Linear Models Homework 8

Joshua Ingram

Problem 1

1.

$$\begin{bmatrix} sales_1 \\ sales_2 \\ \vdots \\ sales_n \end{bmatrix} = \begin{bmatrix} 1 & TV_1 & radio_1 & newspaper_1 \\ 1 & TV_2 & radio_2 & newspaper_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & TV_n & radio_n & newspaper_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \vec{\epsilon} \sim N(\vec{0}, \sigma^2 I_n)$$

2.

Note: This design matrix input should NOT include the column of 1's. This is added in the function.

```
my.MLR <- function(X, y){
  # adding the column of 1's to the design matrix
  intercept <- rep(1, nrow(y))
  X <- cbind(intercept, X)
  X <- as.matrix(X)

  # calculation of useful values and matrices like X'X inverse, X', etc
  n <- nrow(y)
  k <- ncol(X) - 1
  df <- n - (k + 1)
  X_trans <- as.matrix(t(X))
  X_t_X <- X_trans %*% X
  inv_mat <- as.matrix(solve(X_t_X))

  # calculating our beta estimates from the matrix form formula
  estimates <- inv_mat %*% X_trans %*% as.matrix(y)
  # predictions
  y_hat <- estimates[1]*X[,1] + estimates[2]*X[,2] + estimates[3]*X[,3] + estimates[4]*X[,4]

  # RSS
  RSS <- sum((y-y_hat)^2)

  # sigma hat
  sigma_est <- sqrt(RSS/df)

  #v_ii's (the sqrt of the diagonals of the inverse matrix)
  sqrt_v <- sqrt(as.vector(diag(inv_mat)))

  # standard errors
```

```
  SE <- sigma_est * sqrt_v

  # t-values
  t_values <- estimates/SE

  # p-values
  p_values <- pt(abs(t_values), df = df, lower.tail = FALSE) * 2

  # data frame of output
  output <- data.frame(estimates, SE, t_values, p_values)
  colnames(output) <- c("Estimate", "Std. Error", "t value", "p value")
  return(output)

}
```

3.

Problem 2

We have a categorical predictor for both parts of this problem, type, with 3 levels (professional, white collar, and blue collar). We will need 2 dummy variables with our baselines category as blue collar. D_{prof} (column 5 of design matrix) is the dummy variable that takes on 1 when type = prof, 0 otherwise. D_{wc} (column 6 of design matrix) is the dummy variable that takes on 1 when type = wc, 0 otherwise. Additionally (for all design matrices to follow), column 1 is just all ones for the intercept, column 2 is the education column, and column 3 is the women column.

1.

a.

$$prestige_i = \beta_0 + \beta_1(education_i) + \beta_2(income_i) + \beta_3(women_i) + \beta_4(D_{prof}) + \beta_5(D_{wc}) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

b.

$$\begin{bmatrix} prestige_1 \\ prestige_2 \\ prestige_3 \\ prestige_4 \\ prestige_5 \\ prestige_6 \\ prestige_7 \end{bmatrix} = \begin{bmatrix} 1 & 13.11 & 12351 & 11.16 & 1 & 0 \\ 1 & 12.26 & 25879 & 4.02 & 1 & 0 \\ 1 & 9.84 & 7482 & 17.04 & 0 & 1 \\ 1 & 11.13 & 8780 & 3.16 & 0 & 1 \\ 1 & 10.05 & 2594 & 67.82 & 0 & 1 \\ 1 & 8.37 & 4753 & 0 & 0 & 0 \\ 1 & 10 & 6462 & 13.58 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}, \vec{\epsilon} \sim N(\vec{0}, \sigma^2 I_n)$$

2.

a.

$$prestige_i = \beta_0 + \beta_1(education_i) + \beta_2(income_i) + \beta_3(women_i) + \beta_4(D_{prof}) + \beta_5(D_{wc}) + \beta_6(D_{prof})(income_i) + \beta_7(D_{wc})(income_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

b.

Interaction between income and the prof dummy variables is column 7 in the design matrix. Column 8 is the interaction between income and the wc dummy variable.

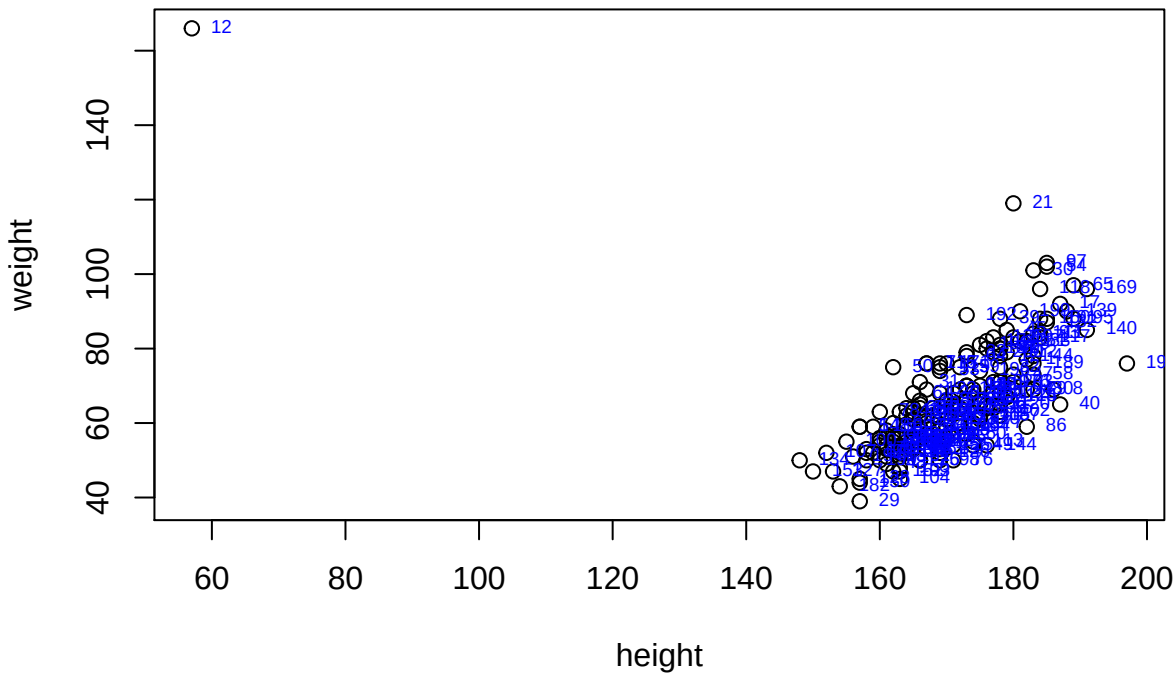
$$\begin{bmatrix} prestige_1 \\ prestige_2 \\ prestige_3 \\ prestige_4 \\ prestige_5 \\ prestige_6 \\ prestige_7 \end{bmatrix} = \begin{bmatrix} 1 & 13.11 & 12351 & 11.16 & 1 & 0 & 12351 & 0 \\ 1 & 12.26 & 25879 & 4.02 & 1 & 0 & 25879 & 0 \\ 1 & 9.84 & 7482 & 17.04 & 0 & 1 & 0 & 7482 \\ 1 & 11.13 & 8780 & 3.16 & 0 & 1 & 0 & 8780 \\ 1 & 10.05 & 2594 & 67.82 & 0 & 1 & 0 & 2594 \\ 1 & 8.37 & 4753 & 0 & 0 & 0 & 0 & 0 \\ 1 & 10 & 6462 & 13.58 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}, \vec{\epsilon} \sim N(\vec{0}, \sigma^2 I_n)$$

Problem 2

1.

a.

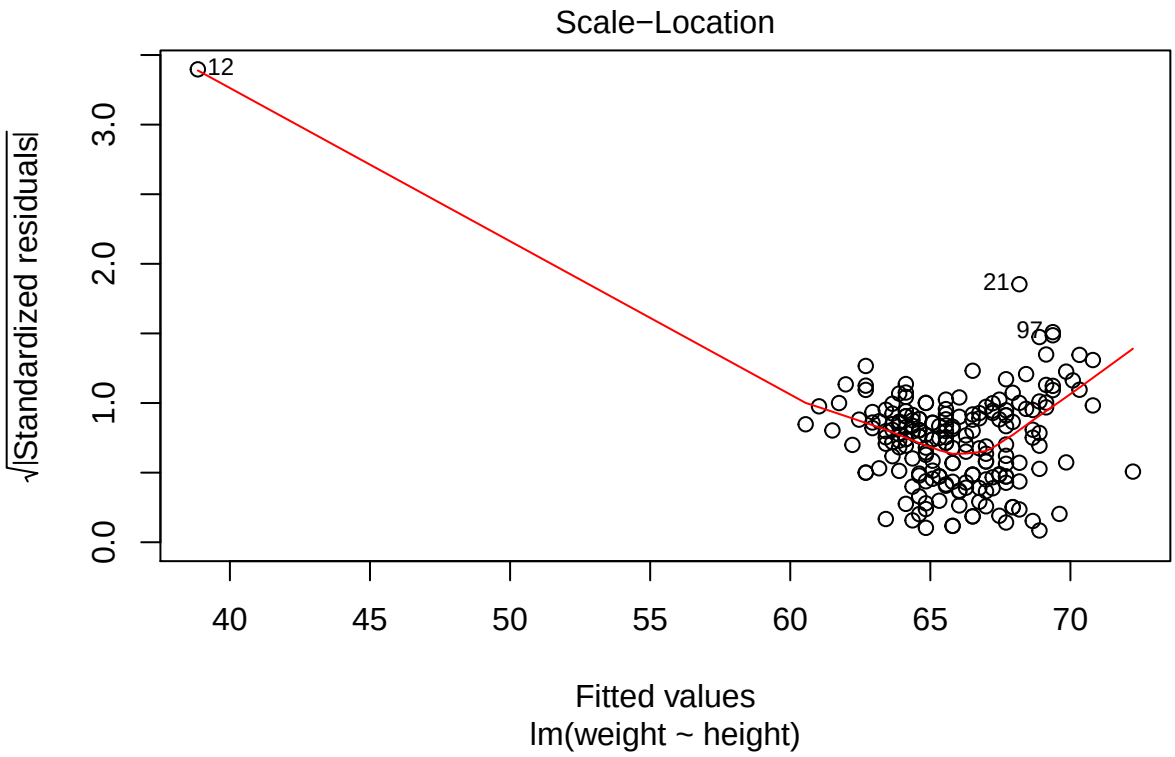
```
plot(weight ~ height, data = davis)
text(x=davis$height,
     y =davis$weight,
     rownames(davis),
     cex =0.6, pos=4, col="blue")
```



b.

21 and 12 seem to be the most likely outliers. 12 is extremely low in the range of heights and it also has a very high weight value relative to the other data points. 21 is in the normal range of height, but it's weight value is somewhat high. To get a numerical measure, we could calculate the standardized residuals.

```
lm_1 <- lm(weight ~ height, data = davis)
stand_1 <- rstandard(lm_1)
# returns sqrt of the standardized residuals
plot(lm_1, which = 3)
```




```
stand_1[12]
```

```
##       12
## 11.54007
```

```
stand_1[21]
```

```
##       21
## 3.434905
```

```
mean(stand_1)
```

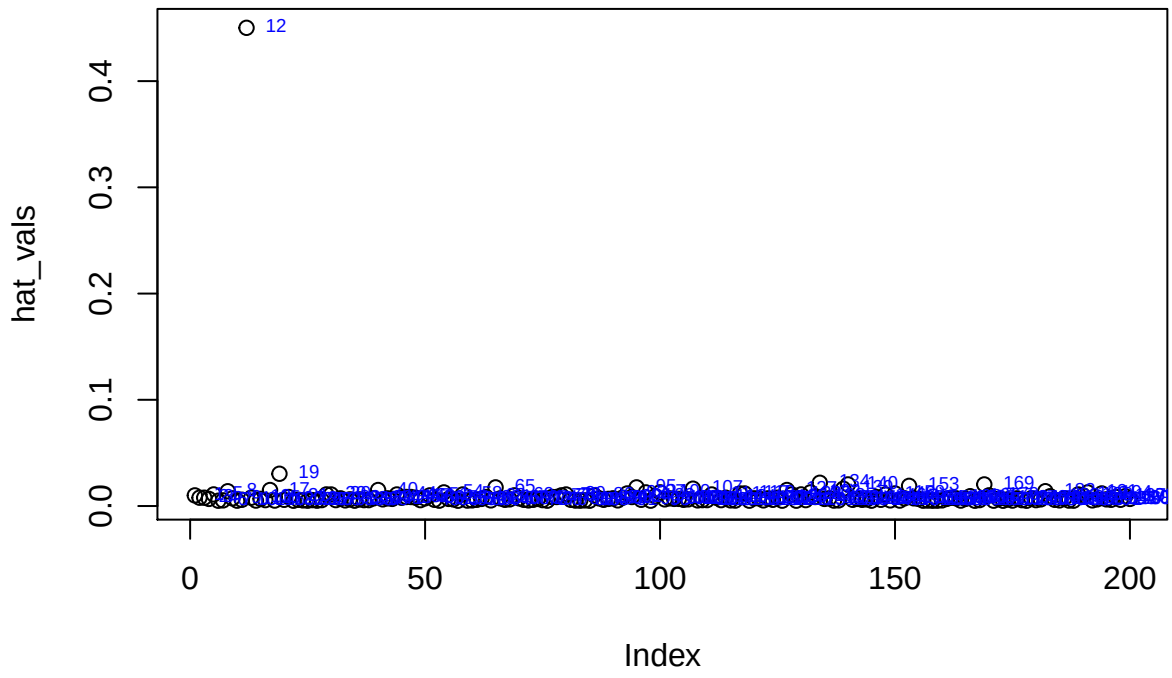
```
## [1] 0.01510447
```

According to our standardized residual plot (of the square roots of the standardized resiudals), 12 and 21 are the most obvious outliers. 12 is definitely the most noticeable outlier relative to the rest. case 12 has standardized residual of 11.54 and 21 has a standardized residual of 3.434905. The mean standardized residual is 0.0151

c.

12 and possibly 19 possibly look to have high leverage (19 less so). We could calculate the hat-values to assess leverage.

```r
hat_vals <- hatvalues(lm_1)
plot(hat_vals)
text(1:length(hat_vals),
     hat_vals,
     rownames(davis),
     cex=0.6, pos=4, col="blue")
```




```r
cutoff <- 2 * (1 +1)/nrow(davis)
cutoff
```

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

```r
hat_vals[12]
```

```
##        12
## 0.4501647
```

12 is really the only outlier with high leverage relative to our other cases. To compare to our cutoff, 12 is greater than 0.02.

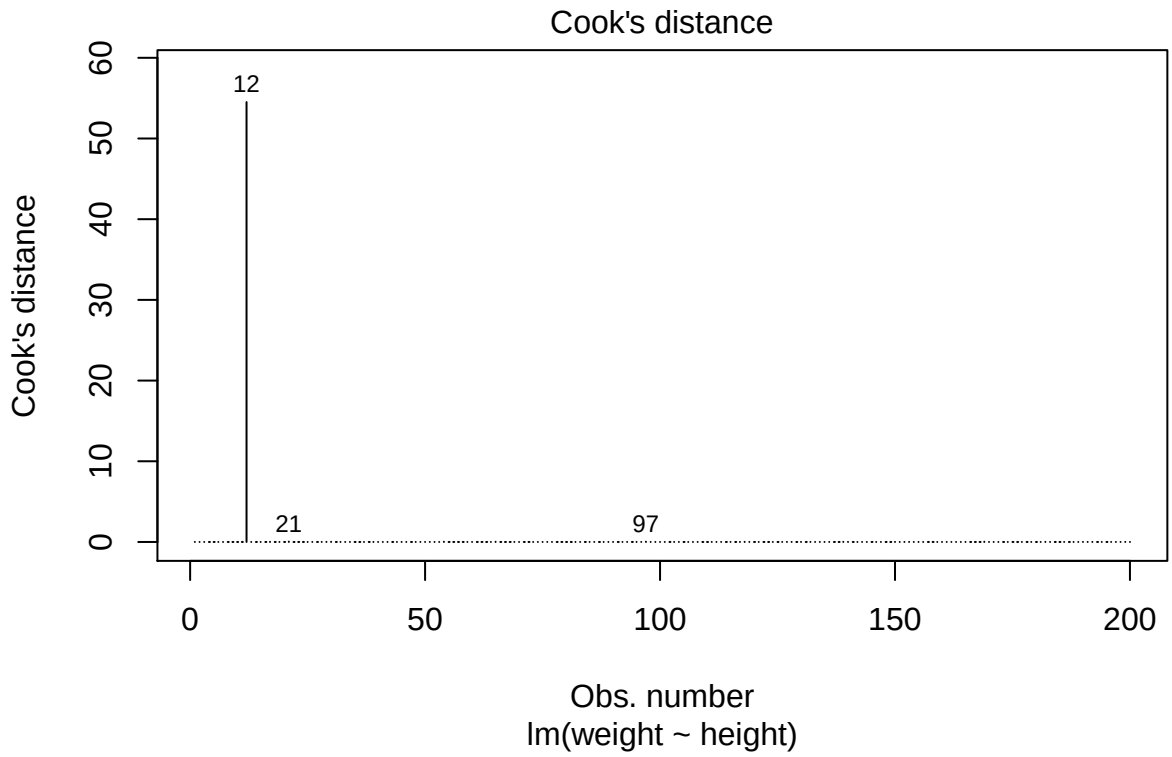
d.

Visually, observation 12 seems to be the only influential outlier since it has high leverage and discrepancy. We will use Cook's distance to measure influence.

```
influence <- cooks.distance(lm_1)
cutoff <- 4/(nrow(davis) - 1 - 1)
cutoff
```

```
## [1] 0.02020202
```

```
plot(lm_1, which=4)
```



According to our graph and values (relative to the cutoff), observation 12 is the only outlier with high influence.

e.

```
lm_1_wo <- lm(weight ~ height, data = davis[-12,])
summary(lm_1)
```

```
##
## Call:
## lm(formula = weight ~ height, data = davis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.696  -9.506  -2.818   6.372 127.145
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.26623   14.95042   1.690  0.09260 .
## height       0.23841    0.08772   2.718  0.00715 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.86 on 198 degrees of freedom
## Multiple R-squared:  0.03597,    Adjusted R-squared:  0.0311
## F-statistic: 7.387 on 1 and 198 DF,  p-value: 0.007152
```

```
summary(lm_1_wo)
```

```
##
## Call:
## lm(formula = weight ~ height, data = davis[-12, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.650  -5.419  -0.576   4.857  42.887
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -130.74698   11.56271  -11.31   <2e-16 ***
## height         1.14922    0.06769   16.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.523 on 197 degrees of freedom
## Multiple R-squared:  0.594,  Adjusted R-squared:  0.592
## F-statistic: 288.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

According to the output from the summary function on our two models (one without the outlier and the other with it), we receive noticeably different coefficient for the model without the outlier (1.15, as opposed to .24). Our R squared is now much larger, being .594 instead of 0.036. This means our model does much better in explaining the variance of our data. The residual standard error is now only 8.523 as opposed to 14.86, so our model is more "stable" and the RSE is lowered significantly by the removal of a single observation.

f.

Based on what was done above, I would suggest removing the observation from our data. It seems like it could have been an entry error (considering how different it is from our normal data) and all of our metrics show that it has a noticeable effect on our model.

2.

a.

No. We now have sales regressed on 3 predictors. To graph this, it would need to be in the 4th dimension... which isn't possible to visualize in an understandable way.

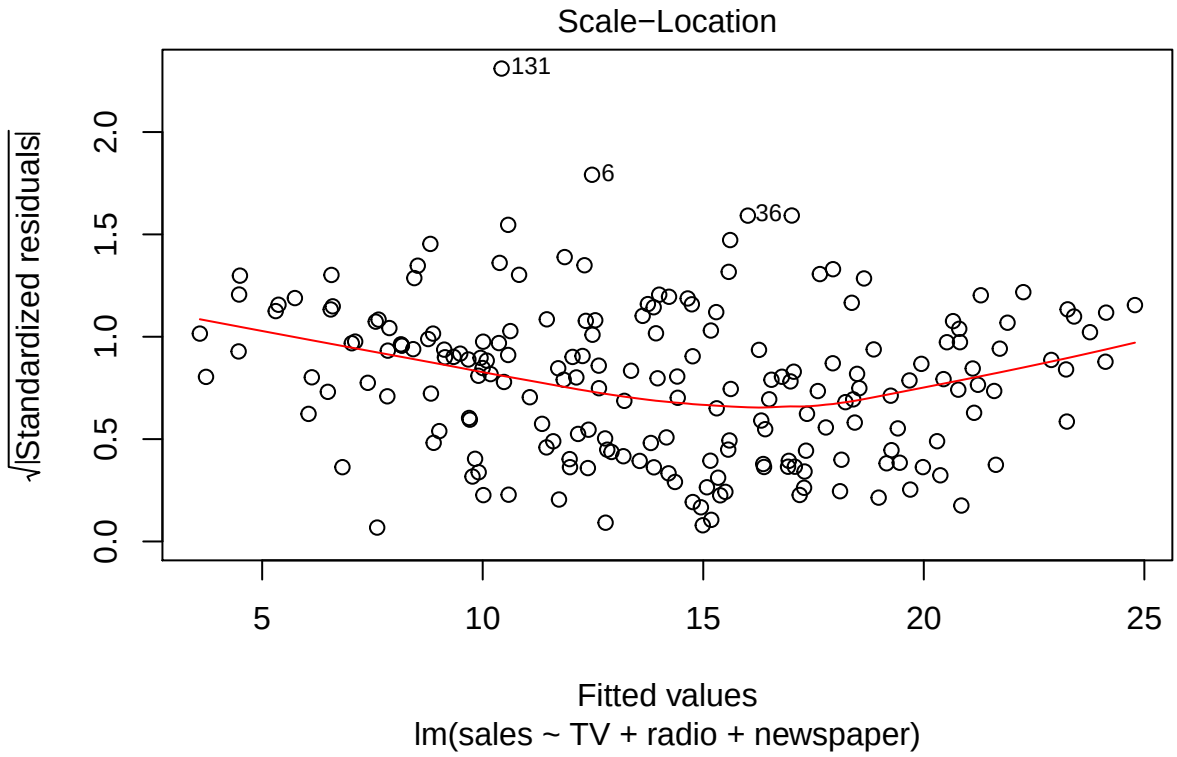
b.

Standardized Residuals

```
lm_2 <- lm(sales ~ TV + radio + newspaper, data = advert)
stand_2 <- rstandard(lm_2)
max(abs(stand_2))
```

```
## [1] 5.336838
```

```
plot(lm_2, which = 3)
```



Observation 131 has the greatest standardized residual of -5.337

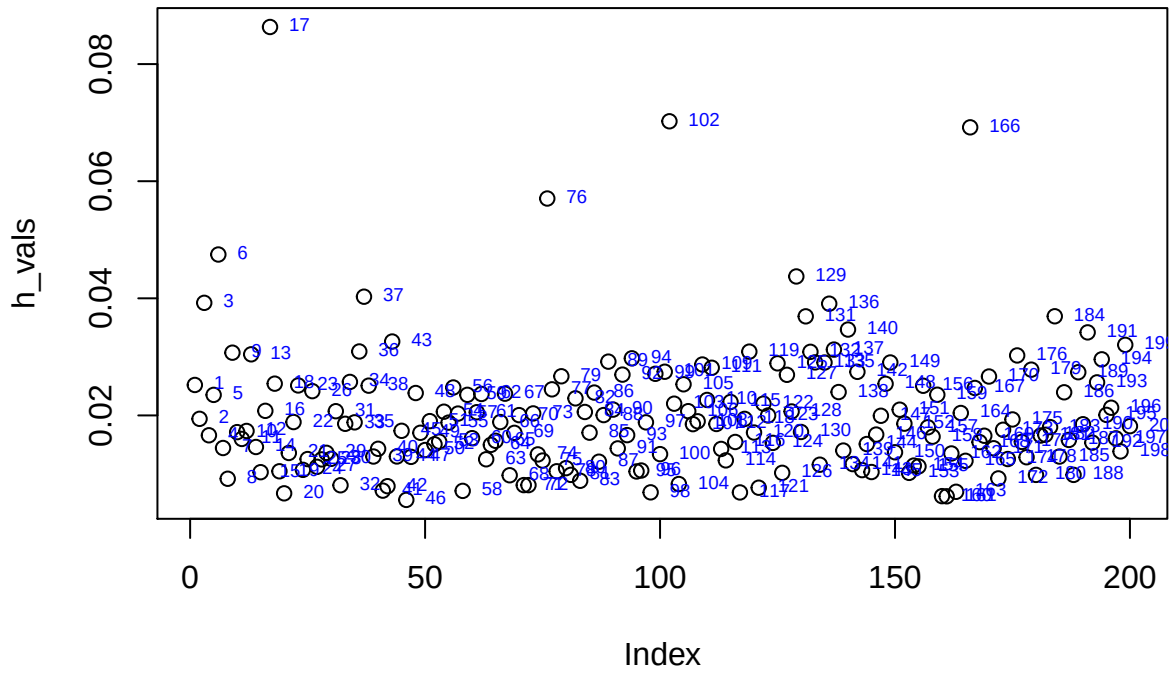
c.

Hat-Values

```
h_vals <- hatvalues(lm_2)
max(h_vals)
```

```
## [1] 0.08633414
```

```
plot(h_vals)
text(1:length(h_vals), h_vals, rownames(advert), cex=0.6, pos=4, col="blue")
```



observation 17 has the gretest hat value, which is 0.0863

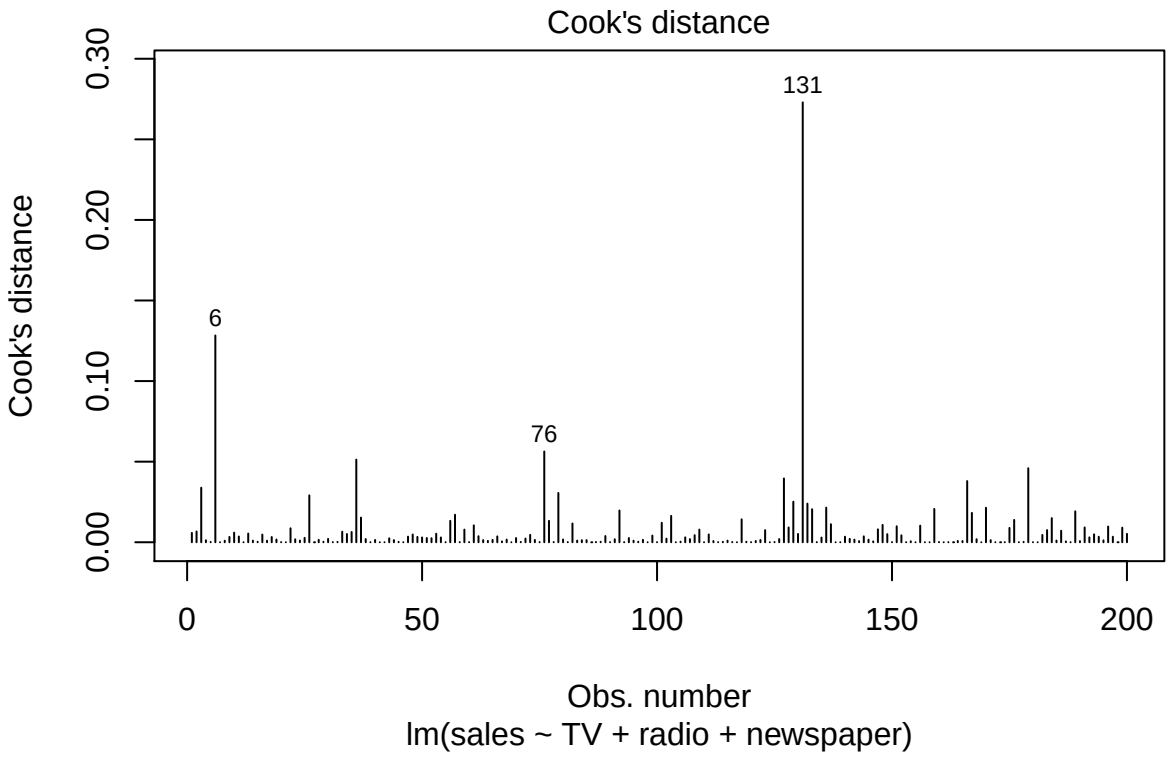
d.

Cook's Distance

```
dist <- cooks.distance(lm_2)
max(dist)
```

```
## [1] 0.2729561
```

```
plot(lm_2, which = 4)
```




```
cutoff <- 4/(200 - 3 - 1)
cutoff
```

```
## [1] 0.02040816
```

Observation 131 has the greatest Cook's distance value of .273.

e.

```
lm_2_wo <- lm(sales ~ TV + radio + newspaper, data = advert[-131,])
summary(lm_2)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = advert)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
```

```
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
summary(lm_2_wo)
```

```
##
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = advert[-131,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4463 -0.9256  0.1983  1.1750  2.7885
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.093065   0.290332  10.654   <2e-16 ***
## TV           0.044845   0.001303  34.425   <2e-16 ***
## radio        0.193905   0.008036  24.130   <2e-16 ***
## newspaper   -0.004252   0.005470  -0.777    0.438
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.562 on 195 degrees of freedom
## Multiple R-squared:  0.9096, Adjusted R-squared:  0.9082
## F-statistic: 653.7 on 3 and 195 DF,  p-value: < 2.2e-16
```

The coefficient estimates are roughly the same (they all retain the same sign as well). There p-values still indicate the same significance as well. The R^2 for the model without observation 131 is .91, which is only marginally better than the model with the outlier. The residual standard error only goes down by about .12 when the observation is removed (from 1.686 to 1.562).

f.

According to our metrics and the outcome of the two models (one without the outlier and one with), I would make the decision to retain the outlier in our data and move forward with making our model with it included. This is because it does not have a noticeable effect on our model (only marginally changing our R^2 when it is removed) and the metrics do not indicate that the outlier has significant leverage.