

Worksheet on Multiple Regression and Residual Diagnostics

Solutions

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

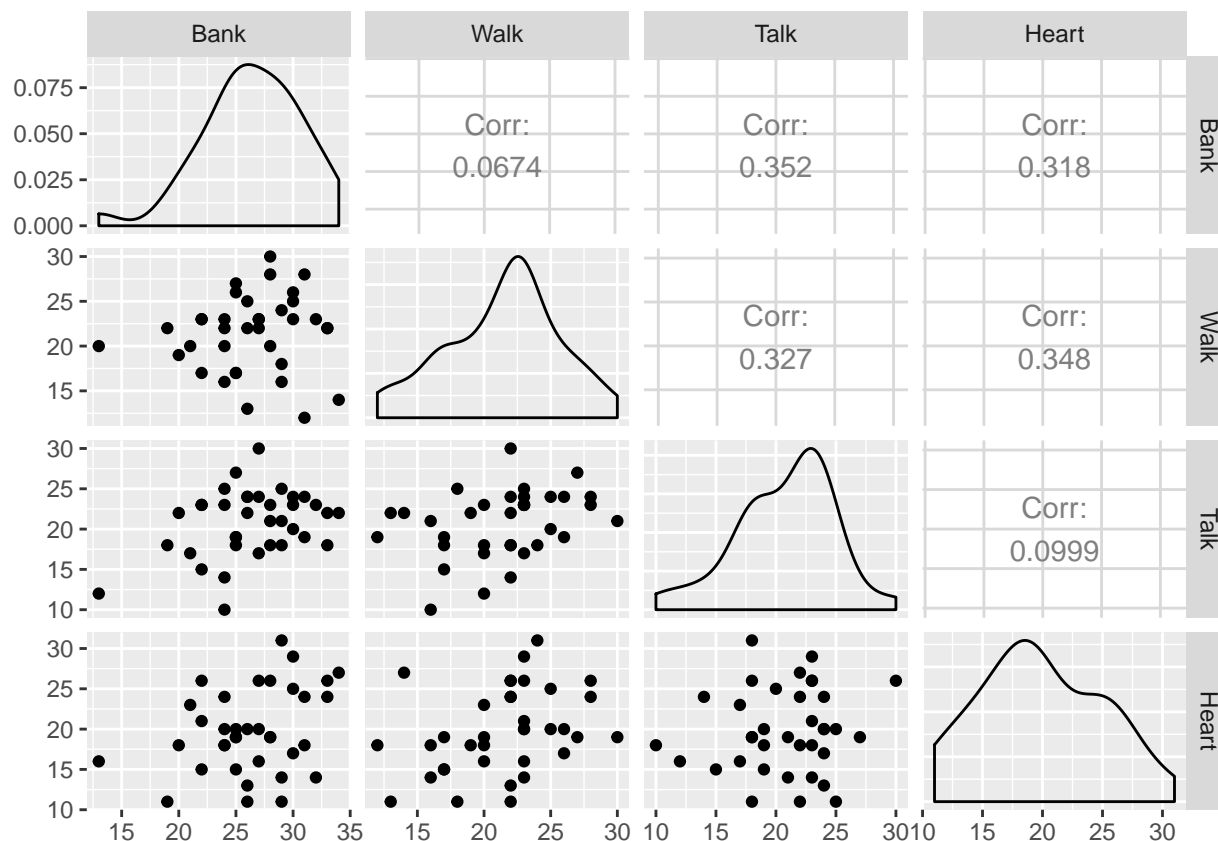
```
pace <- read_csv("http://www.evanlray.com/data/sleuth3/ex0914_pace_of_life.csv")
head(pace)
```

```
## # A tibble: 6 x 4
##   Bank Walk Talk Heart
##   <dbl> <dbl> <dbl> <dbl>
## 1    31    28    24    24
## 2    30    23    23    29
## 3    29    24    18    31
## 4    28    28    23    26
## 5    27    22    30    26
## 6    26    25    24    20
```

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. Here is a pairs plot of the data.

```
ggpairs(pace)
```



2. Based on the 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 appear to be weak linear relationships between the three potential explanatory variables and the response, so linearity appears to be satisfied. Identifying outliers/leverage points may improve this.

Equal variance: This looks pretty reasonable. Focusing on the plots in the last row (versus heart), the points are similarly spread out across the different x values, particularly towards the middle of the data (which is where we expect this behavior - the higher and lower values often have few observations). None of the variables look highly skewed (density plots), which supports the conclusion that equal variance is OK here.

No outliers/high leverage observations: There is a potential leverage point around bank=13 for the plot of heart versus bank; this point looks like it would make the regression coefficient (the slope) less steep. There are two potential leverage points around talk=10 and talk=12 for the plot of heart versus talk; these points would also likely make the regression coefficient less smaller (in magnitude).

3. Here is a summary of a model that has Heart as the response and the other three variables in the data set as explanatory variables. 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
```

If we interpret this question as: is there evidence that *any* of these explanatory variables are associated with the rate of deaths due to heart disease, we should think about an F test here.

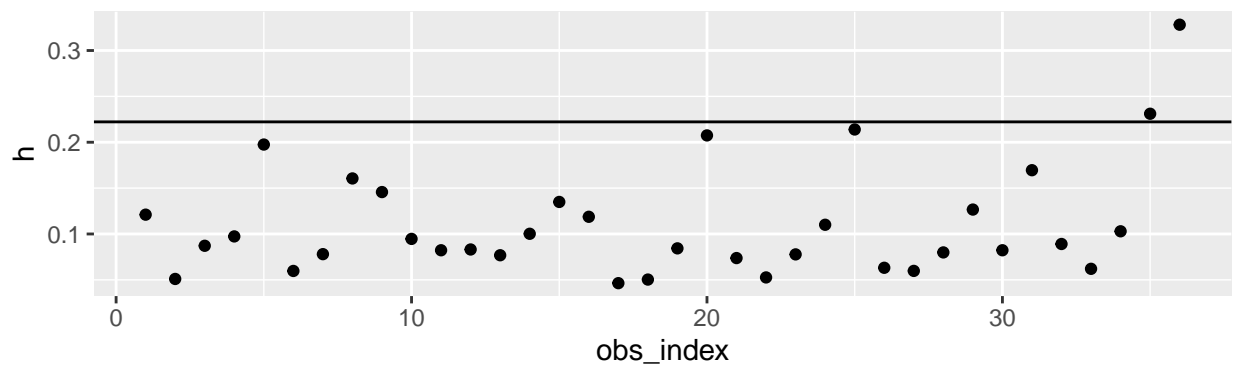
$H_0 : \beta_{bank} = \beta_{walk} = \beta_{talk} = 0$ (the intercept-only model is sufficient) versus $H_A : \text{at least one of } \beta_{bank}, \beta_{walk}, \beta_{talk} \neq 0$ (the intercept-only model is not sufficient)

The F-statistic for this test is 3.073 on 3 and 32 degrees of freedom with a p-value of 0.0416. There is moderate evidence that at least one of these explanatory variables are associated with the rate of deaths due to heart disease.

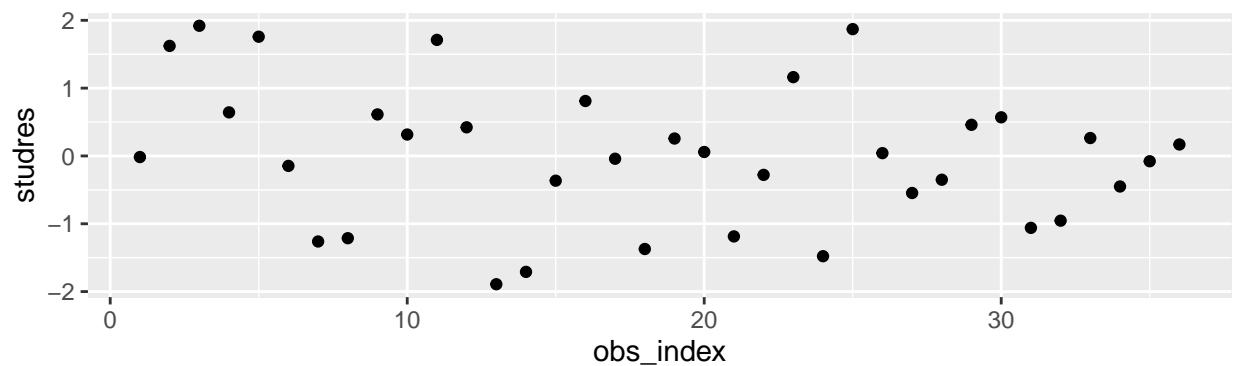
Since we found this moderate evidence, we would continue with individual t -tests for the three coefficients to see if there is evidence that they are different from 0. R does this for us; there is moderate evidence that β_{bank} and β_{walk} are different from 0. We haven't accounted for multiple comparisons here, however, so we should be cautious with this inference.

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

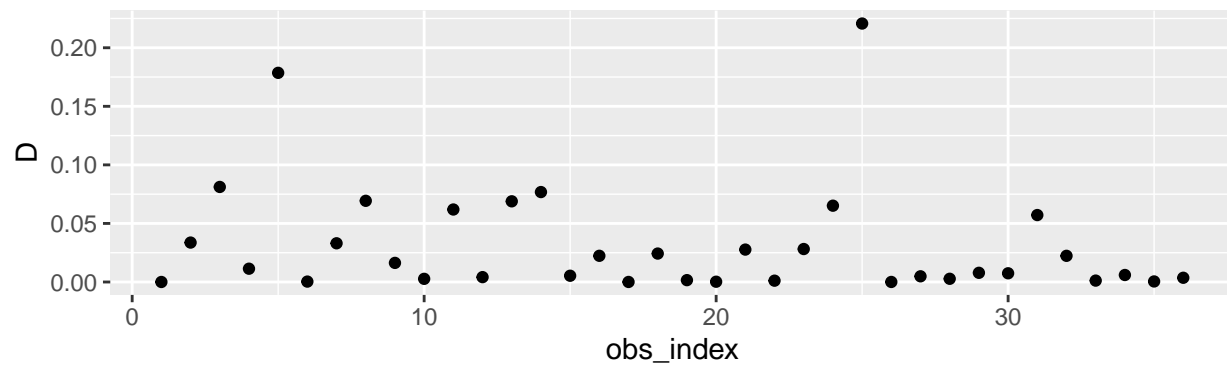
```
pace <- pace %>%  
  mutate(  
    obs_index = row_number(),  
    h = hatvalues(lm_fit),  
    studres = rstudent(lm_fit),  
    D = cooks.distance(lm_fit)  
  )  
ggplot(data = pace, mapping = aes(x = obs_index, y = h)) +  
  geom_hline(yintercept = 2*4/ nrow(pace)) +  
  geom_point()
```



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



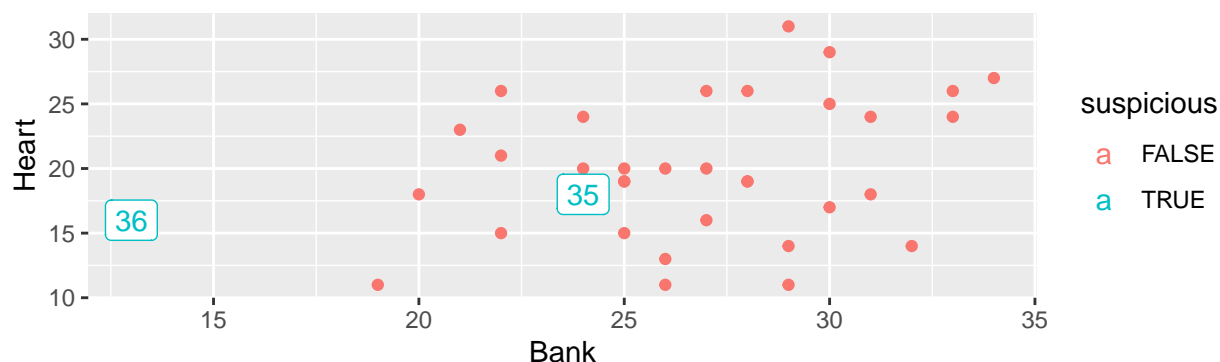
```
ggplot(data = pace, mapping = aes(x = obs_index, y = D)) +  
  geom_point()
```



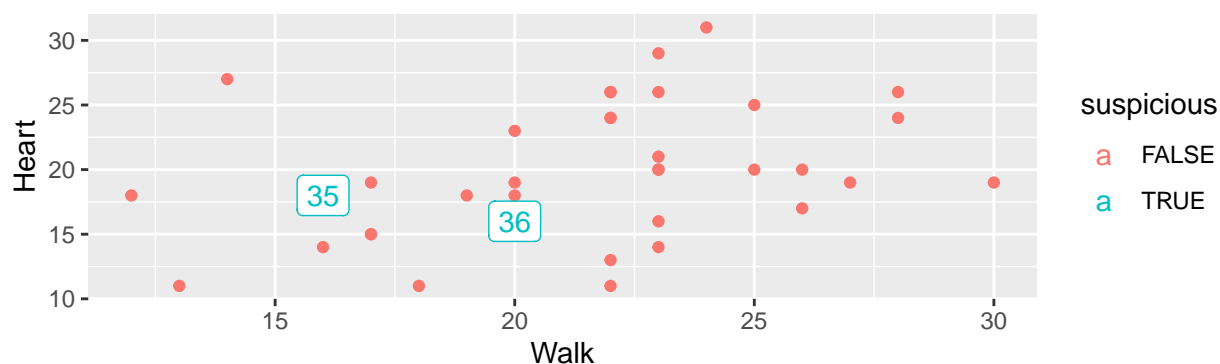
Yes, there are two observations (35 and 36) that have hat values greater than $2 \times p/n$. There are also two values that have larger Cook's distances (5 and 25), but given the scale here, these are actually not that large (recall we had a rule of thumb that we should pay attention to Cook's distance larger than 1).

Here are scatter plots of each quantitative explanatory variable vs. the response, highlighting observation numbers 35 and 36 (these are the two I identified as being worth further attention from the diagnostic plots above).

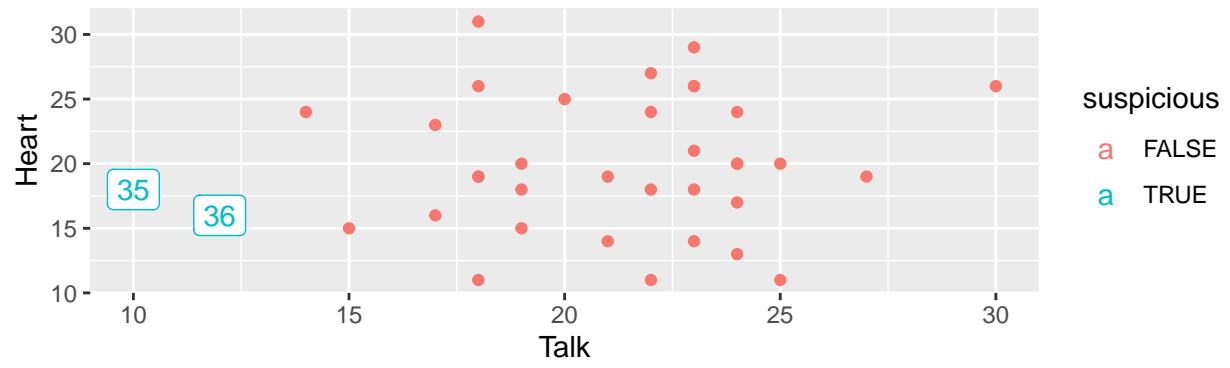
```
obs_to_investigate <- c(35, 36)
pace <- pace %>%
  mutate(
    suspicious = row_number() %in% obs_to_investigate
  )
ggplot(data = pace, mapping = aes(x = Bank, y = Heart, color = suspicious)) +
  geom_point() +
  geom_label(data = pace %>% filter(suspicious), mapping = aes(label = obs_index))
```



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



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



Here is a model fit to a version of the data set that does not include the two suspect observations.

```
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
```

5. What is the interpretation of the coefficient estimate for “Walk” in the model fit without the suspicious observations?

We expect that the mean increase in age adjusted death rate due to heart disease will be 0.449 percent higher in a population in a metropolitan area where pedestrian walking speed is one unit greater than for population in another metropolitan area, assuming these metropolitan areas have the same postal clerk talking speed and the same bank clerk speed (this is equivalent to keeping bank speed and talk speed constant).

6. 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?

The finding of moderate evidence that at least one of these three explanatory variables was related to the response (done in 3.) was dependent on the details of the analysis. We see that excluding observations 35 and 36 results in an F test with a p-value of 0.067. For the individual *t*-tests, we had to consider multiple comparisons, so this evidence was not particularly compelling, either. Generally, I would say that there is weak evidence that these explanatory variables are associated with the (increased) age adjusted rate due to heart disease. This is encouraging for those of us who live in cities!