Regression, Homework 3 Solutions

Jun Lu, Mingfan Meng

Fall 2023

Contents

Total: 35 pts

1. Complete Lab 1

2. Start Lab 2

3. Reading: Ch.2.3, Ch.2.10.

Load Copier Maintenance data and run regression analysis for the rest of the assignment.

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

```
##
## Call:
## lm(formula = minute ~ copier, data = cm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## copier       15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

4. Problem 2.24 (c, d) Copier maintenance (10 pts)

Recall 2.24(b), the ANOVA table, from the previous homework.

```
anova(cm.SLR)
```

```
## Analysis of Variance Table
##
## Response: minute
##           Df Sum Sq Mean Sq F value    Pr(>F)
## copier     1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pr.2.24.c. By how much, relatively, is the total variation in number of minutes spent on a call- reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

- From the ANOVA table, $R^2 = SS_{Regression}/SS_{Total} = 76960/(76960 + 3416) = 0.9575$ which equal to the "Multiple R-squared:" in the output of `summary(cm.SLR)`.
- R^2 is the "coefficient of determination".
- It is a relatively large reduction.

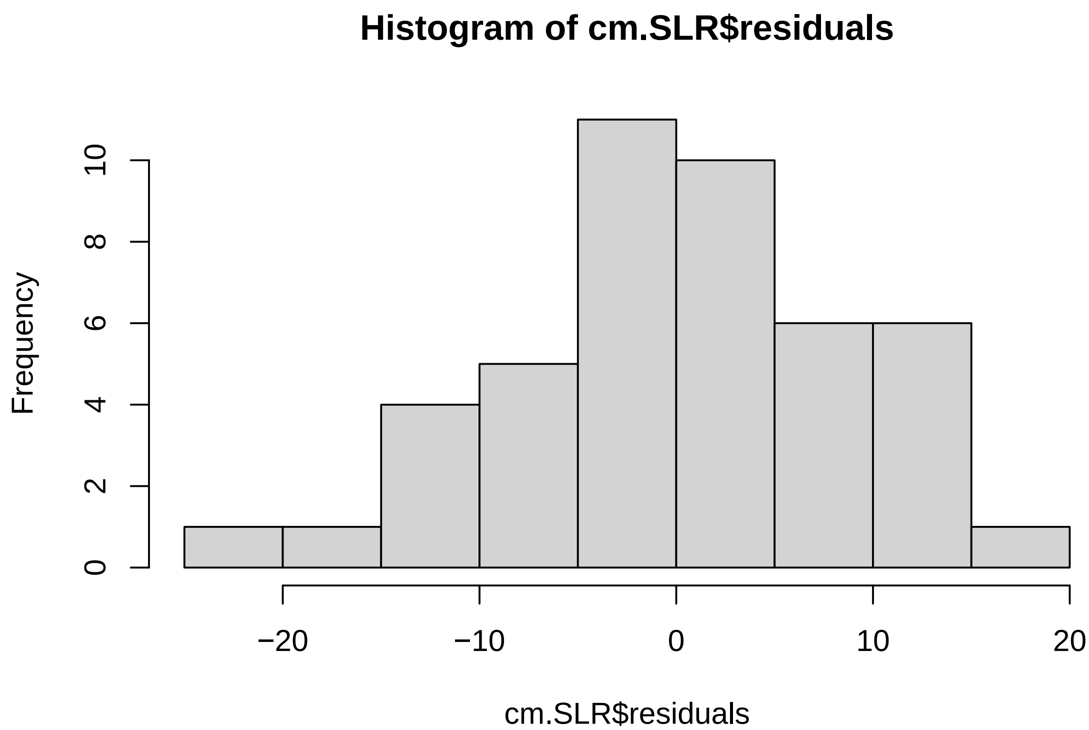
Pr.2.24.d. Calculate r and attach the appropriate sign.

- In general, if $R^2 = (r)^2 = 0.9575$, $r = \pm\sqrt{R^2}$. The sign of r is determined by the direction of the linear association.
- In this case, the estimated slope (15.03) is positive, meaning there is a positive association (i.e., Y increases as X increases) between the variables. Hence, $r = +\sqrt{0.9575} = +0.9785$.
- r is the correlation coefficient between Y and X.

4. Problem 3.4.(c, d, h, e) Copier maintenance (20 pts)

a. Omitted.
b. Omitted.
c. **Prepare a stem-and-leaf plot of the residuals. Are there any noteworthy features in this plot?**

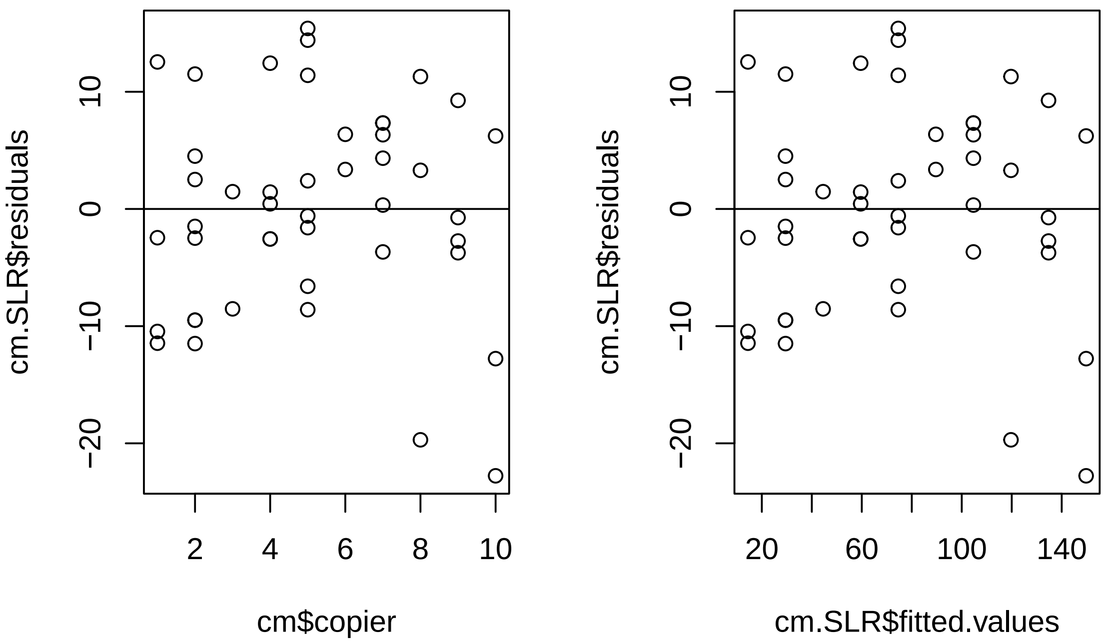
```
hist(cm.SLR$residuals)
```



- Histogram plot of the residual show that the residuals appear to be symmetric and bell-shaped.

d. **Prepare residual plots of e_l versus $\hat{Y}$; and e_l versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings.**

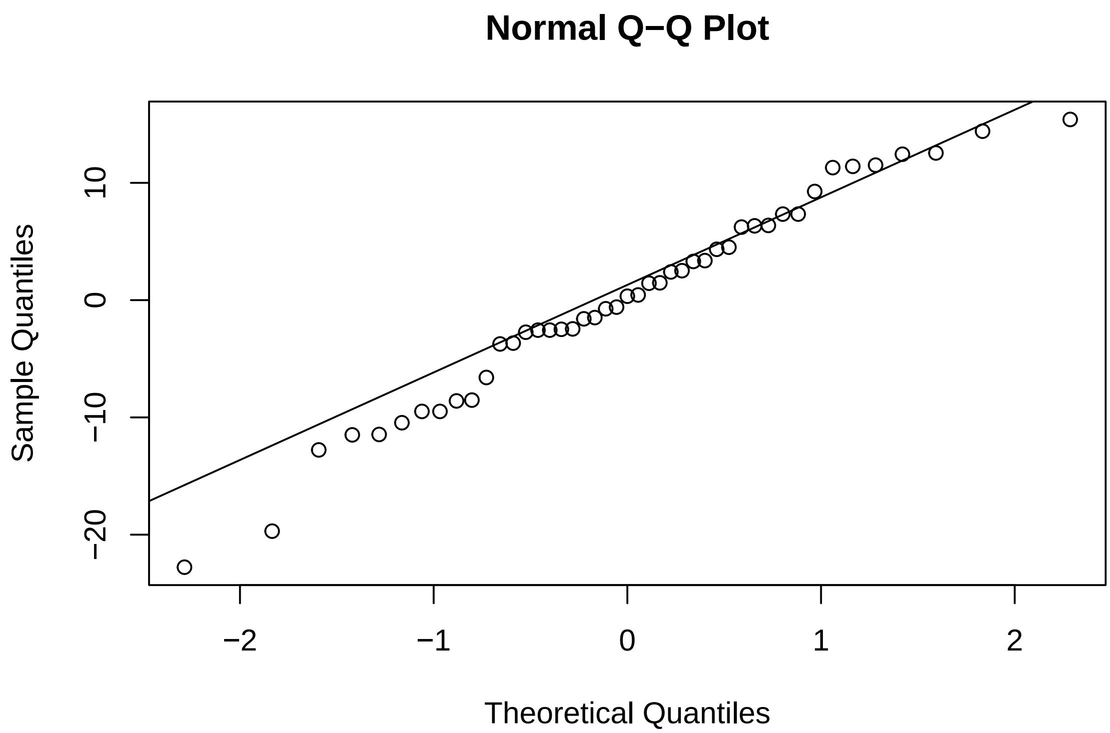
```
par(mfrow = c(1,2))
plot(cm$copier, cm.SLR$residuals)
abline(0,0)
plot(cm.SLR$fitted.values, cm.SLR$residuals)
abline(0,0)
```



- Both residual plots will provide the same information. This is because in SLR, $\hat{Y}$ is just a linear function of x. There might be a curved trend based on the residual plots, though it doesn't appear to be significant. One may suggest there might be outliers (not severe though) near the lower-right corner of the plot. Mentioned one of them for full credit.

e. **(Revised the question in the textbook.) Prepare a Normal Q-Q plot for the residuals and comment. Use statistical software to conduct test(s) to check the normality assumption of the residuals at $\alpha = 0.10$.**

```
qqnorm(cm.SLR$residuals)
qqline(cm.SLR$residuals)
```



- The Q-Q plot from SPSS/R looks reasonably good: the dots are roughly scattered around the reference line randomly. There is minor deviation near the tails, but it is not severe. We consider the residuals are Normally distributed.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  cm.SLR$residuals
## W = 0.97583, p-value = 0.4614
```

- The Shapiro test resulted in a p-value 0.4615, greater than any common choice of signficance level. Hence, we do NOT reject H_0. We can assume the residuals follow Normal distribution.

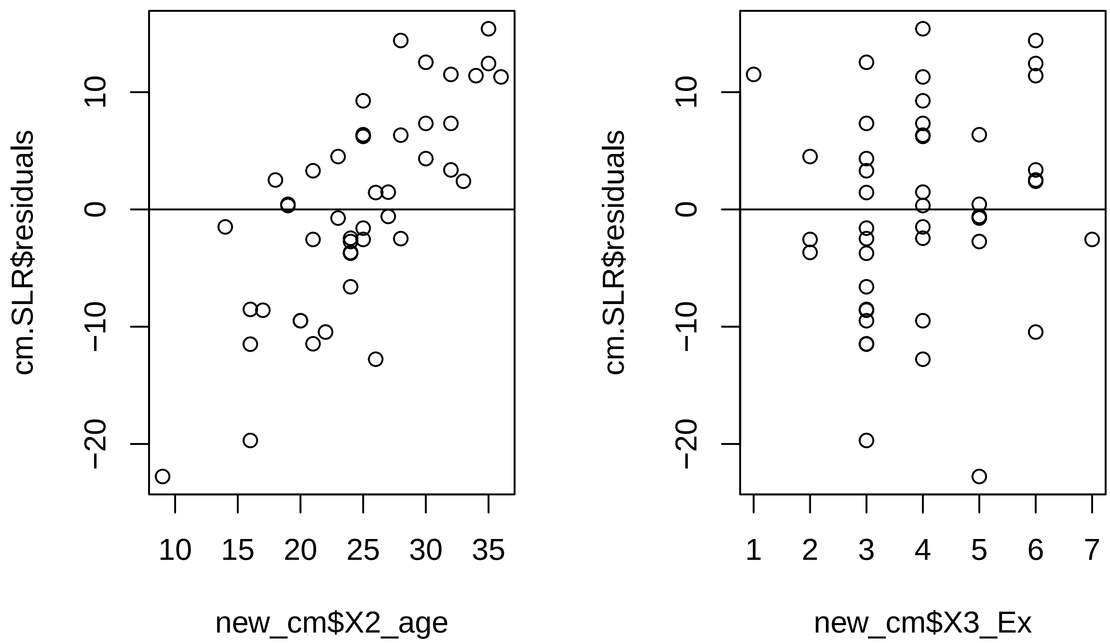
f. Omitted.

g. **Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call (X_2in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?**

```
new_cm <- read.table("./CH03PR04.txt", header = F)
colnames(new_cm) <- c("minute", "X1_copier", "X2_age", "X3_Ex")

par(mfrow = c(1,2))
plot(new_cm$X2_age, cm.SLR$residuals)
abline(0,0)
plot(new_cm$X3_Ex, cm.SLR$residuals)
```

```r
abline(0,0)
```



- The plots show that the residuals have a strong positive correlation with the age of the copiers (X_2). However, there is no evidence of correlation between the residual and X_3. Hence, including X_2 can help to improve the model and X_3 may not be helpful.

6. Test of Linearity (Lack-of-fit F-test using expanded ANOVA table) (15 pts, 5 for calculation, 5 for software, 5 for setup and conclusion.)

Following is the ANOVA table for the Copier Maintenance data we analyzed last time. Use the facts that: (a) $SSPE = 2797.658$, and (b) the predictor (Copier) has 10 distinct values, to further develop the ANOVA table so that we can conduct F-test for lack-of-fit. State the hypothesis, rejection rule and your conclusion for the test (at Type I error rate $= 0.05$). Then, use R or other statistical software to confirm your result.

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	76960.423	1	76960.423	968.657	.000(a)
	Residual	3416.377	43	79.451		
	Total	80376.800	44			
a Predictors: (Constant), X_COPIER						
b Dependent Variable: Y_TIME						

- Hypothesis
  - $H_0 : E(Y) = \beta_0 + \beta_1 X$
  - $H_0 : E(Y) \neq \beta_0 + \beta_1 X$
- Calculation
  - $SSE = 3416.377$, $SS_{PE} = 2797.658$
  - $SS_{LE} = SSE - SS_{PE} = 618.719$
  - $df_{LE} = 10 - 2 = 8$, $df_{PE} = 45 - 10 = 35$
  - $F_{LF} = \frac{(SS_{LE}/df_{LE})}{(SS_{PE}/df_{PE})} = \frac{77.33988}{79.93309} = 0.9676$
  - $p - value = P(F_{(df_1 = df_{LF},\ df_2 = df_{PE})} > F_{LF}) = P(F_{(df_1 8,\ df_2 = 35)} > 0.9676) = 0.4766$. (R code `1 - pf(0.9676, 8, 35)`)
  - If you prefer to use the critical value in the rejection rule, $F(1-\alpha,\ c-2,\ n-c) = 2.217$. (R code `qf(0.95, 8, 35)`)
- Rejection rules
  - If $p - value < \alpha$, reject H_0 and conclude H_a.
  - If $p - value > \alpha$, do not reject H_0.
  - Alternatively, one can use the critical value for the rejection rule. If $F_{LF} > F(1-\alpha; c-2, n-c)$, reject H_0 and conclude H_a. If $F_{LF} \leqslant F(1-\alpha; c-2, n-c)$, do not reject H_0.
- Conclusion
  - Because the $p - value = 0.4766$ (see below) is greater than $\alpha = 0.05$, we do not reject $H_0 :\ E(Y) = \beta_0 + \beta_1 X$\$. The linearity (straight-line) assumption of the model appears to be valid.
  - Alternatively, you can use $F_{LF} = 0.9676 < 2.23$. Hence, we do not reject $H_0 :\ E(Y) = \beta_0 + \beta_1 X$. We assume the straight-line model fits the data.
- R code verification:

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

cm.lof <- lm(minute ~ as.factor(copier), data=cm)

anova(cm.SLR, cm.lof)
```

```
## Analysis of Variance Table
##
## Model 1: minute ~ copier
## Model 2: minute ~ as.factor(copier)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     43 3416.4
## 2     35 2797.7  8    618.72 0.9676 0.4766
```

- At the p-value of 0.4766, we do not reject the null hypothesis. The linearity assumption of the model is appropriate.

—— This is the end of Homework 3. ——