Problem Set 3

Daniel Lee

February 24, 2017

Problem 3

Part a

```
coat1 <- c(2.34, 2.46, 2.83, 2.04, 2.69)
coat2 <- c(2.64, 3.00, 3.19, 3.83)
coat3 <- c(2.61, 2.07, 2.80, 2.58, 2.98, 2.30)
shirt1 <- c(1.32, 1.62, 1.92, 0.88, 1.50, 1.30)
shirt2 <- c(0.41, 0.83, 0.53, 0.32, 1.62)

Y <- c(coat1, coat2, coat3, shirt1, shirt2)
X <- c(rep(1, length(coat1)), rep(2, length(coat2)), rep(3, length(coat3))
       , rep(4, length(shirt1)), rep(5, length(shirt2)))
summary(aov(Y ~ X))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## X            1 13.011  13.011   35.98 3.42e-06 ***
## Residuals   24  8.679   0.362                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the F-test, there seems to be a difference. We can reject the null hypothesis that there is no significant differences in the sturdiness of these three coats and two shirts. The p-value is very small.

Part b

```
coats <- c(coat1, coat2, coat3)
shirts <- c(shirt1, shirt2)
t.test(coats, shirts, var.equal = FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  coats and shirts
## t = 7.7814, df = 19.242, p-value = 2.317e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.153203 2.000858
## sample estimates:
## mean of x mean of y 
##  2.690667  1.113636
```

The results of the 2-sample t-test tells us that we can reject the null hypothesis that there is no significant difference between the sturdiness of the coats and t-shirts.

Part c

The three orthogonal contrasts are the following:

Contrast 1: Shirt 1 is not different from shirt 2. Contrast 2: Coat 2 is not different from coat 1 and coat 3. Contrast 3: Coat 1 is not different from coat 3. Contrast used in part b: All the coats are not different in their means.

```r
# Setting three contrasts to represent

# Contrast 1: Shirt 1 is not different from shirt 2.
contrast1 <- c(0, 0, 0, 1, -1)

# Contrast 2: Coat 2 is not different from coat 1 and coat 3.
contrast2 <- c(1/2, -1, 1/2, 0, 0)

# Contrast 3: Coat 1 is not different from coat 3.
contrast3 <- c(-1, 0, 1, 0, 0)

# Contrast from part b
contrastB <- c(1/3, 1/3, 1/3, -1/2, -1/2)

# Check to make sure contrasts are mutually orthogonal
contrastB %*% contrast1
```

```
##      [,1]
## [1,]    0
```

```r
contrastB %*% contrast2
```

```
##      [,1]
## [1,]    0
```

```r
contrastB %*% contrast3
```

```
##      [,1]
## [1,]    0
```

```r
contrast1 %*% contrast2
```

```
##      [,1]
## [1,]    0
```

```r
contrast1 %*% contrast3
```

```
##      [,1]
## [1,]    0
```

```r
contrast2 %*% contrast3
```

```
##      [,1]
## [1,]    0
```

Calculate sums of squares for all four contrasts.

$$SS(\lambda^T\beta) = \left(\sum_{i=1}^{t} \lambda_i \bar{y}_{i.}\right)^2 \Big/ \left(\sum_{i=1}^{t} \lambda_i^2 / N_i\right)$$

```r
y_bar <- c(mean(coat1),
           mean(coat2),
           mean(coat3),
           mean(shirt1),
           mean(shirt2))

y_var <- c(var(coat1),
               var(coat2),
               var(coat3),
               var(shirt1),
               var(shirt2))

n_j <- c(length(coat1),
                   length(coat2),
                   length(coat3),
                   length(shirt1),
                   length(shirt2))
#sum of suqares for contrasts:
# Contrast 1
(contrast1 %*% y_bar)^2 / sum(contrast1^2/n_j)
```

```
##          [,1]
## [1,] 1.266041
```

```r
# Contrast 2
(contrast2 %*% y_bar)^2 / sum(contrast2^2/n_j)
```

```
##          [,1]
## [1,] 1.239123
```

```r
# Contrast 3
(contrast3 %*% y_bar)^2 / sum(contrast3^2/n_j)
```

```
##           [,1]
## [1,] 0.0195503
```

```r
# Contrast in part b
(contrastB %*% y_bar)^2 / sum(contrastB^2/n_j)
```

```
##          [,1]
## [1,] 16.96621
```

Part d

Two possible confidence intervals are possible:

1. Calculate sample variance using data from only the two shirt brands. This is because we don't assume the variance is equal among the groups.
2. Calculate sample variance using data from all five groups. This is because we assume all the groups have equal

I will construct a 95% confidence interval using both of these estimates of variance.

```r
# 95% CI using only samples from Shirt 1 and Shirt 2
```

```
se_hat <- sqrt((length(shirt1 - 1) * var(shirt1) + length(shirt2 - 1) * var(shirt2))/
                 (length(shirt1) + length(shirt2) - 2))
ci <- sum(y_bar * contrast1) + qt(c(0.025, 0.975),
                                   df = length(shirt1) + length(shirt2) - 2) * se_hat

# 95% CI using all the samples from five groups

se_hat <- sqrt(sum((n_j - 1) * y_var) / (sum(n_j) - 5))
ci <- sum(y_bar * contrast1) + qt(c(0.025, 0.975),
                                   df = sum(n_j) - 5) * se_hat
```

Since both of these confidence intervals include zero, the two brands are not significantly sturdier than the other.

Problem 4

```
rm(list = ls())
brand1 <- c(3.41, 1.83, 2.69, 2.04, 2.83, 2.46, 1.84, 2.34)
brand2 <- c(3.58, 3.83, 2.64, 3.00, 3.19, 3.57, 3.04, 3.09)
brand3 <- c(3.32, 2.62, 3.92, 3.88, 2.50, 3.30, 2.28, 3.57)
brand4 <- c(3.22, 2.61, 2.07, 2.58, 2.80, 2.98, 2.30, 1.66)

jeans <- data.frame(
  wear = c(brand1, brand2, brand3, brand4)
  , brand = c(rep(1, length(brand1)), rep(2, length(brand2)), rep(3, length(brand3)), rep(4, length(bra

jeans$brand <- as.factor(jeans$brand)
```

Part a

```
model <- aov(wear ~ brand, data = jeans)
summary(model)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## brand        3  4.313  1.4376   5.225 0.00543 **
## Residuals   28  7.703  0.2751
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the F-test, the p-value is very small. So, we reject the null hypothesis that there is no significant differences in the wear of the four different jeans brands.

Part b

Possible orthogonal contrasts:

Contrast 1: Are brands 1 and 4 equal to brands 2 and 3?

Contrast 2: Is brand 1 equal to brand 4?

Contrast 3: Is brand 2 equal to brand 3?

```r
# Setting three contrasts to represent

# Contrast 1: Are brands 1 and 4 equal to brands 2 and 3?

contrast1 <- c(1/2, -1/2, -1/2, 1/2)

# Contrast 2: Is brand 1 equal to brand 4?

contrast2 <- c(1, 0, 0, -1)

# Contrast 3: Is brand 2 equal to brand 3?

contrast3 <- c(0, 1, -1, 0)

# Check to make sure contrasts are mutually orthogonal
contrast1 %*% contrast2
```

```
##      [,1]
## [1,]    0
```

```r
contrast1 %*% contrast3
```

```
##      [,1]
## [1,]    0
```

```r
contrast2 %*% contrast3
```

```
##      [,1]
## [1,]    0
```

Calculate sums of squares for the contrasts.

```r
y <- cbind(brand1, brand2, brand3, brand4)
y_bar <- apply(y, 2, mean)

y_var <- apply(y, 2, var)

n_j <- apply(y, 2, length)

#sum of sugares for contrasts:
# Contrast 1
(contrast1 %*% y_bar)^2 / sum(contrast1^2/n_j)
```

```
##          [,1]
## [1,] 4.255903
```

```r
# Contrast 2
(contrast2 %*% y_bar)^2 / sum(contrast2^2/n_j)
```

```
##          [,1]
## [1,] 0.038025
```

```r
# Contrast 3
(contrast3 %*% y_bar)^2 / sum(contrast3^2/n_j)
```

```
##            [,1]
## [1,] 0.01890625
```

Part C

Scheffe's method, alpha = 0.01,

```r
contrast12 <- c(1, -1, 0, 0)
contrast13 <- c(1, 0, -1, 0)
contrast14 <- c(1, 0, 0, -1)
contrast23 <- c(0, 1, -1, 0)
contrast24 <- c(0, 1, 0, -1)
contrast34 <- c(0, 0, 1, -1)

n <- sum(n_j)

df_error <- df.residual(model)
MSerror <- deviance(model)/df_error
J <- 4
s <- J - 1

scheffe_test <- function(contrast, mse, df_s, df_e, n_total, y_ave, alpha){
  scheffe_statistic <- (contrast %*% y_ave)^2 / sum(contrast^2/n_total) / df_s / mse
  scheffe_critical_value <- qf(1 - alpha, df1 = df_s, df2 = df_e)
  if(scheffe_statistic > scheffe_critical_value){
    cat('Reject Null (Statistic=', scheffe_statistic, ')',
        '(Critical Value=', scheffe_critical_value, ')', sep = ' ')
  }else{
    cat('Do Not Reject Null (Statistic=', scheffe_statistic, ')',
        '(Critical Value=', scheffe_critical_value, ')', sep = ' ')
  }
}

scheffe_test(contrast12, MSerror, s, df_error, n, y_bar, 0.01)
```

```
## Reject Null (Statistic= 12.79723 ) (Critical Value= 4.568091 )
```

```r
scheffe_test(contrast13, MSerror, s, df_error, n, y_bar, 0.01)
```

```
## Reject Null (Statistic= 10.72317 ) (Critical Value= 4.568091 )
```

```r
scheffe_test(contrast14, MSerror, s, df_error, n, y_bar, 0.01)
```

```
## Do Not Reject Null (Statistic= 0.1842802 ) (Critical Value= 4.568091 )
```

```r
scheffe_test(contrast23, MSerror, s, df_error, n, y_bar, 0.01)
```

```
## Do Not Reject Null (Statistic= 0.09162517 ) (Critical Value= 4.568091 )
```

```r
scheffe_test(contrast24, MSerror, s, df_error, n, y_bar, 0.01)
```

```
## Reject Null (Statistic= 9.910178 ) (Critical Value= 4.568091 )
```

```r
scheffe_test(contrast34, MSerror, s, df_error, n, y_bar, 0.01)
```

```
## Reject Null (Statistic= 8.096 ) (Critical Value= 4.568091 )
```

According to the Scheffe's method, Brands 1 and 4 are not different. Brands 2 and 3 are not different. But brands 1 and 4 are different from brands 2 and 3, pairwise. This is what we expected.

LSD method, alpha = 0.01

```r
summary(aov(wear ~ brand, data = jeans))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)   
## brand        3  4.313  1.4376   5.225 0.00543 **
## Residuals   28  7.703  0.2751                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis. So we proceed to do pairwise test. We assume that the variances of all the groups are different. So we don't use all the samples in the four groups to estimate variance.

```r
t.test(brand1, brand2, var.equal = FALSE) #reject
```

```
## 
##  Welch Two Sample t-test
## 
## data:  brand1 and brand2
## t = -3.4465, df = 12.685, p-value = 0.004483
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3230934 -0.3019066
## sample estimates:
## mean of x mean of y 
##    2.4300    3.2425
```

```r
t.test(brand1, brand3, var.equal = FALSE) #do not reject
```

```
## 
##  Welch Two Sample t-test
## 
## data:  brand1 and brand3
## t = -2.523, df = 13.673, p-value = 0.0247
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3774397 -0.1100603
## sample estimates:
## mean of x mean of y 
##   2.43000   3.17375
```

```r
t.test(brand1, brand4, var.equal = FALSE) #do not reject
```

```
## 
##  Welch Two Sample t-test
## 
## data:  brand1 and brand4
## t = -0.37239, df = 13.928, p-value = 0.7152
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.6593315  0.4643315
## sample estimates:
## mean of x mean of y 
##    2.4300    2.5275
```

```r
t.test(brand2, brand3, var.equal = FALSE) #do not reject
```

```
## 
##  Welch Two Sample t-test
```

```
##
## data:  brand2 and brand3
## t = 0.26171, df = 11.608, p-value = 0.7981
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5057676  0.6432676
## sample estimates:
## mean of x mean of y
##   3.24250   3.17375
```

```
t.test(brand2, brand4, var.equal = FALSE) #reject
```

```
##
##  Welch Two Sample t-test
##
## data:  brand2 and brand4
## t = 3.1767, df = 13.138, p-value = 0.007203
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2292734 1.2007266
## sample estimates:
## mean of x mean of y
##    3.2425    2.5275
```

```
t.test(brand3, brand4, var.equal = FALSE) #do not reject
```

```
##
##  Welch Two Sample t-test
##
## data:  brand3 and brand4
## t = 2.257, df = 13.332, p-value = 0.04139
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.02923518 1.26326482
## sample estimates:
## mean of x mean of y
##   3.17375   2.52750
```

The pairwise comparison using the t-test indicates that brands 1 and 2 are different from each other. Also brands 2 and 4 are different. The rest of the pairwise comparisons are not significant enough to reject the null hypothesis at alpha = 0.01.

Bonferroni method, alpha = 0.012

```
bonf_test <- function(contrast, mse, df_e, n_total, y_ave, alpha, total_tests){
  bonf_statistic <- (contrast %*% y_ave)^2 / sum(contrast^2/n_total) / mse
  bonf_critical_value <- qf(1 - alpha/total_tests, df1 = 1, df2 = df_e)
  if(bonf_statistic > bonf_critical_value){
    cat('Reject Null (Statistic=', bonf_statistic, ')',
        '(Critical Value=', bonf_critical_value, ')', sep = ' ')
  }else{
    cat('Do Not Reject Null (Statistic=', bonf_statistic, ')',
        '(Critical Value=', bonf_critical_value, ')', sep = ' ')
  }
}

bonf_test(contrast12, MSerror, df_error, n, y_bar, 0.012, 6) #reject
```

```
## Reject Null (Statistic= 38.3917 ) (Critical Value= 11.61552 )
```

```r
bonf_test(contrast13, MSerror, df_error, n, y_bar, 0.012, 6) #reject
```

```
## Reject Null (Statistic= 32.16952 ) (Critical Value= 11.61552 )
```

```r
bonf_test(contrast14, MSerror, df_error, n, y_bar, 0.012, 6) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.5528405 ) (Critical Value= 11.61552 )
```

```r
bonf_test(contrast23, MSerror, df_error, n, y_bar, 0.012, 6) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.2748755 ) (Critical Value= 11.61552 )
```

```r
bonf_test(contrast24, MSerror, df_error, n, y_bar, 0.012, 6) #reject
```

```
## Reject Null (Statistic= 29.73053 ) (Critical Value= 11.61552 )
```

```r
bonf_test(contrast34, MSerror, df_error, n, y_bar, 0.012, 6) #reject
```

```
## Reject Null (Statistic= 24.288 ) (Critical Value= 11.61552 )
```

According to the Bonferroni method, brands 1 and 4 are not significantly different. Brands 2 and 3 are also not significantly different. All other pairwise comparisons are significantly different enough so we reject the null hypothesis.

Tukey's HSD method, alpha = 0.01,

```r
tukey_test <- function(brandi, brandj, mse){
  tukey_critical_value <- qtukey(0.99, nmeans = 4, df = 28)*
                                  sqrt(mse/length(brandi))
  mean_diff <- abs(mean(brandi) - mean(brandj))
  if(mean_diff > tukey_critical_value){
    cat('Reject Null (Statistic=', mean_diff, ')',
        '(Critical Value=', tukey_critical_value, ')', sep = ' ')
  }else{
    cat('Do Not Reject Null (Statistic=', mean_diff, ')',
        '(Critical Value=', tukey_critical_value, ')', sep = ' ')
  }
}

tukey_test(brand1, brand2, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.8125 ) (Critical Value= 0.8956333 )
```

```r
tukey_test(brand1, brand3, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.74375 ) (Critical Value= 0.8956333 )
```

```r
tukey_test(brand1, brand4, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.0975 ) (Critical Value= 0.8956333 )
```

```r
tukey_test(brand2, brand3, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.06875 ) (Critical Value= 0.8956333 )
```

```r
tukey_test(brand2, brand4, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.715 ) (Critical Value= 0.8956333 )
```

```r
tukey_test(brand3, brand4, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.64625 ) (Critical Value= 0.8956333 )
```

According to Tukey's tests, we do not reject any of the pairwise comparisons.

Newman-Keuls method, alpha = 0.01

```r
sort(y_bar)
```

```
##  brand1  brand4  brand3  brand2
## 2.43000 2.52750 3.17375 3.24250
```

Minimum mean is brand 1 while maximum mean is brand 2. I perform Tukey's method on these two groups.

```r
tukey_test(brand1, brand2, MSerror) #do not reject
```

```
## Do Not Reject Null (Statistic= 0.8125 ) (Critical Value= 0.8956333 )
```

According to Newman-Keuls method, there is no signficant differences among the groups.

Problem 5

Part a

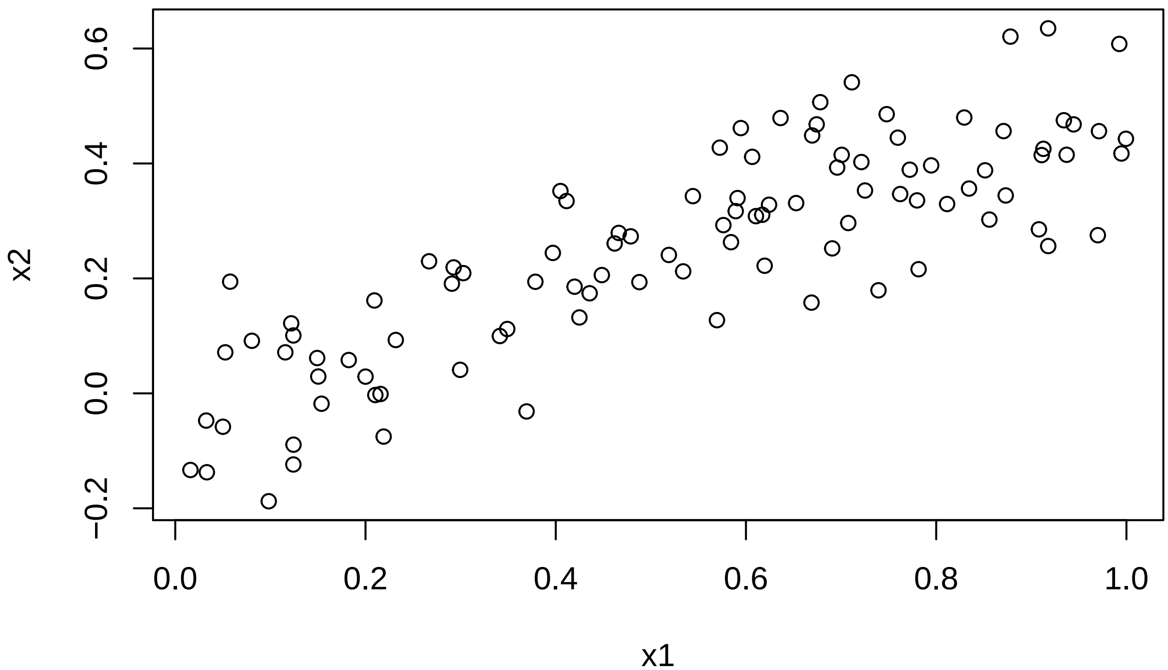
```r
set.seed(240)
x1 <- runif(100)
x2 <- 0.5*x1 + rnorm(100)/10
y <- 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

Part b

```r
cor(x1, x2)
```

```
## [1] 0.835556
```

```r
plot(x1, x2)
```



Part c

```r
fit <- lm(y ~ x1 + x2)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05592 -0.70231 -0.02194  0.75459  3.15141
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.969709   0.218532   9.013 1.81e-14 ***
## x1          2.035884   0.647079   3.146   0.0022 **
## x2          0.005801   1.017236   0.006   0.9955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 97 degrees of freedom
## Multiple R-squared:  0.2532, Adjusted R-squared:  0.2378
## F-statistic: 16.45 on 2 and 97 DF,  p-value: 7.068e-07
```

$\hat{\beta}_0 = 1.969$, $\hat{\beta}_1 = 2.035$, $\hat{\beta}_2 = 0.0058$. $\hat{\beta}_1$ and $\hat{\beta}_2$ were significant.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are very close to the true values β_0 and β_1, while $\hat{\beta}_2$ is not close to the true value of β_2.

I can reject the null hypotheses that $\beta_1 = 0$ and $\beta_2 = 0$ the p-values associated with these tests were significant.

Part d

```r
summary(lm(y ~ x1))
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.05494 -0.70239 -0.02164  0.75511  3.15114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.9695     0.2151   9.154 8.28e-15 ***
## x1            2.0390     0.3537   5.765 9.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 98 degrees of freedom
## Multiple R-squared:  0.2532, Adjusted R-squared:  0.2456
## F-statistic: 33.23 on 1 and 98 DF,  p-value: 9.492e-08
```

The results are very close to the previous results obtained in part c for β_1. We can reject the null hypothesis that $\beta_1 = 0$.

Part e

```r
summary(lm(y ~ x2))
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5851 -0.6310 -0.0088  0.6724  3.0686
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3820     0.1826  13.041  < 2e-16 ***
## x2            2.6800     0.5837   4.591 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.066 on 98 degrees of freedom
```

```
## Multiple R-squared:  0.177,  Adjusted R-squared:  0.1686
## F-statistic: 21.08 on 1 and 98 DF,  p-value: 1.308e-05
```

The results are not close to what we originally saw in part c for β_2. We can reject the null hypothesis that $\beta_2 = 0$.

Part f

There seems to be a seeming contradiction because of the $\hat{\beta}_2$ estimate being influenced by whether or not $\hat{\beta}_1$ was included in the model. This is due to the collinearity between x_2 and x_1 in the data generating process.

Part g

```r
x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)
y <- c(y, 6)

summary(lm(y ~ x1 + x2))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3985 -0.7134 -0.0901  0.6590  3.1700
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1154     0.2209   9.577    1e-15 ***
## x1            0.9522     0.5510   1.728   0.0871 .
## x2            1.8120     0.8397   2.158   0.0334 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.06 on 98 degrees of freedom
## Multiple R-squared:  0.2354, Adjusted R-squared:  0.2198
## F-statistic: 15.09 on 2 and 98 DF,  p-value: 1.939e-06
```

The results from this model are quite a bit different from the true values $\hat{\beta}_1$ and $\hat{\beta}_2$ but similar to the estimates obtained in part c. The intercept and $\hat{\beta}_0$ was significant. Also, $\hat{\beta}_2$ is significant at 5% significance level.

```r
summary(lm(y ~ x1))
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1177 -0.7514 -0.0038  0.7910  3.7041
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   2.1116     0.2249   9.389 2.36e-15 ***
## x1            1.8431     0.3715   4.961 2.92e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.079 on 99 degrees of freedom
## Multiple R-squared:  0.1991, Adjusted R-squared:  0.191
## F-statistic: 24.61 on 1 and 99 DF,  p-value: 2.917e-06
```

The results from this model are consistent with the previous exercise done in part d and $\hat{\beta}_1$ is similar to true value. All the coefficient estimates are significant.

```r
summary(lm(y ~ x2))
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6200 -0.6158 -0.0234  0.6339  3.1045
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.3397     0.1805  12.964  < 2e-16 ***
## x2            2.8993     0.5616   5.163 1.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 99 degrees of freedom
## Multiple R-squared:  0.2121, Adjusted R-squared:  0.2042
## F-statistic: 26.65 on 1 and 99 DF,  p-value: 1.258e-06
```

The results from this model is pretty similar from the estimate obtained in part e. However, the values differ from the true value β_2. Leverage outliers are below.

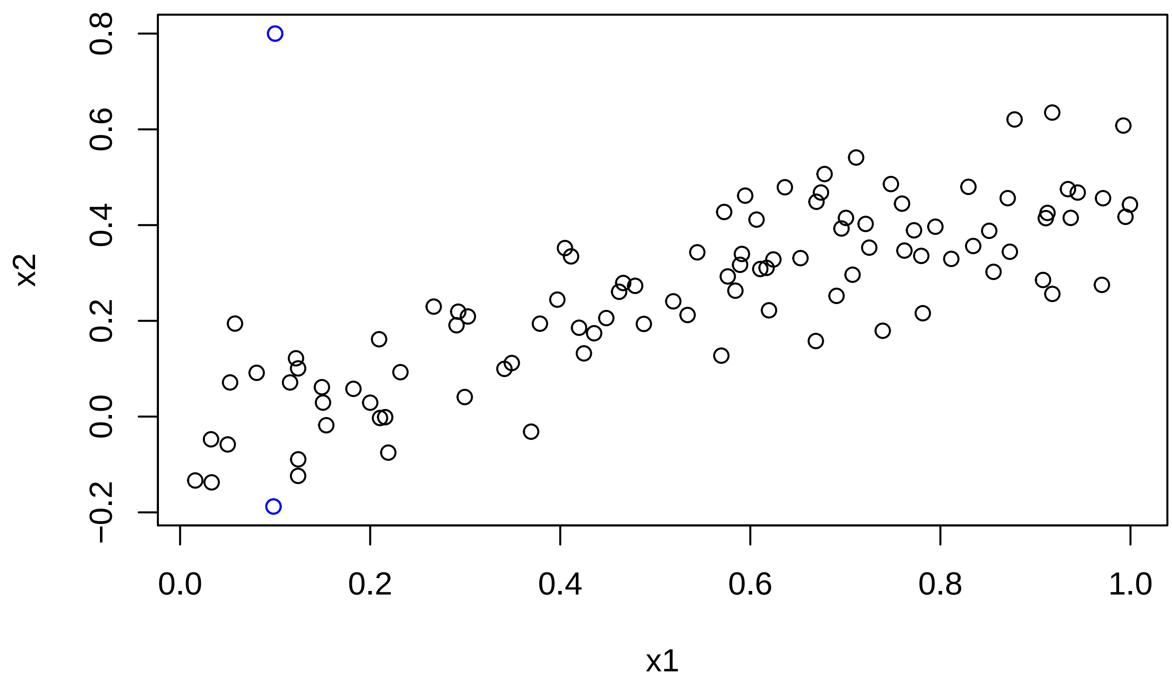
```r
plot(x1, x2)

# Calculate leverage
X <- cbind(1, cbind(x1, x2))
H <- X %*% solve(t(X) %*% X) %*% t(X)
H_ii <- diag(H)

# Determine high leverage points
H_ii >= 2 * (NCOL(X)) / NROW(X)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [45] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [56] FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [67] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [78] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [89] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [100] FALSE  TRUE
```

```r
subs <- H_ii >= 2 * (NCOL(X)) / NROW(X)
points(x1[subs], x2[subs], col = "blue")
```




```r
# This new point is 4.8 times greater than the next highest leverage point
(H_ii[order(H_ii)[101]] - H_ii[order(H_ii)[100]]) / H_ii[order(H_ii)[100]]
```

```
## [1] 4.827788
```

The point `x1[subs]` seems to be an outlier whereas `x2[subs]` seems to be a leverage point. However, the criteria for being a leverage point as opposed to an outlier is hard to define.

Problem 6

```r
rm(list = ls())

load('Carseats.RData')
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
```

```
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

```r
nrow(Carseats)
```

```
## [1] 400
```

Part a

```r
fit_6a <- lm(Sales ~ Price + Urban + US, data = Carseats)
```

Part b

```r
summary(fit_6a)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Price: Sales decreases by 0.05449 for every increase in one unit of Price (holding all else constant).

UrbanYes: No clear effect on carseat sales based on whether store is in urban location or not.

USYes: Holding all else constant, stores in the US can expect to sell 1.2 units more than stores not in US.

Part C

Sales ~ Price + Urban + US, data = Carseats)

$$y_i = \beta_0 + \beta_1 * price_i + \beta_2 * urban_i + \beta_3 * US_i + \epsilon_i$$

Price is a continuous variable. Urban and US are indicator variables.

Part d

We can reject the null hypothesis for Price and US since they are significant based on the results in part b.

Part e

```
fit_6e <- lm(Sales ~ Price + US, data = Carseats)
summary(fit_6e)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

The F-stat is 62.43, which is significant and also larger than the previous F-statistic value.

Part f

Based on the R^2 values, these models explain approximately 24% of the variance in the data. This is not a good fit to the data.

Part g

```
n <- NROW(Carseats)
p <- 3
se_hat <- as.matrix(coef(summary(fit_6e))[,2])
beta_hat <- as.matrix(coef(summary(fit_6e))[,1])
t_crit <- qt(0.975, df = n - p)
cbind(beta_hat - t_crit * se_hat, beta_hat + t_crit * se_hat)
```

```
##                    [,1]        [,2]
## (Intercept) 11.79032020 14.27126531
## Price       -0.06475984 -0.04419543
## USYes        0.69151957  1.70776632
```

Part h

```
X <- cbind(1, cbind(Carseats[,c('Price', 'US')]))
X$US <- as.numeric(X$US == 'Yes')
X <- as.matrix(X)

H <- X %*% solve(t(X) %*% X) %*% t(X)
H_ii <- diag(H)

# Determine outliers
H_ii >= 2 * (NCOL(X)) / NROW(X)
```

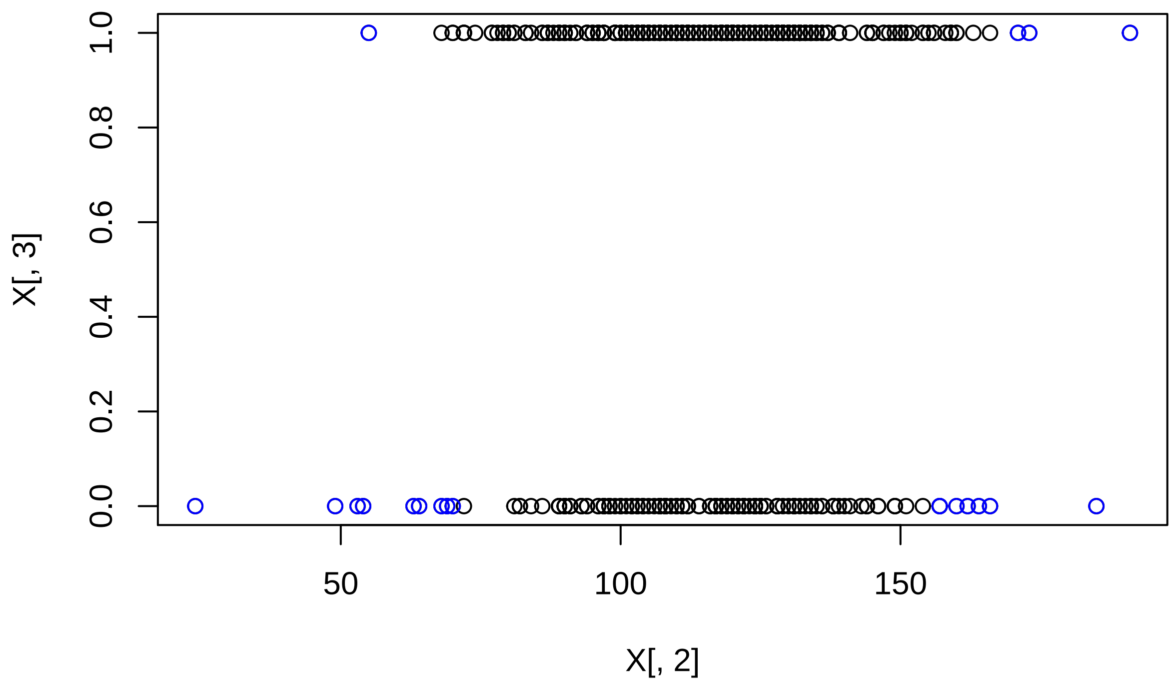
```
##     1     2     3     4     5     6     7     8     9    10    11    12
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    13    14    15    16    17    18    19    20    21    22    23    24
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    25    26    27    28    29    30    31    32    33    34    35    36
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    37    38    39    40    41    42    43    44    45    46    47    48
## FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##    49    50    51    52    53    54    55    56    57    58    59    60
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    61    62    63    64    65    66    67    68    69    70    71    72
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    73    74    75    76    77    78    79    80    81    82    83    84
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    85    86    87    88    89    90    91    92    93    94    95    96
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    97    98    99   100   101   102   103   104   105   106   107   108
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   109   110   111   112   113   114   115   116   117   118   119   120
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   121   122   123   124   125   126   127   128   129   130   131   132
## FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
##   133   134   135   136   137   138   139   140   141   142   143   144
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   145   146   147   148   149   150   151   152   153   154   155   156
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
##   157   158   159   160   161   162   163   164   165   166   167   168
##  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##   169   170   171   172   173   174   175   176   177   178   179   180
## FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##   181   182   183   184   185   186   187   188   189   190   191   192
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
##   193   194   195   196   197   198   199   200   201   202   203   204
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
##   205   206   207   208   209   210   211   212   213   214   215   216
## FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   217   218   219   220   221   222   223   224   225   226   227   228
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   229   230   231   232   233   234   235   236   237   238   239   240
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   241   242   243   244   245   246   247   248   249   250   251   252
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##   253   254   255   256   257   258   259   260   261   262   263   264
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   265   266   267   268   269   270   271   272   273   274   275   276
## FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
##   277   278   279   280   281   282   283   284   285   286   287   288
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   289   290   291   292   293   294   295   296   297   298   299   300
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   301   302   303   304   305   306   307   308   309   310   311   312
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   313   314   315   316   317   318   319   320   321   322   323   324
## FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   325   326   327   328   329   330   331   332   333   334   335   336
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   337   338   339   340   341   342   343   344   345   346   347   348
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   349   350   351   352   353   354   355   356   357   358   359   360
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
##   361   362   363   364   365   366   367   368   369   370   371   372
## FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
##   373   374   375   376   377   378   379   380   381   382   383   384
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
##   385   386   387   388   389   390   391   392   393   394   395   396
## FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##   397   398   399   400
## FALSE FALSE FALSE FALSE
```

```r
subs <- H_ii >= 2 * (NCOL(X)) / NROW(X)
H_ii[subs]
```

```
##         43        126        156        157        160        166
## 0.04333766 0.02596614 0.01610616 0.01535558 0.01570737 0.02856661
##        172        175        192        204        209        270
## 0.02101401 0.02968672 0.01803910 0.01535558 0.01823472 0.01919494
##        273        314        316        357        366        368
## 0.01868734 0.02316470 0.01704881 0.01827894 0.01739884 0.02370705
##        384        387
## 0.01651393 0.01655462
```

```r
plot(X[,2], X[,3])
points(X[subs,2], X[subs,3], col = "blue")
```



Yes there are outliers. All of the points that are outliers are also (by definition) high leverage. The points that are potentially outliers are marked with blue. Points with leverage values greater than $2p/n$ are thought to be outliers here.