Statistical Calculation and Software

Assignment 2

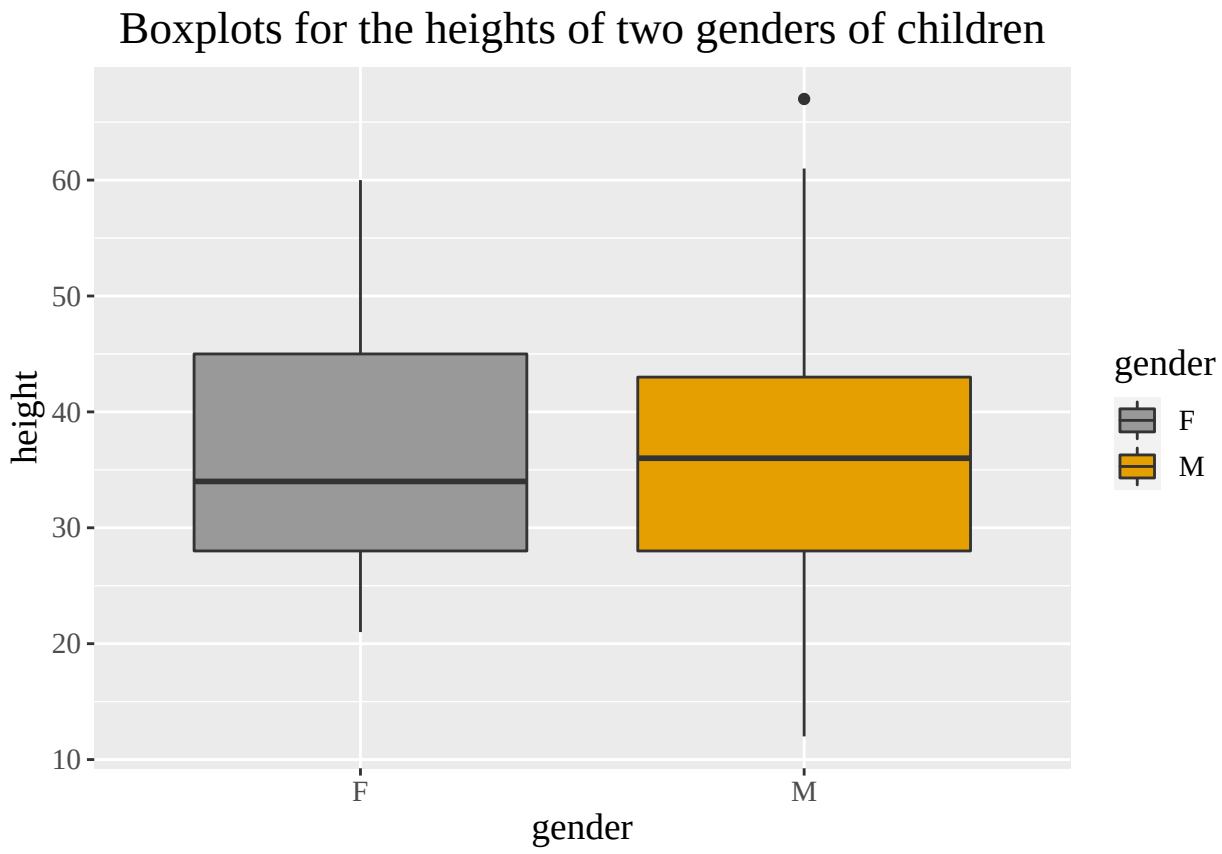
Hanbin Liu 11912410

2.1

```
library(UsingR)
data(kid.weights)
```

(a)

```
ggplot(kid.weights, aes(x = gender, y = height, fill = gender)) +
  geom_boxplot() + xlab("gender") + ylab("height") +
  ggtitle("Boxplots for the heights of two genders of children") +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 14, family = "serif")) +
  scale_fill_manual(values = c("#999999", "#E69F00"))
```




```
# unknown variance
x <- kid.weights$height
n <- length(x)
t <- (sqrt(n) * (mean(x) - 36)) / sd(x)
pval <- 2 * pt(t, n - 1, lower.tail = F)
name <- c("Test statistics: t", "df", "t(0.025,249)", "p-value")
value <-
  c((sqrt(n) * (mean(x) - 36)) / sd(x), n - 1, qt(0.975, n - 1), pval)
data.frame(name, value)
```

```
##                 name       value
## 1 Test statistics: t   0.7740114
## 2                 df 249.0000000
## 3       t(0.025,249)   1.9695369
## 4            p-value   0.4396583
```

We cannot reject H_0 with unknown variance.

```
# known variance
pval <- 2 * pnorm((sqrt(n) * (mean(x) - 36)) / sd(x), lower.tail = F)
name <- c("Test statistic: z", "z_0.025", "p-value")
value <- c((sqrt(n) * (mean(x) - 36)) / sd(x), qnorm(0.975), pval)
data.frame(name, value)
```

```
##                name     value
```

```
## 1 Test statistic: z 0.7740114
## 2           z_0.025 1.9599640
## 3           p-value 0.4389240
```

We cannot reject H_0 if the variance is known to be the current sample variance.

(b)

We have

$$\sup_{\theta\in\Theta_0} L(\theta \mid \mathbf{X}) = \frac{1}{(2\pi\hat{\sigma}_1^2)^{\frac{n}{2}}} \exp\Big\{ -\frac{1}{2\hat{\sigma}_1^2}\sum_{i=1}^{n}(X_i-\mu_0)^2\Big\},$$

and

$$\sup_{\theta\in\Theta} L(\theta \mid \mathbf{X}) = \frac{1}{(2\pi\hat{\sigma}_2^2)^{\frac{n}{2}}} \exp\Big\{ -\frac{1}{2\hat{\sigma}_2^2}\sum_{i=1}^{n}(X_i-\bar{X})^2\Big\},$$

where

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2, \quad \hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Then the likelihood ratio is given by

$$\lambda(\mathbf{X}) = \frac{\sup_{\theta\in\Theta_0} L(\theta \mid \mathbf{X})}{\sup_{\theta\in\Theta} L(\theta \mid \mathbf{X})} = (\frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2})^{\frac{n}{2}} = \left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\mu_0)^2}\right]^{\frac{n}{2}}.$$

Note that

$$\sum_{i=1}^{n}(X_i-\mu_0)^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu_0)^2.$$

It then follows that

$$\lambda(\mathbf{X}) = \left(\frac{1}{1+\frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)^{\frac{n}{2}}.$$

Then

$$\left(\frac{1}{1+\frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}\right)^{\frac{n}{2}} \le k$$

is equivalent to

$$\frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \ge c \iff \frac{n(\bar{X}-\mu_0)^2}{S^2} \ge (n-1)c = c^*.$$

Under H_0, $T = \frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \sim t(n-1)$. Hence,

$$\alpha = \Pr(T^2 \ge c^* \mid H_0),$$

which implies that $\sqrt{c^*} = t_{\frac{\alpha}{2}}(n-1)$. Therefore, the testing procedure is

$$\text{reject } H_0 \text{ if} \left\{ \left| \frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \right| \ge t_{\frac{\alpha}{2}}(n-1) \right\}$$

```r
x <- kid.weights$height[kid.weights$gender == 'M']
n <- length(x)
# likelihood ratio test
pval <-
  2 * pt((sqrt(n) * (mean(x) - 36)) / sd(x), n - 1, lower.tail = F)
name <- c("Test statistic: t", "df", "t(0.025,120)", "p-value")
value <-
  c((sqrt(n) * (mean(x) - 36)) / sd(x), n - 1, qt(0.975, n - 1), pval)
data.frame(name, value)
```

```
##                name       value
## 1 Test statistic: t   1.0187423
## 2                df 120.0000000
## 3      t(0.025,120)   1.9799304
## 4           p-value   0.3103748
```

We cannot reject H_0 that the mean is 36 at $\alpha = 0.05$ level of significance.

```r
# one-sample t test
t.test(x, mu = 36)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 1.0187, df = 120, p-value = 0.3104
## alternative hypothesis: true mean is not equal to 36
## 95 percent confidence interval:
##  35.01751 39.06514
## sample estimates:
## mean of x
##  37.04132
```

We cannot reject H_0 that the mean of male is equal to the mean of female at $\alpha = 0.05$ level of significance. The results are the same.

(c)

Normality assumption:

```r
t.test(height ~ gender, data = kid.weights, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  height by gender
## t = -0.73708, df = 241.67, p-value = 0.2309
## alternative hypothesis: true difference in means between group F and group M is less than 0
## 95 percent confidence interval:
##      -Inf 1.243351
## sample estimates:
## mean in group F mean in group M
##        36.03876        37.04132
```

We cannot reject H_0 that the height of the male and female have the same mean value at $\alpha = 0.05$ level of significance.

Non-parametric Test:

```
wilcox.test(height ~ gender, data = kid.weights)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  height by gender
## W = 7397, p-value = 0.4758
## alternative hypothesis: true location shift is not equal to 0
```

We cannot reject H_0 that the height of the male and female have the same mean value at $\alpha = 0.05$ level of significance.

(d)

```
male <- kid.weights$height[kid.weights$gender == 'M']
female <- kid.weights$height[kid.weights$gender == 'F']
# Siegel-Tukey test (the medians of two samples are close to each other)
library(ACSWR)
siegel.tukey(male, female)
```

```
## [1] 0.3736492 0.6456674
```

```
# Kolmogorov-Smirnov Test
ks.test(male, female)
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  male and female
## D = 0.091678, p-value = 0.6703
## alternative hypothesis: two-sided
```

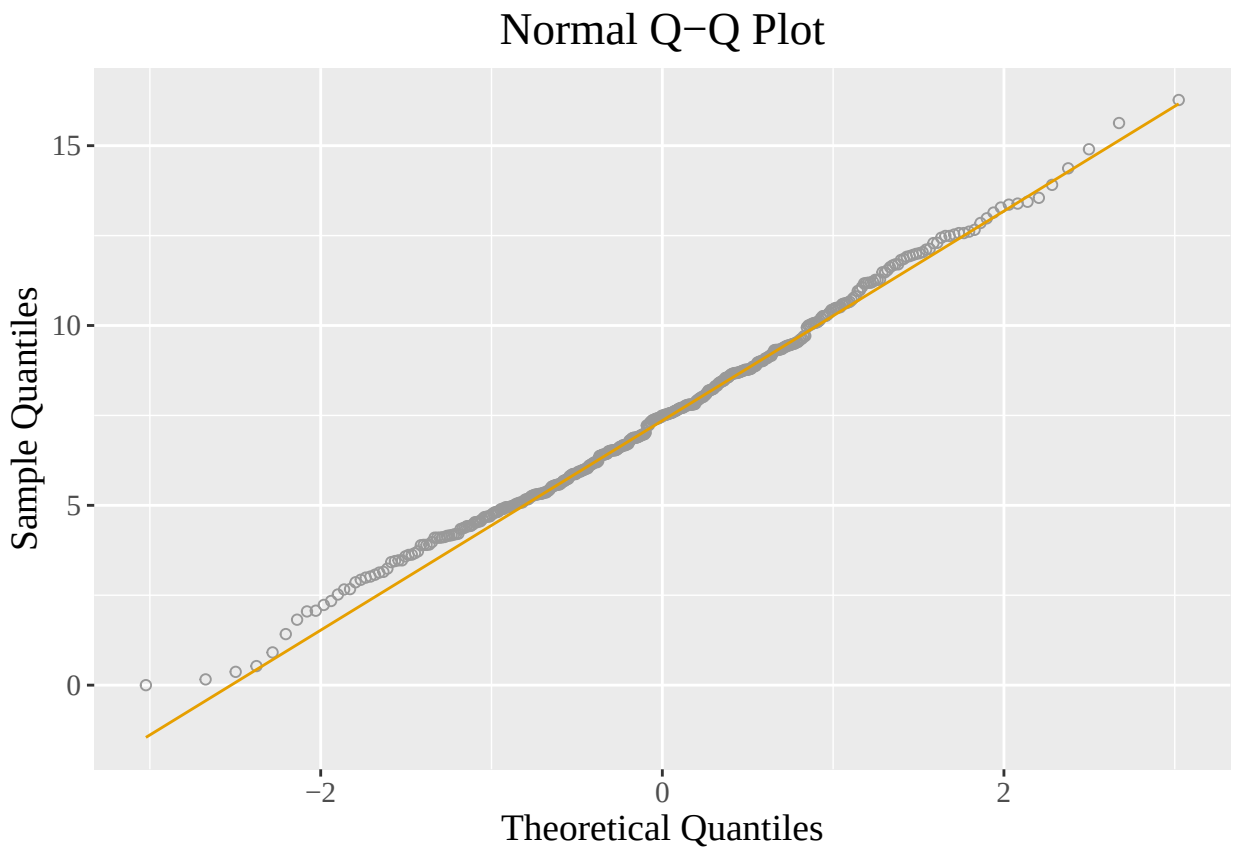
We cannot reject H_0 that the spread of height for the male and female are the same.

2.2

```
Carseats <- read.csv("Carseats.csv")
```

(a)

```
ggplot(Carseats, aes(sample = Sales)) +
  stat_qq(colour = '#999999', pch = 1) +
  stat_qq_line(colour = '#E69F00') +
  ggtitle("Normal Q-Q Plot") +
  theme(plot.title = element_text(hjust = 0.5),
        text = element_text(size = 14, family = "serif")) +
  xlab("Theoretical Quantiles") +
  ylab("Sample Quantiles")
```




```
x <- Carseats$Sales
ks.test(x, "pnorm", mean = mean(x), sd = sd(x))
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.032533, p-value = 0.791
## alternative hypothesis: two-sided
```

We cannot reject H_0 that Sales follow normal distribution.

(b)

```
model <- lm(Sales ~ Price + Advertising + Age + Urban, data = Carseats)
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age + Urban, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.630 -1.534  0.019  1.516  6.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.992823   0.731610  21.860  < 2e-16 ***
## Price       -0.058047   0.004839 -11.997  < 2e-16 ***
## Advertising  0.123051   0.017130   7.183 3.41e-12 ***
## Age         -0.048865   0.007060  -6.921 1.82e-11 ***
## UrbanYes     0.020186   0.249659   0.081    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.271 on 395 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3531
## F-statistic: 55.44 on 4 and 395 DF,  p-value: < 2.2e-16
```

(c)

```
coefficients(model)
```

```
## (Intercept)       Price Advertising         Age    UrbanYes
## 15.99282333 -0.05804694  0.12305077 -0.04886540  0.02018579
```

The coefficient for the intercept is equal to 15.99282333. This means that for a sample with 'Price', 'Advertising', 'Age' are 0 and 'Urban' is 'No', the average expected sales is 15.99282333. It's important to note that the regression coefficient for the intercept is only meaningful if it's reasonable that all of the predictor variables in the model can actually be equal to zero.

The coefficient for 'Price' is -0.05804694. This means that, on average, each additional 'Price' is associated with a decrease of 0.05804694 for the 'Sales', assuming other predictor variables are held constant.

The coefficient for 'Advertising' is 0.12305077. This means that, on average, each additional 'Advertising' is associated with an increase of 0.12305077 for the 'Sales', assuming other predictor variables are held constant.

The coefficient for 'Age' is -0.04886540. This means that, on average, each additional 'Age' is associated with a decrease of 0.04886540 for the 'Sales', assuming other predictor variables are held constant.

The coefficient for 'UrbanYes' is 0.02018579. This means that, on average, a sample whose 'Urban' is 'Yes' has a higher 'Sales' compared to a sample whose 'Urban' is 'No' and the difference is 0.02018579, assuming other predictor variables are held constant.

(d)

The equation form for the model is

$$\text{Sales} = \begin{cases} 15.992823 - 0.058047 * \text{Price} + 0.123051 * \text{Advertising} - 0.048865 * \text{Age}, & \text{if Urban} = \text{No}, \\ 16.013009 - 0.058047 * \text{Price} + 0.123051 * \text{Advertising} - 0.048865 * \text{Age}, & \text{if Urban} = \text{Yes}. \end{cases}$$

(e)

```
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age + Urban, data = Carseats)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.630 -1.534  0.019  1.516  6.306
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.992823   0.731610  21.860  < 2e-16 ***
## Price       -0.058047   0.004839 -11.997  < 2e-16 ***
## Advertising  0.123051   0.017130   7.183 3.41e-12 ***
## Age         -0.048865   0.007060  -6.921 1.82e-11 ***
## UrbanYes     0.020186   0.249659   0.081    0.936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.271 on 395 degrees of freedom
## Multiple R-squared:  0.3596, Adjusted R-squared:  0.3531
## F-statistic: 55.44 on 4 and 395 DF,  p-value: < 2.2e-16
```

We can reject the null hypothesis $H_0 : \beta_j = 0$ for j = 0,1,2,3.

(f)

```
model2 <- lm(Sales ~ Price + Advertising + Age, data = Carseats)
summary(model2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Age, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6247 -1.5288  0.0148  1.5220  6.2925
##
## Coefficients:
```

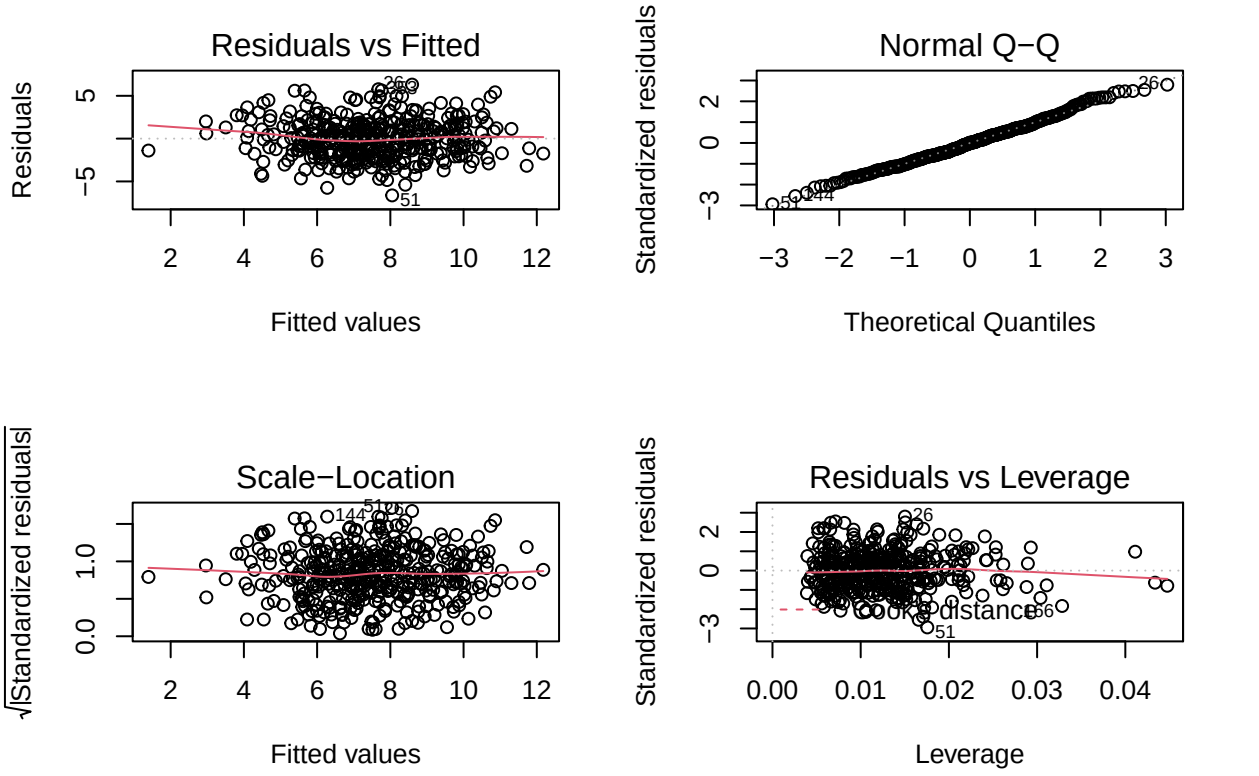
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.003472   0.718754  22.266  < 2e-16 ***
## Price       -0.058028   0.004827 -12.022  < 2e-16 ***
## Advertising  0.123106   0.017095   7.201 3.02e-12 ***
## Age         -0.048846   0.007047  -6.931 1.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.269 on 396 degrees of freedom
## Multiple R-squared:  0.3595, Adjusted R-squared:  0.3547
## F-statistic:  74.1 on 3 and 396 DF,  p-value: < 2.2e-16
```

(g)

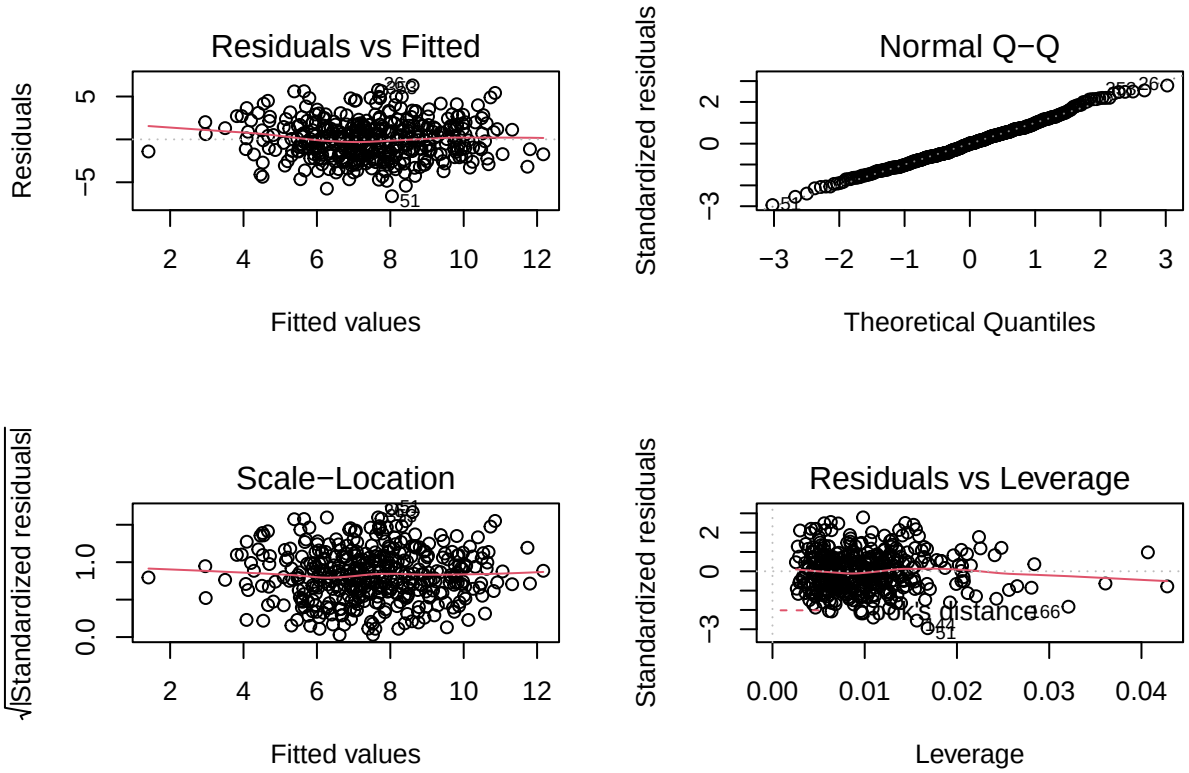
Model in (b) has a $R^2 = 0.3596$, and adjusted $R^2 = 0.3531$ while model in (f) has a $R^2 = 0.3595$ and adjusted $R^2 = 0.3547$. Their R-squared and adjusted R-squared value are close to each other, however, model in (f) involves fewer variables and has a larger adjusted R-squared value. Therefore, model in (f) is better than the model in (b). But since their adjusted R-squared values are all less than 0.36, these two models both have relatively non-ideal performance.

Model diagnostic:

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



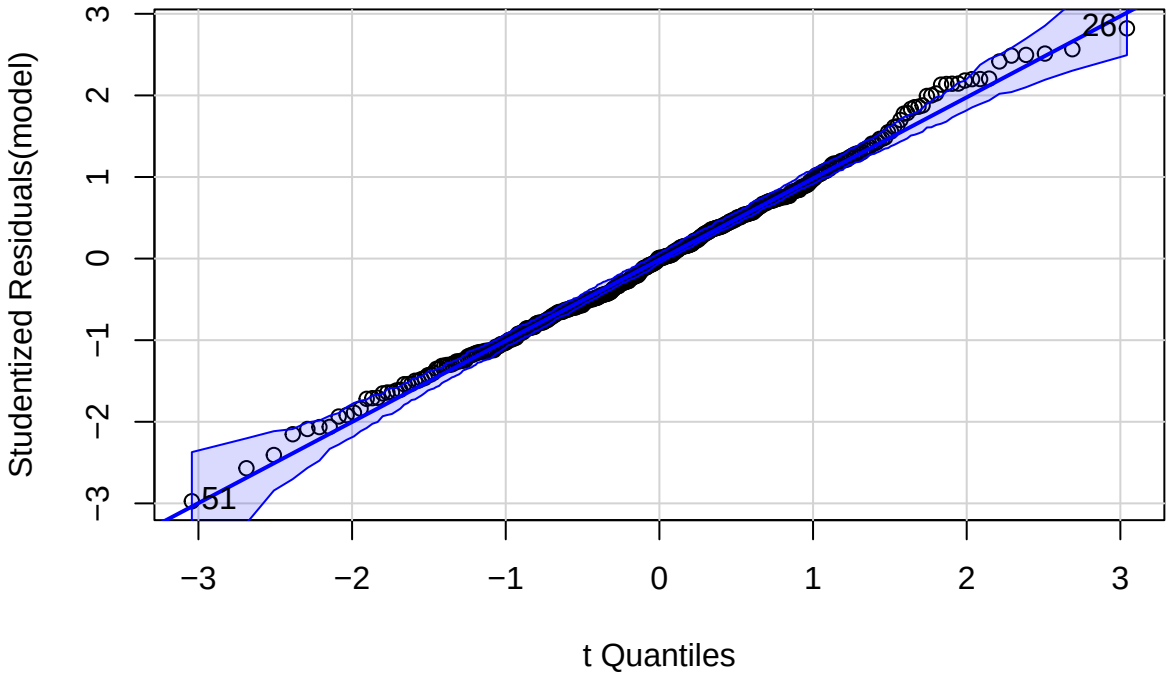
```
plot(model2)
```




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

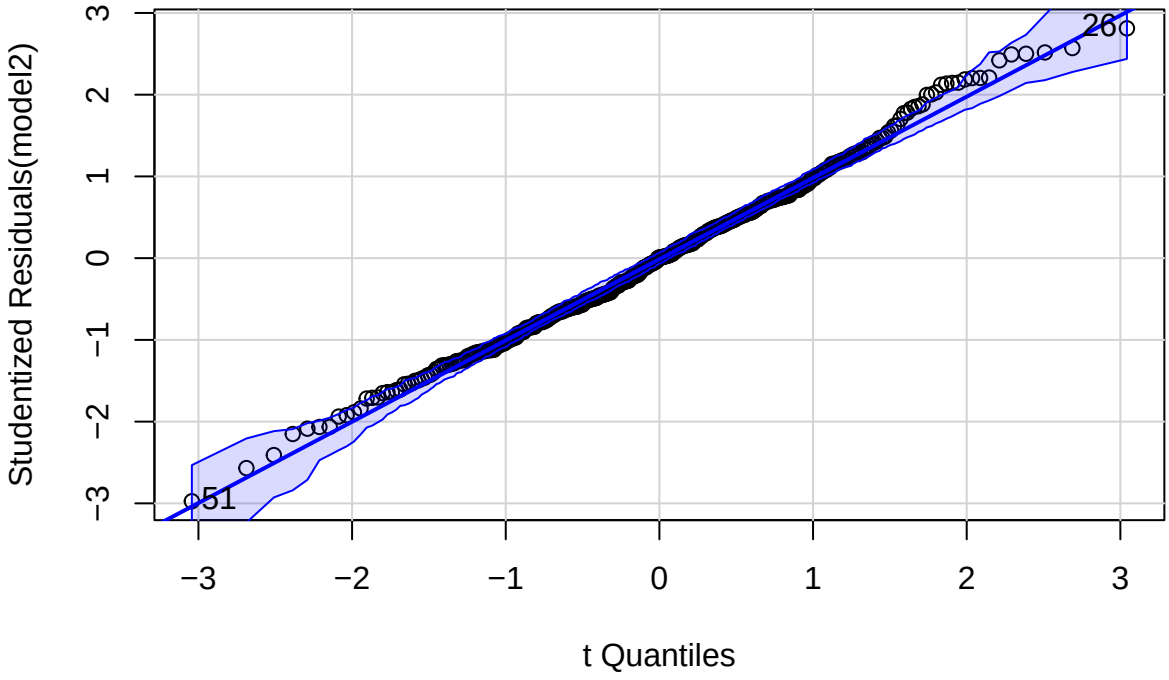
Normality:

```
qqPlot(model)
```




```
## [1] 26 51
```

```
qqPlot(model2)
```




```
## [1] 26 51
```

Independence:

```
durbinWatsonTest(model)
```

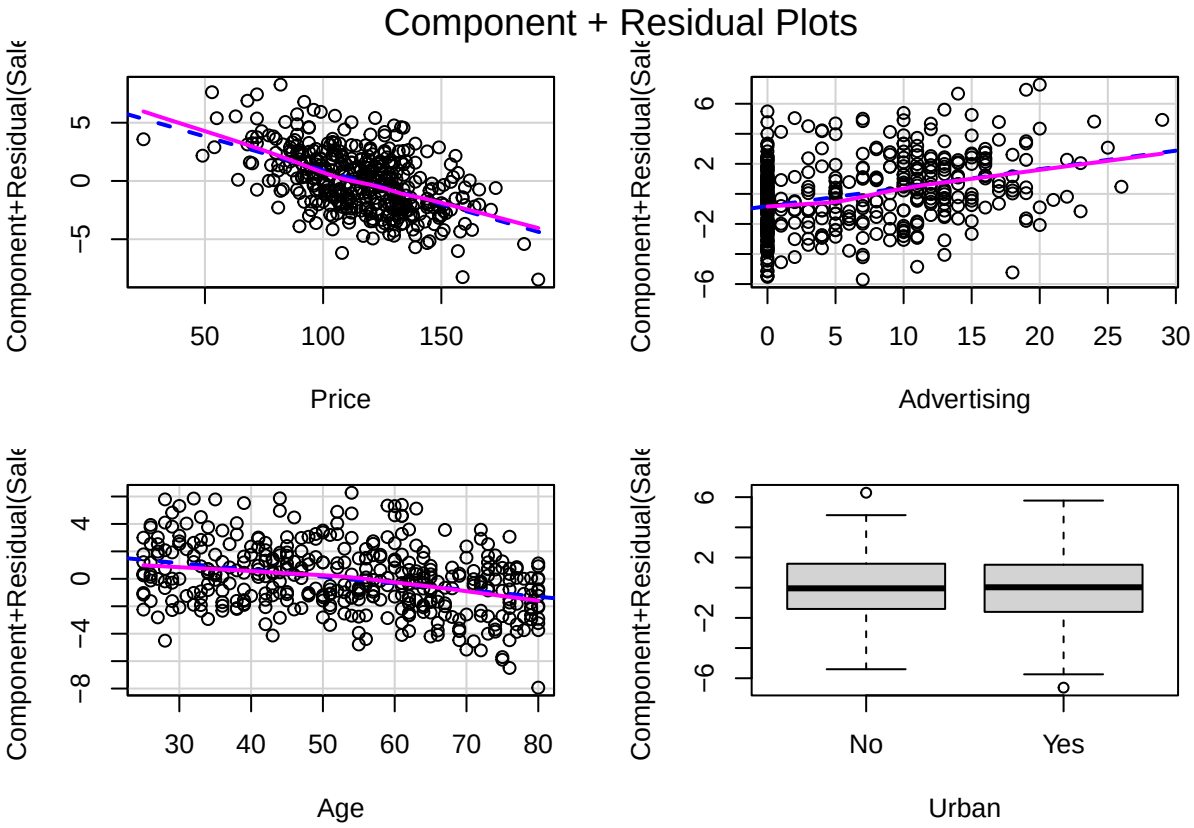
```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02041399      1.953936   0.658
##  Alternative hypothesis: rho != 0
```

```
durbinWatsonTest(model2)
```

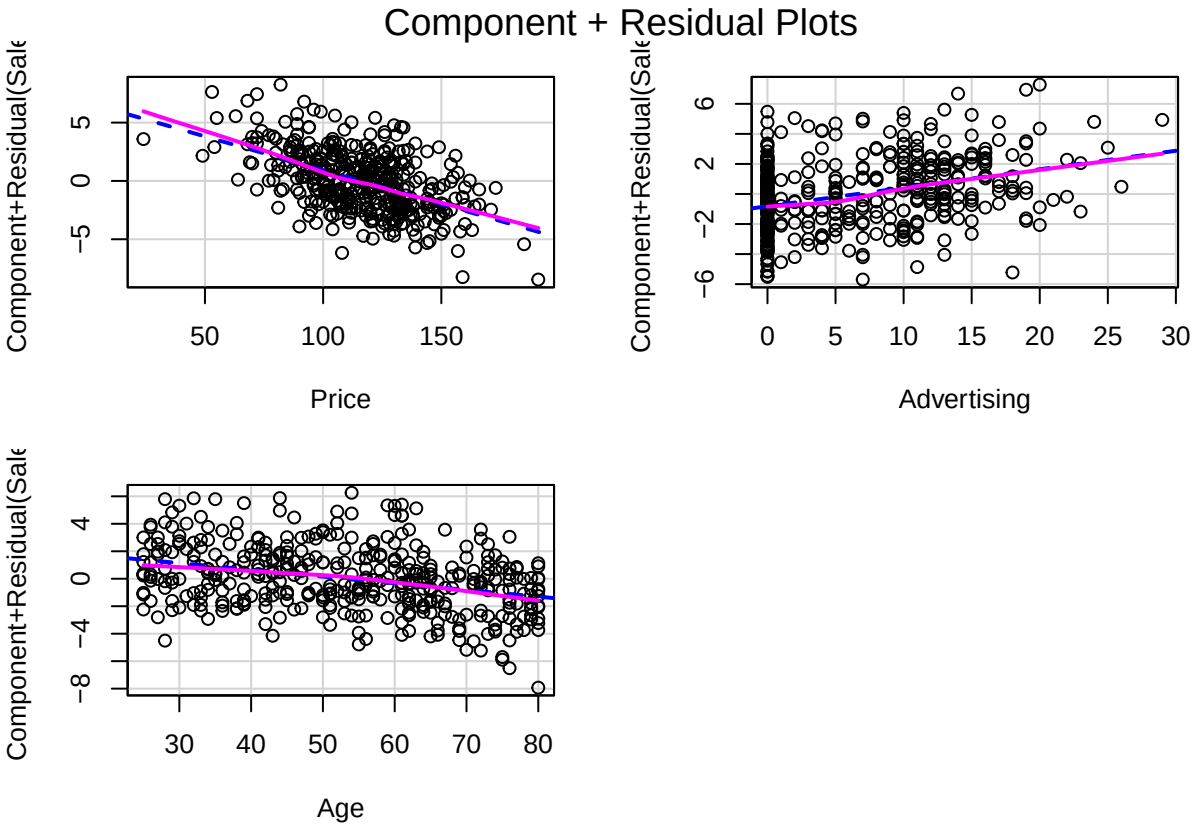
```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02017267      1.954393    0.58
##  Alternative hypothesis: rho != 0
```

Linearity:

```
crPlots(model)
```




```
crPlots(model2)
```

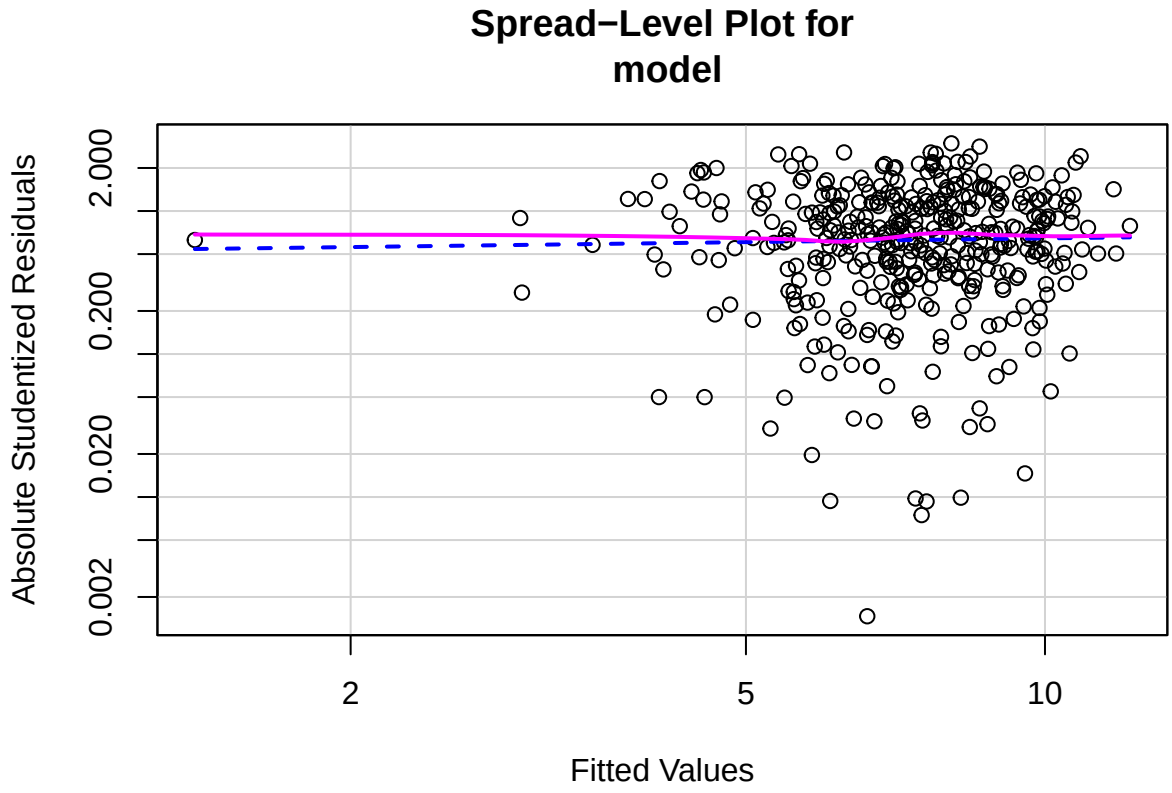


Homoscedasticity:

```
ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.004894055, Df = 1, p = 0.94423
```

```
spreadLevelPlot(model)
```

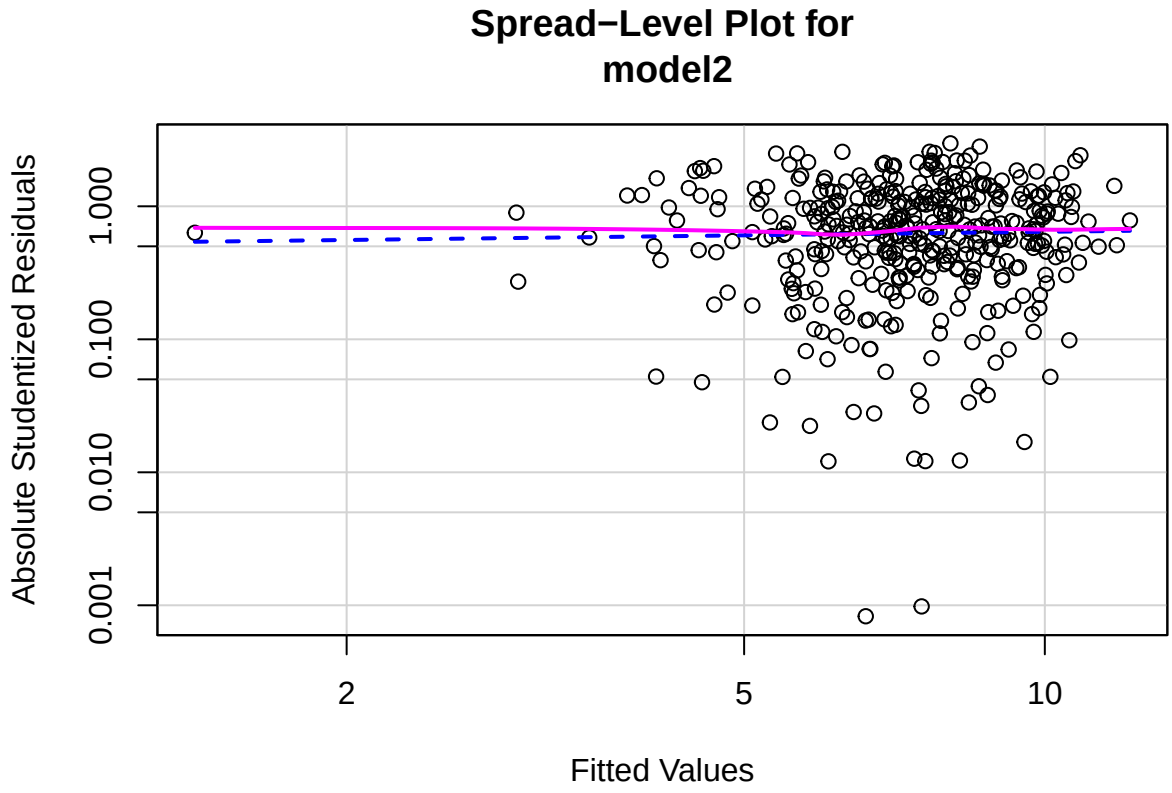



```
##
## Suggested power transformation:  0.9129871
```

```
#
ncvTest(model2)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.004737092, Df = 1, p = 0.94513
```

```
spreadLevelPlot(model2)
```




```
##
## Suggested power transformation:  0.9118642
```

(h)

```
confint(model2)
```

```
##                   2.5 %      97.5 %
## (Intercept) 14.59042068 17.41652325
## Price       -0.06751743 -0.04853857
## Advertising  0.08949838  0.15671410
## Age         -0.06270141 -0.03499112
```

(i)

```
# outilier
outlierTest(model2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 51 -2.974015          0.0031196           NA
```

```r
# high leverage observations
hat.plot <- function(fit) {
  p <- length(coefficients(fit))
  n <- length(fitted(fit))
  plot(hatvalues(fit), main = "Index Plot of Hat Values")
  abline(h = c(2, 3) * p / n,
         col = "red",
         lty = 2)
  identify(1:n, hatvalues(fit), names(hatvalues(fit)))
}
hat.plot(model2)
```


```
## integer(0)
```

There are no outliers but high leverage observations in the model.

(j)

```r
attach(Carseats)
x <- Sales[Urban == 'Yes']
y <- Sales[Urban == 'No']

group <- c("Urban", "Non-urban")
```

```r
mean <- c(mean(x), mean(y))
data.frame(group, mean)
```

```
##       group     mean
## 1     Urban 7.468191
## 2 Non-urban 7.563559
```

Likelihood ratio test:

```r
# method 1
n1 <- length(x)
n2 <- length(y)
Sp <- sqrt(((n1 - 1) * var(x) + (n2 - 1) * var(y)) / (n1 + n2 - 2))
t <- abs(mean(x) - mean(y)) / (sqrt(1 / n1 + 1 / n2) * Sp)
pval <- 2 * pt(t, n1 + n2 - 1, lower.tail = F)
name <- c("Test statistic: t", "df", "t(0.025,398)", "p-value")
value <- c(t, n1 + n2 - 2, qt(0.975, n1 + n2 - 2), pval)
data.frame(name, value)
```

```
##                name       value
## 1 Test statistic: t   0.3076535
## 2                df 398.0000000
## 3      t(0.025,398)   1.9659423
## 4           p-value   0.7585066
```

```r
# method 2: set var.equal = TRUE in t.test()
t.test(Sales ~ Urban, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Sales by Urban
## t = 0.30765, df = 398, p-value = 0.7585
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -0.5140440  0.7047797
## sample estimates:
##  mean in group No mean in group Yes
##          7.563559          7.468191
```

We cannot reject H_0 that the two samples have the same mean.

Mann-Whitney test:

```r
wilcox.test(Sales ~ Urban, Carseats)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sales by Urban
## W = 17225, p-value = 0.5784
## alternative hypothesis: true location shift is not equal to 0
```

We cannot reject H_0 that the two samples have the same mean.

```
paste(c("n1:", "n2:"), c(n1, n2))
```

```
## [1] "n1: 282" "n2: 118"
```

We cannot use the Wilcoxon's Signed-Rank test since they have different size.

(k)

```
# stepwise methods
fit <- lm(Sales ~ ., data = Carseats)
stepAIC(fit, direction = "both")
```

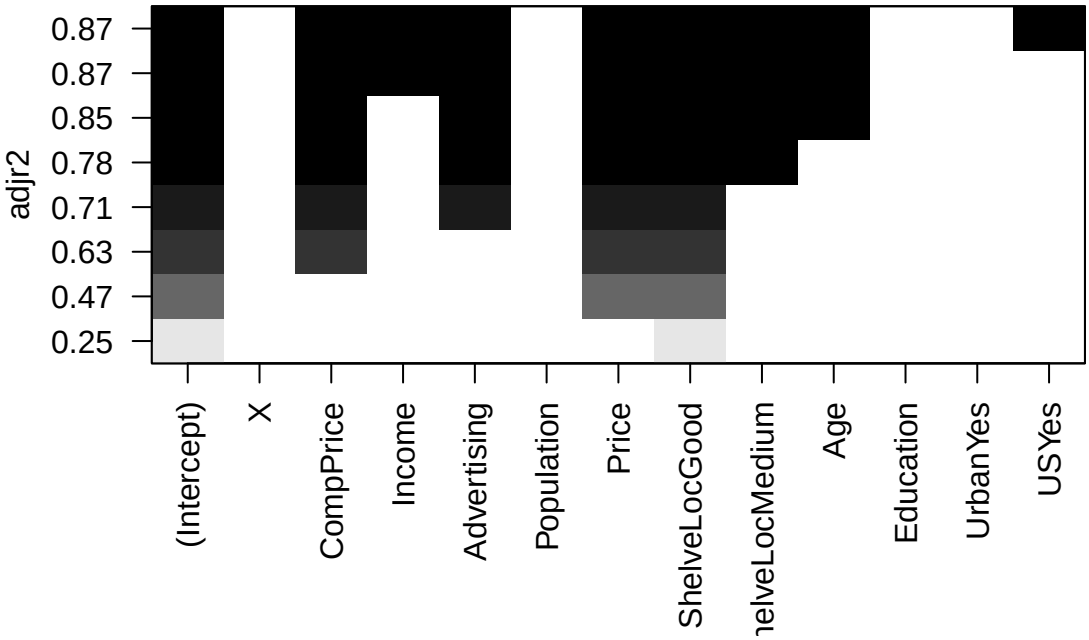
```
## Start:  AIC=28.28
## Sales ~ X + CompPrice + Income + Advertising + Population + Price +
##     ShelveLoc + Age + Education + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Population   1      0.35  402.64  26.63
## - X            1      0.54  402.83  26.82
## - Urban        1      1.32  403.61  27.60
## - Education    1      1.33  403.62  27.61
## - US           1      1.59  403.88  27.86
## <none>                      402.29  28.28
## - Income       1     73.74  476.03  93.60
## - Advertising  1    127.58  529.86 136.46
## - Age          1    217.96  620.25 199.46
## - CompPrice    1    519.98  922.27 358.15
## - ShelveLoc    2   1053.70 1455.98 538.79
## - Price        1   1323.18 1725.47 608.72
##
## Step:  AIC=26.63
## Sales ~ X + CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - X            1      0.52  403.16  25.15
## - Urban        1      1.24  403.88  25.86
## - Education    1      1.52  404.16  26.14
## - US           1      1.93  404.57  26.54
## <none>                      402.64  26.63
## + Population   1      0.35  402.29  28.28
## - Income       1     73.53  476.17  91.73
## - Advertising  1    145.54  548.18 148.06
## - Age          1    219.00  621.64 198.36
## - CompPrice    1    521.79  924.43 357.08
## - ShelveLoc    2   1053.64 1456.28 536.87
## - Price        1   1323.41 1726.05 606.85
##
## Step:  AIC=25.15
```

```
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + Urban + US
##
##               Df Sum of Sq     RSS    AIC
## - Urban        1      1.15  404.31  24.29
## - Education    1      1.36  404.52  24.49
## - US           1      1.89  405.05  25.02
## <none>                      403.16  25.15
## + X            1      0.52  402.64  26.63
## + Population   1      0.33  402.83  26.82
## - Income       1     75.94  479.10  92.18
## - Advertising  1    145.38  548.54 146.32
## - Age          1    218.52  621.68 196.38
## - CompPrice    1    521.69  924.85 355.27
## - ShelveLoc    2   1053.18 1456.34 534.89
## - Price        1   1323.51 1726.67 605.00
##
## Step:  AIC=24.29
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + Education + US
##
##               Df Sum of Sq     RSS    AIC
## - Education    1      1.44  405.76  23.72
## - US           1      1.85  406.16  24.12
## <none>                      404.31  24.29
## + Urban        1      1.15  403.16  25.15
## + X            1      0.43  403.88  25.86
## + Population   1      0.25  404.06  26.04
## - Income       1     76.64  480.96  91.73
## - Advertising  1    146.03  550.34 145.63
## - Age          1    217.59  621.91 194.53
## - CompPrice    1    526.17  930.48 355.69
## - ShelveLoc    2   1053.93 1458.25 533.41
## - Price        1   1322.80 1727.11 603.10
##
## Step:  AIC=23.72
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age + US
##
##               Df Sum of Sq     RSS    AIC
## - US           1      1.63  407.39  23.32
## <none>                      405.76  23.72
## + Education    1      1.44  404.31  24.29
## + Urban        1      1.24  404.52  24.49
## + Population   1      0.41  405.35  25.32
## + X            1      0.28  405.47  25.44
## - Income       1     77.87  483.62  91.94
## - Advertising  1    145.30  551.06 144.15
## - Age          1    217.97  623.73 193.70
## - CompPrice    1    525.25  931.00 353.92
## - ShelveLoc    2   1056.88 1462.64 532.61
## - Price        1   1322.83 1728.58 601.44
##
## Step:  AIC=23.32
```

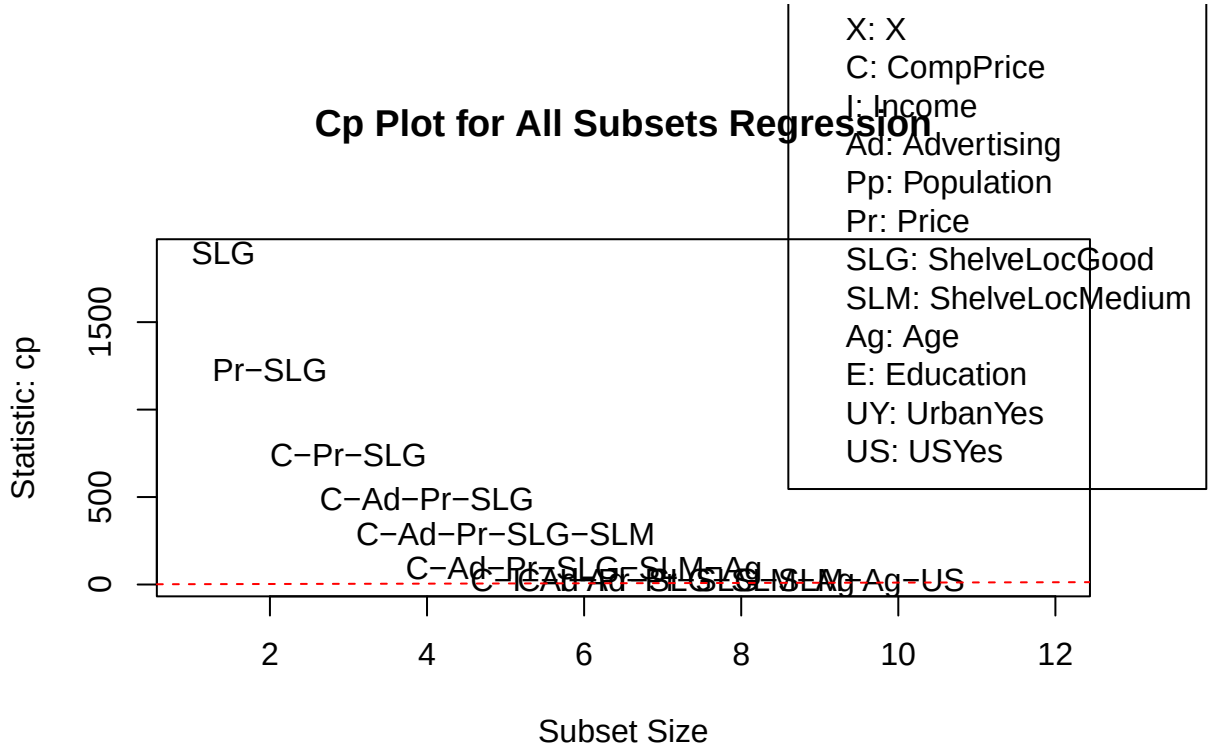
```
## Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
##     Age
##
##               Df Sum of Sq     RSS    AIC
## <none>                     407.39  23.32
## + US           1      1.63  405.76  23.72
## + Education    1      1.22  406.16  24.12
## + Urban        1      1.19  406.20  24.15
## + Population   1      0.72  406.67  24.62
## + X            1      0.27  407.12  25.05
## - Income       1     76.68  484.07  90.30
## - Age          1    219.12  626.51 193.48
## - Advertising  1    234.03  641.42 202.89
## - CompPrice    1    523.83  931.22 352.01
## - ShelveLoc    2   1055.51 1462.90 530.68
## - Price        1   1324.42 1731.81 600.18
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
##     ShelveLoc + Age, data = Carseats)
##
## Coefficients:
##     (Intercept)        CompPrice           Income      Advertising
##         5.47523          0.09257          0.01578          0.11590
##           Price    ShelveLocGood  ShelveLocMedium              Age
##        -0.09532          4.83567          1.95199         -0.04613
```

```r
# all-subsets regression
library(leaps)
leaps <- regsubsets(Sales ~ ., data = Carseats)
plot(leaps, scale = "adjr2")
```




```
library(car)
subsets(
  leaps,
  statistic = "cp",
  main = "Cp Plot for All Subsets Regression",
  legend = c(8.6, 3400)
)
abline(1, 1, lty = 2, col = "red")
```



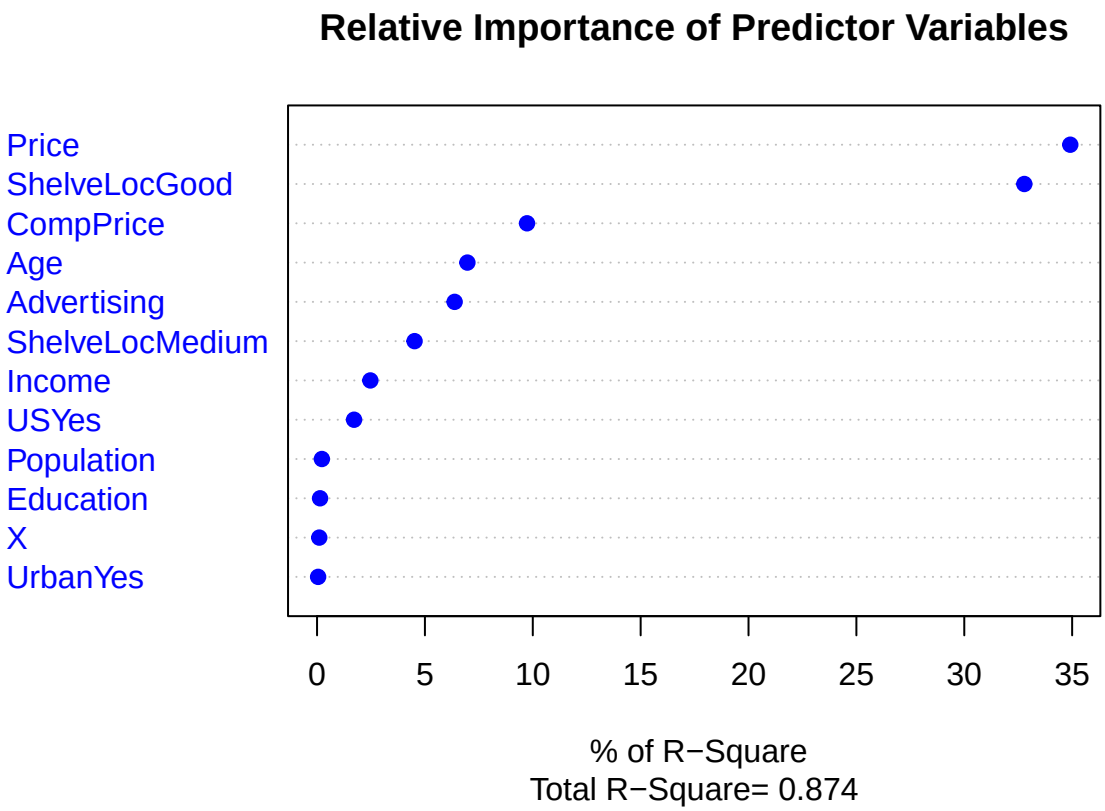
(1)

```
relweights <- function(fit, ...) {
  R <- cor(fit$model)
  nvar <- ncol(R)
  rxx <- R[2:nvar, 2:nvar]
  rxy <- R[2:nvar, 1]
  svd <- eigen(rxx)
  evec <- svd$vectors
  ev <- svd$values
  delta <- diag(sqrt(ev))
  lambda <- evec %*% delta %*% t(evec)
  lambdasq <- lambda ^ 2
  beta <- solve(lambda) %*% rxy
  rsquare <- colSums(beta ^ 2)
  rawwgt <- lambdasq %*% beta ^ 2
  import <- (rawwgt / rsquare) * 100
  import = as.data.frame(import)
  row.names(import) = names(fit$model[2:nvar])
  names(import) = "weights"
  import = import[order(import), 1, drop = FALSE]
  dotchart(
    import$weights,
    labels = row.names(import),
```

```
    xlab = " % of R-Square",
    pch = 19,
    main = "Relative Importance of Predictor Variables",
    sub = paste("Total R-Square=", round(rsquare, digits = 3)),
    ...
  )
  return(import)
}
```

```
temp <- Carseats
temp$ShelveLocGood <- 0
temp$ShelveLocMedium <- 0
temp$UrbanYes <- 0
temp$USYes <- 0
temp$ShelveLocGood[temp$ShelveLoc == "Good"] <- 1
temp$ShelveLocMedium[temp$ShelveLoc == "Medium"] <- 1
temp$UrbanYes[temp$Urban == "Yes"] <- 1
temp$USYes[temp$US == "Yes"] <- 1
```

```
fit <-
  lm(
    Sales ~ X + CompPrice + Income + Advertising + Population + Price + Age +
      Education + ShelveLocGood + ShelveLocMedium + UrbanYes + USYes,
    data = temp
  )
relweights(fit, col = "blue")
```




```
##                    weights
## UrbanYes          0.04963064
## X                 0.10236202
## Education         0.13986689
## Population        0.22348750
## USYes             1.71611269
## Income            2.47061015
## ShelveLocMedium   4.52056314
## Advertising       6.37738776
## Age               6.96848000
## CompPrice         9.73297280
## ShelveLocGood    32.78481178
## Price            34.91371463
```

The results are compatible with the results in (k).

```
detach(Carseats)
```

2.3

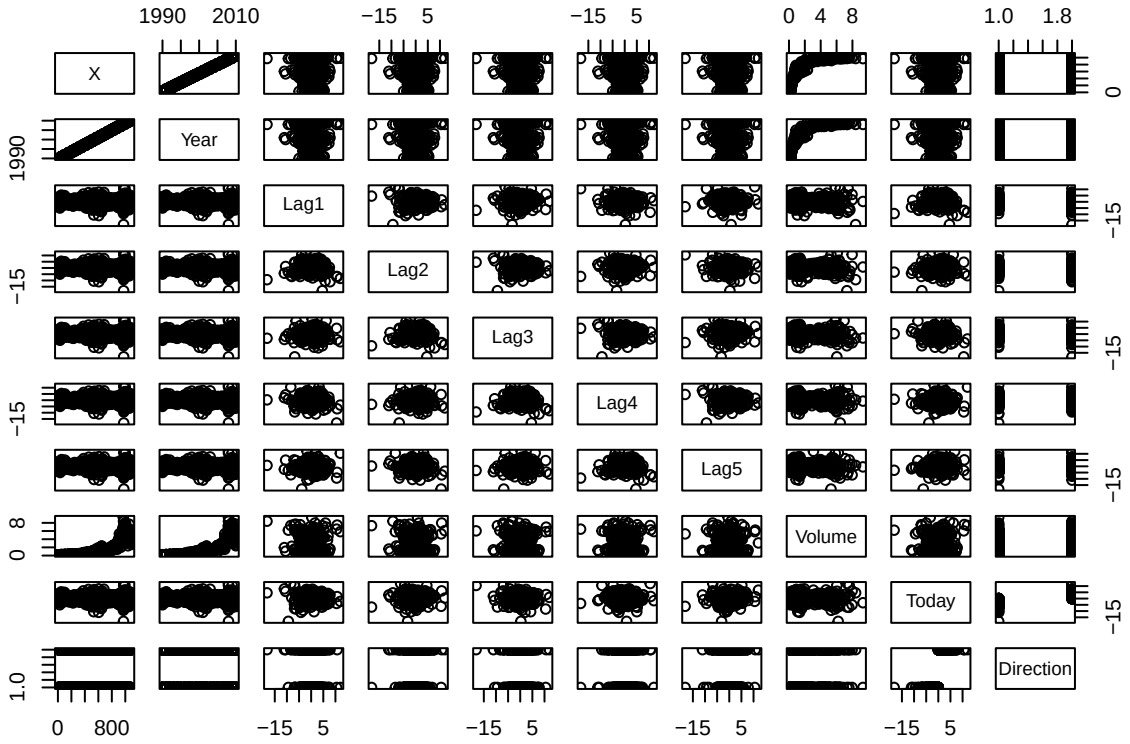
```
weekly <- read.csv("weekly.csv")
```

(a)

```
summary(weekly)
```

```
##        X              Year           Lag1               Lag2
##  Min.   :   1   Min.   :1990   Min.   :-18.1950   Min.   :-18.1950
##  1st Qu.: 273   1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540
##  Median : 545   Median :2000   Median :  0.2410   Median :  0.2410
##  Mean   : 545   Mean   :2000   Mean   :  0.1506   Mean   :  0.1511
##  3rd Qu.: 817   3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090
##  Max.   :1089   Max.   :2010   Max.   : 12.0260   Max.   : 12.0260
##       Lag3               Lag4               Lag5              Volume
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747
##  1st Qu.: -1.1580   1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202
##  Median :  0.2410   Median :  0.2380   Median :  0.2340   Median :1.00268
##  Mean   :  0.1472   Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462
##  3rd Qu.:  1.4090   3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373
##  Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821
##      Today           Direction
##  Min.   :-18.1950   Length:1089
##  1st Qu.: -1.1540   Class :character
##  Median :  0.2410   Mode  :character
##  Mean   :  0.1499
##  3rd Qu.:  1.4050
##  Max.   : 12.0260
```

```
plot(weekly)
```



We can see that some statistics of 'Lags' as well as 'Today' are very similar to each other. And most data should be concentrated in a small interval except for some outliers. No. of 'Down' days is slightly smaller than 'Up' days. Volume is increasing by the year.

(b)

```
# Up -- 1; Down -- 0
weekly$y[weekly$Direction == 'Up'] <- 1
weekly$y[weekly$Direction == 'Down'] <- 0
weekly$y <- factor(weekly$y, levels = c(0, 1), labels = c(0, 1))
fit <-
  glm(
    y ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
    family = binomial(link = "logit"),
    data = weekly
  )
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
##     family = binomial(link = "logit"), data = weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

There is a predictor that appears to be statistically significant. Lag2

(c)

```
a <- exp(predict(fit))
pai <- a / (1 + a)
pai[pai > 0.5] <- 1
pai[pai < 0.5] <- 0
temp <- data.frame(weekly$y, pai)
# assume 'Up' -- 1 is positive
TP <- length(which(temp[1] == 1 & temp[2] == 1))
FN <- length(which(temp[1] == 1 & temp[2] == 0))
FP <- length(which(temp[1] == 0 & temp[2] == 1))
TN <- length(which(temp[1] == 0 & temp[2] == 0))
# confusion matrix
data.frame(
  predict.up = c(TP, FP),
  predict.down = c(FN, TN),
  row.names = c('true.up', 'true.down')
)
```

```
##           predict.up predict.down
## true.up          557           48
## true.down        430           54
```

```
# overall fraction of correct predictions
frac <- (TP + TN) / (TP + FN + FP + TN)
sprintf('overall fraction of correct predictions:   %f', frac)
```

```
## [1] "overall fraction of correct predictions:   0.561065"
```

Assume 'Up' is 'positive', associating with 1 and 'Down' is 'negative', associating with 0. Then from the confusion matrix, we have

Type I error rate = FP/(FP+TN) = 430/(430+54) = 88.84%

Type II error rate = FN/(FN+TP) =48/(48+557) = 7.93%

There are more Type I mistakes. However, this conclusion is opposite if 'Down' is 'positive'.

(d)

```
fit <-
  glm(y ~ Lag2, data = weekly[weekly$Year <= 2009, ], family = binomial())

a <- exp(predict(fit, newdata = weekly[weekly$Year == 2010, ]))
pai <- a / (1 + a)
pai[pai > 0.5] <- 1
pai[pai < 0.5] <- 0
temp <- data.frame(weekly[weekly$Year == 2010, ]$y, pai)
# assume 'Up' -- 1 is positive
TP <- length(which(temp[1] == 1 & temp[2] == 1))
FN <- length(which(temp[1] == 1 & temp[2] == 0))
FP <- length(which(temp[1] == 0 & temp[2] == 1))
TN <- length(which(temp[1] == 0 & temp[2] == 0))
# confusion matrix
data.frame(
  predict.up = c(TP, FP),
  predict.down = c(FN, TN),
  row.names = c('true.up', 'true.down')
)
```

```
##           predict.up predict.down
## true.up           32            0
## true.down         17            3
```

```
# overall fraction of correct predictions
frac <- (TP + TN) / (TP + FN + FP + TN)
sprintf('overall fraction of correct predictions:   %f', frac)
```

```
## [1] "overall fraction of correct predictions:   0.673077"
```