NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 1

1. The data presented in the Table below was published in the March 1, 1984 issue of the Wall Street Journal. It relates to the advertising budget (in millions of dollars) of 21 firms for 1983 and millions of impressions retained per week by the viewers of the products of these firms. The data are based on a survey of 4000 adults in which users of the products were asked to cite a commercial they had seen for the product category in the past week.

IMPACT OF ADVERTISING EXPENDITURE

| Firm | Impressions, millions | Expenditure, millions of 1983 dollars |
|---|---|---|
| 1. Miller Lite | 32.1 | 50.1 |
| 2. Pepsi | 99.6 | 74.1 |
| 3. Stroh's | 11.7 | 19.3 |
| 4. Fed'l Express | 21.9 | 22.9 |
| 5. Burger King | 60.8 | 82.4 |
| 6. Coca Cola | 78.6 | 40.1 |
| 7. McDonald's | 92.4 | 185.9 |
| 8. MCI | 50.7 | 26.9 |
| 9. Diet Cola | 21.4 | 20.4 |
| 10. Ford | 40.1 | 166.2 |
| 11. Levi's | 40.8 | 27.0 |
| 12. Bud Lite | 10.4 | 45.6 |
| 13. ATT/Bell | 88.9 | 154.9 |
| 14. Calvin Klein | 12.0 | 5.0 |
| 15. Wendy's | 29.2 | 49.7 |
| 16. Polaroid | 38.0 | 26.9 |
| 17. Shasta | 10.0 | 5.7 |
| 18. Meow Mix | 12.3 | 7.6 |
| 19. Oscar Meyer | 23.4 | 9.2 |
| 20. Crest | 71.1 | 32.4 |
| 21. Kibbles 'N Bits | 4.4 | 6.1 |

*Source:* http://lib.stat.cmu.edu/DASL/Datafiles/tvadsdat.html

(i) Prepare a `.txt` data file containing the millions of impressions and millions of dollars spent for advertising.

| | impre | adver |
|---|---|---|
| 1 | 32.1 | 50.1 |
| 2 | 99.6 | 74.1 |
| 3 | 11.7 | 19.3 |
| 4 | 21.9 | 22.9 |
| 5 | 60.8 | 82.4 |
| 6 | 78.6 | 40.1 |
| 7 | 92.4 | 185.9 |
| 8 | 50.7 | 26.9 |
| 9 | 21.4 | 20.4 |
| 10 | 40.1 | 166.2 |
| 11 | 40.8 | 27.0 |
| 12 | 10.4 | 45.6 |
| 13 | 88.9 | 154.9 |
| 14 | 12.0 | 5.0 |
| 15 | 29.2 | 49.7 |
| 16 | 38.0 | 26.9 |
| 17 | 10.0 | 5.7 |
| 18 | 12.3 | 7.6 |
| 19 | 23.4 | 9.2 |
| 20 | 71.1 | 32.4 |
| 21 | 4.4 | 6.1 |

The `.txt` file is prepared as follows:

Note: The first column for index is optional. For the other two columns, each should have a heading specifying the variable name. The created data file is named `AdverImpact.txt`.

(ii) Write R codes to do the following:

(iia) Input the data into a data frame.

(iib) Make the variables in the data frame available in the R console.

(iic) Plot millions of impressions on the vertical axis and advertising expenditure on the horizontal axis.

(iid) Fit a simple linear regression model to the data.

(iie) Plot the residual against the fitted values.

```
q1.dat=read.table("D://Rsession/AdverImpact.txt",header=TRUE) # (iia)
y=q1.dat$impre; x=q1.dat$adver                                # (iib)
plot(x, y, xlab='Expenditure', ylab='Impression')             # (iic)
q1.fit = lm(y~x); summary(q1.fit)                             # (iid)
r = q1.fit$resid
fitted = q1.fit$fitted
plot(fitted, r, xlab='Fitted values', ylab='Residuals')       # (iie)
```

(iii) From the results obtained by implementing your R codes, answer the following questions:

(iiic) What is the fitted regression function? Do you think it pays to advertise?

*The fitted results are given below:*

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.16269    7.08948   3.126  0.00556 **
x            0.36317    0.09712   3.739  0.00139 **
```

*From the fitted results, we obtain the fitted regression function as*

$$\hat{y} = 22.1627 + 0.3632x.$$

*(Distinguish between a regression model and a regression function, and between a regression function and an estimated regression function).*

*From the estimated regression function, when $x$ increases to $x+1$, the estimated expected impressions increase from $22.1627 + 0.3632x$ to $22.1627 + 0.3632(x + 1)$. Hence, an increase of one million dollars in advertisement has an expected 0.3532 million of increase in impressions per week. It seems worthy of having the advertisement expenditure.*

2. For 120 graduate students, their ACT and GPA are recorded. The data is given in the file `GPA.txt` on Canvas. Suppose we want to predict GPA based on ACT by a simple linear regression model.

(i) Obtain the fitted regression function.

*The model is fitted by the following codes:*

```
q2.dat = read.table("D://Rsession/GPA.txt"); attach(q2.dat)
q2.fit = lm(GPA~ACT,data=q2.dat); summary(q2.fit)
```

*The fitted results are as follows.*

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
ACT          0.03883    0.01277   3.040  0.00292 **
Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,   Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

*From the fitted results, we have the fitted regression function:*

$$\widehat{GPA} = 2.11405 + 0.03883ACT.$$

3. A criminologist studying the relationship between population density and robbery rate in medium-sized U.S. cities collected the following data for a random sample of 16 cities; $X$ is the population density of the city (number of people per unit area), and $Y$ is the robbery rate last year (number of tobberies per 100,000 people).

| i | 1 | 2 | $\cdots$ | 16 |
|---|---|---|---|---|
| $X_i$ | 59 | 49 | $\cdots$ | 70 |
| $Y_i$ | 209 | 180 | $\cdots$ | 204 |

The full data is provided in the file `robbery.txt` on Canvas.

(i) Assume that a simple linear regression model is appropriate. Obtain the estimated regression function. Plot the estimated regression function and the data. Does the linear regression function appear to give a good fit here? Discuss.

*The R codes for the implementation:*

```
q3.dat = read.table("D://Rsession/robbery.txt")
Y = q3.dat$V1; X = q3.dat$V2
q3.fit = lm(Y~X); summary(q3.fit)
plot(X,Y,xlab="Population density",ylab="Robbery rate")
abline(q3.fit,col="blue")
```

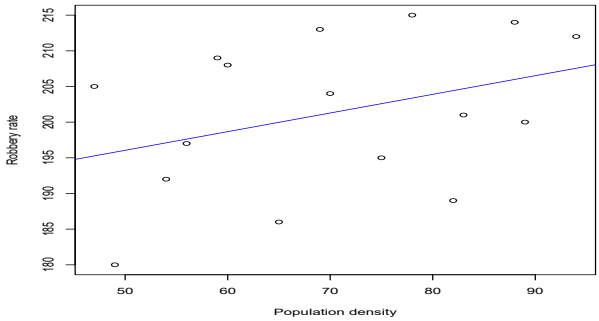
*The fitted results are as follows:*

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 182.9725    12.7223  14.382 8.87e-10 ***
X             0.2616     0.1783   1.467    0.164
Residual standard error: 10.29 on 14 degrees of freedom
Multiple R-squared:  0.1332,    Adjusted R-squared:  0.07132
F-statistic: 2.152 on 1 and 14 DF,  p-value: 0.1645
```

*The estimated regession function is*

$$\hat{Y} = 182.9725 + 0.2616X.$$

*The plot is as follows.*




*The plot shows that the linear relationship is appropriate.*

(ii) Obtain point estimates of the following: (1) the difference in the mean robbery rate in cites that differ by one unit in population density, (2) the mean robbery rate last year in cities with population density $X = 60$, (3) $\epsilon_{10}$, (4) $\sigma^2$.

*(1) The difference is $\beta_1$, the estimate is 0.2616.*

*(2) The estimate is given by* $\hat{\beta}_0 + \hat{\beta}_1 \times 60 = 182.9725 + 0.2616 \times 60 = 198.6685$.

*(3) The value is extracted by* `r = q3.fit$resid; r[10]` *as* $-3.683162$.

*(4)* $\hat{\sigma}^2 = 10.29^2 = 105.8841$.

4. Consider the following regression model:

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}, \qquad i = 1, \ldots, n;\ j = 1, \ldots, n_i.$$

Let $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$. Show that the least squares estimates of $\beta_0$ and $\beta_1$ minimize the following weighted sum of squares:

$$\tilde{Q}(\beta_0, \beta_1) = \sum_{i=1}^{n} n_i (\bar{y}_i - \beta_0 - \beta_1 x_i)^2.$$

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 2

1. The Tri-City Office Equipment Corporation sells an imported desk calculator on a franchise basis and performs preventive maintenance and repair sevice on this calculator. the data have been collected from 18 calls on users to perform routine preventive maintenance service; for each call, $X$ is the number of machines serviced and $Y$ is the total number of minutes spent by the service person. The data is provided in the file `calculator.txt` on Canvas.

   (a) Fit a simple linear regression model to the data and estimate the change in the mean service time when the number of machines serviced increase by one. Construct a 90% confidence interval for the mean change and interpret your confidence interval.

   *The linear regression model is*

   $$Y = \beta_0 + \beta_1 X + \epsilon.$$

   *The model is fitted by the following R codes:*

```
q1.dat = read.table("D://Rsession/calculator.txt",header=TRUE)
y = q1.dat$service.time; x=q1.dat$machine.num
q1.fit = lm(y~x)
summary(q1.fit)
```

   *The fitted results are:*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3221     2.5644  -0.906    0.379
x            14.7383     0.5193  28.383  4.1e-15 ***
Residual standard error: 4.482 on 16 degrees of freedom
Multiple R-squared:  0.9805.
```

*The change in the mean service time when the number of machines serviced increase by one is indeed $\beta_1$. The estimate is $\hat{\beta}_1 = 14.7383$. The confidence interval is computed as*

$$\hat{\beta}_1 \pm t_{16}(0.05) * s_{\hat{\beta}_1} = 14.7383 \pm 1.745664 * 0.5193 = [13.83178,\ 15.64482],$$

*where $t_{16}(0.05)$ is obtained by* `qt(0.95,16)`
*We have 90% confidence that the above interval covers the true value of $\beta_1$.*

(b) Conduct a $t$-test to determine whether or not there is a linear association between $X$ and $Y$, control the type I error rate $\alpha$ at 0.1. State the alternatives, decision rules and conclusion. What is the $p$-value of your test?

*Whether or not there is a linear association between $X$ and $Y$ is equivalent to whether or not $\beta_1 = 0$. The hypotheses to be tested are $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.*
*The test statistic is $T = \hat{\beta}_1 / s_{\hat{\beta}_1} = 28.383$ which can be read from the fitted results.*
*The decision rule is to reject $H_0$ if $|T| > t_{16}(0.05) = 1.7459$; otherwise, do not reject $H_0$.*
*Conclusion: reject $H_0$ since $|T| = 28.383 > 1.7459$.*
*The p-value is 4.1e-15 which can be read from fitted results.*

(c) The manufacturer has given the standard that the mean required time does not increase by more than 14 minutes for each additional machine that is serviced on a service call. Conduct an appropriate test to decide whether this standard is being violated by Tri-City. Control the type I error rate $\alpha$ at 0.05. State the null and alternative hypotheses, decision rule and conclusion. What is the $p$-value of the test?

*The mean required time for each additional machine that is serviced on a service call is $\beta_1$. The standard is equivalent to that $\beta_1 \leq 14$. Since we want to find whether the Tri-city company violates the standard, the hypotheses should be formulated as $H_0 : \beta_1 \leq 14$ vs. $H_1 : \beta_1 > 14$.*
*The test statistic is computed as*

$$T = \frac{\hat{\beta}_1 - 14}{s_{\hat{\beta}_1}} = \frac{14.7383 - 14}{0.5193} = 1.421722.$$

*The decision rule: Reject $H_0$ if $T \geq t_{16}(0.05) = 1.7459$; do not reject $H_0$ otherwise.*
*Conclusion: Do not reject $H_0$ since $T = 1.421754 < 1.7459$. That is, there is no strong evidence to support that the Tri-City company did not meet the standard.*
*p-value is obtained by* `1-pt(1.421754,16)` *which yields 0.08715.*

(d) Obtain a 90% confidence interval for the mean service time on calls in which six machines are serviced.

*The mean service time on calls in which six machines are serviced is* $E(Y|X=6) = \beta_0 + \beta_1 * 6$. *The confidence interval can be computed either by*

```
predict(q1.fit, list(x=6), interval="confidence",level=0.9)
```

*which yields the interval*

$$[83.81593,\ 88.39883],$$

*or by direct computation using the formula:*

$$\hat{Y} \pm s.d(\hat{Y}) t_{16}(0.05),$$

*where* $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * 6 = -2.3221 + 14.7383 * 6 = 86.10738$ *and* $s.d(\hat{Y}) = \sqrt{Var(\hat{Y})}$ *with*

$$Var(\hat{Y}) = Var(\hat{\beta}_0 + \hat{\beta}_1 * 6) = Var(\hat{\beta}_0) + 2 * 6 * Cov(\hat{\beta}_0, \hat{\beta}_1) + 6^2\, Var(\hat{\beta}_1).$$

*The variance matrix of the estimated regression coefficients can be extracted by* `v = vcov(q1.fit)` *as*

```
             (Intercept)          x
(Intercept)    6.575915 -1.2133237
x             -1.213324  0.2696275
```

*Thus*

$$Var(\hat{Y}) = 6.575915 - 12 * 1.2133237 + 36 * 0.2696275 = 1.722621.$$

$t_{16}(0.05)$ *is computed by* `qt(0.95,16)` *as 1.745884.*
*The interval is then*

$$86.10738 \pm \sqrt{1.722621} * 1.745884 = [83.81593,\ 88.39883].$$

(e) Obtain a 90% prediction interval for the mean service time on the next call in which six machines are serviced.

*As in part (d), the interval can be computed either by*

```
predict(q1.fit, list(x=6), interval="prediction",level=0.9)
```

*which yields the interval*

$$[77.95392, 94.26084],$$

*or by direct computation using the formula:*

$$\hat{Y} \pm s.d(Y_{new})t_{16}(0.05),$$

*where* $s.d(Y_{new}) = \sqrt{Var(Y_{new})}$ *with*

$$Var(Y_{new}) = Var(\hat{Y}) + \hat{\sigma}^2 = 1.722621 + 4.482^2 = 21.81095.$$

*The prediction interval is then computed as*

$$86.10738 \pm \sqrt{21.81095} * 1.745884 = [77.95372, 94.26104].$$

*The difference between the two intervals are due to rounding errors.*

(f) Management wishes to estimate the expected service time per machine on calls in which six machines are serviced. Obtain an appropriate 90% confidence interval by converting the interval obtained in part (d).

*The interval is obtained by simply divide the interval in (d) with 6, which gives*

$$[13.96932,\ 14.73314].$$

2. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data for 10 shipments were collected on the number of times the carton was transferred from one aircraft to another over the shipment route ($X$) and the number of ampules found to be broken upon arrival ($Y$). The data is provided in the file `airfreight.txt` on Canvas.

(a) Fit a simple linear regression model, obtain the estimated regression function and the scatter plot of $Y$ against $X$ with the estimated regression function imposed.

*The regression model is*

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

*The estimation and plot are done by the following R code:*

```
q2.dat = read.table("D://Rsession/airfreight.txt",header=TRUE)
y = q2.dat$break.num; x=q2.dat$tran.times
q2.fit=lm(y~x); summary(q2.fit)
plot(x,y); abline(q2.fit)
```

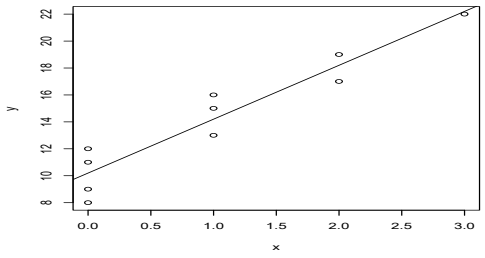
*The fitted summary results are as follows:*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2000     0.6633  15.377 3.18e-07
x             4.0000     0.4690   8.528 2.75e-05
```

*The estimated regression function is*

$$\hat{Y} = 10.2 + 4X.$$

*The plot is as follows:*




(b) Estimate the slope of the regression function with a 95% confidence interval.

*The estimated slope is 4 with standard deviation 0.469, which can be read from the summary fitted results. The 95% confidence interval is computed as*

$$4 \pm 0.469 * 2.306004 = [2.918484, 5.081516],$$

*where* $2.306004 = t_8(0.025)$ *is computed by* `qt(0.975,8)`.

(c) It has been established as a standard that the mean number of broken ampules should not exceed 9 when no transfers are made. Answer whether or not this standard is violated in the shipments by conducting an appropriate test, using $\alpha = 0.025$. State

the null and alternative hypotheses, decision rule, and conclusion. What is the $p$-value of the test?

*From the regression function $EY = \beta_0 + \beta_1 X$, the mean number of broken ampules when no transfers are made is $\beta_0$. Thus, it is equivalent to test $H_0 : \beta_0 \leq 9$ vs. $H_1 : \beta_0 > 9$.*
*The test statistic is*

$$T = \frac{\hat{\beta}_0 - 9}{s_{\hat{\beta}_0}} = (10.2 - 9)/0.6633 = 1.8091,$$

*where $\hat{\beta}_0$ and $s_{\hat{\beta}_0}$ are obtained from the fitted results given earlier.*
*Decision rule: Reject $H_0$ if $T > t_8(0.025) = 2.3060$; do not reject $H_0$ otherwise.*
*Conclusion: Do not reject $H_0$ since $T = 1.8091 < 2.306$.*
*The p-value is given by* `1-pt(1.8091,8)` *which is 0.054.*

3. Shown below are the number of galleys for a manuscript ($X$) and the dollar cost of correcting typographical errors ($Y$) in a random sample of recent orders handled by a firm specializing in technical manuscripts. Assume that the regeression model $Y_i = \beta_1 X_i + \epsilon_i$ is appropriate, with normally distributed independent error terms whose variance is $\sigma^2 = 16$.

| $i$ : | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X_i$ : | 7 | 12 | 4 | 14 | 25 | 30 |
| $Y_i$ | 128 | 213 | 75 | 250 | 446 | 540 |

(a) Give the log likelihood function of $\beta_1$ based on the six observations.

*Under the assumption of normality and the model, the likelihood function is given by*

$$L(\beta_1) = \left(\frac{1}{\sqrt{2\pi 16}}\right)^6 \exp\{-\frac{1}{2 * 16}[(128 - 7\beta_1)^2 + \cdots + (540 - 30\beta_1)^2]\}.$$

*The log likelihood function is given by*

$$\ell(\beta_1) = -(6/2)\ln(2\pi * 16) - (1/32)[(128 - 7\beta_1)^2 + \cdots + (540 - 30\beta_1)^2].$$

(b) Evaluate the likelihood function (ignoring the constant component and factor, i.e., evaluate only $-[(128-7\beta_1)^2+\cdots+(540-30\beta_1)^2]$) for $\beta_1 = 17, 18$ and 19. For which of these $\beta_1$ values is the log likelihood function largest?

*Define the likelihood function in R as*

```
lh.f =function(beta, x, y){
    x=c(7,12,4,14,25,30)
    y=c(128,213,75,250,446,540)
    - sum((y-beta*x)^2)
}
```

*Evaluate the function at 17, 18 and 19 yields*

```
c(lh.f(17); lh.f(18); lh.f(19)) =(-1696, -42, -2248)
```

*The maximum attains at 18.*

(c) Find the maximum likelihood estimate of $\beta_1$.

*The maximum likelihood estimate is the same as the LSE and can be obtained by the R codes:*

```
x=c(7,12,4,14,25,30)
y=c(128,213,75,250,446,540)
q3.fit = lm(y~-1+x); summary(q3.fit)
```

*The fitted results are*

```
  Estimate Std. Error t value Pr(>|t|)
x  17.9285     0.0577   310.7 6.56e-12
```

*The estimate of $\beta_1$ is 17.9285.*

(d) Plot the log likelihood function ignoring the constant component, i.e., $-[(128 - 7\beta_1)^2 + \cdots + (540 - 30\beta_1)^2]$, for values of $\beta_1$ at points equally spaced with distance 0.01 apart in the interval [17, 19]. Does the point at which the log likelihood function is maximized correspond to the maximum likelihood estimate found in (c)?

*The plot is generated by the R codes:*

```
l=NULL
b=seq(17,19,by=0.01)
n=length(b)
for (i in 1:n) {
     l[i]=lh.f(b[i])
}
plot(b,l,type="l")
```

*The maximum point appears around 18, which does correspond to the MLE found in (c).*

Obtain the estimated regression function.

*The model is fitted by the codes below:*

```
q3.dat = read.table("D://Rsession/ChemShip.txt",header=TRUE)
y=q3.dat$time; x1=q3.dat$number; x2=q3.dat$weight
q3.fit=lm(time~number+weight,x=TRUE,data=q3.dat)
summary(q3.fit)
```

*The Fitted results are:*

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3243     3.1108   1.069      0.3
x1            3.7681     0.6142   6.135 1.10e-05
x2            5.0796     0.6655   7.632 6.89e-07
Residual standard error: 5.618 on 17 degrees of freedom
Multiple R-squared: 0.9869,  Adjusted R-squared: 0.9854
```

*The estimated regession function is*

$$\hat{Y} = 3.3243 + 3.7681x_1 + 5.0796x_2.$$

(b) Give the R codes for extracting the following quantities from the fitted object: (1) fitted values, (2) residuals, (3) the estimated regression coefficients, (4) the estimated covariance matrix of the estimated regression coefficients.

*The codes are*

```
yhat=q3.fit$fitted
r = q3.fit$resid
b = q3.fit$coef
V = vcov(q3.fit)
```

(c) Compute the squared Pearson's correlation coefficient between $\boldsymbol{y}$, the observed response values, and $\hat{\boldsymbol{y}}$, the fitted values.

*It is computed by* `cor(y,yhat)^2` *as 0.9869259.*

(d) Compute the squared sample multiple correlation coefficient between $Y$ and $Z = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ by using the formula

$$\text{MCORR}(\boldsymbol{y}, Z)^2 = \frac{\boldsymbol{y}^\top (I - \frac{\mathbf{1}\mathbf{1}^\top}{n}) Z [Z^\top (I - \frac{\mathbf{1}\mathbf{1}^\top}{n}) Z]^{-1} Z^\top (I - \frac{\mathbf{1}\mathbf{1}^\top}{n}) \boldsymbol{y}}{\boldsymbol{y}^\top (I - \frac{\mathbf{1}\mathbf{1}^\top}{n}) \boldsymbol{y}}.$$

Note that $Z^\top(I-\frac{\mathbf{1}\mathbf{1}^\top}{n})Z$ , $Z^\top(I-\frac{\mathbf{1}\mathbf{1}^\top}{n})\boldsymbol{y}$ and $\boldsymbol{y}^\top(I-\frac{\mathbf{1}\mathbf{1}^\top}{n})\boldsymbol{y}$ are the sample variance matrix of $Z$, sample covariance vector between $Z$ and $\boldsymbol{y}$ and the sample variance of $\boldsymbol{y}$ respectively. They can be computed by the R function `var` and `cov`.

*It is computed by the codes:*

```
Z=cbind(x1,x2); v.z=var(Z)
v.zy=cov(Z,y);  v.y=var(y)
mcor2=t(v.zy)%*%solve(v.z)%*%v.zy/v.y
```

inverse of var of Z

*The value* `mcor2` $= 0.9869259$.

(e) Are what you obtained in (d) and (e) the same as $R^2$?

*They are the same as* $R^2$.

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 4 (Solution)

1. In Question 1, Tutorial 3, a small scale experiment for the study of the relation between degree of brand liking ($Y$) and the two variables, moisture content ($X_1$) and sweetness ($X_2$) of the product, was considered. The data was fitted to the multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

The fitted results are given below.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.6500     2.9961  12.566 1.20e-08
x1            4.4250     0.3011  14.695 1.78e-09
x2            4.3750     0.6733   6.498 2.01e-05
Residual standard error: 2.693 on 13 degrees of freedom
Multiple R-squared: 0.9521,  Adjusted R-squared: 0.9447
```

(a) Formulate a hypothesis testing to test whether or not the moisture content of the product has a significant effect on the degree of brand liking at level $\alpha = 0.05$. Give the null and alternative hypotheses, test statistic, and decision rule. Draw your conclusion based on the above fitted results.

*Whether or not the moisture content of the product has a significant effect on the degree of brand liking is equivalent to whether or not $\beta_1$ is not zero.*

*The hypotheses are* **[3 marks]***:*

$$H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0.$$

*Test statistic* **[2 marks]***:* $T = \hat{\beta}_1 / s.d(\hat{\beta}_1)$.

*Decision rule* **[3 marks]** *: if* $|T| \geq t_{13}(0.025)$ *or* $P(|t_{13}| \geq |T|) < 0.05$*, reject* $H_0$*; otherwise, do not reject.*

*Conclusion* **[2 marks]***: From the fitted the results, the p-value of the test is* `1.78e-09`*, which is smaller than 0.05, the* $H_0$ *is rejected.*

(b) The experimenter wishes to find out that the customers would prefer the product more if it is sweeter. Formulate an appropriate one-sided test to see whether this can be confirmed. State your null and alternative hypotheses. Give the $p$-value of the test. Draw your conclusion.

**[5 marks]** *That the customers would prefer the product more if it is sweeter is equivalent to that $\beta_2 > 0$. Since this is what is to be discovered, it should be the alternative hypothesis. So the hypotheses are:*

$$H_0 : \beta_2 \leq 0, \ H_1 : \beta_2 > 0.$$

**[5 marks]** *The p-value of the test is $p = P(t_{13} > 6.498)$ which is computed by* `1-pt(6.498,13)` *as* `1.004918e-05` *which can also be obtained from the p-values of the two-sided test given in the summary table. The null hypothesis is rejected because of the small p-value, and hence it can be confirmed that the customers would prefer the product more if it is sweeter.*

(c) Construct a 95% confidence lower bound for $\beta_2$.

**[5 marks]** *The confidence bound is given by*

$$\beta_2 \geq 4.375 - 0.6733 \times 1.771 = 3.1826,$$

*where* `1.771 = qt(0.95, 13)`.

2. In Question 3, Tutorial 3, we analyzed the data taken on 20 incoming shipments of chemicals in drums arriving at a warehouse. The data contains the number of drums in shipment ($X_1$), total weight of shipment ($X_2$, in hundred pounds), and number of minutes required to handle shipment ($Y$). The data given in the file `ChemShip.txt` was fitted to the following regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, i = 1, \ldots, 20.$$

The Fitted results are given below:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3243     3.1108   1.069      0.3
x1            3.7681     0.6142   6.135 1.10e-05
x2            5.0796     0.6655   7.632 6.89e-07
Residual standard error: 5.618 on 17 degrees of freedom
Multiple R-squared: 0.9869,  Adjusted R-squared: 0.9854
```

(a) Without looking at the data, what would you suspect on the relationship between the time required to handle shipment, $Y$, and the total weight of shipment, $X_2$? In other words, do they have a positive or an negative relationship? Formulate an appropriate hypothesis testing to test the relationship. State your null and alternative hypotheses and draw your conclusion.

**[6 marks]** *More weight needs more time to handle. It can be suspected that the relationship between $Y$ and $X_2$ is positive, that is, $\beta_2 > 0$. Since this is what we need to discover, the hypotheses should be formulated as follows:*

$$H_0 : \beta_2 \leq 0, \ H_1 : \beta_2 > 0.$$

**[6 marks]** *The value of the test statistic $T$ for this test can be obtained from the fitted results as $T = 7.632$.. The p-value is computed as:* `1-pt(7.632,17)=3.4442e-07`*, which can also be obtained from the p-value of the two-sided test given in the summary table..*

*The null hypothesis is rejected because of the small p value.*

(b) The shipment company wanted to find out that, given the total weight of the shipment, the time needed for handling an additional drum should not exceed 5 minutes. Formulate the hypotheses properly to test whether this can be confirmed at level $\alpha = 0.05$. Provide your null and alternative hypotheses, test statistic, and decision rule. Draw your conclusion.

**[6 marks]** *When the total weight of the shipment is fixed, the time required for handling an additional drum is $\beta_1$. It is to be found that $\beta_1 < 5$, the hypotheses are formulated as*

$$H_0 : \beta_1 \geq 5, \ H_1 : \beta_1 < 5.$$

**[2 marks]** *The test statistic:*

$$T = \frac{\hat{\beta}_1 - 5}{\hat{\sigma}_1} = \frac{3.7681 - 5}{0.6142} = -2.0057.$$

**[3 marks]** *Decision rule: either $T < -t_{17}(0.05)$ or $P(t_{17} < T) < 0.05$, reject $H_0$, otherwise, do not reject.*

**[2 marks]** *The p value of the test is* `pt(-2.0057,17)= 0.0305`*. Since the p value is smaller than 0.05, $H_0$ is rejected.*

3. A consumer organization studied the effect of age and gender of automobile owner on size of cash offer for a used car by using six persons in each of age-gender groups who acted as the owner of a used car. A medium price, six-year-old car was selected for the experiment, and the "owners" solicited cash offers for this car from 36 dealers selected at random from the dealers in the region. Randomization was used in assigning the dealers to the "owners". The offers (in hundred dollars) are given in the following table.

| | Age group | | | Gender |
|---|---|---|---|---|
| Gender | Young | Middle | Elderly | mean |
| Male | 21 | 30 | 25 | |
| | 23 | 29 | 22 | |
| | 19 | 26 | 23 | |
| | 22 | 28 | 21 | |
| | 22 | 27 | 22 | |
| | 23 | 27 | 21 | |
| Cell mean | 21.67 | 27.83 | 22.33 | 23.94 |
| Female | 21 | 26 | 23 | |
| | 22 | 29 | 19 | |
| | 20 | 27 | 20 | |
| | 21 | 28 | 21 | |
| | 19 | 27 | 20 | |
| | 25 | 29 | 20 | |
| Cell mean | 21.33 | 27.67 | 20.50 | 23.17 |
| Age mean | 21.5 | 27.75 | 21.42 | |
| Total mean | | | | 23.56 |

(a) Compute the ANOVA table.

*A mini R program is provided in the file* `ST3131hw4.R` *for the direct computation of the ANOVA table. The ANOVA table is as follows:*

| *Source* | *d.f.* | *SS* | *MS* | *F* |
|---|---|---|---|---|
| *Age* | *2* | *316.722* | *158.361* | *66.291* |
| *Gender* | *1* | *5.444* | *5.444* | *2.279* |
| *Interaction* | *2* | *5.056* | *2.528* | *1.058* |
| *Error* | *30* | *71.667* | *2.389* | |

(b) At level $\alpha = 0.05$, test whether or not the interaction between gender and age group is significant. Provide the $p$-value of the test.

*The test statistic*

$$F = \frac{MSAG}{MSE} = 1.058.$$

*The p-value of the test is* `1-pf(1.058,2,30) = 0.3597`. *The interaction is not significant at level* $\alpha = 0.05$.

(c) At level $\alpha = 0.05$, test whether or not the main effect of each of the two factors, gender and age group, is significant. Provide the $p$-values of the tests.

*The test statistics for the main effect of Age group and Gender are, respectively,*

$$F_A = \frac{MSA}{MSE} = 66.291, \text{ and } F_A = \frac{MSG}{MSE} = 2.279.$$

*The p-values are, respectively,* `1-pf(66.291,2,30) = 9.789e-12` *and* `1-pf(2.279,1,30)` *= 0.1416. The main effect of age group is significant, but that of gender is not at level* $\alpha = 0.05$.

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 5

1. In studying the effect of age and gender of automobile owner on size of cash offer for a used car considered in Question 3, Tutorial 4, the data originally collected is corrupted and two records are missing, as shown below:

| | Age group | | |
|---|---|---|---|
| Gender | Young | Middle | Elderly |
| Male | 21 | 30 | 25 |
| | 23 | 29 | 22 |
| | 19 | 26 | 23 |
| | 22 | 28 | |
| | 22 | 27 | 22 |
| | 23 | 27 | 21 |
| Female | 21 | 26 | 23 |
| | 22 | 29 | 19 |
| | 20 | 27 | 20 |
| | 21 | 28 | 21 |
| | | 27 | 20 |
| | 25 | 29 | 20 |

**(i)** Create a data file containing the variables `offer`, `age` and `gender` where `offer` is the size of cash offer. Using the data, fit an interaction model and test whether or not there is significant interaction between age and gender at level 0.05.

*The contents of the data file is as follows:*

```
offer     age gender
1     21   Young      M
2     23   Young      M
3     19   Young      M
4     22   Young      M
5     22   Young      M
6     23   Young      M
```

```
7     21   Young      F
8     22   Young      F
9     20   Young      F
10    21   Young      F
11    25   Young      F
12    30  Middle      M
13    29  Middle      M
14    26  Middle      M
15    28  Middle      M
16    27  Middle      M
17    27  Middle      M
18    26  Middle      F
19    29  Middle      F
20    27  Middle      F
21    28  Middle      F
22    27  Middle      F
23    29  Middle      F
24    25 Elderly      M
25    22 Elderly      M
26    23 Elderly      M
27    22 Elderly      M
28    21 Elderly      M
29    23 Elderly      F
30    19 Elderly      F
31    20 Elderly      F
32    21 Elderly      F
33    20 Elderly      F
34    20 Elderly      F
```

*The interaction model yields the following anova table:*

```
Df  Sum Sq Mean Sq F value    Pr(>F)
gender      1   4.235   4.235  1.8824    0.1810
age         2 294.790 147.395 65.5089 2.754e-11 ***
gender:age  2   8.092   4.046  1.7983    0.1842
Residuals  28  63.000   2.250
```

*The p-value for the test of interaction is 0.1842. The interaction is not significant at level 0.05.*

**(ii)** Based on properly fitted models, make inference on the overall main effcts of `age`

and gender.

*To make inference on the4 overall main effects, we need to fit the interaction model twice, each time with a different order of the factors, to get the ANOVA table with valid F-statistics.*

*The ANOVA table obtained by fitting the model with order:* age, gender, *is as follows:*

```
           Df  Sum Sq Mean Sq F value    Pr(>F)
age         2 294.959 147.479 65.5463 2.736e-11 ***
gender      1   4.067   4.067  1.8074    0.1896
age:gender  2   8.092   4.046  1.7983    0.1842
Residuals  28  63.000   2.250
```

unbalanced data need fit in both order
balanced data order doesn't matter

*From this table, we can make inference on the main effect of* gender. *The p-value of the F-test, 0.1896, is large and we can not conclude that the main effect of* gender *is significant.*

*The ANOVA table obtained by fitting the model with order:* gender, age, *is as follows:*

```
           Df  Sum Sq Mean Sq F value    Pr(>F)
gender      1   4.235   4.235  1.8824    0.1810
age         2 294.790 147.395 65.5089 2.754e-11 ***
gender:age  2   8.092   4.046  1.7983    0.1842
Residuals  28  63.000   2.250
```

*From this table, we can make inference on the main effect of* age. *The p-value of the F-test,* 2.754e-11. *With this small p-value, we conclude that the main effect of* age *is significant.*

2. Consider the problem dealt with in Question 1. A main-effect model was fitted to the data, which produces the following results:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.1393     0.5228  40.433  < 2e-16
genderM       0.6935     0.5294   1.310    0.200
ageMiddle     6.2639     0.6430   9.741 8.39e-11
ageYoung      0.2097     0.6582   0.319    0.752
Residual standard error: 1.539 on 30 degrees of freedom
Multiple R-squared: 0.8079,  Adjusted R-squared: 0.7887
F-statistic: 42.06 on 3 and 30 DF,  p-value: 7.248e-11
```

Variance matrix extracted from the main effect model:

```
             (Intercept)     genderM   ageMiddle    ageYoung
(Intercept)   0.2733431 -0.12740569 -0.20964027 -0.20384910
genderM      -0.1274057  0.28029252 -0.01274057 -0.02548114
ageMiddle    -0.2096403 -0.01274057  0.41348938  0.21658967
ageYoung     -0.2038491 -0.02548114  0.21658967  0.43317935
```

The residual standard error of the interaction model is `sigma.i = 1.5`.

**(i)** Test at level $\alpha = 0.05$ whether or not middle age group and elderly age group have a significant difference in cash offer ignoring their genders.

*Let the model be denoted by*

$$Y = \beta_0 + \alpha_2 U_2 + \beta_2 V_2 + \beta_3 V_3 + \epsilon,$$

*where* $U_2$ *is the dummy variable for* `gender`, $V_2$ *and* $V_3$ *are the dummy variables for* `age`. *Note that* `elderly` *is the first level of* `age`.
*The difference between middle age group and elderly age group is* $\beta_2$. *The hypotheses of the test are:*

$$H_0 : \beta_2 = 0, \ H_1 : \beta_2 \neq 0.$$

*The test statistic is given by*

$$T = \hat{\beta}_2 / sd(\hat{\beta}_2),$$

SE*sigma.l/sigma.M

*where* $\hat{\beta}_2 = 6.2639$, $sd(\hat{\beta}_2) = 0.643 * 1.5/1.539 = 0.6267$. *Hence* $T = 6.2639/0.6267 = 9.9951$.
*The p-value of the test is given by* $2P(t_{28} \geq 9.9951) =$`9.729518e-11` $< 0.05$. *The null hypothesis is rejected.*

**(ii)** Test at level $\alpha = 0.05$ whether or not middle age group and young age group have a significant difference in cash offer ignoring their genders.

*The difference if given by* $\beta_2 - \beta_3$. *The hypotheses:*

$$H_0 : \beta_2 - \beta_3 = 0, \ H_1 : \beta_2 - \beta_3 \neq 0.$$

*The test statistic:*

$$T = (\hat{\beta}_2 - \hat{\beta}_3) / sd(\hat{\beta}_2 - \hat{\beta}_3).$$

interaction model: effect of factor A and B
main effect model: effect of each level of factor A and B (need fit interaction model to get adjusted SD)
use main effect if want get effect of A while ignoring effect of B

*The variance of* $\hat{\beta}_2 - \hat{\beta}_3$ *from the main-effect model is computed as*

$$V.m = 0.4135 + 0.4332 - 2 * 0.21166 = 0.4234.$$

*The sd adjusted by the residual standard error of the interaction model is then*

$$sd(\hat{\beta}_2 - \hat{\beta}_3) = \sqrt{0.4234} * 1.5/1.539 = 0.6342.$$

*Now,* $T = (6.2639 - 0.2097)/0.6342 = 9.5462$.
*The p-value of the test is given by* $2P(t_{28} \geq 9.5462)$ =`2.656055e-10` $< 0.05$. *The null hypothesis is rejected.*

3. A clinical trial for investigating the effects of two operations, A and B, yields the following data:

| A | B | Measurements | $n_{ij}$ |
|---|---|---|---|
| no | no | 31.02 24.03 28.37 26.57 25.83 25.87 26.08 24.89 28.32 29.98 20.41 20.31 23.45 | 13 |
| yes | no | 35.71 34.83 35.13 38.30 33.12 42.07 38.69 43.16 33.11 42.47 40.45 31.95 38.69 42.26 | 14 |
| no | yes | 25.85 39.77 32.88 46.29 33.07 29.23 31.39 28.28 22.43 38.42 | 10 |
| yes | yes | 49.47 42.29 48.60 47.10 33.63 32.00 49.31 50.81 | 8 |

**(i)** Test whether or not there is a significant interaction effect between A and B at level $\alpha = 0.05$. Give the $p$-value of the test.

**(ii)** Test whether or not there are significant main effects of A and B at level $\alpha = 0.05$. Give the $p$-values of the tests.

*Create the response vector* `y` *and factors* `A` *and* `B`. *Fit the interaction model twice, first by specifying the formula as* `y~A*B` *and second by* `y~B*A`. *The interaction effect can be inferred from both ANOVA tables generated by the* `anova` *function. The main effect of* `B` *can be inferred from the ANOVA table of the first fitting and that of* `A` *from the second fitting.*
*The ANOVA tables of the model fitted with different orders of A and B are given as follows:*

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
A          1 1443.03 1443.03  50.754 1.101e-08
B          1  476.22  476.22  16.750  0.000195
A:B        1    1.25    1.25   0.044  0.834902
Residuals 41 1165.70   28.43
```

```
           Df  Sum Sq Mean Sq F value    Pr(>F)
B           1  361.27  361.27  12.707 0.0009424
A           1 1557.98 1557.98  54.797  4.48e-09
B:A         1    1.25    1.25   0.044 0.8349023
Residuals 41 1165.70   28.43
```

*From the above ANOVA tables, we obtained the F statistics for the interaction and main effects:*

$$F_{AB} = 0.044,\ F_A = 54.797,\ F_B = 16.750.$$

*Their corresponding p-values:*

$$p_{AB} = 0.8349023,\ p_A = \texttt{4.48e-09},\ p_B = 0.000195.$$

*At level 0.05, the interaction effect is not significant and the two main-effects are significant by comparing the p-values with 0.05.*

**(iii)** At level $\alpha = 0.05$, test whether the effect of operation A is the same when operation B is not applied and when operation B is applied together with operation A. How is this test related to that in part (i)?

*The model can be expressed as*

$$Y = \beta_0 + \alpha_2 A_2 + \beta_2 B_2 + \gamma_{22} A_2 B_2 + \epsilon.$$

*The difference of the effect of A when B is not applied and when B is applied is $\gamma_{22}$. It amounts to test*

$$H_0 : \gamma_{22} = 0,\ H_1 : \gamma_{22} \neq 0.$$

*It can be tested by the t-statistic $\hat{\gamma}_{22}/sd(\hat{\gamma}_{22})$. The value of the t-statistic and the p-value of the test can be obtained from the fitted results:*

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.7792     1.4789  17.432  < 2e-16 ***
AY           12.0736     2.0537   5.879 6.43e-07 ***
BY            6.9818     2.2428   3.113  0.00337 **
AY:BY        -0.6834     3.2581  -0.210  0.83490
```

*The p-value is 0.8349. The null hypothesis is not rejected.*

*This test is equivalent to the test in (i), since $\gamma_{22}$ is the only independent interaction contrast.*

4. An engineer designed a battery by using three types of materials and he tested the life of the battery under three temperatures in an experiment. The following are the data from the experiment:

Life (in hours) data for the Battery Design

| Material Type | Temperature | | | | | |
|---|---|---|---|---|---|---|
| | 15 | | 70 | | 125 | |
| 1 | 130 | 155 | 34 | 40 | 20 | 70 |
| | 74 | 180 | 80 | 75 | 82 | 58 |
| 2 | 150 | 188 | 136 | 122 | 25 | 70 |
| | 159 | 126 | 106 | 115 | 58 | 45 |
| 3 | 138 | 110 | 174 | 120 | 96 | 104 |
| | 168 | 160 | 150 | 139 | 82 | 60 |

**(i)** Create the vector of life time and factor vectors for material type and temperature in R such that they can be used in the function `lm`.

```
y=c(130,74,150,159,138,168,155,180,188,126,110,
    160,34,80,136,106,174,150,40,75,122,115,120,
    139,20,82,25,58,96,82,70,58,70,45,104,60)
Type=factor(rep(c(1,1,2,2,3,3),6))
Temp=factor(c(rep(15,12),rep(70,12),rep(125,12)))
```

**(ii)** The following model is assumed for the data:

$$y = \mu_0 + \sum_{i=2}^{3} \alpha_i u_i + \sum_{j=2}^{3} \beta_j v_j + \sum_{i=2}^{3}\sum_{j=2}^{3} \gamma_{ij} u_i v_j + \epsilon,$$

where $u_i$'s and $v_j$'s are dummy variables defined for material type and temperature respectively in the conventional way.

**(iia)** Interpret the meaning of $\gamma_{ij}$, $i = 2, 3, j = 2, 3$.

$$\gamma_{ij} = (\mu_{ij} - \mu_{i1}) - (\mu_{1j} - \mu_{11}),$$

*which is the difference between the effect of factor B when its level change from 1 to j while factor A is fixed at level i and that while factor A is fixed at level 1.*

**(iib)** Let $\mu_{ij}$ denote the cell mean when the material type is at level $i$ and temperature is at level $j$. Express the following interaction effects in terms of $\gamma_{ij}$'s:

$$\mu_{23} - \mu_{21} - \mu_{33} + \mu_{31},$$
$$\mu_{23} - \mu_{22} - \mu_{33} + \mu_{32}.$$

$$\mu_{23} - \mu_{21} - \mu_{33} + \mu_{31} = \gamma_{23} - \gamma_{33},$$
$$\mu_{23} - \mu_{22} - \mu_{33} + \mu_{32} = \gamma_{23} - \gamma_{22} - \gamma_{33} + \gamma_{32}.$$

**(iii)** Make inference on the overall interaction effect and the main effect of material type and temperature. Draw your conclusion based on $p$-values by your own judgment.

*The fitted interaction model has the following ANOVA table:*

```
           Df Sum Sq Mean Sq F value    Pr(>F)
Type        2  10684  5341.9  7.9114  0.001976 **
Temp        2  39119 19559.4 28.9677 1.909e-07 ***
Type:Temp   4   9614  2403.4  3.5595  0.018611 *
Residuals  27  18231   675.2
```

*Since the data is balanced, both the interaction effect and the two main effects can be inferenced by using this table. By the small p-values, all these effect are significant.*

**(iv)** Test whether or not there is a significant difference in life for batteries made with material type 3 and 2 under the temperature 125.

*Note that 125 is the third level of temperature. The difference in life for batteries made with material type 3 and 2 under the temperature 125 is $\mu_{33} - \mu_{23}$. In terms of the regression parameters,*

$$\mu_{33} = \beta_0 + \alpha_3 + \beta_3 + \gamma_{33},\ \mu_{23} = \beta_0 + \alpha_2 + \beta_3 + \gamma_{23}.$$

*Hence*

$$\mu_{33} - \mu_{23} = \alpha_3 - \alpha_2 + \gamma_{33} - \gamma_{23}.$$

*where $\alpha_2, \alpha_3, \gamma_{23}, \gamma_{33}$ correspond to the components 2,3,8,9 of the regression coefficients in the fitted results of* `lm(y~Type*Temp)` *as shown below:*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    134.75      12.99  10.371 6.46e-11 ***
Type2           21.00      18.37   1.143 0.263107
```

```
Type3            9.25      18.37   0.503 0.618747
Temp70         -77.50      18.37  -4.218 0.000248 ***
Temp125        -77.25      18.37  -4.204 0.000257 ***
Type2:Temp70    41.50      25.98   1.597 0.121886
Type3:Temp70    79.25      25.98   3.050 0.005083 **
Type2:Temp125  -29.00      25.98  -1.116 0.274242
Type3:Temp125   18.75      25.98   0.722 0.476759
```

*Let* $b = (\hat{\alpha}_2, \hat{\alpha}_3, \hat{\gamma}_{23}, \hat{\gamma}_{33})^\top$ *and* $V$ *be the estimated variance matrix of* $b$*. Let* $d = (-1, 1, -1, 1)^\top$*. The test statistic is given by*

$$T = d^\top b / \sqrt{d^\top V d},$$

*where* $b$ *and* $V$ *are extracted from* `q4.fit` *as*

```
b=q4.fit$coet[c(2,3,8,9)]
V=vcov(q4.fit)[c(2,3,8,9), c(2,3,8,9)]
```

*It is computed that* $T = 1.9593$*. The p-value is computed as* $2P(t_{27} > 1.9593) = 0.06048$*. We can claim that the life for batteries made with material type 3 and 2 under the temperature 125 is not significantly different.*

**(v)** Test whether or not there is a significant difference in average life across all temperatures for batteries made with material type 3 and 2. ignoring other factor

*To answer this question, a main-effect model is fixed giving the following results:*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   122.47      11.17  10.965 3.39e-12 ***
Type2          25.17      12.24   2.057  0.04819 *
Type3          41.92      12.24   3.426  0.00175 **
Temp70        -37.25      12.24  -3.044  0.00472 **
Temp125       -80.67      12.24  -6.593 2.30e-07 ***
Residual standard error: 29.97 on 31 degrees of freedom
Multiple R-squared:  0.6414,    Adjusted R-squared:  0.5951
F-statistic: 13.86 on 4 and 31 DF,  p-value: 1.367e-06

vcov(main.fit)
            (Intercept)    Type2    Type3   Temp70  Temp125
(Intercept)    124.7515 -74.8509 -74.8509 -74.8509 -74.8509
Type2          -74.8509 149.7018  74.8509   0.0000   0.0000
```

```
Type3        -74.8509  74.8509 149.7018   0.0000   0.0000
Temp70       -74.8509   0.0000   0.0000 149.7018  74.8509
Temp125      -74.8509   0.0000   0.0000  74.8509 149.7018
```

*The test statistic in terms of the parameters in the main effect model is given by*

$$\begin{aligned} T &= (\hat{\alpha}_3 - \hat{\alpha}_3)/\sqrt{Var(\hat{\alpha}_2) + Var(\hat{\alpha}_3) - 2Cov(\hat{\alpha}_2, \hat{\alpha}_3)} \\ &= (41.92 - 25.17)/(\sqrt{149} * \sqrt{675.2}/29.97) \\ &= 1.5827. \end{aligned}$$

*The p-value is computed as* `2*pt(1.5827,27,lower.tail=FALSE) =  0.1251`. *There is no significant difference in average life across all temperatures for batteries made with material type 3 and 2 at level* $\alpha < 0.1251$.

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 6

**Solution to Question 1 & 2 are to be submitted for grading. The submission deadline is 2:00pm, Wednesday, October 12.**

1. For the clinical trial considered in Tutorial 5, Question 3, treat the operation as a single factor and the four combinations of operation A and B as the levels, i.e., treat (no,no), (yes, no), (no, yes) and (yes, yes) respectively as level 1, 2, 3 and 4. The data is re-described as follows:

| Level | Measurements | $n_i$ |
|---|---|---|
| 1 | 31.02 24.03 28.37 26.57 25.83 25.87 26.08 24.89 28.32 29.98 20.41 20.31 23.45 | 13 |
| 2 | 35.71 34.83 35.13 38.30 33.12 42.07 38.69 43.16 33.11 42.47 40.45 31.95 38.69 42.26 | 14 |
| 3 | 25.85 39.77 32.88 46.29 33.07 29.23 31.39 28.28 22.43 38.42 | 10 |
| 4 | 49.47 42.29 48.60 47.10 33.63 32.00 49.31 50.81 | 8 |

**(i)** [**5 marks**] Write R-codes to fit the following model:

$$y = \beta_0 + \beta_2 u_2 + \beta_3 u_3 + \beta_4 u_4 + \epsilon,$$

where $u_j$'s are the dummy variable defined in the conventional way for the factor operation. Provide the summary table and the estimated variance matrix of the estimated regression parameters.

*The R codes and the results are as follows:*

```
y1=c(31.02,24.03,28.37,26.57,25.83,25.87,26.08,24.89,28.32,29.98,20.41,20.31,23.45)
y2=c(35.71,34.83,35.13,38.30,33.12,42.07,38.69,43.16,33.11,42.47,40.45,31.95,38.69,42.26)
y3=c(25.85,39.77,32.88,46.29,33.07,29.23,31.39,28.28,22.43,38.42)
y4=c(49.47,42.29,48.60,47.10,33.63,32.00,49.31,50.81)
y = c(y1,y2,y3,y4); Operation=factor(rep(c(1,2,3,4), c(13,14,10,8)))
q1.fit =lm(y~Operation); summary(q1.fit); vcov(q1.fit)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   25.779      1.479  17.432  < 2e-16
Operation2    12.074      2.054   5.879 6.43e-07
Operation3     6.982      2.243   3.113  0.00337
Operation4    18.372      2.396   7.668 1.91e-09
```

```
            (Intercept) Operation2 Operation3 Operation4
(Intercept)    2.187047  -2.187047  -2.187047  -2.187047
Operation2    -2.187047   4.217877   2.187047   2.187047
Operation3    -2.187047   2.187047   5.030209   2.187047
Operation4    -2.187047   2.187047   2.187047   5.740999
```

**(ii)** [**10 marks**] Do a general exploring to find out whether there are any significant contrasts among the four treatment means. Identify at least FOUR apparently significant contrasts. Give reasons why you identify them. Test their significance by controlling the family-wise type I error rate at level $\alpha = 0.05$.

*From the summary table, it can be identified that $\beta_2, \beta_3$ and $\beta_4$ are individually significant, and it can also be guessed that $\beta_3 - \beta_4$ is individually significant since according to the significance of $\beta_3$ a difference of amount 6.982 between two levels is individually significant and the difference between level 3 and 4 exceeds this amount. Thus, the following contrasts are apparently significant:*

$$\mu_2 - \mu_1,\ \mu_3 - \mu_1,\ \mu_4 - \mu_1,\ \mu_3 - \mu_4.$$

take largest gap btw estimates as contrast

*The test statistics are:*

$$\begin{aligned} T_{i1} &= \frac{\hat{\beta}_i}{s.d.(\hat{\beta}_i)},\ i = 2, 3, 4; \\ T_{34} &= \frac{\hat{\beta}_3 - \hat{\beta}_4}{\sqrt{Var(\hat{\beta}_3)) + Var(\hat{\beta}_4) - 2Cov(\hat{\beta}_3, \hat{\beta}_4)}}. \end{aligned}$$

*The values of the statistics are computed by using the information given in the summary table and the estimated variance matrix as:*

$$T_{21} = 5.8788,\ T_{31} = 3.1130,\ T_{41} = 7.6677,\ T_{34} = -4.5034.$$

*Since it is a general exploring, Scheffe's criterion is to be used. The Scheffe's criterion at $\alpha = 0.05$ is $\sqrt{3f_{3,41}(0.05)} = 2.9152$. Comparing the absolute values of the test statistics with Scheffe's criterion, all the identified contrasts are simultaneously significant.*

**(iii)** [**5 marks**] Construct 95% simultaneous confidence intervals for the contrasts identified in (ii).

*Let $L$ be the estimated contrast, $s_L$ its estimated standard deviation, the simultaneous confidence interval is given by*

$$[L - 2.9152 s_L,\quad L + 2.9152 s_L].$$

*For the four contrasts identified in (ii), the simultaneous c.i. are:*

$C_{21}$ [6.0866, 18.0607]
$C_{31}$ [0.4436, 13.5200]
$C_{41}$ [11.3872, 25.3569]
$C_{34}$ [−18.7635, −4.0170].

**(iv)** The clinician was only concerned with whether or not operation A and B have significant main effect and whether or not there is significant interaction between the two operations. These correspond to whether or not the following contrasts are significantly nonzero: (a) $\mu_1 + \mu_3 - (\mu_2 + \mu_4)$, (b) $\mu_1 + \mu_2 - (\mu_3 + \mu_4)$ and (c) $\mu_1 - \mu_2 - (\mu_3 - \mu_4)$.

no interaction -> incr in A parallel across levels of B

**(iva)** [**5 marks**] Express the contrasts in terms of the regression parameters.

*In terms of the regression coefficients, the contrasts are expressed as*

$$\beta_3 - (\beta_2 + \beta_4),\ \beta_2 - (\beta_3 + \beta_4),\ -\beta_2 - (\beta_3 - \beta_4).$$

**(ivb)** [**5 marks**] Test the significance of the contrasts using a proper criterion to control the overall type I error rate at $\alpha = 0.05$.

*For each of the contrast, the test statistic is computed as*

$$\boldsymbol{c}^\top \boldsymbol{b}/\sqrt{\boldsymbol{c}^\top V \boldsymbol{c}},$$

*where* $\boldsymbol{c}$ *is the vector of the coefficients in the linear combination* $c_2\beta_2+c_3\beta_3+c_4\beta_4$ *for the contrast,* $\boldsymbol{b} = (\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)^\top$ *and* $V$ *is the estimated variance matrix of* $\boldsymbol{b}$*. The values for the test statistics of the three contrasts are:*

$$T_1 = -7.2018,\ T_2 = -4.0761,\ T_3 = -0.2097.$$

*Since the contrasts are pre-specified, Bonferroni's, method is to be used to control the overall type I error rate. The Bonferroni critical value is* $t_{41}(0.05/6) = 2.4962$*. The two main effect contrasts are significant and that of the interaction is not significant by comparing the* $T$*-values with the critical value.*

2. The problem dealt with in Question 1, Tutorial 5, is re-considered here. The results from fitting a main-effect model are given below:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.1393     0.5228  40.433  < 2e-16
```

```
genderM       0.6935     0.5294   1.310    0.200
ageMiddle     6.2639     0.6430   9.741 8.39e-11
ageYoung      0.2097     0.6582   0.319    0.752
Residual standard error: 1.539 on 30 degrees of freedom

Variance matrix extracted from the main effect model:
            (Intercept)     genderM   ageMiddle    ageYoung
(Intercept)   0.2733431 -0.12740569 -0.20964027 -0.20384910
genderM      -0.1274057  0.28029252 -0.01274057 -0.02548114
ageMiddle    -0.2096403 -0.01274057  0.41348938  0.21658967
ageYoung     -0.2038491 -0.02548114  0.21658967  0.43317935
```

The residual standard error of the interaction model is `sigma.i = 1.5`.

**(i)** [**10 marks**] Make a pairwise comparison of the age group means by using Tukey's criterion to find out the pairs which are significantly different at level $\alpha = 0.05$. Construct 95% simultaneous confidence intervals for all the pairs using Tukey's criterion.

*The test statistics for the three pairs are computed as follows:*

$$\begin{aligned} T_{21} &= 6.2039/0.6430/(1.5/1.539) = 9.741/1.5 * 1.539 = 9.9943, \\ T_{31} &= 0.319/1.5 * 1.539 = 0.3273; \\ T_{23} &= (6.2639 - 0.2097)/\sqrt{0.4135 + 0.4332 - 2 * 0.2166}/1.5 * 1.539 \\ &= 9.6598. \end{aligned}$$

*From the studentized range distribution table,* $q_{3,28}(0.05) = 3.499$. *The critical value which the t-statistics should be compared with is* $3.499/\sqrt{2} = 2.4742$. *The difference between Middle age and elderly, between Middle age and Young are significant, but not that between Young and elderly.*

*The confidence intervals are computed as*

$$\begin{aligned} (2,1): &\quad [4.7133,\ 7.8145]; \\ (3,1): &\quad [-1.3775,\ 1.7969]; \\ (2,3): &\quad [4.5036,\ 7.6048]. \end{aligned}$$

**(ii)** [**10 marks**] For the problem above, in addition to Tukey's criterion, discuss whether or not Scheffe's criterion and Bonferroni method can also control the family-wise

type I error rate, provide your reasons. Among the criteria which can control the family-wise type I error rate, which one is the best? and why?

*Both criteria can control the family-wise type I error rate. For Scheffe's criterion, it controls the overall type I error rate for all possible contrasts which include the pairwise contrasts. For the Bonferroni method, it is because the pairwise contrasts are the only ones of concern.*

*For pairwise comparison, Tukey's criterion is always smaller than Scheffe's criterion and hence is better.*

*The Bonferroni's critical value in this problem is* $t_{28}(0.05/6) = 2.5465$ *which is larger than 2.4742. So Bonferroni is worse than Tukey.*

*The Tukey's criterion is the best.*

3. Consider a study comparing the means of $g = 6$ groups, with each group having a sample size of five. The overall type I error rate is to be controlled at $\alpha = 0.05$.

**(i)** Confirm that the critical value of Scheffe's criterion is 3.6198, and that of Tukey's criterion applied to $|\bar{X}_i - \bar{X}_j|/se(\bar{X}_i - \bar{X}_j)$ is 3.0922. Which criterion is better when only pairwise comparisons are of interest?

*The Scheffe's critical value is computed using* `sqrt(5*qf(0.95,5,24))` *which yields 3.6198.*

*The quantile of the studentized range distribution* $q_{6,24}(0.05) = 4.373$ *which divided by* $\sqrt{2}$ *yielding 3.0922.*

*Tukey's is better since its critical value is smaller.*

**(ii)** Confirm that, for $k = 4$ and $k = 8$ contrasts, the critical value for controlling the overall type I error rate according to Bonferroni's criterion are, respectively, 2.7002 and 2.9970. If the contrasts of concern are only pairwise ones, among the three criteria, which one is the best?

*The two critical values are obtained by*

```
qt(0.05/8,24,lower.tail=FALSE)
qt(0.05/16,24,lower.tail=FALSE)
```

*In each of the cases, compared with Scheffe's critical value 3.6198 and Tukey's critical value 3.0922, the Bonferroni's critcal value is the smallest, and hence Bonferroni's approach is the best.*

**(iii)** Confirm that, for $k = 15$ pairwise contrasts, which is the total number of the pairwise contrasts in this problem, the critical value of Bonferroni's criterion for controlling

the overall type I error rate is 3.2584. Is Tukey's criterion better than Bonferroni's criterion?

*The Bonferroni's critical value is obtained by* `qt(0.05/30,24,lower.tail=FALSE)`. *It becomes larger than Tukey's critical value 3.0922. The Tukey's criterion is better.*

**(iv)** If $k = 20$ contrasts, which are not confined to pairwise ones, are considered. Does Tukey's criterion still apply? Can Scheffe's and Bonferroni's be applied for controlling overall type I error rate? Which criterion is better?

*When the contrasts are not confined to pairwise ones, Tukey's criterion does not apply anymore. But both Scheffe and Bonferroni's criterion can still be applied. Bonferroni's critical value is obtained by* `qt(0.05/40,24,lower.tail=FALSE)` *as 3.3761, which is still smaller than Scheffe's critical value* 3.6198*, and hence is better.*

**(v)** Obtain the critical values of Bonferroni's criterion for $k = 30, 40$ contrasts. In each of the two cases, which of Bonferroni's and Scheffe's criterion is better?

*The Bonferroni's critical values are obtained by*

```
qt(0.05/60,24,lower.tail=FALSE)
qt(0.05/80,24,lower.tail=FALSE)
```

*as 3.5405 and 3.6561. In the case of $k = 30$, Bonferroni is still better than Scheffe, but in the case of $k = 40$, Bonferroni becomes worse than Scheffe.*

**(vi)** Prove that in general, for controlling the family-wise type I error rate for pairwise comparison, the critical value of Scheffe's criterion is larger than that of Tukey's criterion.

*Let $\mathcal{C}$ denote the set of all contrasts and $\mathcal{P}$ denote the contrasts of pairwise comparison. We have $\mathcal{P} \subset \mathcal{C}$, and*

$$P(\max_{c\in\mathcal{C}} |T_c| \geq d) \geq P(\max_{c\in\mathcal{P}} |T_c| \geq d),$$

*for any constant $d$. Let $q_\alpha$ and $c_\alpha$ denote the critical values of Tukey's and Scheffe's criteria respectively. We have*

$$P(\max_{c\in\mathcal{C}} |T_c| \geq q_\alpha) \geq P(\max_{c\in\mathcal{P}} |T_c| \geq q_\alpha) = \alpha = P(\max_{c\in\mathcal{C}} |T_c| \geq c_\alpha).$$

*Hence*

$$P(\max_{c\in\mathcal{C}} |T_c| \geq q_\alpha) \geq P(\max_{c\in\mathcal{C}} |T_c| \geq c_\alpha),$$

*which implies that $c_\alpha \geq q_\alpha$.*

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 7

1. For the **SMSA data set** considered in Tutorial 3.

**(i)** Suppose that people are concerned with the differences in serious crimes ($y$) among the regions $z$ and think that the following variables might affect the rate of serious crimes: population density ($x_1$, total population divided by land area), total personal income ($x_2$) and percent high school graduates ($x_3$).

(ia) Suggest an appropriate regression model for dealing with the concern.

*Here the prime purpose of the study is to compare the serious crimes in different regions, the variables $x_1, x_2, x_3$ are concomitant variables. The appropriate model is as follows:*

$$y = \beta_0 + \sum_{j=1}^{3} \alpha_j x_j + \sum_{k=2}^{4} \beta_k z_k + \epsilon,$$

*where $z_k$'s are the dummy variable for regions.*

(ib) A model is fitted to the data, which yields the following results:

```
Estimated coefficients:

 (Intercept)          x1          x2          x3          z2          z3          z4
-7747.027762  2443.184486    8.605784  -131.693527  9142.365366 17408.660144 26821.270021

Estimated variance matrix:

(Intercept)           x1       x2           x3           z2           z3           z4
(Intercept)  1.662488e+08 -5156333.6561 169.1947 -2895352.2202 -2064060.3821 -2.026821e+07 17186007.9838
x1          -5.156334e+06  2008009.2788 -99.3373    62356.9546  1384486.9157  1.852863e+06  1139626.9088
x2           1.691947e+02      -99.3373   0.0265       -5.0192      -24.1182 -7.380800e+00       -5.2891
x3          -2.895352e+06    62356.9546  -5.0192    55077.8296  -198314.2332  1.378800e+05  -566625.7660
z2          -2.064060e+06  1384486.9157 -24.1182  -198314.2332 20924969.5983  1.167900e+07 14189529.1919
z3          -2.026821e+07  1852862.9274  -7.3808   137880.0487 11678998.8811  1.843041e+07 10881673.2113
z4           1.718601e+07  1139626.9088  -5.2891  -566625.7660 14189529.1919  1.088167e+07 28370530.6398


Analysis of Variance Table
       Df     Sum Sq    Mean Sq   F value    Pr(>F)
x1      1 1.9396e+11 1.9396e+11  676.3277 < 2.2e-16
x2      1 8.3193e+11 8.3193e+11 2900.8761 < 2.2e-16
x3      1 2.6892e+08 2.6892e+08    0.9377    0.3346
z       3 9.7021e+09 3.2340e+09   11.2768 1.208e-06
Residuals 134 3.8429e+10 2.8679e+08
```

Give the model which was fitted.

*The model is as given in (ia).*

(ic) From the results given in (ib), test whether there is any difference in serious crimes among the regions: give the $p$-value of the test and draw your conclusion.

*The question amounts to test*

$$H_0 : \boldsymbol{\beta} = (\beta_2, \beta_3, \beta_4)^\top = 0, \ v.s. \ H_1 : \boldsymbol{\beta} \neq 0.$$

*It can be tested by the $F$-value corresponding to $z$ in the ANOVA table which is 11.2768 with a p-value* `1.208e-06`.
*From the small p-value, we can reject the null hypothesis and conclude that the difference among the regions are significant.*

(id) Test at level $\alpha = 0.05$ whether or not there is significant difference (1) between region 2 and 4, (2) between region 1 and 3, and (3) between region 1 and 4.

*For (1), it amounts to test*

$$H_0 : \beta_2 - \beta_4 = 0, \ H_1 : \beta_2 - \beta_4 \neq 0.$$

*From the estimated coefficients and the estimated variance matrix, the $t$-statistic is computed as*

$$T_1 = \frac{9142.3654 - 26821.2700}{\sqrt{20924969.5983 + 28370530.6398 - 2 * 14189529.1919}} = -3.8656.$$

*The p-value is computed as* `2*pt(3.8656,134,lower.tail=FALSE) = 0.000172`.

*For (2) and (3), it is equivalent ot test:*

$$H_{0j} : \beta_j = 0, \ v.s. \ H_{1j} : \beta_j \neq 0, j = 3, 4.$$

*The $t$-statistics are computed as*

$$\begin{aligned} T_2 &= 17408.6601/\sqrt{18430410} = 4.055, \\ T_3 &= 26821.270/\sqrt{28370530} = 5.036. \end{aligned}$$

*The p-values are computed as* `2*pt(abs(T2),134,lower.tail=FALSE) = 8.46e-05`, `2*pt(abs(T2),134,lower.tail=FALSE) = 1.51e-06`.
*When the overall type I error rate is to be controlled, the individual p-values must be multiplied by 3. The Bonforroni p-values of the three tests are less than 0.05, all the null hypotheses are rejected.*

**(ii)** Suppose that people want to find out how the three variables, $x_1, x_2$ and $x_3$, affect the serious crimes and how their effects differ in different regions.

(iia) What is the appropriate model for this purpose?

*The model should be able to compare both the intercepts and slopes of the regression function within the four regions. A model including all the main-effect terms as well as the interaction between the x's and z is appropriate. The model is as follows:*

$$y = \beta_0 + \sum_{j=1}^{3} \alpha_j x_j + \sum_{k=2}^{4} \beta_k z_k + \sum_{k=2}^{4}\sum_{j=1}^{3} \gamma_{kj} z_k x_j + \epsilon.$$

*The regression function expressed separately for each of the regions are:*

$$y = \begin{cases} \beta_0 + \sum_{j=1}^{3} \alpha_j x_j + \epsilon, & \textit{region 1}; \\ (\beta_0 + \beta_2) + \sum_{j=1}^{3} (\alpha_j + \gamma_{2j}) x_j + \epsilon, & \textit{region 2}; \\ (\beta_0 + \beta_3) + \sum_{j=1}^{3} (\alpha_j + \gamma_{3j}) x_j + \epsilon, & \textit{region 3}; \\ (\beta_0 + \beta_4) + \sum_{j=1}^{3} (\alpha_j + \gamma_{4j}) x_j + \epsilon, & \textit{region 4}. \end{cases}$$

(iib) People believe that the population density ($x_1$) and total personal income ($x_2$) have the same effects among different regions and that they would like to find out specifically whether the effect of percent high school graduates ($x_3$) is different among the regions. State an appropriate model for this purpose. Fit the model and test the significance of the difference: (1) state the hypotheses in terms of the parameters in the model, (2) provide a method for testing the hypotheses and draw your conclusion.

*The appropriate model is*

$$y = \beta_0 + \sum_{j=1}^{3} \alpha_j x_j + \sum_{k=2}^{4} \beta_k z_k + \sum_{k=2}^{4} \gamma_{k3} z_k x_3 + \epsilon.$$

*(1) Let* $\boldsymbol{\gamma} = (\gamma_{23}, \gamma_{33}, \gamma_{43})^\top$. *The hypotheses are:*

$$H_0 : \boldsymbol{\gamma} = 0, \ H_1 \boldsymbol{\gamma} \neq 0.$$

*(2) To test the above hypotheses, the test statistic is given by*

$$F = \hat{\boldsymbol{\gamma}}^\top \hat{\Sigma}_\gamma \boldsymbol{\gamma} / 3,$$

*which follows a F-distribution with d.f. 3 and 131.*
*The R codes for obtaining* $\hat{\boldsymbol{\gamma}}$, $\hat{\Sigma}_\gamma$, $F$ *and the p-value are as follows:*

```
p.int = lm(y~x1+x2+x3*z); summary(p.int); b=p.int$coef; V=vcov(p.int)
gm = b[8:10]; v.gm= V[8:10, 8:10]; gm; v.gm
F= t(gm)%*%solve(v.gm)%*%gm/3
p.value = pf(F,3,131,lower.tail=FALSE)
```

*The computed results are:*

```
Estimated gamma's:

    x3:z2     x3:z3     x3:z4
584.97348 117.00861 -98.95316

Estimated variance matrix:

         x3:z2    x3:z3    x3:z4
x3:z2 615301.5 353163.2 361453.9
x3:z3 353163.2 482326.4 362571.2
x3:z4 361453.9 362571.2 666665.6

F-value:
 0.3358275

P-value:
 0.7994507
```

*With the big p-value, the null hypothesis is not rejected. It is concluded that there is no evidence to show that $x_3$ has different effects across the regions.*

2. In a Study on the Efficacy of Nosocomial Infection Control (SENIC), it is to determine whether infection surveillance and control programs have reduced the rate of nosocomial (hospital acquired ) infection in the United States hospitals. The SENIC data provided in the file: `SENIC.txt`, on Canvas, consists of a random sample of 113 hospitals. The variables of the data set are as follows:

```
y: Infection risk;             x1:Average days of patients stayed;
x2:Average age of patients;    x3:Routine culturing ratio;
x4:Routine chest X-ray ratio;  x5:Number of beds;
x6:Medical school affiliation; x7:Region (1,2,3,4);
x8:Average daily census;       x9:Number of nurses;
x10:Available facilities and services.
```

**(i)** Consider the relationship between infection risk ($y$, the response variable) and the

following three predictor variables: patient's average age ($x_2$), average daily census ($x_8$) and available facilities and services ($x_{10}$).

(ia) For each of the predictor variables, conduct an appropriate test at level $\alpha = 0.05$ to find out whether or not the variable has the same effect in the four regions.

*An appropriate model for dealing with this problem is the model including all main effect terms of the x's and the region (R) as well as interaction terms between the x's and R. Whether or not a x-variable has the same effect in the four regions is equivalent to the regression parameter corresponding to the interaction terms between that x and R. The test is the F-test. The F-statistic is computed though the Wald statistic approach. The R codes for the computation are as follows:*

```
q2.d = read.table("D://Rsession/SENIC.txt",header =TRUE)

y = q2.d$y; z1=q2.d$x2; z2=q2.d$x8; z3=q2.d$x10; R=factor(q2.d$x7)
                                  particular interaction effect
q2.fit=lm(y~z1+z2+z3+R+z1:R+z2:R+z3:R)
                                  z1:R        z2:R        z3:R
b=q2.fit$coef; b1=b[8:10]; b2=b[11:13]; b3=b[14:16]
V=vcov(q2.fit);V1=V[8:10,8:10]; V2=V[11:13,11:13]; V3=V[14:16,14:16]
F1=t(b1)%*%solve(V1)%*%b1/3; F2=t(b2)%*%solve(V2)%*%b2/3;
F3=t(b3)%*%solve(V3)%*%b3/3;
```

*The F-values corresponding to $x_2$, $x_8$ and $x_{10}$ are respectively 1.4789, 0.7530 and 0.1092. The p-values are: 0.2251, 0.5233 and 0.9546, respectively. With these p-values, there is no significant evidence to show that these variables have different effect across the regions.*

(ib) Test at level $\alpha = 0.05$ whether or not there are any differences among the four regions after adjusting for the effect of the three predictor variables.

*To deal with this problem, the appropriate model is the one with only the main-effect terms. The anova table from fitting the model is given below:*

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) | |
|---|---|---|---|---|---|---|
| z1 | 1 | 0.000 | 0.0002 | 0.0002 | 0.98982 | |
| z2 | 1 | 29.393 | 29.3930 | 19.9924 | 1.958e-05 | *** |
| z3 | 1 | 6.842 | 6.8422 | 4.6539 | 0.03324 | * |
| R | 3 | 9.302 | 3.1008 | 2.1091 | 0.10343 | |
| Residuals | 106 | 155.842 | 1.4702 | | | |

*The region effect is tested by the F-value corresponding R in the above table. Since the p-value 0.10343 is larger than 0.05, there are no significant differences among the four regions.*
*The F-statistic can also be computed though the Wald statistic approach by using the R codes below:*

```
m.f = lm(y~z1+z2+z3+R)
```

```
b=m.f$coef[5:7]; V=vcov(m.f)[5:7,5:7]
F=t(b)%*%solve(V)%*%b/3;
```

**(ii)** Fit a regression model with response variable $y$, the infection risk, and predictor variables: patient's average age ($x_2$), average daily census ($x_8$), available facilities and services ($x_{10}$), including all the main-effect terms and all the two-variable interaction terms.

*The model is fitted which yields the following results:*

```
i.f = lm(y~z1*z2+z1*z3+z2*z3); summary(i.f);
```

```
           Df  Sum Sq Mean Sq F value    Pr(>F)
z1          1   0.000  0.0002  0.0002  0.989136
z2          1  29.393 29.3930 22.7522 5.912e-06 ***
z3          1   6.842  6.8422  5.2963  0.023326 *
z1:z2       1   9.331  9.3309  7.2228  0.008360 **
z1:z3       1   6.810  6.8098  5.2712  0.023648 *
z2:z3       1  12.065 12.0652  9.3393  0.002838 **
Residuals 106 136.939  1.2919
```

(iia) Use a $F$-test at level $\alpha = 0.05$ to test whether or not there are any interaction effects.

*The F-statistic for the test is computed by*

```
b=i.f$coef[5:7]; V=vcov(i.f)[5:7,5:7]
F=t(b)%*%solve(V)%*%b/3; p.value = pf(F,3,106,lower.tail=FALSE)
```

*which yields:* $F = 7.2778,\ p = 0.0001743$. *At level 0.05, the interaction effects are significant.*

(iib) If the $F$-test in (iia) is significant, which interaction effects are significant at level $\alpha = 0.05$, controlling the overall type I error rate by Bonferroni criteria.

*Multiply the p-values for the interaction terms in the summary table by 3, the Bonferroni p-values are obtained as*

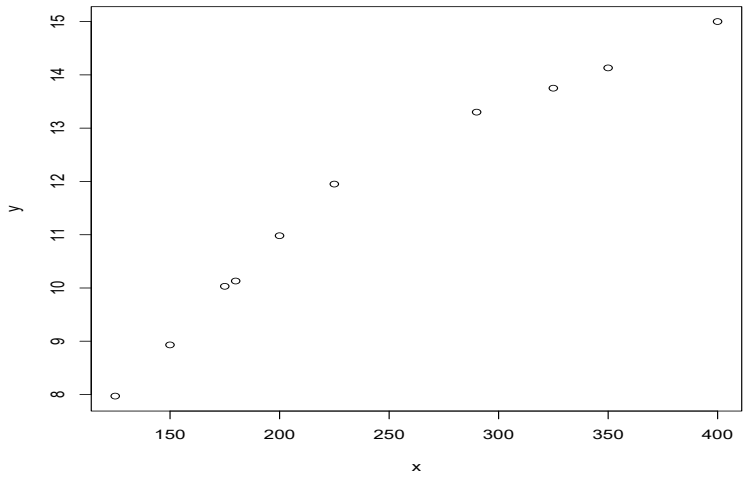
$$0.02508, 0.0709, 0.0085.$$

*At level 0.05, the interaction between $x_2$ and $x_8$, and between $x_8$ and $x_{10}$ are significant, but that between $x_2$ and $x_{10}$ is not.*

3. An electronics company periodically imports shipments of a certain large part used as a component in several of its products. The size of the shipment varies depending upon

production schedules. For handling and distribution to assembly plants, shipments of size 250 thousand parts or less are sent to warehouse A; larger shipments are sent to warehouse B since this warehouse has specialized equipment that provides greater economies of scale for large shipments. The data set contains information on the cost ($y$) of handling the shipment in the warehouse (in thousand dollars) and the size of the shipment ($x$, in thousand parts). The data is provided in `shipment.txt` on Canvas.

**(i)** Draw a scatter plot of $y$ against $x$. Does the scatter plot suggest a piecewise linear regression model? Determine the point of $x$ at which there is a change of the slope of the regression line.




*The plot is given above. There is an obvious change of slope around 250. A piecewise linear regression model with a change point at $x = 250$ could be appropriate. This also complies with the background information.*

**(ii)** Fit a piecewise linear regression model with a change of slope at $x = 250$ and test whether or not the change of slope is significant at level $\alpha = 0.05$.

*The R codes and fitted results are given below:*

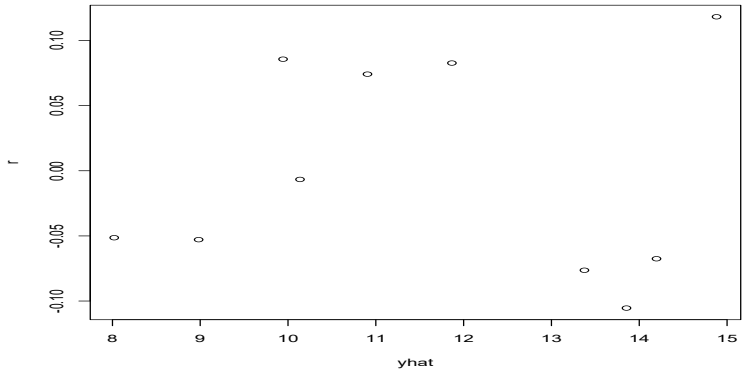
```
q3.dat =read.table("D://Rsession/shipment.txt",header =TRUE)
y=q3.dat$y; x=q3.dat$x1; z=(x-250)*(x>=250)
q3.fit = lm(y~x+z); summary(q3.fit)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.2139257  0.1785735   18.00 4.04e-07 ***
x            0.0384599  0.0009554   40.25 1.52e-09 ***
z           -0.0247734  0.0016460  -15.05 1.37e-06 ***
Residual standard error: 0.09303 on 7 degrees of freedom
```

```
Multiple R-squared:  0.9988,    Adjusted R-squared:  0.9985
F-statistic:  2938 on 2 and 7 DF,  p-value: 5.813e-11
```

*The coefficient of $z$, the auxiliary variable, has a p-value* `1.37e-06` *and is significantly nonzero. Hence the change of slope is significant.*

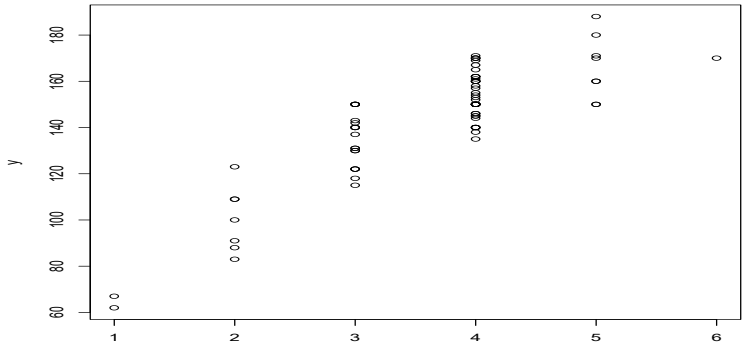
**(iii)** Plot the residuals against the fitted values and comment on the appropriateness of the piecewise linear regression model.

*The plot is given below. No patterns or trends show up in the plot, which indicates the appropriateness of the model.*




4. In 1981, $n = 78$ bluegills were randomly sampled from Lake Mary in Minnesota. The researchers measured and recorded the following data: length of the fish ($y$, in mm) and age of the fish ($x$, in years). The data is given in `bluegills.txt` on Canvas.

**(i)** Draw a scatter plot of $y$ against $x$. Suggest a proper regression model for the data.




*The plot is given above. The plot shows a quadratic curvature. A quadratic regression model can be suggested.*

**(ii)** Fit the model and test whether or not there is a significant relation between $y$ and $x$ at level $\alpha = 0.05$. Is $R^2$ of the model the same as the square of the Pearson's correlation coefficient between $y$ and $x$? Make comment.

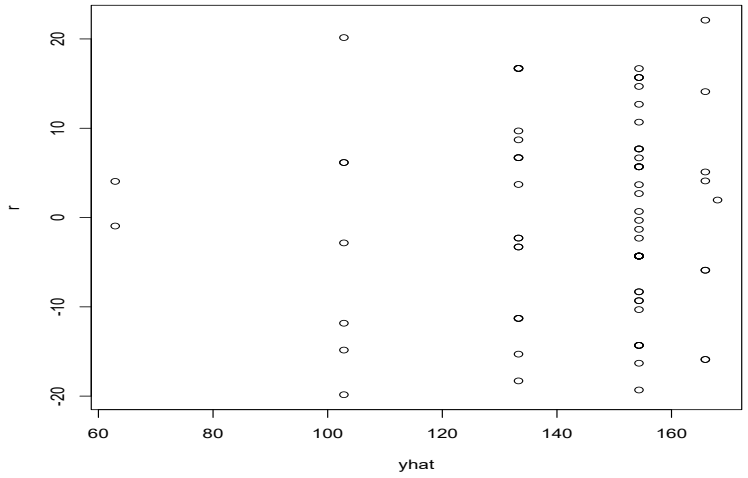
*The fitted results are given below:*

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  147.608      1.472 100.270  < 2e-16 ***
x.c           19.809      1.431  13.845  < 2e-16 ***
I(x.c^2)      -4.719      0.944  -4.999 3.67e-06 ***
Residual standard error: 10.91 on 75 degrees of freedom
Multiple R-squared:  0.8011,    Adjusted R-squared:  0.7958
F-statistic: 151.1 on 2 and 75 DF,  p-value: < 2.2e-16
```

*The significant $F$-test has a p-value less than* `2.2e-16`*. The regression relationship is significant, i.e., there is a significant relation between $y$ and $x$.*

*The $R^2$ is not the same as the square of the Pearson's correlation coefficient between $y$ and $x$. The $R^2$ of a regression model measures the correlation of the response with all covariates in the model. In the current case, the covariates in the model includes $x$ and $x^2$. The $R^2$ equals the square of the multiple correlation coefficient between $y$ and $(x, x^2)$.*

**(iii)** Plot the residuals against the fitted values and comment on the appropriateness of the model.




*The plot is given above. No patterns or trends show up in the plot, which indicates the appropriateness of the model.*

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 8

**Questions 1, 2 and 5 of this tutorial are to be submitted for marking. The deadline of submission is 2:00 pm, Wednesday, Oct. 26**

1. Consider the following regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, i = 1, \ldots, n,$$

where $\epsilon_i$ are i.i.d. $\sim N(0, \sigma^2)$.

**(i)** [**10 marks**] Give the MLE of $\boldsymbol{\beta}$ and $\sigma^2$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

*The MLE of $\boldsymbol{\beta}$ is the same as the LSE which is given by $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}$.*
*The MLE of $\sigma^2$ is obtained by maximizing*

$$-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2,$$

*which yields $\hat{\sigma}^2 = \frac{1}{n}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2 = \frac{n-p-1}{n}\hat{\sigma}^2_{lse}$.*

**(ii)** [**10 marks**] Derive explicitly the expression of AIC and BIC. Confirm that the derived AIC differ from $n\ln(\hat{\sigma}^2) + 2j_M$ by the constant $n(\ln(2\pi) + 1)$.

*For the normal regression model, the maximum likelihood is given by*

$$L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-n/2}\exp\{-\frac{1}{2\hat{\sigma}^2}\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2\} = (2\pi\hat{\sigma}^2)^{-n/2}\exp\{-\frac{n}{2}\}$$

*Hence*

$$\begin{aligned} AIC &= -2\ln L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + 2j_M = n\ln(2\pi\hat{\sigma}^2) + n + 2j_M \\ &= n\ln\hat{\sigma}^2 + 2j_M + n\ln(2\pi) + n. \\ BIC &= n\ln\hat{\sigma}^2 + n\ln(2\pi) + n + \ln n j_M. \end{aligned}$$

2. Consider the insurance example considered in Lecture notes 6. The data set `insurance.txt` is provided on Canvas.

**(i)** [**15 marks**] The data is fitted to the following model

$$log.Y = \beta_0 + \sum_{j=1}^{8} \beta_j X_j + \epsilon.$$

The following provides some fitted results:

```
Residual standard error: 0.2093 on 45 degrees of freedom
Multiple R-squared:  0.8461,  Adjusted R-squared: 0.8188
F-statistic: 30.93 on 8 and 45 DF,  p-value: 7.8e-16
```

Using only the results above, compute the AIC and BIC of the model with $j_M$ replaced by $j_M + 2$. Confirm that your computed AIC and BIC values are the same as those given in Lecture notes 6.

*From the information given we identify that* $n = 45 + 9 = 54$, $\hat{\sigma}^2_{lse} = 0.2093^2 = 0.04380649$ *and hence* $\hat{\sigma}^2_{mle} = \frac{45}{54}0.04380649 = 0.03650541$. $\pi$ *is computed as* $\pi = 4 * (4 * atan(1/5) - atan(1/239))$. *Thus*

$$\begin{aligned} AIC &= n \ln \hat{\sigma}^2 + 2(j_M + 2) + n \ln(2\pi) + n \\ &= 54 \ln(0.03650541) + 2 * 10 + 54 * (\ln(2\pi) + 1) \\ &= -5.510558. \\ BIC &= n \ln \hat{\sigma}^2 + \ln(n)(j_M + 2) + n \ln(2\pi) + n \\ &= 54 \ln(0.03650541) + \ln(54) * 10 + 54 * (\ln(2\pi) + 1) \\ &= 14.37928. \end{aligned}$$

*The AIC and BIC computed by the functions* `AIC` *and* `BIC` *are, respectively,* $-5.525623$ *and* $14.36422$. *The difference is caused by rounding error of* $\hat{\sigma}_{lse}$.

*The more accurate* $\hat{\sigma}_{lse}$ *can be obtained by* `summary(full.fit)$sigma` *which gives* $\hat{\sigma}_{lse} = 0.2092708$. *When this value is used in the above computation, the computed AIC and BIC will be the same as* $-5.525623$ *and* $14.36422$.

**(ii)** [**5 marks**] After you have fitted the model using

```
full.fit = glm(log.Y.~X1+X2+X3+X4+X5+X6+X7+X8,data=insur.dat),
```

implement the following codes repeatedly for four times:

```
CV.1.f = cv.glm(insur.dat,full.fit)$delta[2]
CV.k.f = cv.glm(insur.dat,full.fit,K=11)$delta[2]
c(CV.1.f, CV.k.f)
```

Are the values produced by `CV.1.f` and `CV.k.f` keep constant? If values differ from repetition to repetition, what could be the reason?

*The following codes are used for the computation:*

```
set.seed(12345)
r=1
CV=NULL
while (r<=4) {
CV.1.f = cv.glm(insur.dat,glm.fit)$delta[2]
CV.k.f = cv.glm(insur.dat,glm.fit,K=11)$delta[2]
CV=rbind(CV, c(CV.1.f, CV.k.f) )
r=r+1
}
CV
```

*The results are:*

```
leave-out-one          k-fold
[1]     0.0540974      0.05221861
[2]     0.0540974      0.06043276
[3]     0.0540974      0.05494720
[4]     0.0540974      0.05819644
```

*The values for the k-fold CV differ since each time the partition of the data set is random which yields different partitions.*

3. Consider the pull strength example in Lecture notes 6. The data set `pullstrength.csv` provided on Canvas gives information on pull strength ($y$), die height ($x_1$), post height ($x_2$), loop height ($x_3$), wire length ($x_4$), bond width on the die ($x_5$) and bond width on the post ($x_6$).

   **(i)** Fit sequentially the models:

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon,\ k = 1, 2, \ldots, 6.$$

   Provide the estimated variances of the estimated coefficients in each of these models in a table. For each of the predictors, does the estimated variance of its estimated coefficient increase when more predictors are included in the model? If does not, is it a contradiction to the theory that more predictors in the model increases the variance of estimated coefficients?

*The estimated variances of the estimated parameters for the six models are obtained by the following R codes:*

```
pull.dat=read.csv(file="D://Rsession/pullstrength.csv",header=TRUE)
M1 = lm(y~x1, data=pull.dat)
M2 = lm(y~x1+x2, data=pull.dat)
M3 = lm(y~x1+x2+x3, data=pull.dat)
M4 = lm(y~x1+x2+x3+x4, data=pull.dat)
M5 = lm(y~x1+x2+x3+x4+x5, data=pull.dat)
M6 = lm(y~x1+x2+x3+x4+x5+x6, data=pull.dat)

round(diag(vcov(M1)),4); round(diag(vcov(M2)),4)
round(diag(vcov(M3)),4); round(diag(vcov(M4)),4)
round(diag(vcov(M5)),4); round(diag(vcov(M6)),4)
```

*The estimated variances are given below:*

| | (Intercept) | x1 | x2 | x3 | x4 | x5 | x6 |
|---|---|---|---|---|---|---|---|
| M1 | 20.5357 | 0.6077 | | | | | |
| M2 | 64.9928 | 0.6103 | 0.1082 | | | | |
| M3 | 51.5671 | 0.3575 | 0.0599 | 0.0227 | | | |
| M4 | 52.7866 | 0.2837 | 0.0602 | 0.0192 | 0.0027 | | |
| M5 | 61.6521 | 0.3097 | 0.0710 | 0.0201 | 0.0029 | 6.1709 | |
| M6 | 65.7687 | 0.3468 | 0.0716 | 0.0203 | 0.0032 | 6.7484 | 2.3085 |

*For a given parameter, the estimated variances are not strictly increasing as the number of variables increase in the model. This is due to the inaccuracy of the estimated $\sigma^2$ from each model.*

**(ii)** For each of the models, divide the estimated variances by $\hat{\sigma}^2$ obtained from that model. List them in a similar table to that in (i). Comment on what you find from the table.

*This is done through the following R codes:*

```
s1=summary(M1)$sigma; s2=summary(M2)$sigma; s3=summary(M3)$sigma;
s4 =summary(M4)$sigma; s5=summary(M5)$sigma;s6=summary(M6)$sigma;

round(diag(vcov(M1))/s1^2,4); round(diag(vcov(M2))/s2^2,4)
round(diag(vcov(M3))/s3^2,4); round(diag(vcov(M4))/s4^2,4)
round(diag(vcov(M5))/s5^2,4); round(diag(vcov(M6))/s6^2,4)
```

*After the factor $\hat{\sigma}^2$ is removed from the estimated variances, we get the following:*

```
(Intercept)       x1       x2       x3       x4       x5       x6
11.7332   0.3472
37.1309   0.3487   0.0618
53.2515   0.3692   0.0618   0.0235
68.7022   0.3693   0.0783   0.0250   0.0035
77.2424   0.3880   0.0890   0.0252   0.0037   7.7313
82.2758   0.4339   0.0896   0.0253   0.0040   8.4421   2.8879
```

*Thus, the estimated variances (if $\hat{\sigma}^2$ is replaced by $\sigma^2$ or a common estimate) increase strictly as the number of variables in the model increases.*

**(iii)** For each of the models, compute their AIC, BIC and leave-out-one CV score.

*The AIC, BIC and leave-out-one CV score of the six models are computed by the R codes:*

```
M1 = glm(y~x1, data=pull.dat)
M2 = glm(y~x1+x2, data=pull.dat)
M3 = glm(y~x1+x2+x3, data=pull.dat)
M4 = glm(y~x1+x2+x3+x4, data=pull.dat)
M5 = glm(y~x1+x2+x3+x4+x5, data=pull.dat)
M6 = glm(y~x1+x2+x3+x4+x5+x6, data=pull.dat)

AIC=c(AIC(M1), AIC(M2),AIC(M3),AIC(M4),AIC(M5),AIC(M6))
BIC=c(BIC(M1), BIC(M2),BIC(M3),BIC(M4),BIC(M5),BIC(M6))
CV=c( cv.glm(pull.dat,M1)$delta[2], cv.glm(pull.dat,M2)$delta[2],
cv.glm(pull.dat,M3)$delta[2], cv.glm(pull.dat,M4)$delta[2],
cv.glm(pull.dat,M5)$delta[2],cv.glm(pull.dat,M6)$delta[2])
```

*They are given as follows:*

```
AIC      BIC       CV1
M1   68.44152 71.27484 1.9136461
M2   69.29120 73.06895 1.9874602
M3   58.81759 63.53979 1.1389535
M4   55.11045 60.77709 0.9398758
M5   56.42598 63.03705 1.0730371
M6   56.93381 64.48932 1.1039817
```

*Note: to compute CV scores, the models must be fitted by* `glm`.

4. The **SMSA data** is re-described as follows. This data set provides information for 141 large Standard metropolitan Statistical Areas in the United States. The columns of the data set correspond to:

1. identification number,
2. land area,
3. total populatoin,
4. percent of population in central cities,
5. percent of populaton 65 or older,
6. number of active physicians,
7. number of hospital beds,
8. percent high school graduates,
9. civilian labor force,
10. total personal income,
11. total serious crimes,
12. geographic region.

Take the `number of active physicians` as the response variable and the others as predictor variables.

**(i)** Carry out the following model selection procedures to select a model: (a) Forward selection procedure, (b) Backward selection proceture, and (c) Both upwards and downwards stepwise selection procedures. Give the selected models.

*The procedures are carried out by the following R codes:*

```
SMSA.dat = read.table("D://Rsession/SMSA.txt",header=TRUE)
attach(SMSA.dat)

y.p=doctor; x1=Land; x2=T.p; x3=P.city; x4=p.65; x5=beds; x6=h.sch;
x7=labor; x8=income; x9=crimes; x10=factor(region);
p.dat=data.frame(y.p,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10)

library(MASS)
lower.m = lm(y.p~1, data=p.dat)
upper.m = lm(y.p~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10, data=p.dat)

# Forward selection
forward=stepAIC(lower.m,scope=list(lower=~1,upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10),
direction="forward")
summary(forward)

# Backword selction
backward=stepAIC(upper.m,scope=list(upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,lower=~1),
direction="backward")
```

```
summary(backward)

# Stepwise upwards
stepUp=stepAIC(lower.m,scope=list(lower=~1, upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10),
direction="both")
summary(stepUp)

# Stepwise downwards
stepDown=stepAIC(upper.m,scope=list(lower=~1, upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10),
direction="both")
summary(stepDown)
```

*All the procedures produce the same model as follows:*

`y.p~x1+x2+x5+x8+x9+x10.`

**(ii)** Using the LASSO approach to select variables. Give the selected variables.

*The LASSO approach is carried out by the following codes:*

```
library(iterators)
library(foreach)
library(Matrix)
library(shape)
library(glmnet)

X1 = as.matrix(p.dat[,2:9])
z=p.dat$x10; y =p.dat$y.p
u2 = rep(0,length(z)); u3=u2; u4=u2
u2[z==2] = 1; u3[z==3]=1; u4[z==4]=1

X2=cbind(X1,u2,u3,u4)
penal = cv.glmnet(X2, y, labmda=seq(2,8,by=0.1))
coef(penal, s = penal$lambda.min)
```

z is for factor

*The following are the estimated coefficients:*

```
(Intercept) -899.16300522
x1            -0.02394969
x2             .
x3             5.45413998
x4            12.82775662
x5             0.13677883
x6             5.75406071
x7             .
```

```
x8             0.18272027
u2          -182.11048565
u3            72.63889858
u4           331.60237588
```

*The selected variables are* `x1, x3, x4, x5, x6, x8, x10`*. Note the factor variable* `x10` *is represented by* `u2, u3, u4`*. If any of them is selected, it implies that* `x10` *is selected.*

**(iii)** Fit the model with the predictor variables as those selected by the LASSO approach. Compare the model with those selected in (i) by BIC and leave-out-one CV. Give the best model under each criterion.

*The comparison is done by the following codes:*

```
 library(boot)
m1 = glm(y.p~x1+x2+x5+x8+x9+x10,data=p.dat)
m2 = glm(y.p~x1+x3+x4+x5+x6+x8+x10,data=p.dat)

cv.1=cv.glm(data=p.dat, m1)$delta[2]
cv.2=cv.glm(data=p.dat, m2)$delta[2]
data.frame(model=c("m1", "m2"),BIC=c(BIC(m1),BIC(m2)),CV=c(cv.1,cv.2))
```

```
  model      BIC       CV
     m1 2157.043 367412.4
     m2 2217.457 511153.9
```

*The model selected in (i) is the better model by both criteria.*

5. The following is the result of an intermediate step in the upward stepwise procedure with the SMSA data

```
Step:  AIC=1725.42
y.p ~ x8 + x9 + x5 + x2 + x10 + x1

       Df Sum of Sq      RSS    AIC
<none>              25599578 1725.4
- x1    1    538644 26138223 1726.3
+ x6    1    181085 25418494 1726.4
+ x3    1    154143 25445435 1726.6  *
+ x7    1    110383 25489195 1726.8
+ x4    1     52175 25547403 1727.1
```

```
- x10   3   1666197 27265776 1728.3
- x2    1   1705119 27304698 1732.5  **
- x8    1   1792033 27391611 1733.0
- x5    1  11339255 36938834 1775.1
- x9    1  12719153 38318731 1780.3
```

**(i)** [**5 marks**] Identify the models corresponding * and ** respectively. Which of these two models is better by the criterion AIC?

*The model corresponding to* * *is* $y.p \sim x8 + x9 + x5 + x2 + x10 + x1 + x3$.

*The model corresponding to* ** *is* $y.p \sim x8 + x9 + x5 + x10 + x1$.

*The first model has AIC = 1726.6, the second one has AIC = 1732.5. The first model is better.*

**(ii)** [**5 marks**] Should the procedure continue or stop? Give your reasons.

*The procedure should stop.*

*By removing any of the variables from the current model, the AIC will be bigger than the AIC of the current model, hence none of the variables can be removed. By adding any of the remaining variables to the current model, the AIC will also be bigger than the AIC of the current model, hence none of the remaining variables can be added.*

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 9

1. For the **SMSA data**, consider the `number of active physicians` as the response variable and the others as predictor variables.

   **(i)** Carry out the forward selection procedures for the model selection, taking the lower model as the one containing the factor variable `geographic region` and the upper model as the full model containing all the predictor variables. Is the model selected the same as that selected by the forward selection procedure with the lower model as the null model? If not, select among the two models by using AIC.

   *The procedures are carried out by the R codes:*

```
SMSA.dat = read.table("D://Rsession/SMSA.txt",header=TRUE)

y=SMSA.dat$doctor; x1=SMSA.dat$Land; x2=SMSA.dat$T.p; x3=SMSA.dat$P.city;
x4=SMSA.dat$p.65; x5=SMSA.dat$beds; x6=SMSA.dat$h.sch;x7=SMSA.dat$labor;
x8=SMSA.dat$income; x9=SMSA.dat$crimes;
x10=factor(SMSA.dat$region);
p.dat=data.frame(y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10)


## (i) selecting models
library(MASS)

# Starting models for the selection procedure:
null1 = lm(y~1, data=p.dat); null2 = lm(y~x10, data=p.dat)

# Selection with null model as starting model
f1=stepAIC(null1,scope=list(upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,lower=~1),
           direction="forward")
summary(f1)

# Selection with starting model containg x10
```

```
f2=stepAIC(null2,scope=list(upper=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10,lower=~x10),
           direction="forward")
summary(f2)
```

*The model selected while taking null model as the starting model is*

```
y ~ x8 + x9 + x5 + x2 + x10 + x1
```

*The model selected while taking* `y~x10` *as the starting model is*

```
y ~ x10 + x9 + x5 + x1 + x6
```

*The AIC values of the two models are:*

```
  AIC(f1)   AIC(f2)
 2127.556  2132.503
```

*The model selected while taking null model as the starting model is better and is finally selected.*

**(ii)** For the model selected in (i),

*The raw materials used for the following parts are computed by the R codes:*

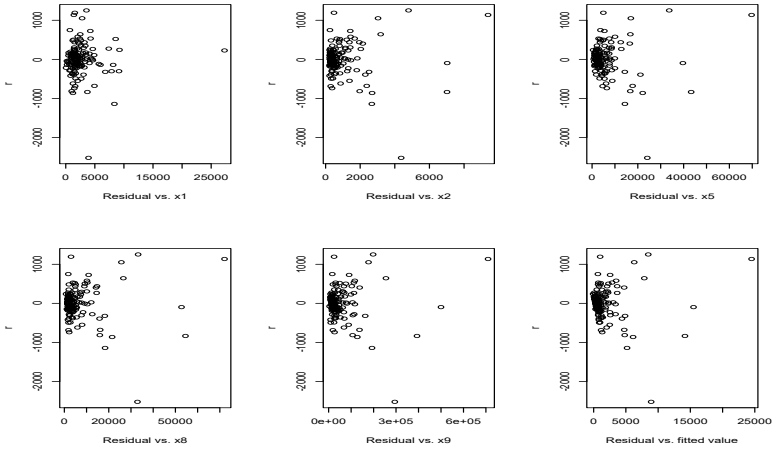
```
fit1=glm(y~x1 + x2 + x5 + x8 + x9 + x10,data=p.dat)
yhat = fit1$fitted.values
r = residuals(fit1,type="pearson")
h = hatvalues(fit1,type="diagonal")
infl = influence(fit1, do.coef = FALSE)
rsta = rstandard(fit1, infl, type = "pearson")
rstu = rstudent(fit1, infl, type = "pearson")
cook =cooks.distance(fit1, infl,res = infl$pear.res,
dispersion = summary(fit1)$dispersion,hat = infl$hat)
```

(iia) Check whether or not the regression function is linear in terms of the numerical predictor variables.

*The linearity is checked by plotting the Pearson's residual against each of the predictors, which are carried out by the R codes:*

```
par(mfrow=c(2,3))
plot(x1, r,xlab="Residual vs. x1")
plot(x2, r,xlab="Residual vs. x2")
plot(x5, r,xlab="Residual vs. x5")
plot(x8, r,xlab="Residual vs. x8")
plot(x9, r,xlab="Residual vs. x9")
plot(yhat,r,xlab="Residual vs. fitted value")
```

*The plots are as follows:*

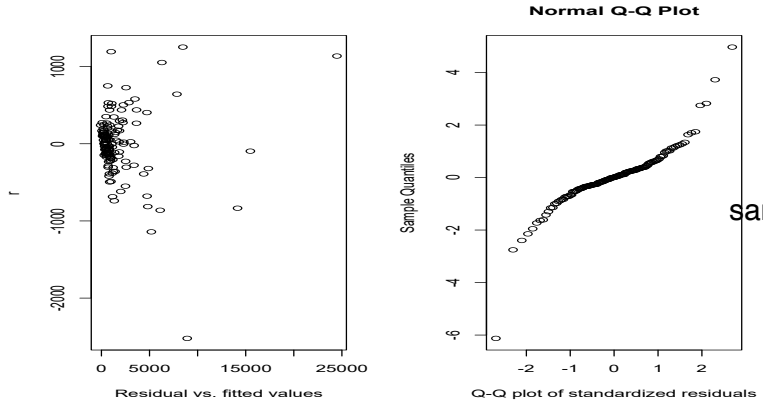



*Since there is no obvious trend showing up in the plots, it can be considered that the linearity assumption holds.*

(iib) Check whether or not the assumption of constant variance is violated.

(iic) Check whether or not the assumption of normality is violated.

*The constancy of variance is checked by the plot of Pearson's residual against the fitted values. The normality is checked by the normal probability plot of the studentized residuals. The plots are shown below:*




sample quantile vs expected,
-> heavier tail

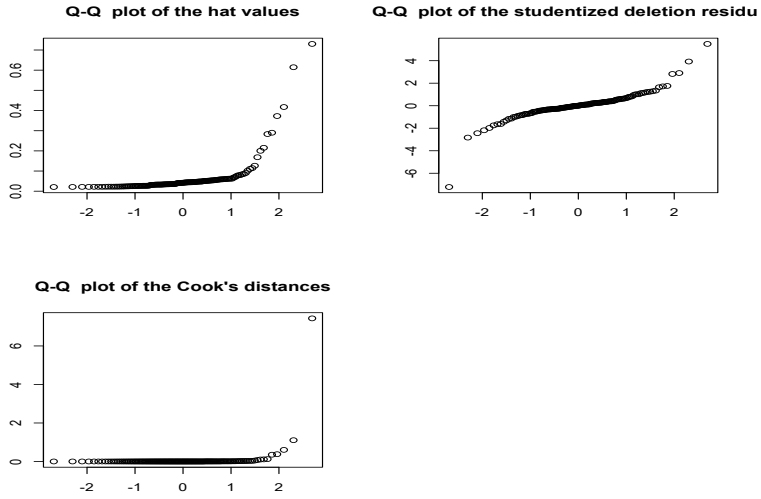
*The plot of Pearson's residuals against the fitted values does not show obvious evidence for non-constancy of variance, but it reveals that there are some outliers. The normal probability plot suggests that the normality is violated and the distribution of the data might be a symmetric distribution with heavier tails than the normal distribution.*

**(iii)** Compute the hat values, studentized deleted residuals and Cook's distances.

(iiia) Draw the Q-Q plot for each of these measures.

*The plots are given below:*

can only check outliers




(iiib) Identify the observations whose hat values exceeds $2p/n$, where $p$ is the number of parameters in the regression function including the dummy variables for the factor `geographic region`.

$n = 141, p = 9, 2p/n = 0.1276596$
*The observations are identified by* `h[h>=2*p/n]` *as*

```
    1     2     3     4     6     7     8    12    27
0.730 0.290 0.373 0.417 0.200 0.283 0.168 0.215 0.615
```

(iiic) Identify the observations whose studentized deleted residual has a absolute value bigger than $2d$ where $d$ is the sample standard deviation of the studentized deleted residuals.

*The sample standard deviation of the residuals is obtained by* `sqrt(var(rstu))` *as* 1.179096, $2d = 2.358$. *Th observations are identified by* `rstu[abs(rstu)>2*d]` *as*

```
    1      3      4      5      7      9     68
5.490 -2.440  3.922 -7.216  2.900 -2.828  2.817
```

(iiid) Identify the observations whose Cook's distance is bigger than 0.1.

*The observations are identified by* `cook[cook>0.1]` *as*

```
    1     3     4     5     7     8     9    27
7.431 0.379 1.104 0.603 0.349 0.104 0.109 0.122
```

**(iv)** Conduct formal tests by creating dummy variables to check simultaneously whether or not the observations identified in (iiid) are outliers, using Bonferroni's criterion to control the overall type I error rate at level 0.05.

*The dummy variables are created by the codes:*

```
u1=u2=u3=u4=u5=u6=u7=u8=rep(0,n)
u1[1]=1; u2[3]=1; u3[4]=1;u4[5]=1; u5[7]=1; u6[8]=1; u7[9]=1; u8[27]=1
```

*The model is re-fitted to the original variables together with the dummy variables:*

```
fit2=glm(y~x1 + x2 + x5 + x8 + x9 + x10 +u1+u2+u3+u4+u5+u6+u7+u8)
summary(fit2)
```

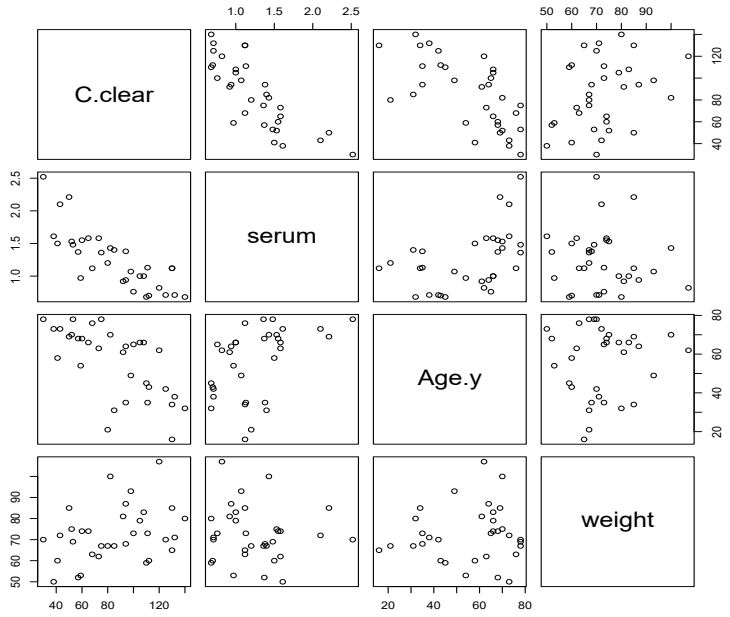
*The summary results are:*

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.894e+01  7.459e+01  -1.327  0.18710
x1          -7.538e-03  1.675e-02  -0.450  0.65350
x2          -9.152e-01  3.994e-01  -2.292  0.02362 *
x5           7.897e-02  1.922e-02   4.109 7.18e-05 ***
x8           2.310e-01  5.140e-02   4.493 1.59e-05 ***
x9           1.346e-02  2.766e-03   4.866 3.38e-06 ***
x102        -3.286e+01  7.855e+01  -0.418  0.67644
x103         5.540e+01  8.119e+01   0.682  0.49633
x104         1.109e+02  1.048e+02   1.058  0.29208
u1           2.628e+03  5.669e+02   4.636 8.88e-06 ***
u2          -1.399e+03  4.375e+02  -3.198  0.00176 **
u3           1.236e+03  4.103e+02   3.013  0.00314 **
u4          -2.909e+03  3.162e+02  -9.198 1.10e-15 ***
u5           5.288e+02  3.750e+02   1.410  0.16095
u6          -6.129e+02  3.406e+02  -1.800  0.07436 .
u7          -1.370e+03  3.128e+02  -4.381 2.49e-05 ***
u8          -1.375e+02  4.700e+02  -0.292  0.77043
```

*The Bonferroni critical values is obtained by* `qt(0.05/(2*8),124,lower.tail=FALSE)` *as* 2.781854. *By controlling the overall type I error rate at 0.05, it can be claimed that the observations 1 (u1), 3 (u2), 4(u3), 5(u4), 9(u7) are influential outliers.*

2. Creatinine clearance is an improrant measure of kidney function but is difficult to obtain in clinical office setting because it requires 24 hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data for 33 males. The data consists of creatinine clearance ($Y$), serum creatinine concentration ($X_1$), age ($X_2$) and weight ($X_3$). The data is provided in the file `kidney.txt` on Canvas.

**(i)** The following is the pairwise scatter plots:

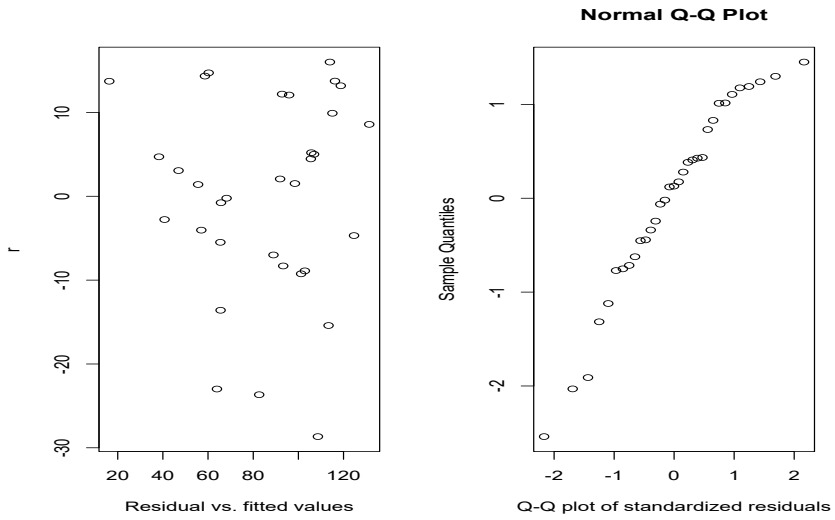



From the plots what relations between the response variable `C.clear` and each of the covariates you can suggest? Provide your reasons.

*There are a slight curvature in the plots of* `C.clear` *($Y$) against* `serum` *($X_1$) and against* `Age.y` *($X_2$), which suggests a non-linear relationship between $Y$ and $X_1$, and between $Y$ and $X_2$. A possible scale for $X_1$ could be $1/X_1$, for $X_2$ could be $\sqrt{X_2}$. The plot of $Y$ against $X_3$ shows a linear trend and there is no obvious curvature, it can be considered that the relation is linear.*

**(ii)** The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ is fitted to the data.

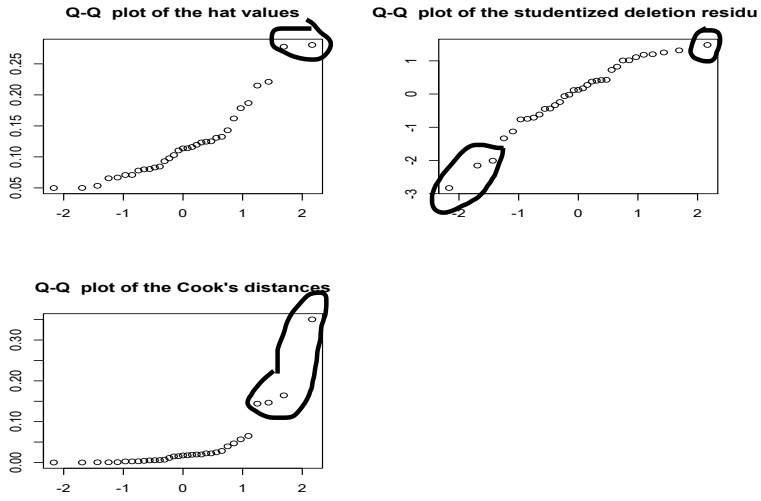
**(iia)** The following are the plot of residuals vs. fitted values (left panel) and the Q-Q plot of the studentized residuals (right panel):




Are the plots "null patterns"? If they are not null patterns, what could be the possible discrepancies? Explain.

*Both plots are not the null pattern. The possible discrepancies indicated by the plot of residuals vs. fitted values could be either the variance is not constant or the regression function is not appropriate.*
*When we draw the line $y = x$ in the second plot, it can be found that there are a few points which fall below the line. It might be caused either by non-normality or by some outliers.*

**(iib)** The following are the Q-Q plots of the three outlier measures: hat value, studentized deletion residual and Cook's distance.




Circle out the possible outliers in terms of the three measures.

**(iic)** A formal test is to be carried out to check formally whether or not the observations 16,20,21,26 and 29 are significant outliers. The following are the R codes and the fitted results:

```
u1=rep(0,n);u2=u1;u3=u1;u4=u5=u1;
u1[16]=1;u2[20]=1;u3[21]=1;u4[26]=1; u5[29]=1
outlier=glm(C.clear~serum+Age.y+weight+u1+u2+u3+u4+u5,
          data=kidney.dat)
summary(outlier)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 143.3637    11.8931  12.054 1.14e-11 ***
serum       -42.4945     4.6435  -9.151 2.70e-09 ***
Age.y        -0.8176     0.1218  -6.714 6.05e-07 ***
weight        0.5890     0.1281   4.597 0.000116 ***
u1           16.2631    10.3110   1.577 0.127827
u2            9.0249    10.1591   0.888 0.383165
u3          -30.2122     9.3382  -3.235 0.003525 **
u4          -34.6656     9.9302  -3.491 0.001884 **
u5          -26.5429     9.0991  -2.917 0.007551 **
```

Using Bonferroni approach to control the overall type I error rate at level $\alpha = 0.05$, which of the five observations are significant outliers?

*Multiplying the individual p-values for each of the dummy variables by 5, the Bonferroni p-values for* $u3, u4, u5$ *are less than 0.05. Therefore, observations 21, 26 and 29 can be claimed as significant outliers.*

NATIONAL UNIVERSITY OF SINGAPORE
Department of Statistics and Applied Probability

ST3131: Regression Analysis

TUTORIAL 10

1. The data set `mtcars` consists of the following variables: `mpg`, `cyl`, `disp`, `hp`, `drat`, `wt`, `qsec`, `vs`, `am`, `gear`, `carb`. The data set is in the form of a data.frame and is available in the R console.

   (i) Consider the regression model with `mpg` as the response variable and all the other variables as the predictor variables. Check multicolinearity (a) by using correlation matrix of the predictor variables, (b) by fitting different models with certain predictor variables removed, and (c) by using variance inflation factor.

   *(a) Correlation matrix:*

```
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

   *High correlations show up between a number of variables, indicating multicolinearity.*

*(b) Fitted coefficient of the full model, the model removing* `wt` *and the model removing* `hp`*:*

```
Full model:
Intercep     cyl   disp     hp  drat     wt  qsec     vs    am  gear   carb
12.303    -0.111  0.013 -0.021 0.787 -3.715 0.821  0.318 2.520 0.655 -0.199

Model with wt removed:
Intercept   cyl   disp     hp  drat         qsec     vs    am  gear   carb
15.571    0.120 -0.014 -0.011 1.327        0.094  0.668 2.901 1.186 -1.329

Model with hp removed:
Intercept   cyl   disp         drat     wt  qsec     vs    am  gear   carb
11.097   -0.301  0.004        0.927 -3.266 0.899 -0.248 2.424 0.524 -0.630
```

*After removing* `wt` *or* `hp`*, the estimated coefficients have significant changes, some even change signs.*

*(c) VIF:*

```
   cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb
15.374 21.620  9.832  3.375 15.165  7.528  4.966  4.648  5.357  7.909
```

*Quite a number of variables have VIF bigger than 5.*

*All the three approaches indicate serious multicolinearity.*

(ii) After removing the predictors with high VIF, carry out a forward selection procedure to select a model. Using the selected model, inference on the significance of the effects of the predictors on the response variable.

*Sequentially removing the variables with VIF larger than 5, the variables removed are in the order:* `disp`,~~`cy1`~~cyl,`wt`,`hp`*. The VIFs for the models with the variables removed in sequence are as follows:*

```
disp removed:
   cyl     hp   drat     wt   qsec     vs     am   gear   carb
14.285  7.123  3.329  6.189  6.914  4.916  4.645  5.324  4.311

cy1 removed:
   hp  drat    wt  qsec    vs    am  gear  carb
6.016 3.112 6.051 5.919 4.271 4.286 4.690 4.290

wt removed:
```

```
   hp  drat  qsec    vs    am  gear  carb
5.075 3.021 4.714 3.792 4.052 4.373 3.642

hp removed:
 drat  qsec    vs    am  gear  carb
2.849 3.754 3.792 3.902 4.335 2.456
```

*The model selected by the forward selection procedure and fitted results are as follows:*

```
lm(formula = mpg ~ drat + carb + gear + am + qsec, data = mtcars)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.3800     8.2855  -0.649 0.521821
drat          2.3524     1.5807   1.488 0.148715
carb         -1.7844     0.4775  -3.737 0.000925 ***
gear          2.0291     1.3759   1.475 0.152290
am            3.9944     2.0202   1.977 0.058702 .
qsec          0.7241     0.4223   1.714 0.098352 .
Residual standard error: 2.845 on 26 degrees of freedom
Multiple R-squared: 0.8132,  Adjusted R-squared: 0.7772
F-statistic: 22.63 on 5 and 26 DF,  p-value: 1.028e-08
```

*The variable* `carb` *has the most significant and a negative effect on* `mpg`*; all the other variables have positive effect on* `mpg`*, among which* `am` *and* `qsec` *are quite significant.*

(iii) Using ridge regression to fit the model with all the predictors.

*The following R codes carry out the ridge regression:*

```
library(MASS)
regRidge = lm.ridge(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,
data=mtcars, lambda = seq(1, 50, 0.01))
plot(regRidge$lambda, regRidge$GCV,type="l")
lambda_best = regRidge$lambda[which(regRidge$GCV==min(regRidge$GCV))]
regRidge_best = lm.ridge(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb,
data=mtcars, lambda = lambda_best)
round(regRidge_best$coef,3)
```

*which yields the following fitted coefficients:*

```
   cyl   disp     hp   drat     wt   qsec     vs     am   gear   carb
-0.656 -0.643 -0.783  0.554 -1.185  0.284  0.383  0.796  0.395 -0.869
```

2. Consider the Trial for children with neurological problems (TCNP) dealt with in Lecture notes 8.

(i) Find a variance stabilization transformation for the response variable.

*Using the regression approach to estimate $\alpha$ in the relation $\sigma = c\mu^{\alpha}$:*

```
fit.a=lm(log(s)~log(y))
summary(fit.a)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.1273     0.2240  -9.499 2.54e-06 ***
log(y)        2.5376     0.6139   4.133  0.00203 **
```

$\alpha \approx 2.5$, $\lambda = 1 - \alpha = -3/2$.

(ii) Fit the regression model with the transformed response variable. Check the adequacy of the fitted model.

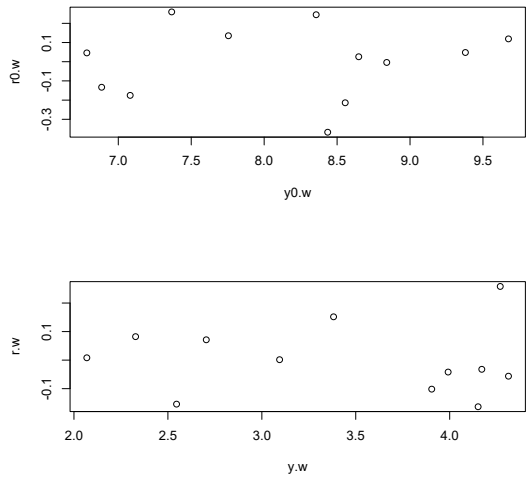
*The original and transformed data are fitted by*

```
fit0=lm(y~S+G+T,weight=n); r0=fit0$resid; y0=fit0$fitted
r0.w=r0*n^(1/2); y0.w=y0*n^(1/2)
y.t=y^(-3/2); fit=lm(y.t~S+G+T,weight=n); r=fit$resid; y.h=fit$fitted
r.w=r*n^(1/2); y.w=y.h*n^(1/2)
par(mfrow=c(2,1)); plot(y0.w,r0.w); plot(y.w,r.w);
```

*Note that in the fitting of both the original and transformed data, the models must be weighted by the number of observations at each cell. Since the original response is the cell mean $\bar{y}_i$ with variance $\sigma_i^2/n_i$, even if the original observations have the same variance, $\bar{y}_i$ won't have the same variance. For the transformation, we have, approximately,*

$$E[h(\bar{y}_i) - h(\mu_i)]^2 = [h(\mu_i)]^2 E(\bar{y}_i - \mu_i)^2 = [h(\mu_i)]^2 \sigma_i^2(\mu_i)/n_i.$$

*The variance stablization transformation $h$ only makes $[h(\mu_i)]^2\sigma_i^2(\mu_i)$ a constant. After the transformation $h$, $h(\bar{y}_i)$ will have approximately the variance $\sigma^2/n_i$ where $\sigma^2$ is constant for all $i$. The residual plots are given below.*




*The original model demonstrates non-constancy of variance. The model with transformed data mitigates the non-constancy of variance problem.*

3. In Question 2, Tutorial 9, the relationship between Creatinine clearance, an measure of kidney function, and the three predictor variables — serum creatinine concentration ($X_1$), age ($X_2$) and weight ($X_3$) — was investigated by the following linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon.$$

**(i)** The diagnostics of the model reveals that the scale of certain predictor variables might not be appropriate or that the variances might not be constant. To rectify the problem, try the following measures:

(ia) Find a variance stablization transformation of $y$, regress the transformed response on $x_1, x_2$ and $x_3$, check whether the problem has been rectified by using the plots of residuals against fitted values and each of the predictors.

*The following R codes are used (1) to fit the original model and use the residuals and the fitted values to estimate the transformation:*

```
fit=lm(y~x1+x2+x3); yhat = fit$fitted;  r=fit$resid
v.tran=lm(log(abs(r)) ~ log(yhat)); summary(v.tran)
```

*The fitting of log absolute residual on log fitted values yields the result:*

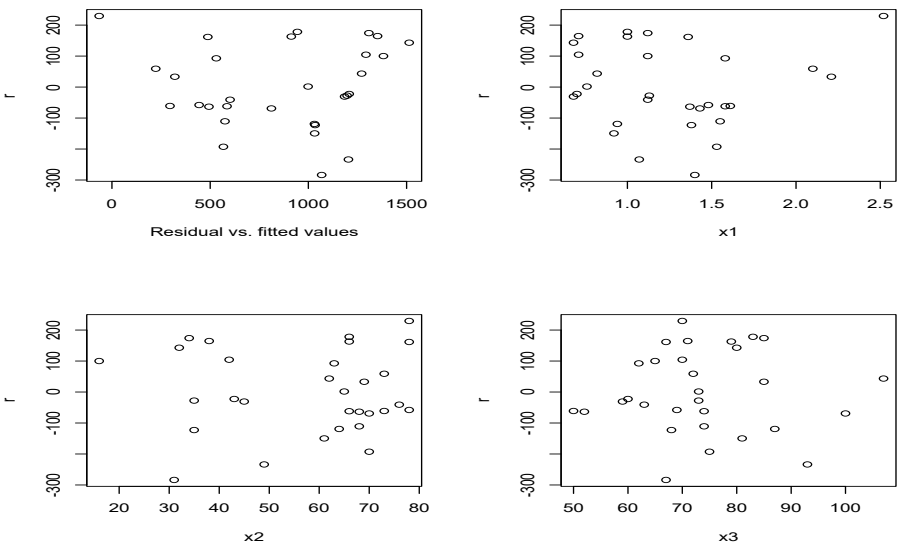
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3890     1.8942   1.789   0.0844 .
log(yhat)    -0.4158     0.4286  -0.970   0.3403
```

*Thus $\alpha$ is estimated as $-0.5$ (since it the closest half integer to $-0.4158$), the estimate of $\lambda$ is $1-(-0.5)=1.5$. The following carry jut the fitting with the transformed data $y^{1.5}$:*

```
y.t=y^(1.5);
fit.v=lm(y.t~x1+x2+x3)
yhat = fit.v$fitted;
r=fit.v$resid

par(mfrow=c(2,2))
plot(yhat, r,xlab="Residual vs. fitted values")
plot(x1,r);
plot(x2,r);
plot(x3,r);
```

*The plots are given below:*



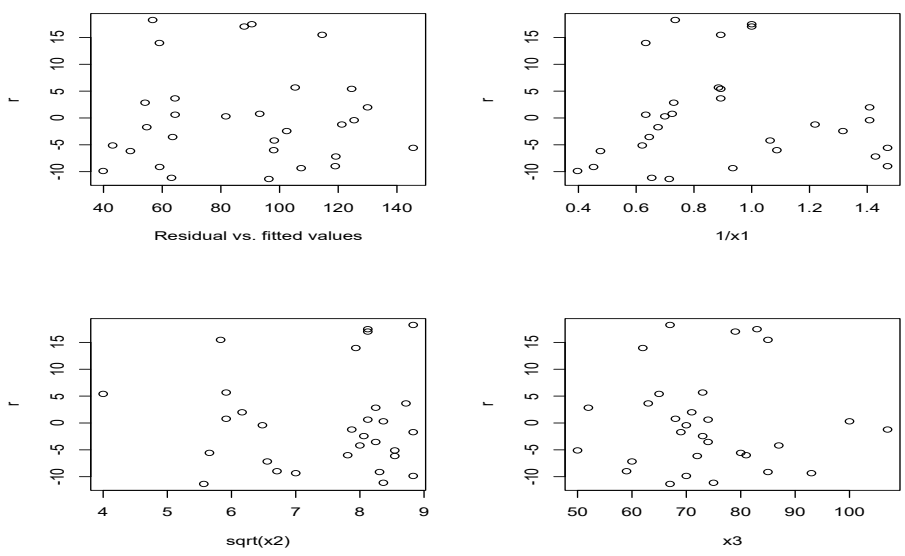

*There is no obvious evidence in the plots to show any violation of constancy of variance and linearity of the regression function.*

(ib) Regression $y$ on $1/x_1$, $\sqrt{x_2}$ and $x_3$, check whether the problem has been rectified by using the same residual plots as in (ia).

*The R codes:*

```
fit1=lm(y~I(1/x1) + I(sqrt(x2)) + x3)
yhat = fit1$fitted;
r = fit1$resid
par(mfrow=c(2,2))
plot(yhat, r,xlab="Residual vs. fitted values")
plot(1/x1,r);
plot(sqrt(x2),r);
plot(x3,r);
```

*The residual plots:*

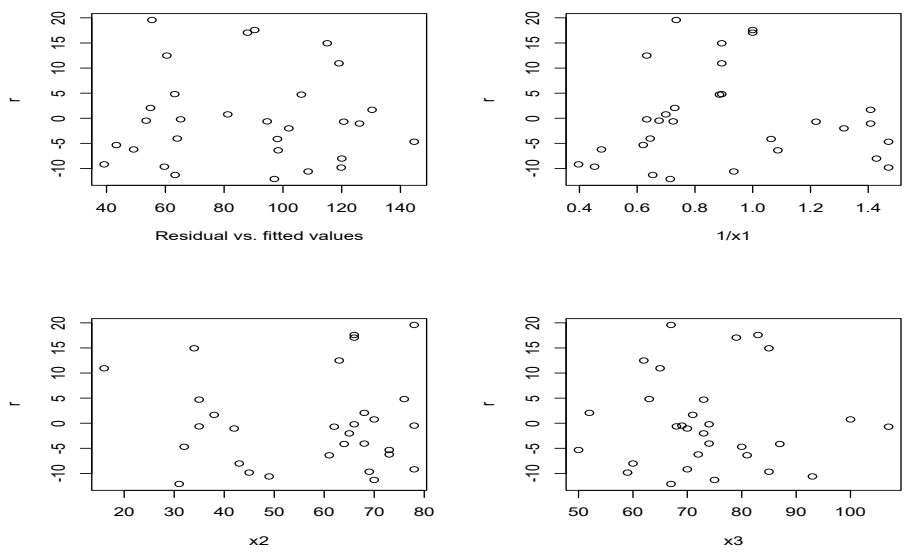



*The residual plot against* $1/x_1$ *still shows a curvature pattern, the problem is not rectified.*

(ic) Repeat (ib) by regressing $y$ on $1/x_1$, $x_2$ and $x_3$. *The R codes:*

```
fit2=lm(y~I(1/x1) + x2 + x3)
yhat = fit2$fitted
r = fit2$resid
par(mfrow=c(2,2))
plot(yhat, r,xlab="Residual vs. fitted values")
plot(1/x1,r);
plot(x2,r);
plot(x3,r);
```

*The residual plots:*

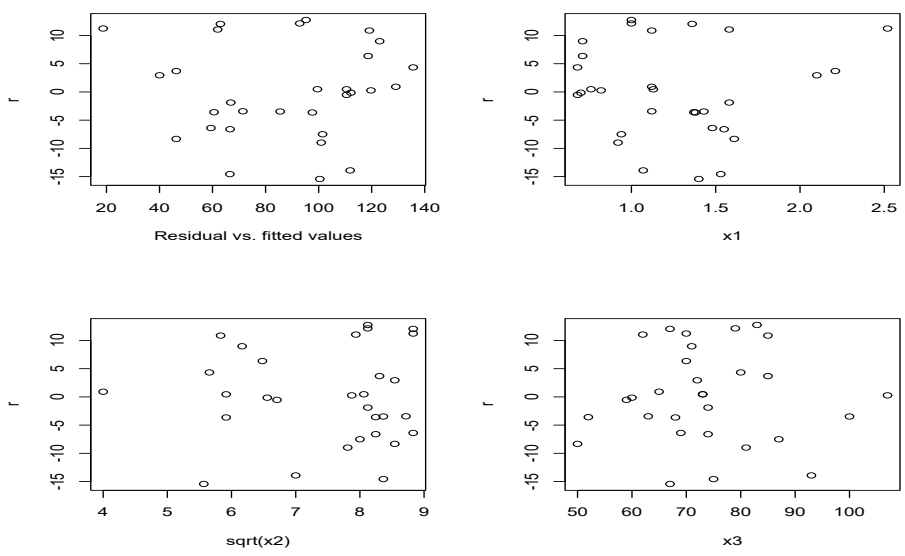



*Still, the residual plot against* $1/x_1$ *still shows a curvature pattern, the problem is not rectified.*

(id) Repeat (ib) by regressing $y$ on $x_1$, $\sqrt{x_2}$ and $x_3$.
*The R codes:*

```
fit3=lm(y~x1 + I(sqrt(x2)) + x3)
yhat = fit3$fitted
r = fit3$resid
par(mfrow=c(2,2))
plot(yhat, r,xlab="Residual vs. fitted values")
plot(1/x1,r);
plot(x2,r);
plot(x3,r);
```

*The residual plots:*




*There is no obvious evidence in the plots to show any violation of constancy of variance and linearity of the regression function.*

**(ii)** For the models considered in (i) which appear to have rectified the problem, find the one with the smallest AIC.

*The models fitted in (ia) and (id) seem to have recitified the problem. Their AIC values are given below:*

```
Model      AIC
   ia 221.9393
   id 220.7207
```

*The model fitted in (id) has the smaller AIC.*