Stat 416/615, Regression Homework 2 Solution

Jun Lu, Mingfan Meng

Fall 2023

Contents

Total: 85 pts

1. Reading assignment

2. Start Lab 3

3. Problem 3.17 (skip b) Sales growth (25 pts, 5 for each sub-question.)

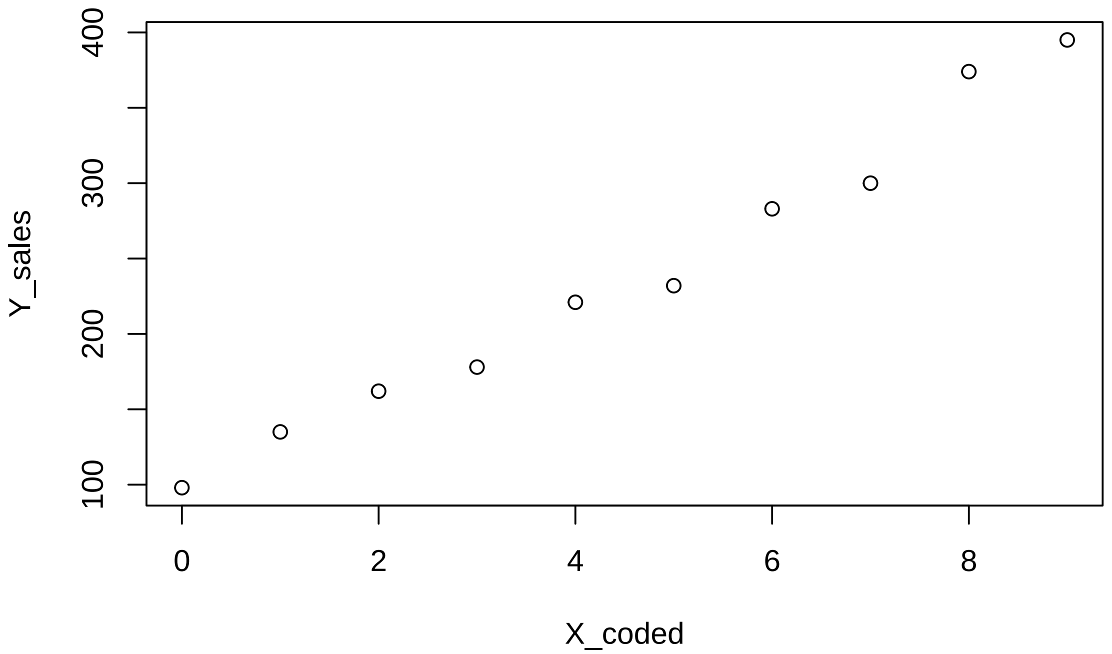
A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands of units.

```
sg <- read.table("./CH03PR17.txt", header = F)
colnames(sg) <- c("Y_sales", "X_coded")
head(sg)
```

```
##   Y_sales X_coded
## 1      98       0
## 2     135       1
## 3     162       2
## 4     178       3
## 5     221       4
## 6     232       5
```

a. Prepare a scatter plot of the data. Does a linear relation appear adequate here?

```
plot(Y_sales ~ X_coded, data=sg)
```



- There appears to be a positive linear association between the variables. The sales (Y) increases as year (X) increases. The variation of Y gets bigger as X increases.
- A linear relationship appears adequate.

b. Skipped.

c. Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

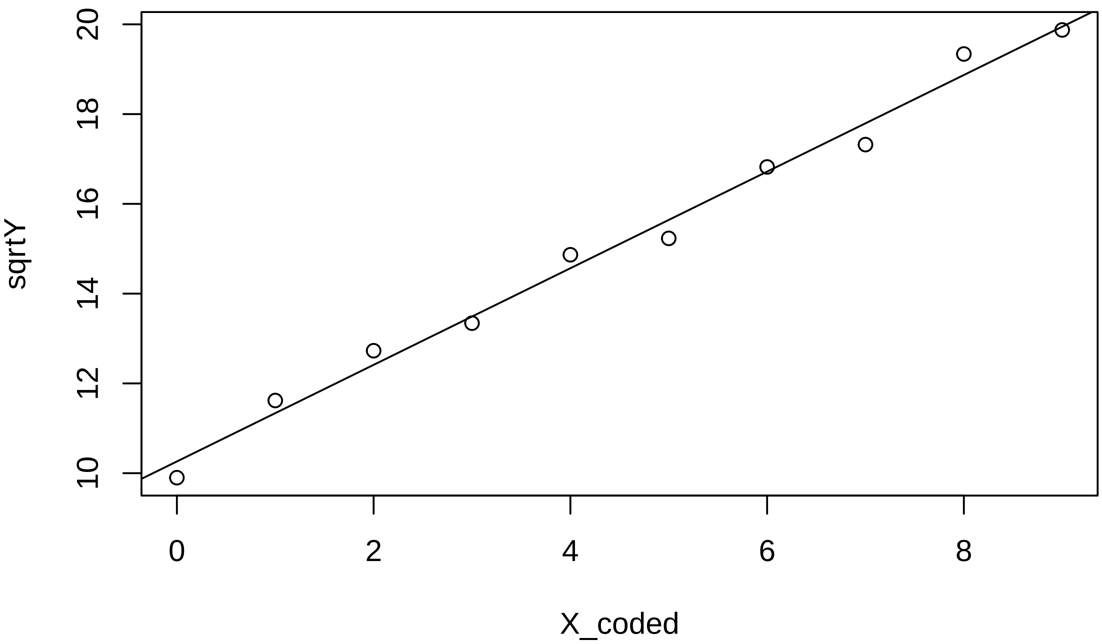
```
sg$sqrtY<-sqrt(sg$Y_sales)
sgSqrtY.SLR<-lm(sqrtY ~ X_coded, data=sg)
sgSqrtY.SLR
```

```
##
## Call:
## lm(formula = sqrtY ~ X_coded, data = sg)
##
## Coefficients:
## (Intercept)      X_coded
##      10.261        1.076
```

- According to above results, the intercept(β_0) is 10.261, and the slope(β_1) is 1.076. The estimated linear regression function is $\hat{Y_i'} = 10.261 + 1.076 * (X_i)$, when $Y' = \sqrt{(Y)}$.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

```
plot(sqrtY ~ X_coded, data=sg)
abline(sgSqrtY.SLR$coefficients)
```



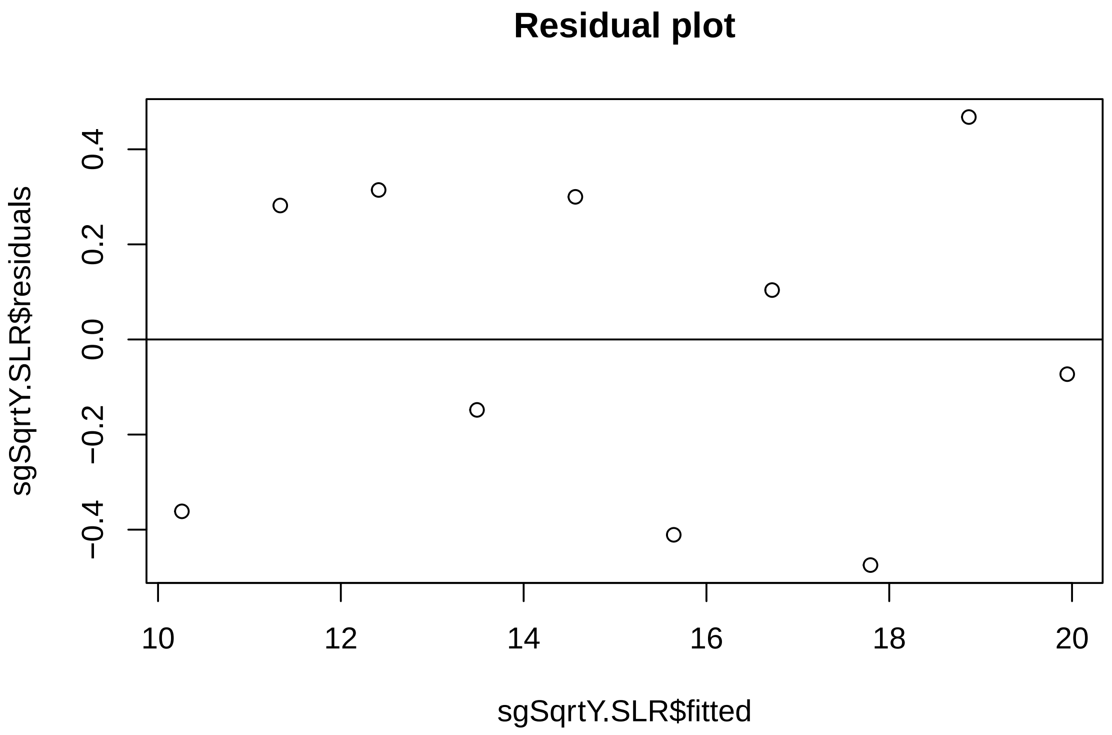
- According to the plot, the dot is around the regression line and there is no outlier. It appears to be a good fit to the transformed data.

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

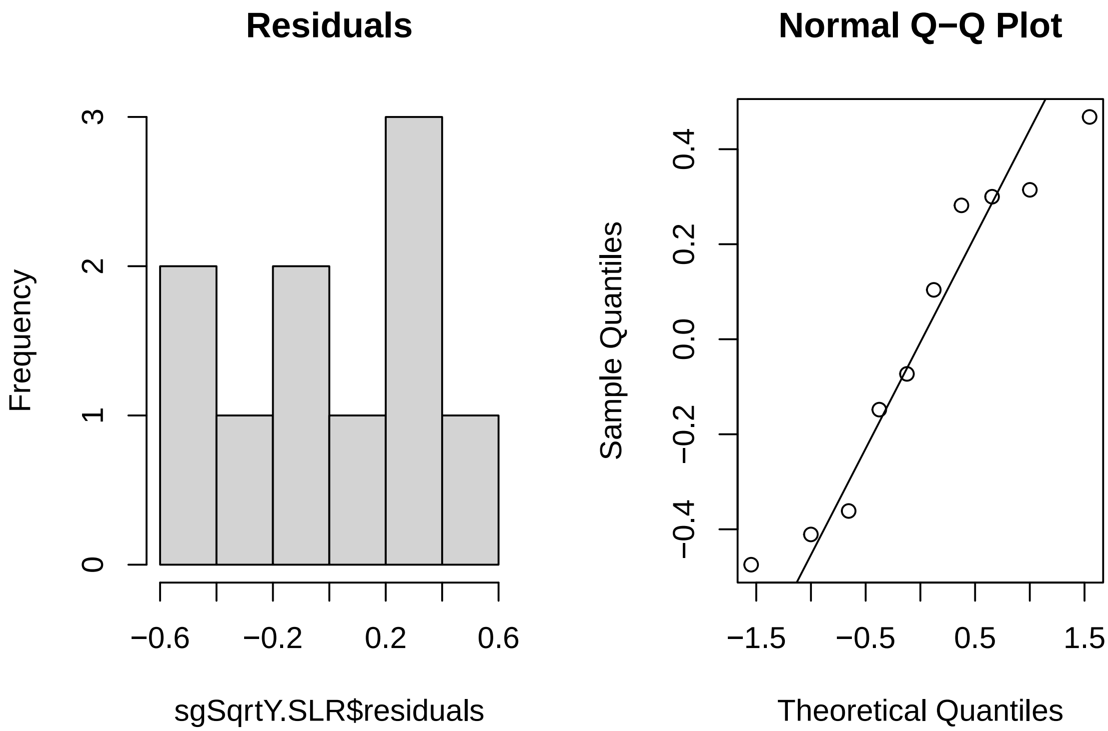
```
head(sgSqrtY.SLR$residuals)
```

```
##          1          2          3          4          5          6 
## -0.3614366  0.2817268  0.3144070 -0.1481427  0.2999702 -0.4108441
```

```
plot(sgSqrtY.SLR$fitted, sgSqrtY.SLR$residuals, main="Residual plot")
abline(0,0)
```



```
par(mfrow=c(1, 2))
hist(sgSqrtY.SLR$residuals, main="Residuals")
qqnorm(sgSqrtY.SLR$residuals)
qqline(sgSqrtY.SLR$residuals)
```



```
shapiro.test(sgSqrtY.SLR$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sgSqrtY.SLR$residuals
## W = 0.91651, p-value = 0.3288
```

- The variance appears to be constant for the residuals.Meanwhile, there might exits a curvilinear pattern. The Q-Q plot shows that the normality assumption might be violated. However, with only 10 observations, it is difficult to make conclusions.
- Note that the F-test of Lack-of-fit (i.e., F-test of linearity) can not be applied to this data set because all observations have unique x-values.

f. Express the estimated regression function in the original units.

- $\hat{Y} = (10.261 + 1.076X)^2$

4. Problem 4.7 Copier maintenance. (20 pts. a-10, b-5, c-5)

```
cm <- read.table("./CH01PR20.txt", header = F)
colnames(cm) <- c("minute", "copier")
cm.SLR <- lm(minute ~ copier, data = cm)
```

a. Estimate the expected number of minutes spent when there are 3, 5, and 7 copiers to be serviced, respectively. Use interval estimates with a 90 percent family confidence coefficient based on the Working-Hotelling procedure.

- The regression function is $\widehat{Totalminutes} = -0.58 + 15.03(copier)$

- The family confidence level is 0.9. I.e., the family $\alpha = 0.1$. There are 3 intervals in this family (g=3).
- Working-Hotelling procedure (in SLR)
  - CI: $\hat{y} \pm W * se(\hat{y}_{mean})$
  - $\hat{y}$ and $se(\hat{y}_{mean})$ are calculated using the `predict()` function in R. (Other software can calculate and save them as well.)
  - $W = \sqrt{2F_{(1-\alpha;\, 2,\, n-2)}} = \sqrt{2F_{(0.9;\, 2,\, 45-2)}} = 2.2047$

```
W<- sqrt(2 * qf(1-0.1, 2, 43))
W
```

```
## [1] 2.204725
```

```
cm.CI <- predict(cm.SLR, newdata = data.frame(copier=c(3, 5, 7)),
                  se.fit = TRUE, interval = "confidence", level = 1- (0.1/3))
cm.CI
```

```
## $fit
##         fit       lwr       upr
## 1  44.52559  40.84285  48.20832
## 2  74.59608  71.67227  77.51989
## 3 104.66658 101.12260 108.21056
##
## $se.fit
##        1        2        3
## 1.675012 1.329831 1.611900
##
## $df
## [1] 43
##
## $residual.scale
## [1] 8.913508
```

```
Yhat <- cm.CI$fit[, 1]  # Y_hat
se <- cm.CI$se.fit  # se(Y_hat_mean)

# R is used as a "calculator" here.
# You do not have to "code" the computation.
lower <- Yhat - W * se
upper<- Yhat + W * se
data.frame(X=c(3,5,7), Yhat, se, WHlower = lower, WHupper = upper)
```

```
##   X      Yhat       se   WHlower   WHupper
## 1 3  44.52559 1.675012  40.83265  48.21853
## 2 5  74.59608 1.329831  71.66417  77.52800
## 3 7 104.66658 1.611900 101.11278 108.22038
```

- Conclusion: At 90% family confidence level, using Working-Hotelling method,
  - When $X = 3$, the CI for mean service time is (40.83, 48.22)
  - When $X = 5$, the CI for mean service time is (71.66, 77.53)
  - When $X = 7$, the CI for mean service time is (101.11, 108.22)
- Bonferroni procedure
  - CI: $\hat{\mu} \pm B * s(\hat{\mu})$
  - $B = t_{(1-\alpha/(2g), df=df_E=(n-2)} = t_{(1-0.1/6, 45-2)} = 2.1986$ (could be 2.202 due to rounding.)

```
B <- qt(1- 0.1/(2*3), 43)
B
```

```
## [1] 2.198632
```

```
lower <- Yhat - B * se
upper<- Yhat + B * se
lower
```

```
##         1         2         3
##  40.84285  71.67227 101.12260
```

```
upper
```

```
##         1         2         3
##  48.20832  77.51989 108.21056
```

```
cbind(lower, upper)
```

```
##       lower     upper
## 1  40.84285  48.20832
## 2  71.67227  77.51989
## 3 101.12260 108.21056
```

- Conclusion. At 90% family confidence level, using Bonferroni method,
  - When $X = 3$, the CI for mean service time is (40.84, 48.21)
  - When $X = 5$, the CI for mean service time is (71.67, 77.52)
  - When $X = 7$, the CI for mean service time is (101.12, 108.21)
- Compare the two procedure, Bonferroni produces narrower confidence interval.
- Note that the Bonferroni method can be directly implemented in R. We can set each member's confidence interval in `level =` in the `predict()` function.

```
predict(cm.SLR, newdata = data.frame(copier=c(3, 5, 7)),
                  interval = "confidence", level = 1-0.1/3)
```

```
##         fit       lwr       upr
## 1  44.52559  40.84285  48.20832
## 2  74.59608  71.67227  77.51989
## 3 104.66658 101.12260 108.21056
```

b. Two service calls for preventive maintenance are scheduled in which the numbers of copiers to be serviced are 4 and 7, respectively. A family of prediction intervals for the times to be spent on these calls is desired with a 90 percent family confidence coefficient. Which procedure, Scheffe or Bonferroni, will provide tighter prediction limits here?

- The family confidence level is 0.9. I.e., the family $\alpha = 0.1$. There are 2 intervals in this family (g=2).
- Note that we will compute the prediction interval because the question asks the times to be spent on these 2 individual service calls. (Not the "mean service time".)
- $B = t_{(1-\alpha/(2g), df=df_E=(n-2))} = t_{(1-0.1/4,\, 45-2)} = 2.0167$

```
B <- qt(1- 0.1/(2*2), 43)
B
```

```
## [1] 2.016692
```

- $S = \sqrt{gF_{(1-\alpha;\, g,\, n-2)}} = \sqrt{2F_{(0.9;\, 2,\, 45-2)}} = 2.2047$

```
S <- sqrt(2*qf(0.9, 2, 43))
S
```

```
## [1] 2.204725
```

- Conclusion : the Bonferroni will provide the tighter prediction limits because it has a smaller critical value.

c. Obtain the family of prediction intervals required in part (b), ~~using the more efficient procedure~~ *using both Scheffe and Bonferroni procedures.*

- Bonferroni procedure
  - Use `predict()` directly. Set `level=` for each interval.

```
cm.PI <- predict(cm.SLR, newdata = data.frame(copier=c(4, 7)),
                  se.fit = TRUE, interval = "prediction", level = 1-0.1/2)
cm.PI
```

```
## $fit
##         fit      lwr       upr
## 1  59.56084 41.35419  77.76748
## 2 104.66658 86.39922 122.93394
##
## $se.fit
##        1        2
## 1.433068 1.611900
##
## $df
## [1] 43
##
## $residual.scale
## [1] 8.913508
```

  - Use formula-based calculation. PI: $\hat{y} \pm B * se(\hat{y}_{new})$. B is worked in part (b). $se(\hat{y}_{new}) = \sqrt{(se(\hat{y}_{mean}))^2 + MSE}$, where $se(\hat{y}_{mean})$ and MSE can be obtained from the above output.

```
Yhat.new <- cm.PI$fit[, 1]
se.new <- sqrt(cm.PI$se.fit^2 + cm.PI$residual.scale^2)
```

```
# Bonferroni
lower <- Yhat.new - B*se.new
upper <- Yhat.new + B*se.new
cbind(c(4, 7), Yhat.new, se.new, lower, upper)
```

```
##      Yhat.new   se.new    lower     upper
## 1 4  59.56084 9.027974 41.35419  77.76748
## 2 7 104.66658 9.058082 86.39922 122.93394
```

  - At 90% family confidence level, using Bonferroni method, the PI for the service time is (41.35, 77.77) for 4 copiers; and (86.40, 122.94) for 7 copiers.

- Scheffe Procedure. (S is worked out in part (b).)
  - $\hat{Y_{new}} \pm S * s(\hat{Y_{new}})$

```
lower <- Yhat.new - S * se.new
upper <- Yhat.new + S * se.new
cbind(lower, upper)
```

```
##      lower     upper
## 1 39.65663  79.46504
## 2 84.69600 124.63716
```

  - At 90% family confidence level, using Scheffe method, the PI for the service time is (39.66, 79.47) for 4 copiers; and (84.70, 124.64) for 7 copiers.

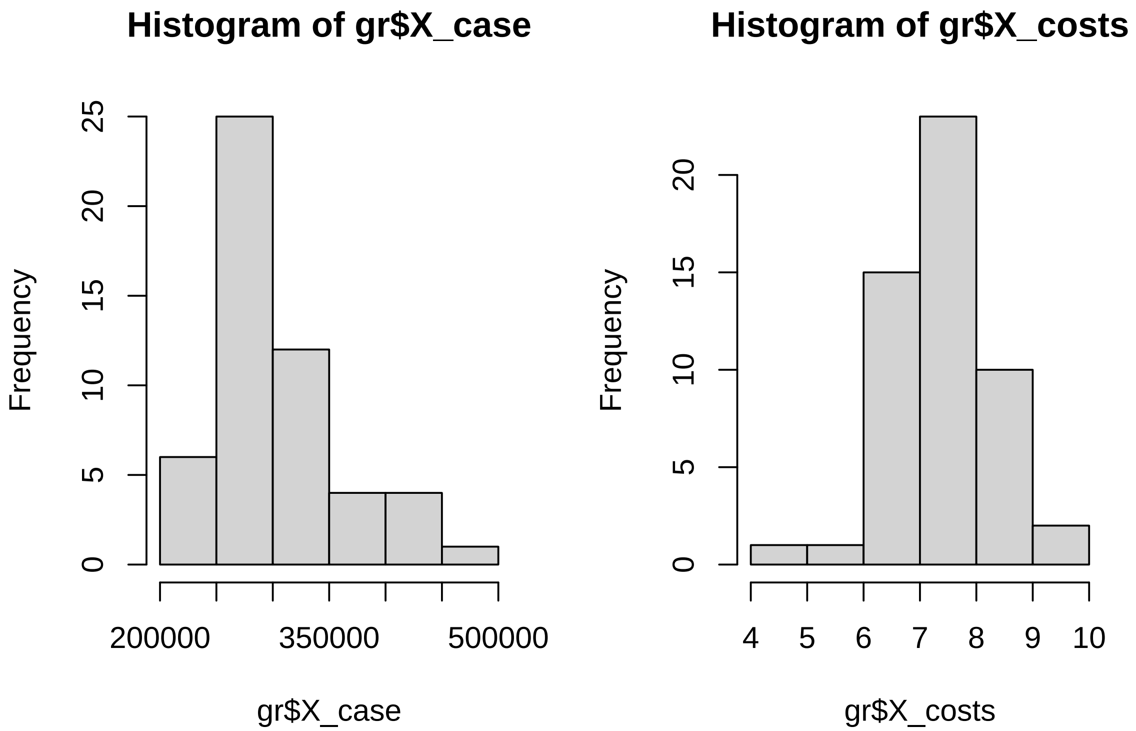
5. Problem 6.9. Grocery retailer (15 pts)

A large, national grocery retailer tracks productivity and costs of its facilities closely. Data below were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1) the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y).

```
gr <- read.table("./CH06PR09.txt", header = F)
colnames(gr) <- c("Y_labor", "X_case", "X_costs", "X_holiday")
```

a. Prepare separate stem-and-leaf plots for the number of cases shipped X_{i1} and the indirect cost of the total hours X_{i2}. Are there any outlying cases present? Are there any gaps in the data?

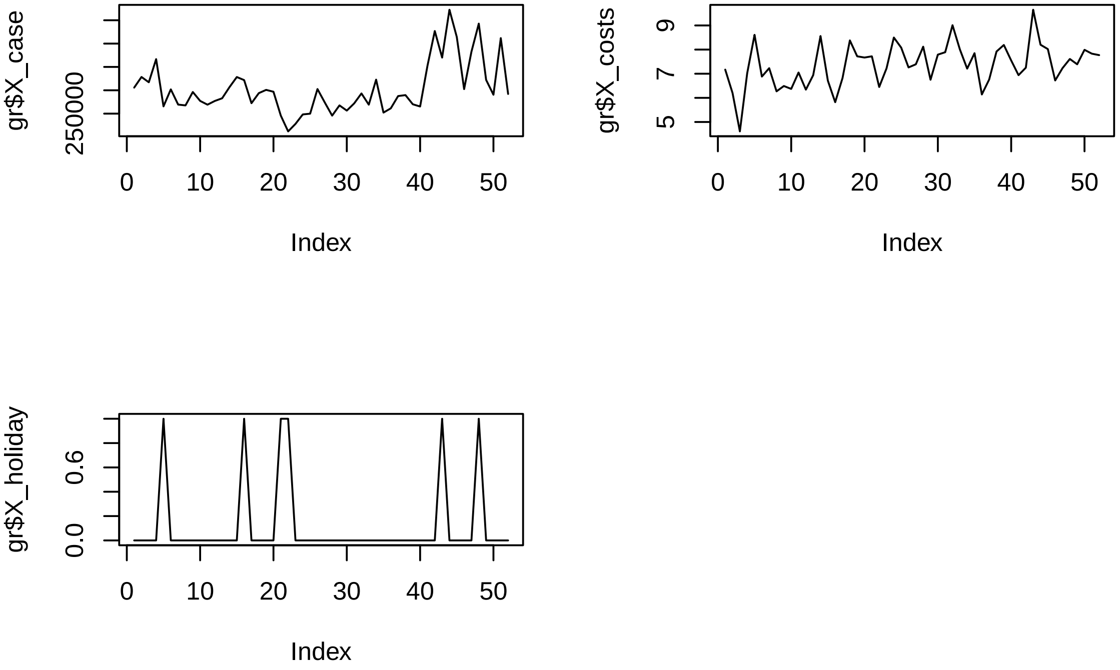
```
par(mfrow=c(1,2))
hist(gr$X_case)
hist(gr$X_costs)
```



- Variable "case shipped" is skewed to the right, and variable "indirect labor hours" is slightly skewed to the left. Neither histogram shows strong indication of outliers.

b. The cases are given in consecutive weeks. Prepare a time plot for each predictor variable. What do the plots show?

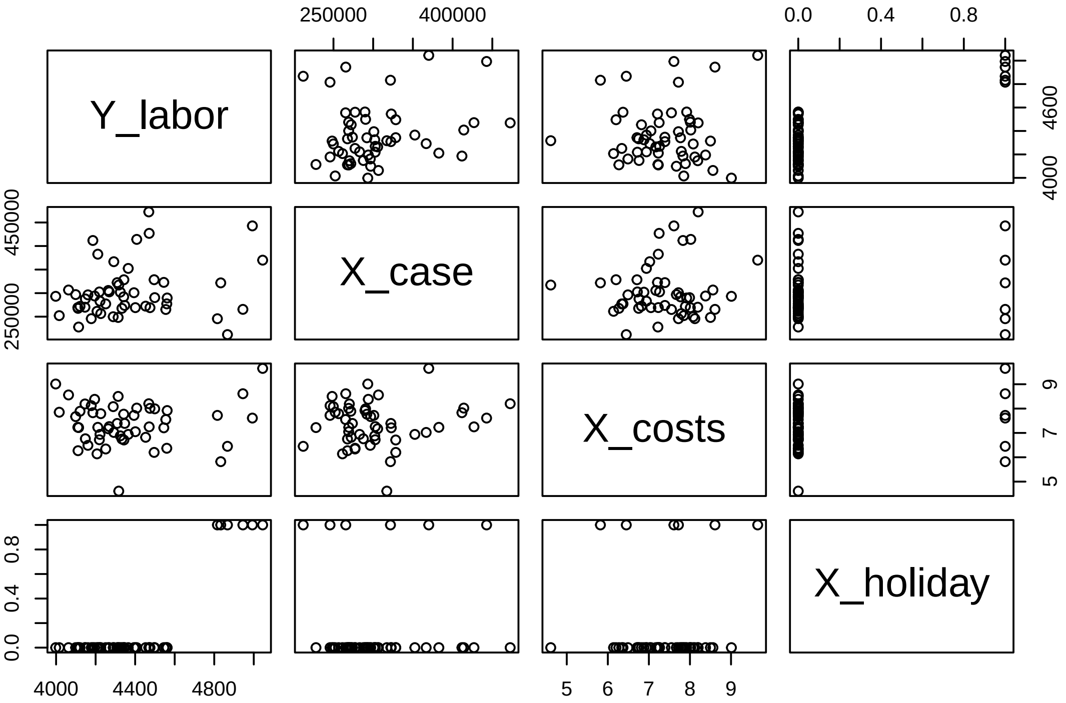
```
par(mfrow=c(2,2))
plot(gr$X_case, type="l")
plot(gr$X_costs, type="l")
plot(gr$X_holiday, type="l")
```



- None of the time plots shows obvious time trend or seasonal change. The spikes in "Holiday" is due to the fact that it is a 2-value (0-1) variable (no deduction if not mentioned). Some may state that (no deduction if the following is not stated):
    - "case shipped" appears to have a "jump" after week 40.
    - "cost" has a (weak) increasing trend.

c. Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?

```
pairs(gr)
```



```r
cor(gr)
```

```
##                Y_labor     X_case    X_costs  X_holiday
## Y_labor   1.0000000 0.20766494 0.06002960 0.81057940
## X_case    0.2076649 1.00000000 0.08489639 0.04565698
## X_costs   0.0600296 0.08489639 1.00000000 0.11337076
## X_holiday 0.8105794 0.04565698 0.11337076 1.00000000
```

- The response variable “Total labor hours” appears to be strongly associated with “Holiday.” Association between any other pairs of variables appears to be weak.

6. Problem 6.10 Grocery retailer (25 pts: a,b,d-5, c-10)

a. Fit regression model (6.5) to the data for three predictor variables. State the estimated regression function. How are b_1 , b_2 , and b_3 interpreted here?

```r
gr.mlr <- lm(Y_labor ~ X_case + X_costs+ X_holiday, data = gr)
gr.mlr
```

```
##
## Call:
## lm(formula = Y_labor ~ X_case + X_costs + X_holiday, data = gr)
##
## Coefficients:
## (Intercept)       X_case      X_costs    X_holiday
##   4.150e+03    7.871e-04   -1.317e+01    6.236e+02
```

- $\hat{Y} = 4150 + 0.000787 * X_1 - 13.166 * X_2 + 623.554 * X_3$
- b_1, b_2, b_3 are all the slope of the each predictor variable.

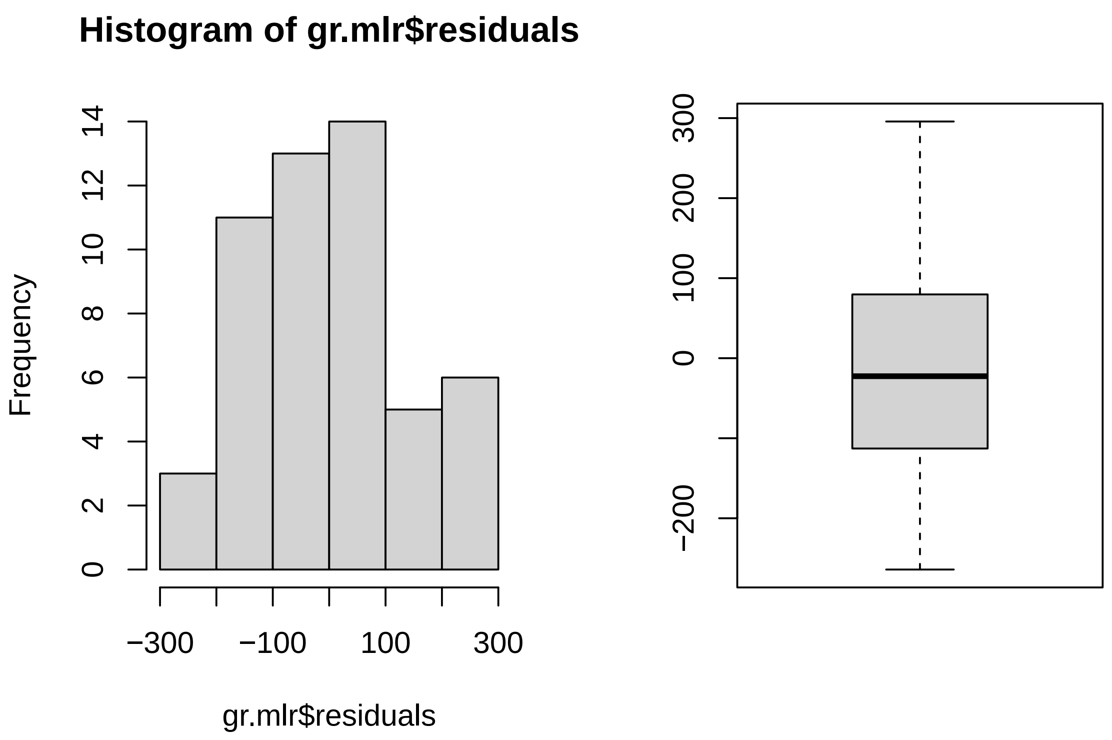
  - For every unit increase in **case shipped** (X_1), the mean **total labor hour** (response, Y) is expected to increase 0.000787, after adjusting for the effects of **indirect cost** (X_2) and **holiday** (X_3).
  - For every unit increase in **indirect cost** (X_2), the mean **total labor hour** (response, Y) is expected to decrease 13.166 (change -13.166), after adjusting for the effects of **case** and **holiday**.
  - Changing **holiday** (X_3) from non-holiday (0) to holiday (1), the mean **total labor hour** (response, Y) is expected to increase 623.554, after adjusting for the effects of **case shipped** and **indirect cost**.

b. Obtain the residuals and prepare a boxplot of the residuals. What information does this plot provide?

```
head(gr.mlr$residuals, 5)
```

```
##         1         2         3         4         5
## -32.06348 169.20509 -21.82543 -54.11955  75.93372
```

```
par(mfrow=c(1,2))
hist(gr.mlr$residuals)
boxplot(gr.mlr$residuals)
```

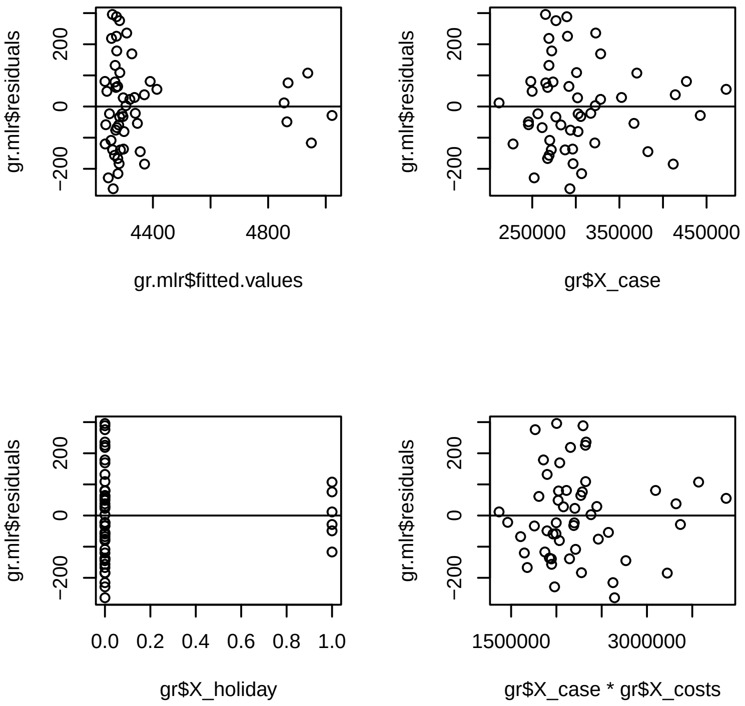


- According to the boxplot and histgram, the mean is closely to 0 and the dot limit in (-300, 300) range. Beside, the variances are nearly equal, which suggest the residuals may follow Normal distribution.

c. Plot the residuals against $\hat{Y}$, X_1, X_2 , X_3, and X_1 X_2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

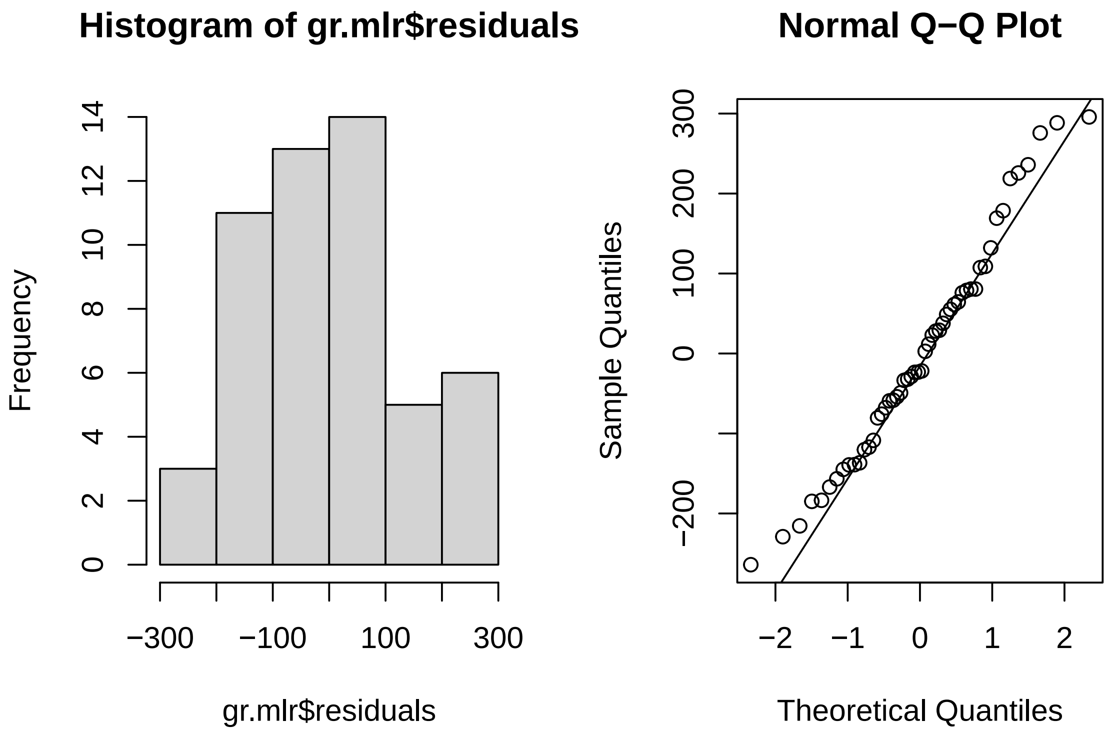
```
par(mfrow=c(2,3))
plot(gr.mlr$fitted.values, gr.mlr$residuals)
abline(c(0,0))
plot(gr$X_case, gr.mlr$residuals)
```

```
abline(c(0,0))
plot(gr$X_costs, gr.mlr$residuals)
abline(c(0,0))
plot(gr$X_holiday, gr.mlr$residuals)
abline(c(0,0))
plot(gr$X_case*gr$X_costs, gr.mlr$residuals)
abline(c(0,0))
```



- According to the separate graphs, the dots in the case numbers residual plots, labor hours costs residuals plots and case numbers times costs residual plot are around the 0 and variances appear to be constant.
- Residuals plots against the predicted values $\hat{Y}$ and the predictors **holiday** reveal that observations from non-holidays may have a larger variance than those from holidays, i.e., the constant variance assumption of the model (residuals) may not be valid.
- Residual vs X_1X_2 does not show any pattern. We do not need to add X_1X_2 to the analysis.

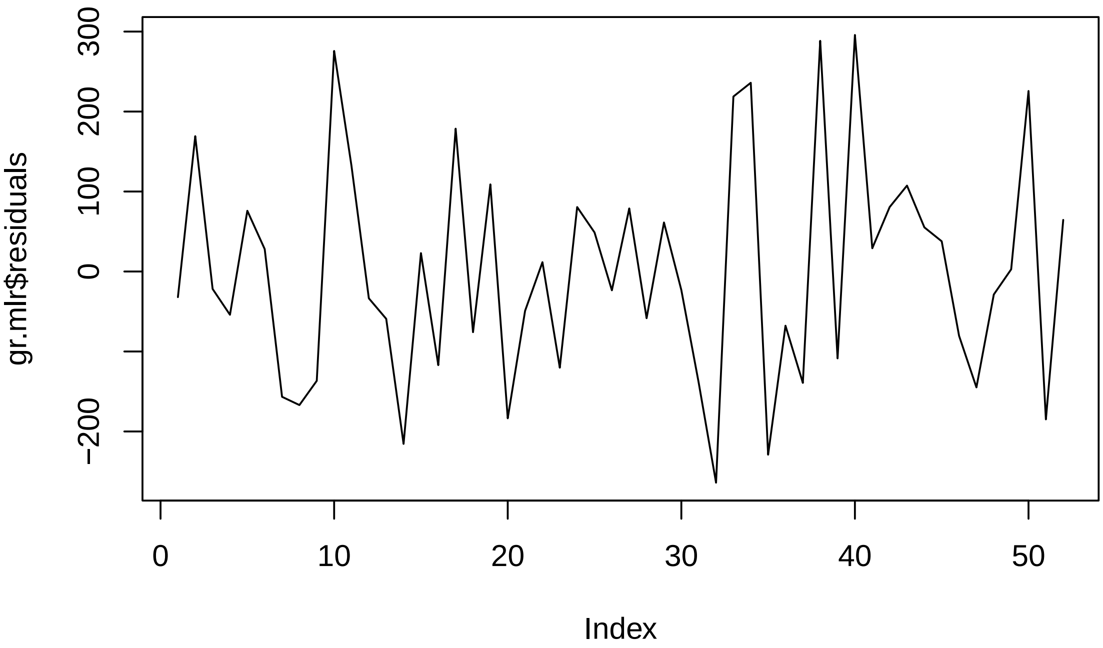
```
par(mfrow=c(1,2))
hist(gr.mlr$residuals)
qqnorm(gr.mlr$residuals)
qqline(gr.mlr$residuals)
```



- According to the histogram plot and normal Q-Q plot, the histogram plot shows the normal distribution and the dots on the normal Q-Q plot are around the regression lines.we can assume the residuals are Normally distributed.

d. Prepare a time plot of the residuals. Is there any indication that the error terms are correlated? Discuss.

```
plot(gr.mlr$residuals, type = "l")
```



- The residual versus time plot did not show any evidence that the error terms were correlated over time.

—— This is the end of this assignment. ——