Homework 1

Each part of the problems 5 points
Due in class or by email before 9am on Thursday January 20

1. The dataset `swiss` in the library `faraway` records standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. Our goal is to specify a linear model that predicts Fertility. When answering the questions below, please provide relevant R code, output and figures, and keep the discussion short.

   (a) Conduct an initial exploration of the dataset. Specifically, provide boxlots of all the variables on the dataset using `boxplot(swiss)`, and check for multicollinearity using `cor(swiss)` and `pairs(swiss).

   **Answer:**
   
   ```r
   boxplot(swiss)
   ```

   ![Box plot of variables](image)

   **Figure 1: Box plot of variables**

   We can see Catholic covers a wide range of values, while infant mortality is condensed.

   ```r
   cor(swiss)
   ```
As all the correlations are less than 0.8, it indicates there are no signs for strong multicollinearity. However, as some correlations are between 0.3 to 0.8, mild multicollinerarity does exist.

pairs(swiss)

Figure 2: Pairwise correlation

The plot also indicates some evidence of multicollinearity, in particular of linear relationships between Agriculture and Examination, and Education and Examination. Quadratic terms of some of the variables in the model may be needed.

(b) Consider a full linear model, where all variables in the dataset are used as predictors. Use the Box-Cox transformation to verify whether a transformation of the response is necessary to improve the quality of fit. Apply the appropriate transformation if necessary.

Answer:

library (MASS)
As $\lambda = 1$ lays in the 95% interval and is convient to use, we could decide that $\lambda = 1$, Therefore, no transformation is actually needed.

(c) Carry out the model diagnostic checks: check for unequal variance, outliers, deviations from Normality, and influential observations. Briefly summarize your results.

**Answer:**
From residual plot, there is no pattern for the residual, the constant variance assumption holds.
From the QQ plot, we don’t see much heavy tail, so normality assumption holds.
Some extreme values in Cook’s distance show that there are a few outliers in data to make the cook’s distance are relatively larger.

```r
full <- lm(
  Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data=swiss
)
plot(full, which = 1)
```
plot(full, which = 2)
plot(full, which = 3)
plot(full, which = 4)
qf(1-0.05, length(full$coef, nrow(swiss) - length(full$coef) )
plot(studres(full))

Figure 4: Checking Normality and Constant Variance Assumptions

(d) Use `step()` to perform stepwise variable selection, starting from the full model. What algorithm is used by the procedure? Which model is selected?

**Answer:**

Using stepwise variable selection, starting from the full model, we obtain the following reduced model.

\[
\hat{\text{Fertility}}_i = 62.1013 - 0.1546\text{Agriculture}_i - 0.9803\text{Education}_i + 0.1247\text{Catholic}_i + 1.0784\text{Mortality}
\]

\[\text{stepwise} <- \text{step(full, trace = FALSE)}\]
\[\text{stepwise}\]
\[\text{Call:}\]
\texttt{lm(formula = Fertility \sim Agriculture + Education + Catholic} \\
\texttt{+ Infant.Mortality, data = swiss)}

<table>
<thead>
<tr>
<th>Coefficients:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)   Agriculture Education Catholic Infant.Mortality</td>
</tr>
<tr>
<td>62.1013       -0.1546    -0.9803     0.1247     1.0784</td>
</tr>
</tbody>
</table>

The final AIC = 189.9

(e) Predictors in the full model are relatively correlated. Why ridge regression is recommended in presence of multicollinearity?

\textbf{Answer:}

Ridge regression produces biased parameter estimation. However adding a biasing constant can reduce the SE of parameter estimates, and also their numerical stability. If the bias is small, the procedure can reduce the overall expected MSE of parameter estimates.

(f) Use \texttt{lm.ridge} in the library \texttt{MASS} to fit ridge regression to the full model. Use 50 biasing parameters between 0 and 50. Plot parameter estimates versus biasing constant, and interpret the results.

\textbf{Answer:}

Run the ridge regression model with 50 biasing parameters between 0 and 50.

\begin{verbatim}
library(MASS)
mylambda <- seq(from = 0, to = 50, length = 50)
ridge <- lm.ridge(Fertility ~ Agriculture + Examination + Education+Catholic + Infant.Mortality, data = swiss,lambda=mylambda)
plot(ridge)
abline(h=0)
\end{verbatim}

The parameter estimates vs. bias graph shows that as $\lambda$ is increasing, the absolute value of each parameter estimate is shrinking. An informal rule is to select the smallest $\lambda$ which produces stabilized traces of the parameters. In this case, one can chose $\lambda < 10$.

(g) Based on the results of the ridge regression, which variables would you include in the model? Do your conclusions agree with the results of stepwise selection? Why or why not?

\textbf{Answer:}

The ridge regression starts by including all the predictors. According to the trace plot, \texttt{Agriculture} has the smallest absolute value of the standardized coefficient for small-to-moderate $\lambda$, it is therefore the first candidate for exclusion. The result disagrees with the stepwise procedure, which excluded \texttt{Examination}. In this dataset (but not always), the predictor removed by the stepwise procedure corresponds to the smallest absolute value of coefficient $\beta$ when $\lambda = 0$. The fitted regression line is

$$
\hat{\text{Fertility}}_i = -3.386071\text{Agriculture}_i - 2.304663\text{Examination}_i - 7.626079\text{Education}_i \\
+ 3.932303\text{Catholic}_i + 3.142519\text{Mortality}_i
$$
Figure 5: Estimation versus Biasing

ridge2 <- lm.ridge (Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data = swiss, lambda = 1.33)
ridge2$coef

Agriculture Examination Education Catholic Infant.Mortality
-3.386071 -2.304663 -7.626079 3.932303 3.142519

(h) Use bootstrap to produce a 95% confidence interval of the parameter associated with Education, while using the biasing parameter 1.4 in ridge regression. Use $B = 100$ resamplings of observations, and reflection method to calculate the confidence interval.

**Answer:**
This confidence interval is calculated based on ridge regression. Since this is an observational study and we do not have a balanced experimental design, the simplest procedure is to resample pairs of $(X, Y)$, as follows:

```r
# function that samples observations and obtains parameter estimate
getc0ef.boot <- function(x) {
  select.id <- sample(1:nrow(x), nrow(x), replace=TRUE)
  x.boot <- x[select.id,]
  lm.boot <- lm.ridge(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data = x.boot, lambda=1.4)
  lm.boot$coef["Education"]
}
```
# perform the bootstrap procedure B times
B<-100
educ.boot <- rep(NA,B)
for ( i in 1:B){
  educ.boot[i] <- getCoef.boot(swiss)
}

# plot the sampling distribution of the estimated coefficient
hist(educ.boot)

# confidence interval by reflection method
low <- ridge$coef["Education"]-  
  (quantile(educ.boot, 1-0.05/2)-ridge$coef["Education"])
high <- ridge$coef["Education"] +  
  ridge$coef["Education"]-(quantile(educ.boot, 0.05/2))
c(low, high)

The 95% confidence interval of the parameter associated with Education is  
[-13.985534, -3.688392]

Note that, although given as part of the homework, this is not an appropriate  
procedure to estimate the confidence interval of a parameter in conjunction with  
ridge regression. An appropriate procedure will include selection of the biasing  
parameter within each iteration of resampling.

(i) Compare the confidence interval above with the confidence interval for the same  
parameter, and from the same bootstrap resampling, obtained by the percentile  
method. Which method performs best at uncovering the parameter associated  
with Education, which was obtained when using the unbiased estimation of the  
full model?

**Answer:**

The 95% confidence interval obtained from percentile method in the same bootstrap  
sampling is [-14.744907, -2.904874]. This confidence interval is slightly shifted  
towards parameters that are larger in absolute values. The difference is due to the  
fact that percentile-based confidence interval does not account for the bias in the  
estimate by ridge regression.

quantile(educ.boot, c(0.05/2, 1-0.05/2))  
  2.5% 97.5%  
-14.744907 -2.904874
2. Methods Qualifying Exam, January 2011 A tire manufacturer studies the relationship between the tread density and the traction of the tire. He considered three different tread densities and tested three tires for each density, however one test result was lost. The data of the experiment are summarized below, where $Y_{ij}$ denotes the traction measured for the $j$th tire with tread density $X_i$, $i = 1,\ldots,3$ and $j = 1,\ldots,n_i$.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$X_i$ (treads/inch)</th>
<th>$n_i$</th>
<th>$\bar{Y}<em>i = \frac{1}{n_i} \sum</em>{j=1}^{n_i} Y_{ij}$</th>
<th>$\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>3</td>
<td>4.0</td>
<td>0.23</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>2</td>
<td>3.8</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>3.0</td>
<td>3</td>
<td>3.0</td>
<td>0.35</td>
</tr>
<tr>
<td>Mean</td>
<td>$\bar{X}_i = 2$</td>
<td></td>
<td>$\bar{Y}<em>i = \sum</em>{i=1}^{3} n_i \bar{Y}<em>i / \sum</em>{i=1}^{3} n_i = 3.575$</td>
<td></td>
</tr>
</tbody>
</table>

(a) State a simple linear regression model that can be used for these data, and interpret the parameters.

**Answer:**

$Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma^2)$. $\beta_0$ is the intercept and $\beta_1$ is the slope of the linear relationship.

(b) Provide least squares estimates for the parameters $\beta_0$ and $\beta_1$ in the model.

**Answer:**

\[
\hat{\beta}_1 = \frac{\sum_{j=1}^{n_i} n_i (X_i - \bar{X}_i) \bar{Y}_i}{\sum_{j=1}^{n_i} n_i (X_i - \bar{X}_i)^2} = \frac{3(-1)(4) + 2(0)(3.8) + 3(1)(3)}{3(1) + 2(0) + 3(1)} = \frac{-3}{6} = -0.5
\]

\[
\hat{\beta}_0 = \bar{Y}_i - \hat{\beta}_1 \bar{X}_i = 3.575 - (-0.5)(2) = 4.575
\]

(c) The manufacturer would like to conduct the lack of fit test for the simple linear model.

i. State the null and the alternative hypotheses in mathematical notation.

**Answer:**

$H_0 : EY_{ij} = \beta_0 + \beta_1 X_i$, $H_a : EY_{ij} = \mu_i$

ii. Interpret the model used under $H_a$ above, and provide its parameter estimates. Are these estimates unbiased if the simple linear regression model holds? Do they have the smallest variance?
Answer:
The model is a one-way ANOVA, where $\hat{\mu}_i = \bar{Y}_i$. The estimates are unbiased even if linear regression under $H_0$ holds, however they do not have the smallest variance.

iii. The sum of squares of the error of the simple linear regression is 0.96. Conduct the lack of fit test at the confidence level of 95%, and state your conclusion.

Answer:
The test statistic is

$$\frac{[SSE(H_0) - SSE(H_a)]/[df_e(H_0) - df_e(H_a)]}{SSE(H_a)/df_e(H_a)} = \frac{0.96 - (0.23 + 0.14 + 0.35)}{(0.23 + 0.14 + 0.35)/5} = 1.66 < F(1 - 0.05, 1, 5) = 6.6$$

We fail to reject $H_0$, and there is no evidence against the linear model.

3. Researchers study the effect of body size (measured by the “quetelet index” quet), age (age), and smoking history (smk=0 if nonsmoker, and smk=1 if current smoker) on the systolic blood pressure (sbp). The SAS output of the candidate model fit is given below.

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>4938.41058</td>
<td>1646.13686</td>
<td>30.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>1487.55817</td>
<td>53.12708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>31</td>
<td>6425.96875</td>
<td>53.12708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td></td>
<td>7.28883</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dependent Mean</td>
<td></td>
<td>144.53125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coeff Var</td>
<td></td>
<td>5.04308</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Standard Squared Partial

| Variable | DF Estimate | Error t Value | Pr > |t| Type I SS | Type II SS | Corr Type II Tolerance |
|----------|-------------|---------------|------|--------|-----------|----------------------|
| Intercept| 1 47.75825  | 10.42695      | 4.58 | <.0001 | 668457    | 1114.54664           |
| age      | 1 1.10769   | 0.31600       | 3.51 | 0.0016 | 3861.63038| 652.79332            | 0.30499 0.35469     |
| smk      | 1 10.20701  | 2.60153       | 3.92 | 0.0005 | 840.69935 | 817.81834            | 0.35474 0.99680     |
| quet     | 1 9.31297   | 4.41790       | 2.11 | 0.0441 | 236.08086 | 236.08086            | 0.13697 0.35536     |

(a) State the linear model that corresponds to the output above, and the assumptions. Interpret the parameter associated with the variable smk.

Answer:

$\text{sbpi} = \beta_0 + \beta_1 \text{agei} + \beta_2 \text{smki} + \beta_3 \text{queti} + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, 32$. $\beta_2$ is the difference between the expected sbp of smokers and non-smokers, when the values of the other variables are held fixed.
(b) Determine and interpret the partial coefficient of determination of quet, given that smk and age are in the model.

**Answer:**

\[ R^2_{sbp, quet | smk, age} = \frac{SSR(quot | smk, age)}{SSE(smk, age)} = \frac{236.08086}{236.08086 + 1487.55817} = 0.136 \]

quet explains 13% of the overall variation in sbp, after smk, age are added to the model.

(c) Below is the added variable plot for quet. Define the axes of the plot. Based on the plot, can you recommend additional modeling steps?

**Answer:**

X axis: residuals of a linear model with quet and response, and smk and age as predictors. Y axis: residuals of a linear model with sbp and response, and smk and age as predictors. The plot helps assess the functional form of quet in multiple regression. The linear predictor is clearly necessary, however the linearity may be due to the observation with a high \( r(quet | smk, age) \), and low \( r(sbp | smk, age) \). It may be helpful to do additional diagnostic tests for this observation, and include a quadratic term of quet to the model.

(d) Is there any evidence of multicollinearity in this model?

**Answer:**

All values of Tol > .1, indicating that VIF < 10 for all the predictors, and there is no strong evidence of multicollinearity.

(e) Ridge regression is frequently used in presence of multicollinearity. Discuss the advantages and disadvantages of ridge regression for parameter estimation and inference.
Answer:
Advantage: In presence of multicollinearity, ridge regression yields estimates with a smaller MSE. Disadvantages: It yields biased parameter estimates; standard inference procedures are not applicable.