Lab 13: Multiple Regression

Pace of Life (Adapted from Sleuth3 Exercise 9.14)

We have observations on indicators of pace of life from 36 different metropolitan regions of different sizes throughout the United States:

- `Bank`: bank clerk speed
- `Walk`: pedestrian walking speed
- `Talk`: postal clerk talking speed
- `Heart`: age adjusted death rate due to heart disease

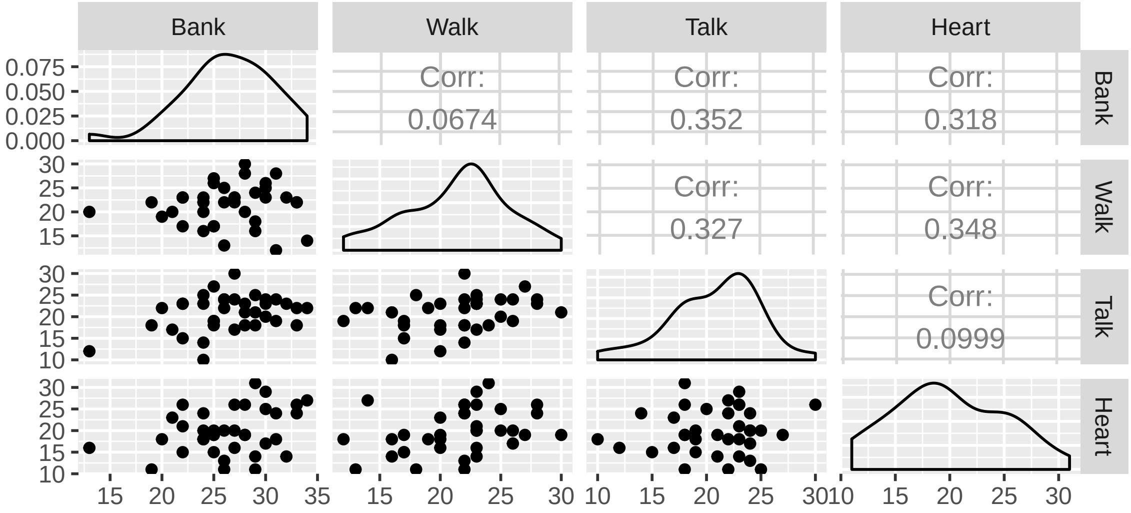
```
pace <- read_csv("http://www.evanlray.com/data/sleuth3/ex0914_pace_of_life.csv")
```

```
## Parsed with column specification:
## cols(
##   Bank = col_integer(),
##   Walk = col_integer(),
##   Talk = col_integer(),
##   Heart = col_integer()
## )
```

Let's model the relationship between the death rate due to heart disease (our response variable) and the other indicators of pace of life.

1. Make a pairs plot of the data.

```
ggpairs(pace)
```



2. Based on your pairs plot, perform an initial check of the conditions of linearity, equal variance, and no outliers/high leverage observations. Do you see any potential causes for concern?

Linearity: There are not strong relationships between any of the explanatory variables and the response, but the relationships are not specifically non-linear. (Looking at the last row of plots, which all have the response variable, Heart, on the vertical axis.)

Equal variance: The vertical spread of the points is fairly consistent across each of the three explanatory variables (looking at the last row of plots.)

High leverage observations: The plots of Bank vs. Walk, Bank vs. Talk, and Bank vs. Heart all point to a high leverage observation with a value of bank that is about 13. None of the other points particularly stand out to me in the plots.

3. Fit a model that has Heart as the response and the other three variables in the data set as explanatory variables. Print a summary of your model fit. Is there any indication of associations between the variables in the model and the rate of deaths due to heart disease?

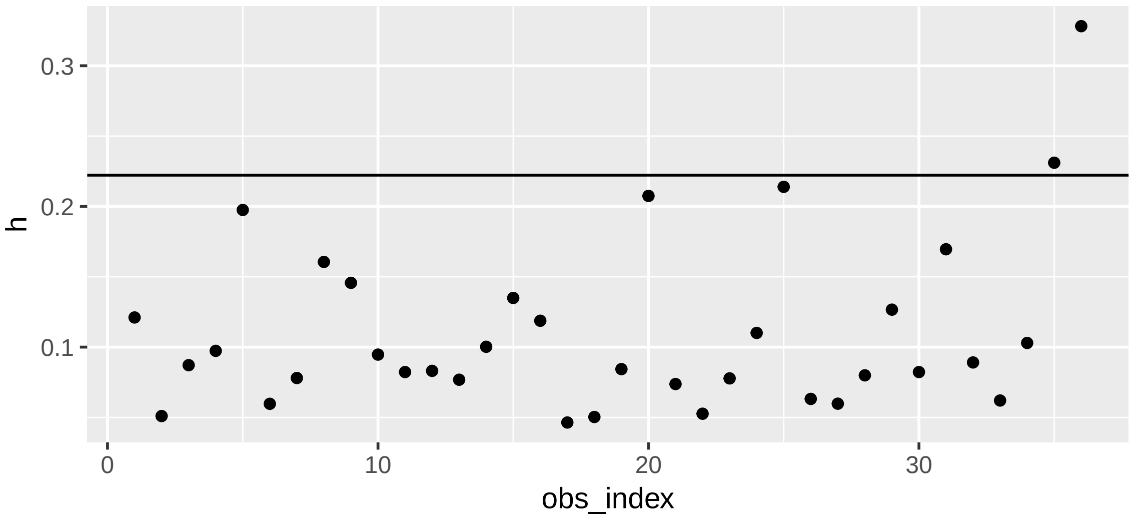
```
lm_fit <- lm(Heart ~ Bank + Walk + Talk, data = pace)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Heart ~ Bank + Walk + Talk, data = pace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4014 -3.0263  0.0602  2.6748  8.4646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1787     6.3369   0.502   0.6194
## Bank          0.4052     0.1971   2.056   0.0480 *
## Walk          0.4516     0.2009   2.248   0.0316 *
## Talk         -0.1796     0.2222  -0.808   0.4249
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.805 on 32 degrees of freedom
## Multiple R-squared:  0.2236, Adjusted R-squared:  0.1509
## F-statistic: 3.073 on 3 and 32 DF,  p-value: 0.04162
```

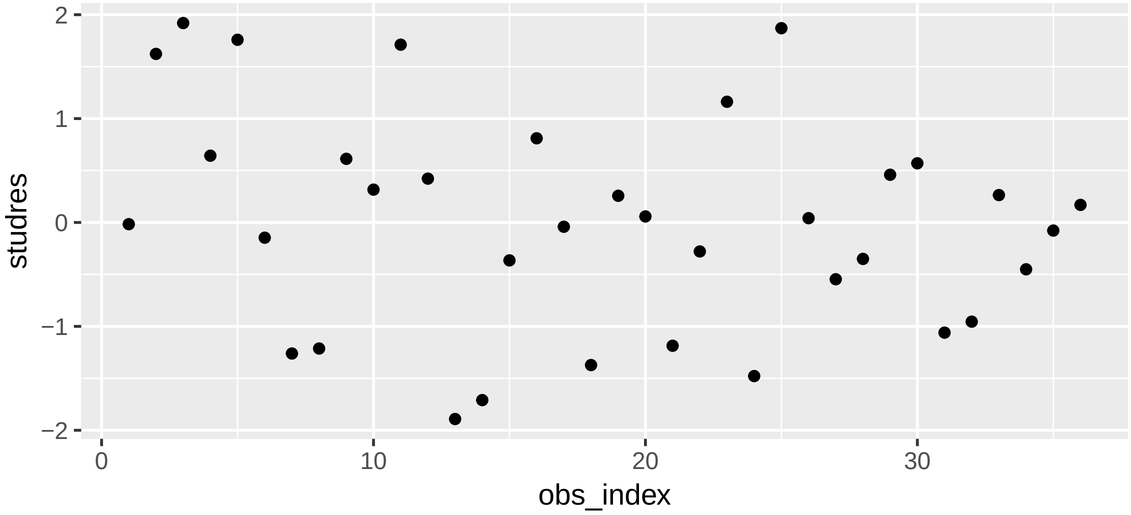
There is a moderate amount of evidence that there is an association between bank clerk speed, walking speed and age adjusted death rate due to heart disease after accounting for talking speed.

4. Make plots showing the leverage, studentized residual, and Cook's distance for each observation. Do these diagnostics suggest that any observations are worth investigating further?

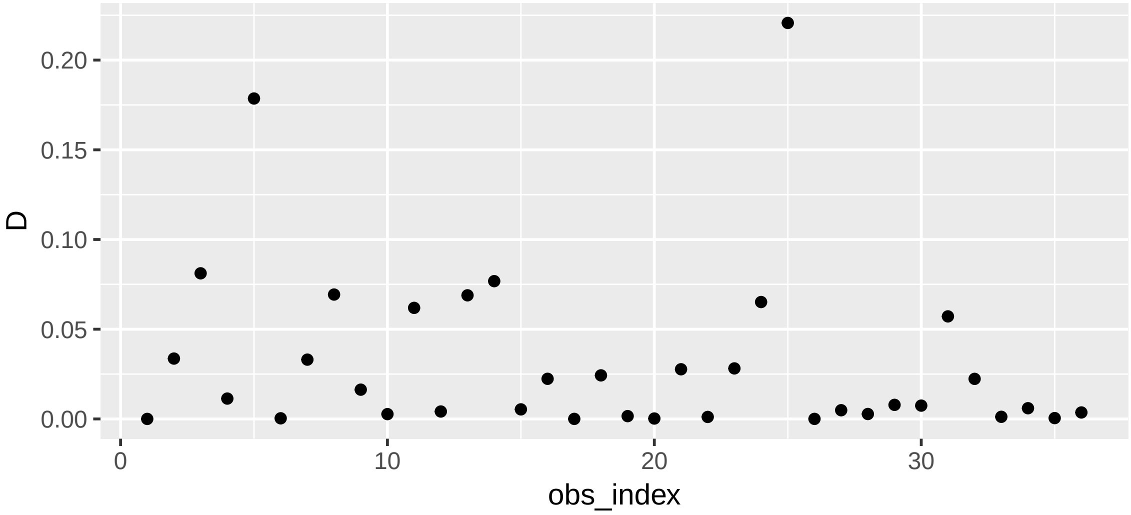
```
pace <- pace %>%
  mutate(
    obs_index = row_number(),
    h = hatvalues(lm_fit),
    studres = rstudent(lm_fit),
    D = cooks.distance(lm_fit)
  )

ggplot(data = pace, mapping = aes(x = obs_index, y = h)) +
  geom_hline(yintercept = 2*4/ nrow(pace))+
  geom_point()
```



```
ggplot(data = pace, mapping = aes(x = obs_index, y = studres)) +
  geom_point()
```



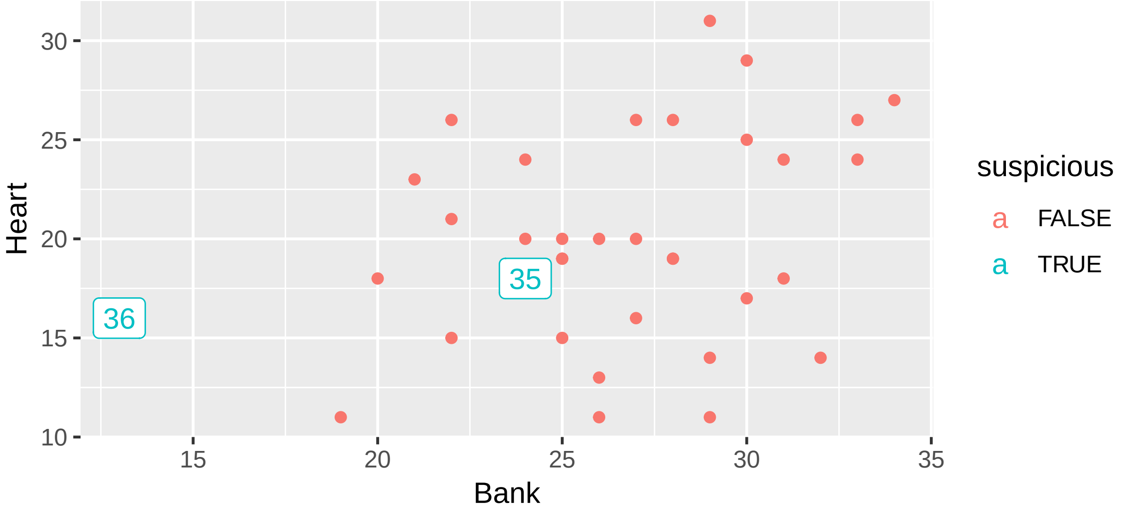
```
ggplot(data = pace, mapping = aes(x = obs_index, y = D)) +
  geom_point()
```



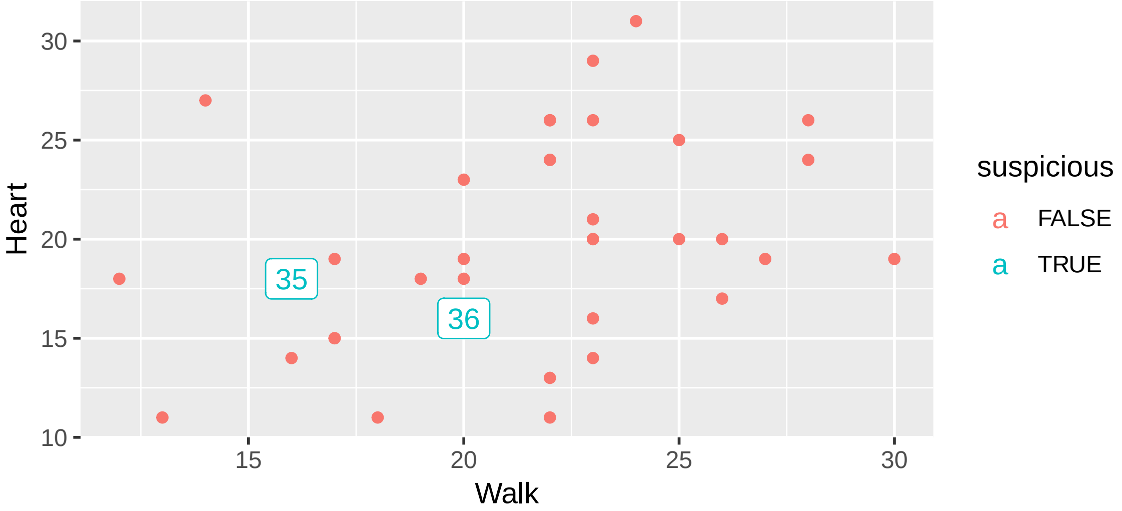
The plot of leverage indicates that we should investigate observations 35 and 36 to see if they are affecting our inferences too much. The plots of studentized residuals and Cook's distances do not indicate any serious problems.

6. Make scatter plots of each quantitative explanatory variable vs. the response, highlighting any observations you identified above as being worth further attention.

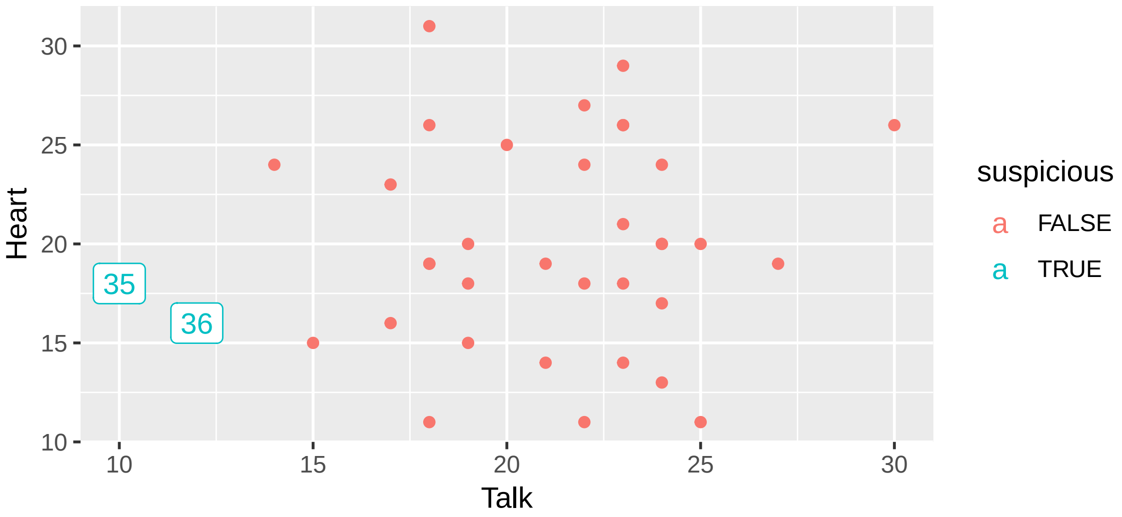
```r
obs_to_investigate <- c(35, 36)

pace <- pace %>%
  mutate(
    suspicious = row_number() %in% obs_to_investigate
  )

ggplot(data = pace, mapping = aes(x = Bank, y = Heart, color = suspicious)) +
  geom_point() +
  geom_label(data = pace %>% filter(suspicious), mapping = aes(label = obs_index))
```



```r
ggplot(data = pace, mapping = aes(x = Walk, y = Heart, color = suspicious)) +
  geom_point() +
  geom_label(data = pace %>% filter(suspicious), mapping = aes(label = obs_index))
```



```r
ggplot(data = pace, mapping = aes(x = Talk, y = Heart, color = suspicious)) +
  geom_point() +
  geom_label(data = pace %>% filter(suspicious), mapping = aes(label = obs_index))
```



7. Create a version of the data set that does not include any suspect observations. Fit the model again, without those suspect observations. Print a summary of your model fit.

```
pace2 <- pace %>%
  filter(
    !suspicious
  )

lm_fit2 <- lm(Heart ~ Bank + Walk + Talk, data = pace2)
summary(lm_fit2)
```

```
##
## Call:
## lm(formula = Heart ~ Bank + Walk + Talk, data = pace2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4811 -3.9503  0.0251  2.7203  8.4580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.6861     8.1503   0.330   0.7440
## Bank          0.4219     0.2276   1.854   0.0736 .
## Walk          0.4492     0.2082   2.158   0.0391 *
## Talk         -0.1755     0.2628  -0.668   0.5095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.96 on 30 degrees of freedom
## Multiple R-squared:  0.2089, Adjusted R-squared:  0.1298
## F-statistic:  2.64 on 3 and 30 DF,  p-value: 0.06749
```

8. How would you sum up what you have learned about the associations between each of the explanatory variables and the response based on this analysis?

After accounting for bank clerk speed and talking speed, the data provide a moderate amount of evidence against the null hypothesis that there is no association between walking speed and the rate of deaths due to heart disease. This effect was present whether or not two high leverage observations were included.

The data also provided some evidence of an association between bank clerk speed and the rate of deaths due to heart disease, after accounting for walking speed and talking speed, but the strength of evidence for this association was reduced after who high leverage observations were removed.

The data did not provide any evidence of an association between talking speed and the rate of deaths due to heart disease after accounting for walking speed and talking speed.

9. What is the interpretation of the coefficient estimate for "Walk" in your model fit?

We estimate that after accounting for the effects of bank clerk speed and talking speed, a 1 unit increase in walking speed is associated with an increase in mean rate of deaths due to heart rate of about 0.45 units, in the population of cities similar to those included in this study.

. . . OR. . . (either is fine)

We estimate that a 1 unit increase in walking speed while holding bank clerk speed and talking speed fixed is associated with an increase in mean rate of deaths due to heart rate of about 0.45 units, in the population of cities similar to those included in this study.