$$

In the function from Part iii, add an additional element to the `return` argument that computes a vector of the *ridge* regression variance inflation factors. Test the updated function with $c = 0$, which should return the regular least squares variance inflation factors form Part ii.

v. Define a vector of tuning parameters that goes from 0 to .2 in steps of .01. Use the `sequence` function to accomplish this. Write a loop that computes the average of the VIF values for each c. Plot the average VIF versus the list `seq(0,.2,.01)`. Choose c based on where the average VIF significantly stops decreasing. Note: We don't want to make c too big because that will increase the bias of our estimators $\hat{\beta}_R$.

vi. Based on the tuning parameter c selected in Part v, run the ridge regression and display the estimated unstandardized ridge coefficients.

vii. To answer the research question we must use a nonparametric technique to construct an approximate sampling distribution of ridge estimator $\hat{\beta}_1^R$ (technically on the unstandardized ridge estimator). Run a bootstrap procedure on this estimated coefficient. Run this using $B = 10,000$ bootstrapped samples. Compute the estimated variance of the approximate sampling distribution.

viii. Use the bootstrap interval from Part vi to answer the original research question. That is, construct a 95% bootstrapped interval for the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant.

Problem 2

Consider a the data set $y_1, y_2, \ldots, y_n$. Using the absolute value loss function ($\psi(u) = |u|$), minimize the objective function

$$Q(b) = \sum_{i=1}^{n} \psi(y_i - b) = \sum_{i=1}^{n} |y_i - b|,$$

with respect to b.

Problem 3

Consider a study investigating the association between marijuana usage in college students and parental usage of drugs and alcohol. The following two-way frequency table summarizes the results. Each of 445 college students was classified according to both frequency of marijuana use and parental use of alcohol and psychoactive drugs.

		Marijuana Use	
		No	Yes
Parental Use of	No	141	94
Alcohol and Drugs	Yes	85	125

Use `R` to test if the **odds ratio** of a college student that uses marijuana who came from a household without parental drug & alcohol use verses a household with parental drug & alcohol use statistically differs from 1. I.e., at 5% significance, test if the odds ratio $\Theta = e^{\beta_1}$ statistically differs from 1.

Problem 4

A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$ and a client who did not receive a flu shot was coded $Y = 0$. In addition. data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate greater awareness. Also included in the data was client gender, where males were coded $X_3 = 1$ and females were coded $X_3 = 0$. The data set `FluShots.txt` is Posted on Canvas.

Use `R` to perform the following tasks:

i. Find the maximum likelihood estimators of $\beta_0, \beta_1, \beta_2, \beta_3$. State the fitted response function.

ii. Obtain $\exp(\hat{\beta}_1), \exp(\hat{\beta}_2), \exp(\hat{\beta}_3)$. Interpret these numbers.

iii. What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? Estimate this probability with 95% confidence.

iv. Construct a ROC plot for the above model. Interpret the ROC plot.

STAT GU4205/GR5205 Homework 6 KEY

Practice

Problem 1

An assistant in the district sales office of a national cosmetics firm obtained data on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Let Y denote sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant, and was told to use *the most appropriate* model based on the covariance structure of explanatory variables X_1, X_2, X_3. The increase in expected sales should be estimated with its uncertainty, i.e., construct the estimate with its 95% confidence interval. The data set `CosmeticsSales.txt` is posted on Canvas.

Use `R` to perform the following tasks:

See the `R` script file `4205HW6` for the complete code for this exercise.

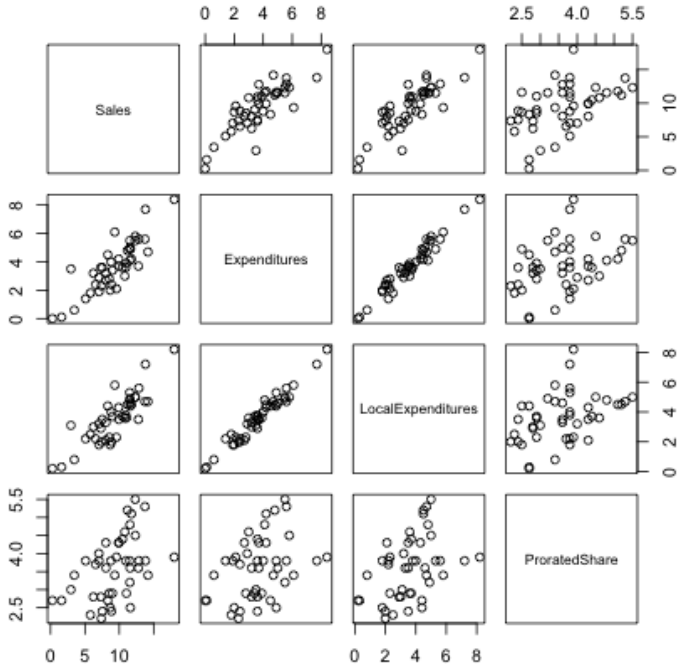
i. Create a scatterplot matrix and a correlation matrix of the variables. Based on the output, do you see any potential problems with using ordinary least squares to test the researcher's claim?

Solution

	Sales	Expenditures	LocalExpenditures	ProratedShare
Sales	1.00	0.84	0.84	0.47
Expenditures	0.84	1.00	0.97	0.38
LocalExpenditures	0.84	0.97	1.00	0.41
ProratedShare	0.47	0.38	0.41	1.00

Based on Figure 1 and the correlation matrix, there does appear to be a problem with multicollinearity, i.e., the two expenditure covariates have a sample correlation coefficient of 0.97. This will cause instability in our parameter estimates if we wanted to use OLS.

Figure 1



ii. Find the variance inflation factors for each covariate. Does there appear to be a problem with multicollinearity?

Solution

	Expenditures	LocalExpenditures	ProratedShare
VIF	20.072031	20.716101	1.217973

Yes there does appear to be a problem with multicollinearity. The maximum VIF is much greater than 10 and the average VIF is much greater than 1.

iii. Write a function in `R` that computes the unstandardized ridge regression estimates using the `CosmeticsSales` data set. The function should have two inputs: the first is a data frame that defaults to the `CosmeticsSales` data set, and the second is a scalar for the the tuning parameter c. The output of the function should be a vector of the unstandardized ridge regression estimates. Test the function with $c = 0$, which should return the regular least squares estimates.

See the `R` script

iv. VIF values for ridge regression coefficients are defined analogously to those for ordinary least squares regression coefficients. Namely, the VIF value for $\hat{\beta}_k^R$ measures how large the variance of $\hat{\beta}_k^R$ is relative to what the variance would be if the predictor variables were uncorrelated. It can be shown that the VIF values for the ridge regression coefficients $\hat{\beta}_k^R$ are the diagonal elements of the following $(p-1) \times (p-1)$ matrix:

$$(\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{XX}(\mathbf{r}_{XX} + c\mathbf{I})^{-1}.$$

In the function from Part iii, add an additional element to the `return` argument that computes a vector of the *ridge* regression variance inflation factors. Test the updated function with $c = 0$, which should return the regular least squares variance inflation factors form Part ii.

See the `R` script

v. Define a vector of tuning parameters that goes from 0 to .2 in steps of .01. Use the `sequence` function to accomplish this. Write a loop that computes the average of the VIF values for each c. Plot the average VIF versus the list `seq(0,.2,.01)`. Choose c based on where the average VIF significantly stops decreasing. Note: We don't want to make c too big because that will increase the bias of our estimators $\hat{\beta}_R$.

Solution

See the `R` script

I chose $c = .11$ because this is the first value that the average VIF dips below 1. This is displayed in Figure 2. You could have chosen a different value for c.

vi. Based on the tuning parameter c selected in Part v, run the ridge regression and display the estimated unstandardized ridge coefficients.

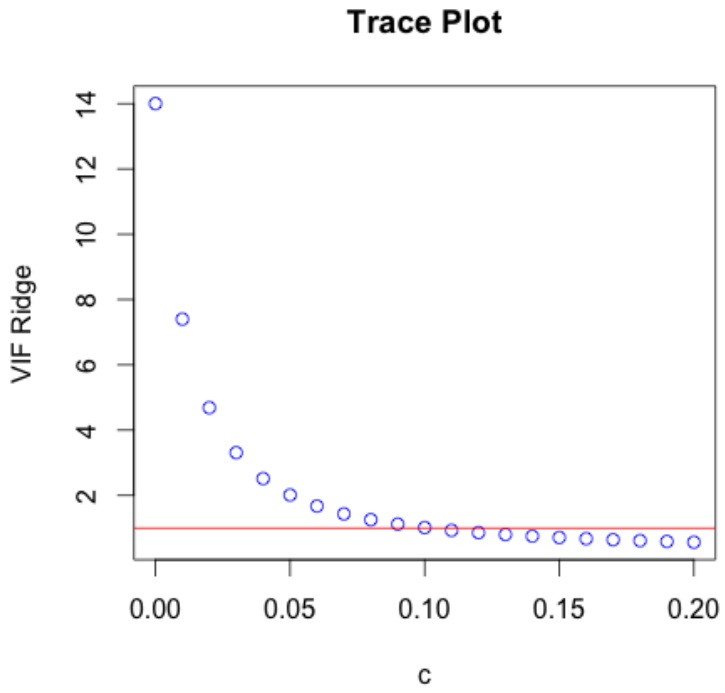
Intercept	Expenditures	LocalExpenditures	ProratedShare
1.363	0.770	0.756	0.655

vii. To answer the research question we must use a nonparametric technique to construct an approximate sampling distribution of ridge estimator $\hat{\beta}_1^R$ (technically on the unstandardized ridge estimator). Run a bootstrap procedure on this estimated coefficient. Run this using $B = 10,000$ bootstrapped samples. Compute the estimated variance of the approximate sampling distribution.

Solution

See the `R` script

Figure 2



The estimated variance and standard error of the approximate sampling distribution are respectively 0.0144 and 0.1200. Here I used `set.seed(0)`.

viii. Use the bootstrap interval from Part vi to answer the original research question. That is, construct a 95% bootstrapped interval for the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant.

Solution

Below I am reporting both the 95% regular bootstrap interval and the percentile based bootstrap interval.

95% Regular Boot Interval	95% Percentile Boot Interval
(0.5152, 0.9908)	(0.7449, 1.2205)

Consider testing the hypothesis $H_0 : \beta_1 = 0$. Zero is not contained in both bootstrapped intervals indicating that X_1 is a significant predictor after holding the other covariates constant. This was not the case when we used OLS.

Problem 2

Consider the data set $y_1, y_2, \ldots, y_n$. Using the absolute value loss function ($\psi(u) = |u|$), minimize the objective function

$$Q(b) = \sum_{i=1}^{n} \psi(y_i - b) = \sum_{i=1}^{n} |y_i - b|,$$

with respect to b.

Solution

Notice we can write Q as

$$Q(b) = \sum_{i:y_i>b} (y_i - b) - \sum_{i:y_i \leq b} (y_i - b) = \sum_{i:y_i>b} y_i - \sum_{i:y_i \leq b} y_i + (n_1 - n_2)b,$$

where n_1 is the number of observations less than b and n_2 is the number of observations greater than b. Taking the derivative and equating the expression equal to zero yields $n_1 = n_2$, which implies that Q is optimized at $\hat{b} = \tilde{y}$ (sample median). Note that the derivative of Q is nonnegative when b is to the right of the median and nonpositive when b is to the left of the median, hence Q is minimized at $\hat{b} = \tilde{y}$.

Problem 3

Consider a study investigating the association between marijuana usage in college students and parental usage of drugs and alcohol. The following two-way frequency table summarizes the results. Each of 445 college students was classified according to both frequency of marijuana use and parental use of alcohol and psychoactive drugs.

		Marijuana Use	
		No	Yes
Parental Use of	No	141	94
Alcohol and Drugs	Yes	85	125

Use `R` to test if the **odds ratio** of a college student that uses marijuana who came from a household without parental drug & alcohol use verses a household with parental drug & alcohol use statistically differs from 1. I.e., at 5% significance, test if the odds ratio $\Theta = e^{\beta_1}$ statistically differs from 1.

Solution

R code and output

```
> Y <- c(rep(1,125+94),rep(0,85+141))
> X <- c(rep(1,125),rep(0,94),rep(1,85),rep(0,141))
>
> model <- glm(Y~X,data=data,family=binomial(link="logit"))
> summary(model)

Call:
glm(formula = Y ~ X, family = binomial(link = "logit"), data = data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.345  -1.011  -1.011   1.019   1.354

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4055     0.1332  -3.045  0.00233 **
X             0.7911     0.1936   4.086  4.4e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 616.79  on 444  degrees of freedom
Residual deviance: 599.77  on 443  degrees of freedom
AIC: 603.77

Number of Fisher Scoring iterations: 4
```

Testing procedure

For this problem, we are testing the null alternative pair $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, which is equivalent to testing $H_0 : \Theta = 1$ versus $H_1 : \Theta \neq 1$. In support of our research question, notice that the sample odds ratio is bigger than one, i.e., $\exp(\hat{\beta}_1) = 2.206 > 1$. The correct Z-statistic and P-value for this procedure are respectively $z_{calc} = 4.086$ and $pvalue = 4.4 * 10^{-1}$. Thus at 5% significance, we reject the null hypothesis and conclude that the **odds ratio** of a college student that uses marijuana who came from a household without parental drug & alcohol use verses a household with parental drug & alcohol use statistically differs from 1.

Problem 4

A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$ and a client who did not receive a flu shot was coded $Y = 0$. In addition. data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate greater awareness. Also included in the data was client gender, where males were coded $X_3 = 1$ and females were coded $X_3 = 0$. The data set `FluShots.txt` is Posted on Canvas.

Use `R` to perform the following tasks:

i. Find the maximum likelihood estimators of $\beta_0, \beta_1, \beta_2, \beta_3$. State the fitted response function.

Solution

R code and output

```
> data <- read.table("FluShots.txt")
> model <- glm(Y~X1+X2+X3,data=data,family=binomial(link="logit"))
> summary(model)

Call:
glm(formula = Y ~ X1 + X2 + X3, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4037  -0.5637  -0.3352  -0.1542   2.9394

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17716    2.98242  -0.395  0.69307
X1           0.07279    0.03038   2.396  0.01658 *
X2          -0.09899    0.03348  -2.957  0.00311 **
X3           0.43397    0.52179   0.832  0.40558
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 105.09  on 155  degrees of freedom
AIC: 113.09

Number of Fisher Scoring iterations: 6
```

The fitted response function is :

$$\hat{p} = \frac{\exp\left(-1.17716 + 0.07279X_1 - 0.09899X_2 + 0.43397X_3\right)}{1 + \exp\left(-1.17716 + 0.07279X_1 - 0.09899X_2 + 0.43397X_3\right)}$$

ii. Obtain $\exp(\hat{\beta}_1), \exp(\hat{\beta}_2), \exp(\hat{\beta}_3)$. Interpret these numbers.

Interpretation

$\exp(\hat{\beta}_1) = 1.0755, \quad \exp(\hat{\beta}_2) = .9058, \quad \exp(\hat{\beta}_3) = 1.5434$

The interpretation is similar for each quantity. For example, take $exp(\hat{\beta}_1) = 1.0755$. We say for a 1 year increase in a client's age, the odds ratio of receiving a flu shot is 1.0755, after holding the heath awareness and gender variables constant. Notice that this quantity is near 1, which appears to support no significant relationship between age and flu shots. After examining the P-value, the relationship is significant at $\alpha = .05$.

iii. What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? Estimate this probability with 95% confidence.

R code and output

```
> # Part iii
>
> pred.logistic <- predict(model,newdata=data.frame(X1=55,X2=60,X3=1),se.fit = T)
>
> # Linear intervals
> L <- pred.logistic$fit-qnorm(.975)*pred.logistic$se.fit
> U <- pred.logistic$fit+qnorm(.975)*pred.logistic$se.fit
>
> # p.hat
> exp(pred.logistic$fit)/(1+exp(pred.logistic$fit))
         1
0.06422197
```

```
>
> # Intervals for estimated probability
> exp(L)/(1+exp(L))
         1
0.02470269
> exp(U)/(1+exp(U))
        1
0.1567997
```

iv. Construct a ROC plot for the above model. Interpret the ROC plot.

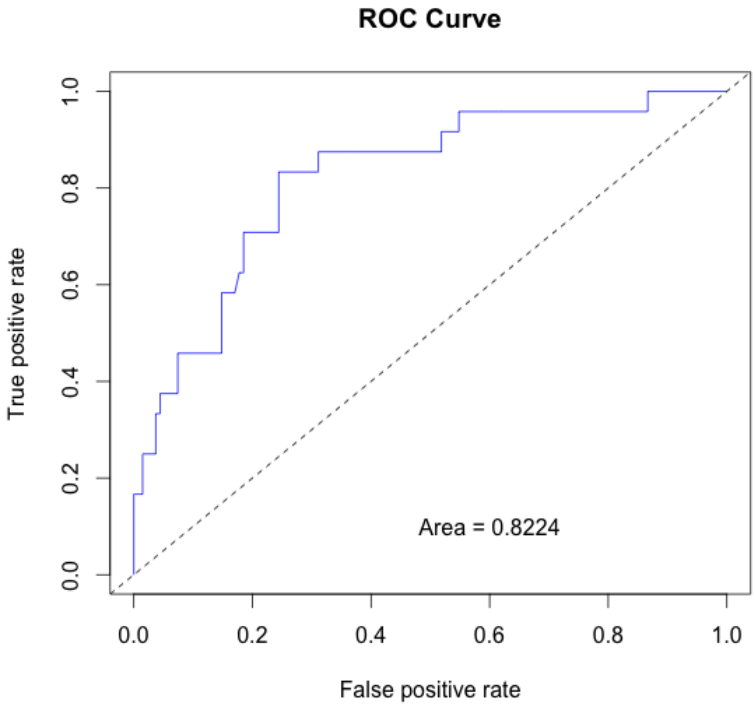
R code and output

```
# Library ROCP
library(ROCR)

p <- predict(model, type="response")
pr <- prediction(p,data$Y)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf,main="ROC Curve",col="blue")
abline(a=0,b=1,lty=2)

# Area under the curve
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
text(.6,.1,paste("Area =",round(auc,4)))
```

Interpretation

The logistic model for the flu shots application shows *good* predictive power because the area under the ROC curve is 0.80. An area higher than .90 would support an excellent predictive model.

Figure 3


ROC Curve
True positive rate
False positive rate
Area = 0.8224
0.0
0.2
0.4
0.6
0.8
1.0