

「-2」，例如：「學號-姓名-Regression-R-Midterm-2.docx」。

Notes:

1. This is an Open Book exam; you are free to use any materials including laptop, tablet and internet.
2. Smart phone and the communication software/APP (e.g., Messenger, IG, LINE, WeChat,..) are prohibited.
3. Copy the R codes and the results from _____ and paste it to this answer sheet.
4. Change the file name of this answer sheet according to your ID and Full Name. Upload the answer sheet to <http://ftp.hmwu.idv.tw:8080/login.html?lang=tchinese>
5. Account: **rege111** , password: classroom number.

(1) **Data file: Grade_Point_Average.csv**

20% **Grade point average.** The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i :	1	2	3	...	118	119	120
X_i :	21	14	28	...	28	16	28
Y_i :	3.897	3.885	3.778	...	3.914	1.860	2.948

- a. Obtain the least squares estimates of β_0 and β_1 , and state the estimated regression function.
- b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?
- c. Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

(1) a.

codes:

```
GPA <- read.csv("C:/Users/User/Desktop/Grade_Point_Average.csv")
colnames(GPA) <- c("Y", "X")
lm(Y ~ X, data = GPA)
```

Outputs:

```
Call:
lm(formula = Y ~ X, data = GPA)
```

Coefficients:

```
(Intercept)      2.11405
X              0.03883
```

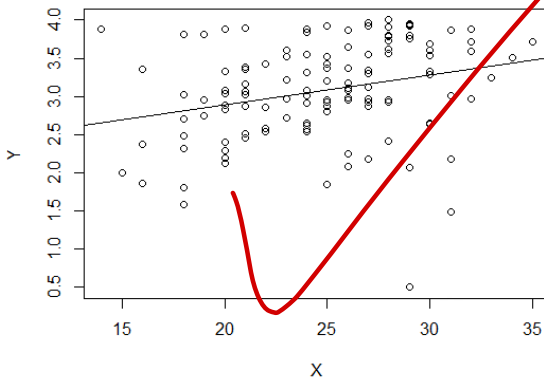
The least squares estimates of β_0 is 2.11405 and β_1 is 0.03883. The estimated regression function is $Y = 2.11405 + 0.03883 X$.

(1) b.

codes:

```
attach(GPA)
plot(X, Y)
abline(lm(Y ~ X, data = GPA))
```

Outputs:



From the above picture, the estimated regression function appear to fit the data well.

(1) c.

codes:

```
y <- 2.11405 + 0.03883 * 30
y
```

Outputs:

3.27895

The point estimate of the mean freshman GPA for students with ACT test score $X = 30$ is 3.27895.

(2) Data file: Grade_Point_Average.csv

20% Refer to **Grade point average**

- a. Set up the ANOVA table.
- b. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.
- c. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model?

(2) a.

codes:

```
GPA.lm <- lm(Y ~ X, data = GPA)
summary.aov(GPA.lm)
```

Outputs:

```
          Df Sum Sq Mean Sq F value Pr(>F)
X             1    3.59   3.588    9.24 0.00292 **
Residuals  118   45.82    0.388
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(2) b.

1. H_1 : Y and X has a linear association.

2. Decision rule:

From (2) a., the p -value of F test is $0.00292 < \alpha = 0.01$, reject H_0 .

3. Conclusion:

Y and X has a linear association.

(2) c

codes:

```
GPA.lm <- lm(Y ~ X, data = GPA)
summary(GPA.lm)
```

Outputs:

```
Call:
lm(formula = Y ~ X, data = GPA)
```

Residuals:
Min 1Q Median 3Q Max
-2.74004 -0.33827 0.04062 0.44064 1.22737

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.11405 0.32089 6.588 1.3e-09 ***
X 0.03883 0.01277 3.040 0.00292 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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 (2) a., $SSR = 3.59$ and from the above result, $R^2 = 0.07262$; coefficient of determination.

(3) Data file: Grade_Point_Average_X.csv

30% Refer to **Grade point average**

- Prepare a box plot for the ACT scores X_i . Are there any noteworthy features in this plot?
- Prepare a dot plot of the residuals. What information does this plot provide?
- Plot the residual e_i against the fitted values \hat{Y}_i . What departures from regression model (2.1) can be studied from this plot? What are your findings?
- Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?
- Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X . Divide the data into the two groups, $X < 26$, $X \geq 26$, and use $\alpha = .01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?
- Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

i :	1	2	3	...	118	119	120
X_2 :	122	132	119	...	140	111	110
X_3 :	99	71	75	...	97	65	85

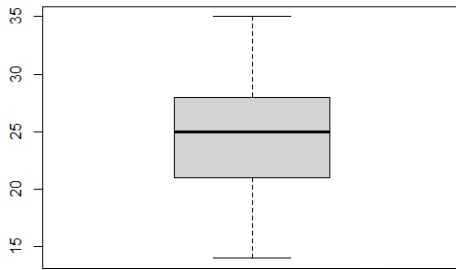
(3) a.
codes:

```
GPA_X <- read.csv("C:/Users/User/Desktop/Grade_Point_Average_X.csv")  
colnames(GPA_X) <- c("Y", "X", "X2", "X3")  
boxplot(GPA_X$X, main = "Boxplot of ACT test score")
```

Outputs:



Boxplot of ACT test score



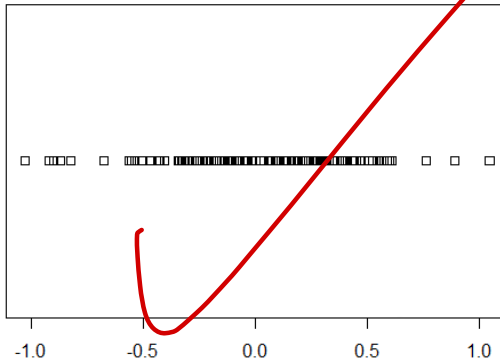
(3) b.

codes:

```
GPA_X.lm <- lm(Y ~ X + X2 + X3, data = GPA_X)
stripchart(GPA_X.lm$residuals, main = "Dot plot of residuals")
```

Outputs:

Dot plot of residuals

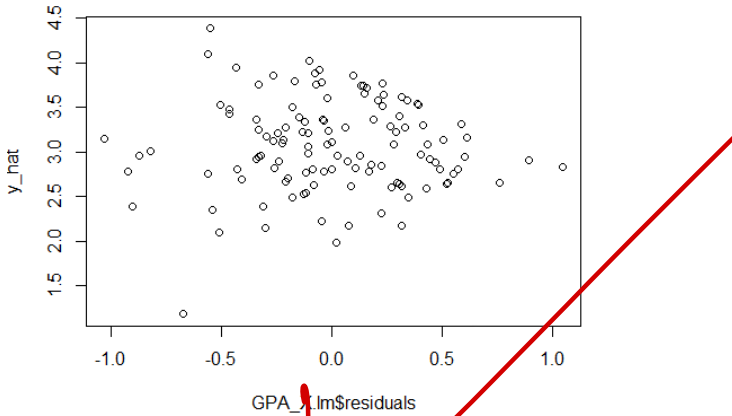


3. (c)

codes:

```
y_hat <- -1.837546 + -0.001509 * GPA_X$X + 0.040314 * GPA_X$X2 + 0.002623 *
GPA_X$X3 plot(GPA_X.lm$residuals, y_hat)
```

Outputs:



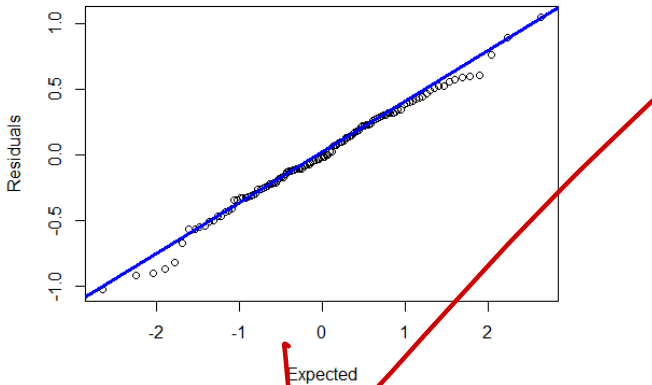
(3) d.
codes:

```
# Normal probability plot
qqnorm(GPA_X.lm$residuals, ylab = "Residuals", xlab = "Expected", main = "Normal
probability plot")
qqline(GPA_X.lm$residuals, col = 'blue', lwd = 3)
# Correlation coefficient
expected <- qqnorm(GPA_X.lm$residuals, ylab = "Residuals", xlab = "Expected",
main = "Normal probability plot")
cor(expected$x, expected$y)
```

Outputs:

Normal probability plot

Normal probability plot



Correlation coefficient:

0.9962909

(3) e.

codes:

```
group <- ifelse(GPA_X$X < 26, 1, 2)
newdata <- data.frame(GPA_X$Y, group)
library(onewaytests)
bf.test(Y ~ as.factor(group), data = newdata, alpha = 0.01)
```

Outputs:

```
Brown-Forsythe Test (alpha = 0.01)
```

```
-----
data : Y and as.factor(group)
statistic : 5.315237
num df : 1
```

```
denom df : 104.0315
```

```
p.value : 0.02312413
```

```
Result : Difference is not statistically significant.
```

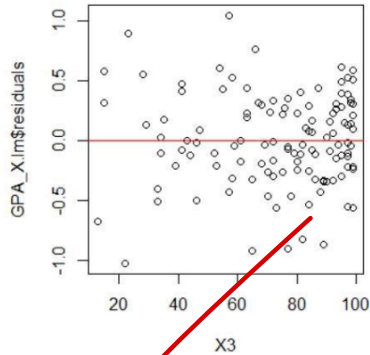
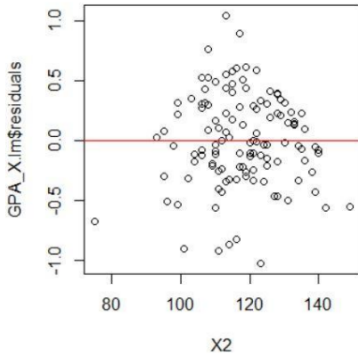
Because the p-value of Brown-Forsythe Test is $0.02312413 > 0.01$, do not reject H_0 : error variance is constant at $\alpha = 0.01$. Hence the error variance is constant.

(3) f.

codes:

```
par(mfrow = c(1, 2))
plot(X2, GPA_X.lm$residuals)
abline(h = 0, col = "red")
plot(X3, GPA_X.lm$residuals) abline(h = 0, col = "red")
```

Outputs:



The model could be improved by adding the X2 to the model because from the graph above, X2 seems to have a linear relationship with the residuals while the rank had no clear trend with the residuals. In addition, after fitting the GPA with the X2 and X3, the summary says that the X2 variable is a statistically significant predictor of the GPA.

(4) [Data file: Solution_concentration.csv](#)

30% **Solution concentration.** A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The results follow.

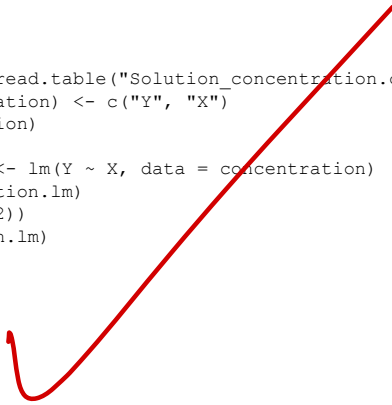
i :	1	2	3	...	13	14	15
X_i :	9	9	9	...	1	1	1
Y_i :	.07	.09	.08	...	2.84	2.57	3.10

- Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?
- Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?
- Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.
- Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?
- Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?
- Express the estimated regression function in the original units.

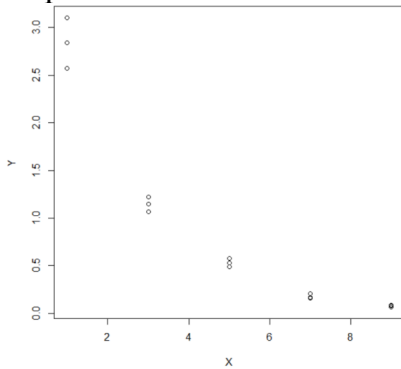
(4) a.

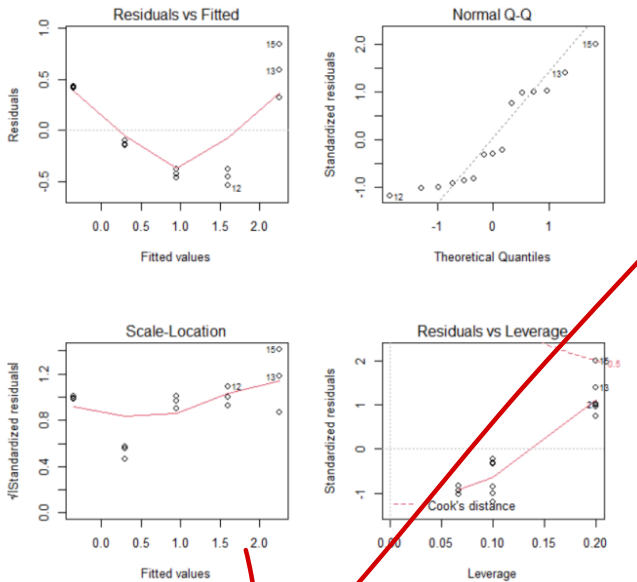
codes:

```
concentration <- read.table("Solution_concentration.csv", sep = ",")
colnames(concentration) <- c("Y", "X")
attach(concentration)
plot(X, Y)
concentration.lm <- lm(Y ~ X, data = concentration)
summary(concentration.lm)
par(mfrow = c(2, 2))
plot(concentration.lm)
```



Outputs:





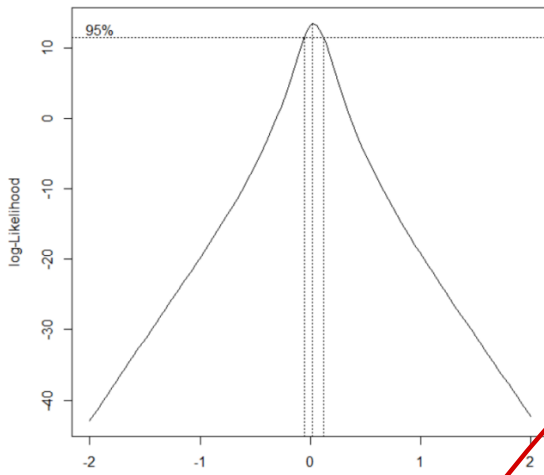
From the output, we can see that unequal error variances and nonnormality of the error terms. To remedy these departures from the simple linear regression model, we need a transformation on Y , since the shapes and spreads of the distributions of Y need to be changed.

(4) b.

codes:

```
library(MASS)
library(ALSM)
boxcox(concentration.lm, lambda = seq(-2, 2, by = 0.1))
boxcox.sse(X, Y, l = seq(-2, 2, by = 1))
```

Outputs:



```
> boxcox.sse(X, Y, l = seq(-2, 2, by = 1))
lambda      SSE
1      -2    68.84280491
2      -1     3.16846767
5       0     0.03897303
3       1     2.92465333
4       2    64.15599383
```

From the output, we take $Y' = \log Y$

(4) c.

codes:

```
concentration.log.lm <- lm(log(Y) ~ X, data = concentration)
summary(concentration.log.lm)
```

Outputs:

```
> summary(concentration.log.lm)
Call:
lm(formula = log(Y) ~ X, data = concentration)
Residuals:
Min 1Q Median 3Q Max
-0.19102 -0.10228  0.01569  0.07716  0.19699
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.50792  0.06028  25.01 2.22e-12 ***
X -0.44993  0.01049  -42.88 2.19e-15 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.115 on 13 degrees of freedom
```

Multiple R-squared: 0.993, Adjusted R-squared: 0.9924
F-statistic: 1838 on 1 and 13 DF, p-value: 2.188e-15

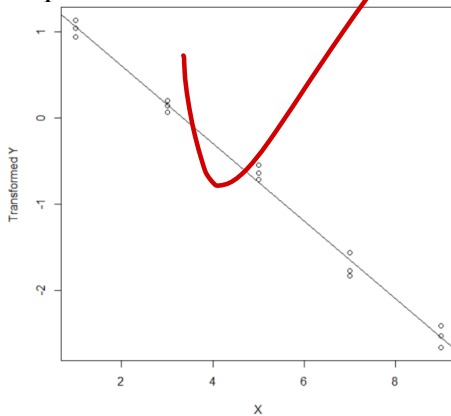
The least squares estimates of β_0 is 1.50792 and β_1 is -0.44993.
The estimated regression function is $\hat{Y} = 1.50792 - 0.44993 X$.

(4) d.

codes:

```
par(mfrow = c(1, 1))  
plot(X, log(Y), xlab = "X", ylab = "Transformed Y")  
abline(concentration.log.lm)
```

Outputs:



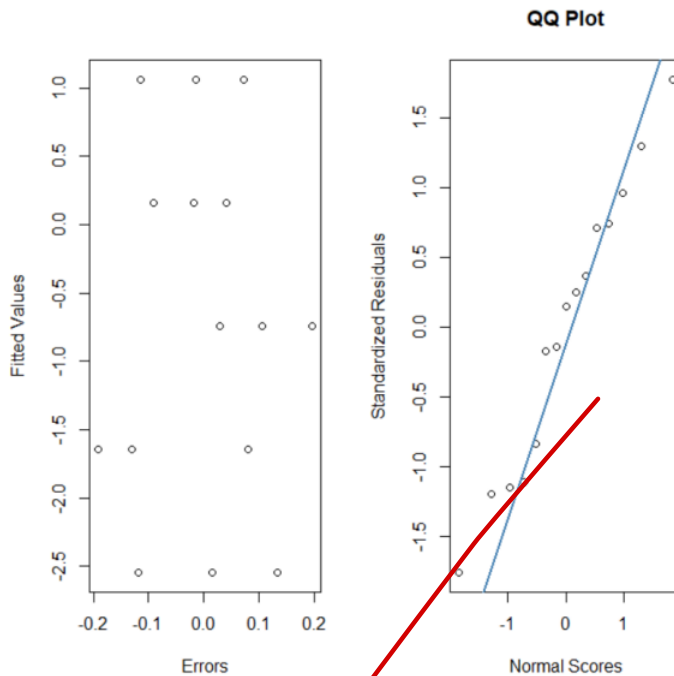
From the results it was a perfect fit.

(4) e.

codes:

```
ei <- concentration.log.lm$residuals  
yhat <- concentration.log.lm$fitted.values  
par(mfrow = c(1, 2))  
plot(ei, yhat, xlab = "Errors", ylab = "Fitted Values")  
stdei <- rstandard(concentration.log.lm)  
qqnorm(stdei, ylab = "Standardized Residuals", xlab = "Normal Scores", main = "QQ  
Plot")  
qqline(stdei, col = "steelblue", lwd = 2)
```

Outputs:



From the output, we see that error variances are constant, errors are approximately normally distributed.

(4) f.

Since $\log_e(Y) = 1.50792 - 0.44993 X$.

Hence $Y = \exp(1.50792 - 0.44993 X)$