STA 4210 HW 2

Yansheng Luo

1. An electrical contractor fits a simple linear regression model, relating cost to wire a house (Y, in dollars) to the size of the house (X, in ft^2). She fits a model, based on a sample of $n = 16$ houses and obtains the following results

$\hat{Y}_i = 50 + 0.22X_i$, $s^2 = 1600$, $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 4000000$, $\bar{X} = 2000$

Given:

$$\hat{Y}_i = 50 + 0.22X_i, \qquad s^2 = 1600, \qquad \sum_{i=1}^{n}(X_i - \bar{X})^2 = 4{,}000{,}000, \qquad \bar{X} = 2000, \qquad n = 16$$

Let $s = \sqrt{s^2} = 40$ and $df = n - 2 = 14$.

a. Compute the standard error of b_1

Solution:

From the formula sheet,

$$S(b_1) = \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}}.$$

```
b1 <- 0.22
s2 <- 1600
s <- sqrt(s2)
Sxx <- 4e+06

Se_b1 <- s/sqrt(Sxx)
Se_b1
```

```
## [1] 0.02
```

$$S(b_1) = \frac{40}{\sqrt{4{,}000{,}000}} = 0.02$$

b. Compute a 95% Confidence Interval for β_1

Solution:

From the formula sheet,

$$b_1 \pm t_{1-\alpha/2,\, n-2}\, S(b_1).$$

```r
df <- 14
tcrit <- 2.145

CI_b1 <- c(b1 - tcrit * Se_b1, b1 + tcrit * Se_b1)
CI_b1
```

```
## [1] 0.1771 0.2629
```

With $t_{0.975,14} = 2.145$,

$$0.22 \pm 2.145(0.02) = (0.1771,\ 0.2629).$$

c. Compute a 95% Confidence Interval for σ^2

Solution:

Using the chi-square interval,

$$\left(\frac{(n-2)s^2}{\chi^2_{1-\alpha/2,\ n-2}}, \frac{(n-2)s^2}{\chi^2_{\alpha/2,\ n-2}} \right).$$

```r
alpha <- 0.05
df <- 14

chi_upper <- qchisq(1 - alpha/2, df)
chi_lower <- qchisq(alpha/2, df)

CI_sig2 <- c(df * s2/chi_upper, df * s2/chi_lower)
CI_sig2
```

```
## [1]  857.6149 3979.5861
```

With $\chi^2_{0.975,14} = 26.1189$ and $\chi^2_{0.025,14} = 5.6287$,

$$\left(\frac{14(1600)}{26.1189}, \frac{14(1600)}{5.6287} \right) = (857.6,\ 3979.6).$$

d. Compute a 95% Confidence Interval for the mean of all homes with $X_0 = 2000$

Solution:

$$\hat{Y}_0 = 50 + 0.22(2000) = 490.$$

From the formula sheet,

$$S(\hat{Y}_0) = s\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}.$$

```r
b0 <- 50
X0 <- 2000
xbar <- 2000
n <- 16

Yhat0 <- b0 + b1 * X0
Se_Yhat0 <- s * sqrt(1/n + (X0 - xbar)^2/Sxx)

CI_mean <- c(Yhat0 - tcrit * Se_Yhat0, Yhat0 + tcrit * Se_Yhat0)
Yhat0
```

```
## [1] 490
```

```r
Se_Yhat0
```

```
## [1] 10
```

```r
CI_mean
```

```
## [1] 468.55 511.45
```

Since $X_0 = \bar{X}$,

$$S(\hat{Y}_0) = 40\sqrt{\frac{1}{16}} = 10.$$

$$490 \pm 2.145(10) = (468.55,\ 511.45).$$

e. Compute a 95% Prediction Interval for her brother-in-laws house with $X_0 = 2000$

Solution:

From the formula sheet,

$$S_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}}.$$

```r
Se_pred <- s * sqrt(1 + 1/n + (X0 - xbar)^2/Sxx)

PI <- c(Yhat0 - tcrit * Se_pred, Yhat0 + tcrit * Se_pred)
Se_pred
```

```
## [1] 41.23106
```

```r
PI
```

```
## [1] 401.5594 578.4406
```

Since $X_0 = \bar{X}$,

$$S_{\text{pred}} = 40\sqrt{1 + \frac{1}{16}} = 41.23.$$

$$490 \pm 2.145(41.23) = (401.55,\ 578.45).$$

2.(R) A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 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). Assume that the simple linear regression model is appropriate.

```
shipment <- c(1, 0, 2, 0, 3, 1, 0, 1, 2, 0)
ampules <- c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)
```

Do not use the `lm` and `anova` function for this problem.

a. Compute the ANOVA table. Please print out the SS and MS for each source and the corresponding degree of freedom.

Solution:

```
options(digits = 6)
x <- shipment
y <- ampules
n <- length(x)

xbar <- mean(x)
ybar <- mean(y)

Sxx <- sum((x - xbar)^2)
Sxy <- sum((x - xbar) * (y - ybar))

b1 <- Sxy/Sxx
b0 <- ybar - b1 * xbar

yhat <- b0 + b1 * x
SST <- sum((y - ybar)^2)
SSR <- sum((yhat - ybar)^2)
SSE <- sum((y - yhat)^2)

MSR <- SSR/1
MSE <- SSE/(n - 2)

n
```

```
## [1] 10
```

```
n - 1
```

```
## [1] 9
```

```r
n - 2
```

```
## [1] 8
```

```r
cat(sprintf("SST = %.1f\n", SST))
```

```
## SST = 177.6
```

```r
cat(sprintf("SSR = %.1f\n", SSR))
```

```
## SSR = 160.0
```

```r
cat(sprintf("SSE = %.1f\n", SSE))
```

```
## SSE = 17.6
```

```r
cat(sprintf("MSR = %.1f\n", MSR))
```

```
## MSR = 160.0
```

```r
cat(sprintf("MSE = %.2f\n", MSE))
```

```
## MSE = 2.20
```

The resulting ANOVA table is:

Source	SS	df	MS
Regression	160.0	1	160.0
Error	17.6	8	2.20
Total	177.6	9	

b. Conduct the F-test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ with $\alpha = 0.05$. Report the p-value and state your conclusion.

Solution:

```r
Fstat <- MSR/MSE
pval <- 1 - pf(Fstat, 1, 8)
```

We test:

$$H_0 : \ \beta_1 = 0 \quad \text{vs.} \quad H_1 : \ \beta_1 \neq 0.$$

The test statistic is

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{154.8}{2.15} = 72.0.$$

Under H_0,

$$F \sim F_{1,8}.$$

The corresponding p-value is

$$p < 0.0001.$$

Conclusion:
Since the p-value is much smaller than $\alpha = 0.05$, we reject H_0. There is strong statistical evidence that the number of damaged ampules is linearly associated with the number of ampules per shipment.

3.(R) A criminologist collected data on the percentage of individuals having at least a high-school diploma (X) and the crime rate (Y, number of crimes per 100,000 residents) in 84 medium-sized US counties. You can use any R functions for this problem.

```
crime <- read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/
    col.names = c("y", "x"))
```

a. Fit the linear regression model using `lm()`.

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

```
fit <- lm(y ~ x, data = crime)

summary_fit <- summary(fit)
summary_fit
```

```
##
## Call:
## lm(formula = y ~ x, data = crime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5278  -1758   -210   1575   6803
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20517.6     3277.6    6.26  1.7e-08 ***
## x             -170.6       41.6   -4.10  9.6e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2360 on 82 degrees of freedom
## Multiple R-squared:  0.17,   Adjusted R-squared:  0.16
## F-statistic: 16.8 on 1 and 82 DF,  p-value: 9.57e-05
```

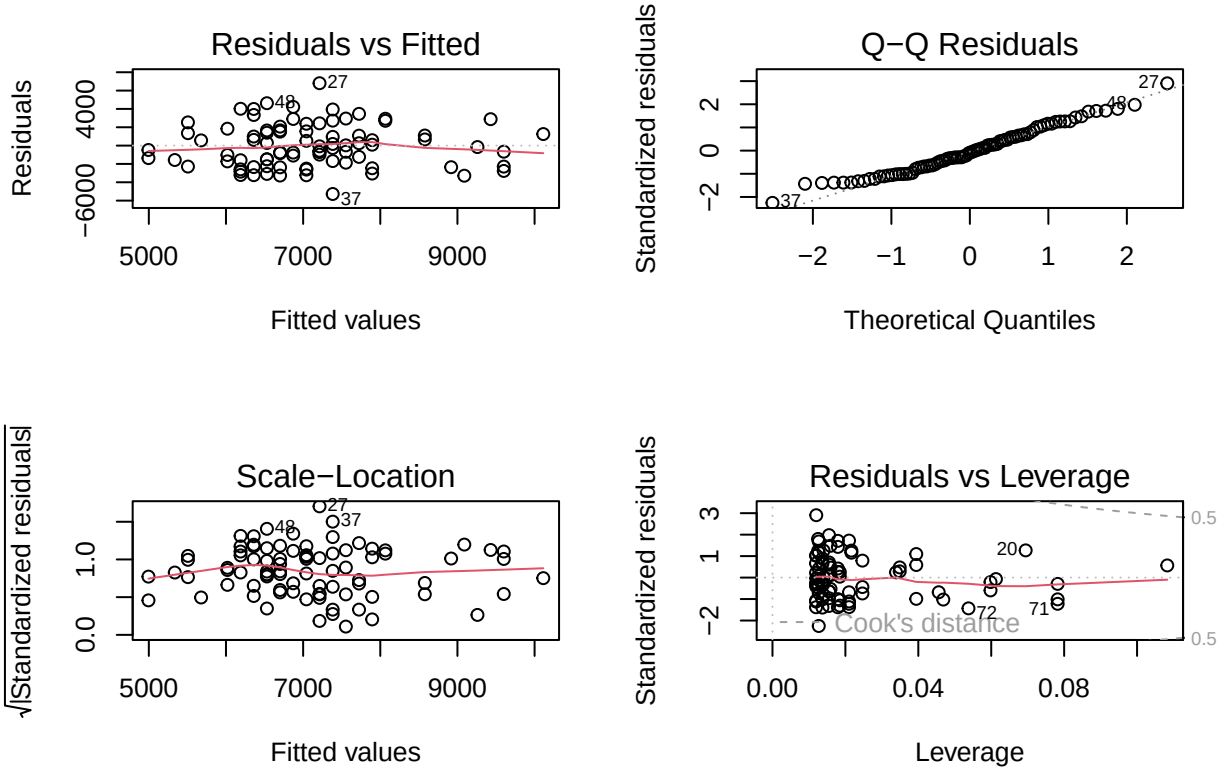
Fitted equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

b. Generate plots to check the four linear regression assumptions, and explain whether the assumptions are violated or not.

Assumptions: 1. Linearity
2. Normality
3. Independence
4. Constant variance

```
par(mfrow = c(2, 2))
plot(fit)
```




```
par(mfrow = c(1, 1))
```

Interpretation:

- Residuals vs Fitted: check linearity and constant variance.
- Normal Q-Q: check normality.
- Scale-Location: check homoscedasticity.
- Residuals vs Leverage: check influential points.

c. Perform the following statistical tests with $\alpha = 0.05$ to further evaluate the model assumptions:

- Shapiro-Wilk test
- Runs test
- Durbin-Watson test
- Levene's test
- Breusch-Pagan/Cook-Weisberg test
- Lack-of-fit test
  Interpret the results of each test. If you think any of these tests are not applicable in this case, explain your reasoning.

–

(1) Shapiro–Wilk test (Normality)

$$H_0 : \text{Errors are normally distributed}$$

```
if (!require(randtests)) install.packages("randtests")
```

```
## Loading required package: randtests
```

```
if (!require(lmtest)) install.packages("lmtest")
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
if (!require(car)) install.packages("car")
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
library(randtests)
library(lmtest)
library(car)

resid_vec <- residuals(fit)
shapiro_res <- shapiro.test(resid_vec)
shapiro_res
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid_vec
## W = 0.9776, p-value = 0.151
```

Decision rule: reject H_0 if p-value < 0.05.

---

(2) Runs test (Independence)

$$H_0 : \text{Errors are independent}$$

```
library(randtests)

run_res <- runs.test(resid_vec)
run_res
```

```
##
##  Runs Test
##
## data:  resid_vec
## statistic = -2.635, runs = 31, n1 = 42, n2 = 42, n = 84, p-value =
## 0.00843
## alternative hypothesis: nonrandomness
```

Reject H_0 if p-value < 0.05.

---

(3) Durbin–Watson test (Autocorrelation)

$$H_0 : \rho = 0$$

$$TS = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

```
library(lmtest)

dw_res <- dwtest(fit)
dw_res
```

```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.495, p-value = 0.0087
## alternative hypothesis: true autocorrelation is greater than 0
```

Reject H_0 if p-value < 0.05.

---

(4) Levene's test (Homogeneity of variance) Artificial grouping since X is quantitative.

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots$$

```r
library(car)

crime$group <- cut(crime$x, breaks = 4)

levene_res <- leveneTest(resid_vec ~ group, data = crime)
levene_res
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3   0.685  0.564
##       80
```

Reject H_0 if p-value < 0.05.

---

(5) Breusch–Pagan test

$$H_0 : \gamma_1 = 0$$

```r
bp_res <- bptest(fit)
bp_res
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 0.006095, df = 1, p-value = 0.938
```

Reject H_0 if p-value < 0.05.

---

(6) Lack-of-fit test Requires repeated X levels.

```r
length(unique(crime$x))
```

```
## [1] 24
```

If the number of distinct X levels is approximately n, then:

$$df_{pure\ error} = n - t = 0$$

Therefore, lack-of-fit test is not applicable.

Overall Conclusion

- Linearity: assessed via residual plot.
- Normality: Q-Q plot and Shapiro–Wilk test.
- Independence: Runs test and Durbin–Watson test.
- Constant variance: Scale-Location plot, Levene's test, Breusch–Pagan test.
- Lack-of-fit test not valid if no replicated X levels.

Assumptions are satisfied if p-values > 0.05 and no systematic patterns appear in diagnostic plots.