

Multiple regression

Sunday, April 5, 2020 4:49 PM



mult_reg

Chapter 11: Multiple Regression, Pairs Plots

Duncan's Occupational Prestige Data

Intro to data

We have a data set with measurements on 45 different U.S. occupations as of 1950 (descriptions from Fox and Weisberg, 2011):

- **type**: Type of occupation. A factor with the following levels: **prof**, professional and managerial; **wc**, white-collar; **bc**, blue-collar.
- **income**: Percentage of occupational incumbents in the 1950 US Census who earned \$3,500 or more per year (about \$37,500 in 2019 US dollars).
- **education**: Percentage of occupational incumbents in 1950 who were high school graduates (this might be more like having a college degree in current times?)
- **prestige**: Percentage of respondents in a social survey who rated the occupation as “good” or better in prestige

```
head(Duncan, 3)
```

```
##           type income education prestige occupation
## accountant prof     62         86      82 accountant
## pilot      prof     72         76      83      pilot
## architect  prof     75         92      90  architect
```

References:

- Fox, J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition, Sage.
- Duncan, O. D. (1961) A socioeconomic index for all occupations. In Reiss, A. J., Jr. (Ed.) Occupations and Social Status. Free Press [Table VI-1].

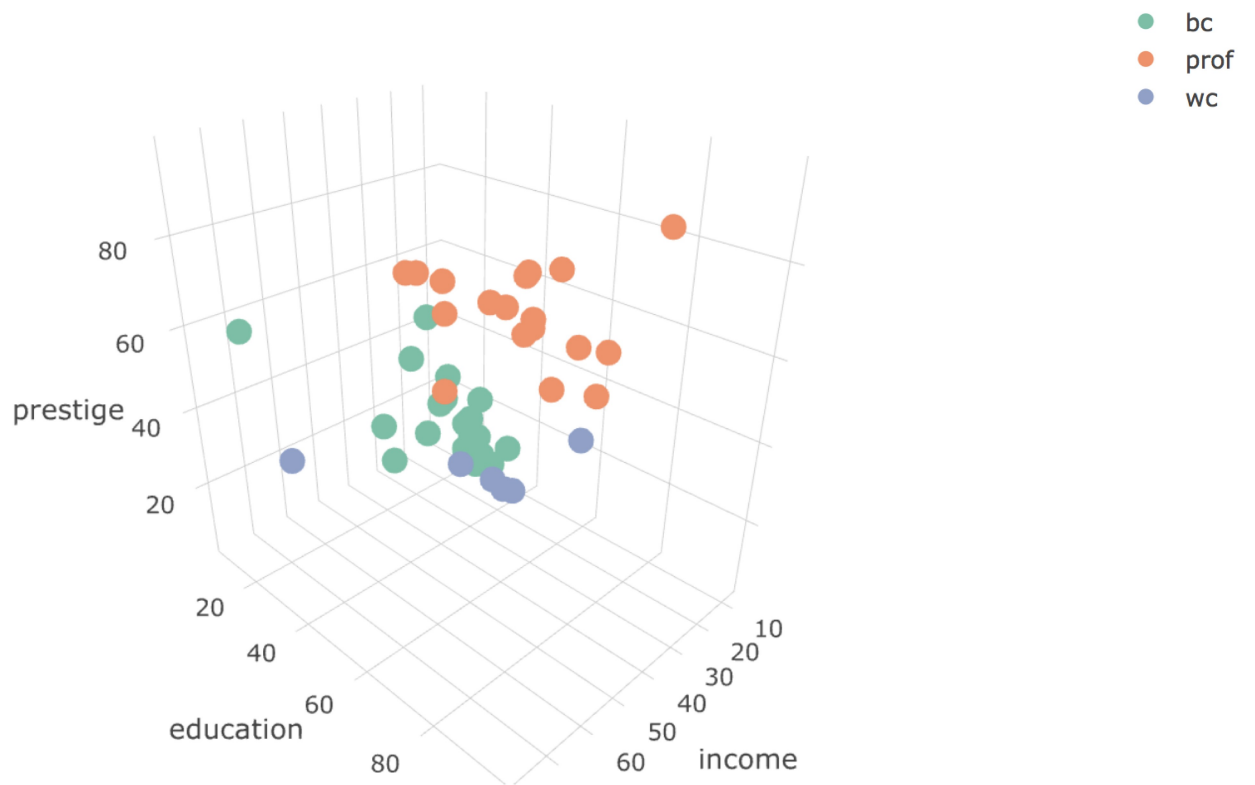
Let's consider a model for occupational prestige as a function of income, education, and type of occupation.

Option 1 for Plots: plotly

- Formatting very similar to, but not exactly the same as, ggplot2
- **Can't show output in pdf, only for html output or interactive use**
- Can't be used for any more variables than we have in this example.
- If plotly code doesn't give you what you want right away, it can be essentially impossible to fix (not a fully developed and functional package).

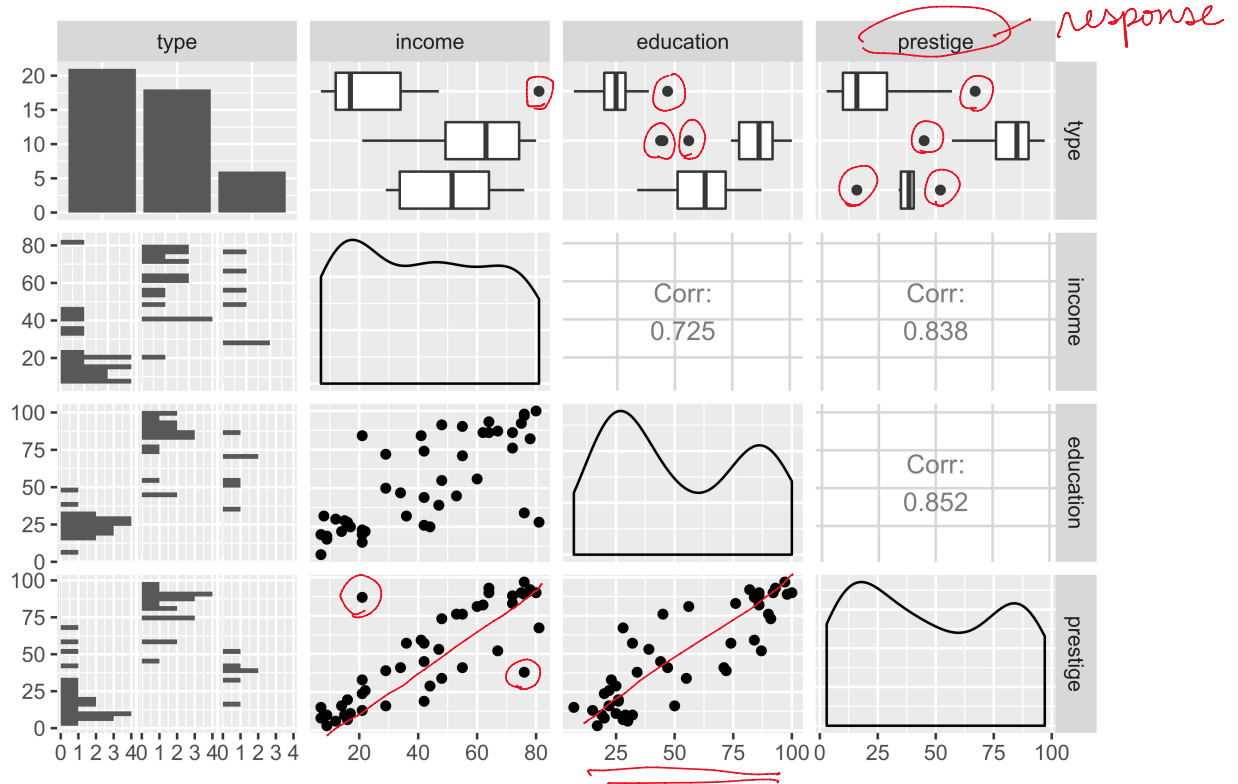
```
library(plotly)
plot_ly(Duncan, x = ~income, y = ~education, z = ~prestige, color = ~type) %>%
  add_markers()
```

Here's a screenshot, will demo live:



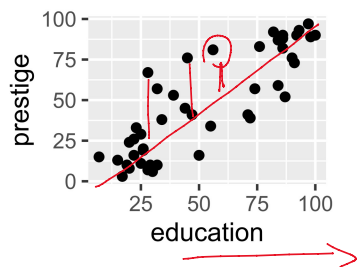
Option 2 for Plots: Pairs Plots

```
library(GGally) # contains the ggpairs function
# I like to plot only variables I'm interested in at the moment, with the response last
ggpairs(Duncan %>% select(type, income, education, prestige))
```



Compare the plot in the third column and fourth row to the following:

```
ggplot(data = Duncan, mapping = aes(x = education, y = prestige)) +
  geom_point()
```



Is there any evidence of outliers or influential observations?

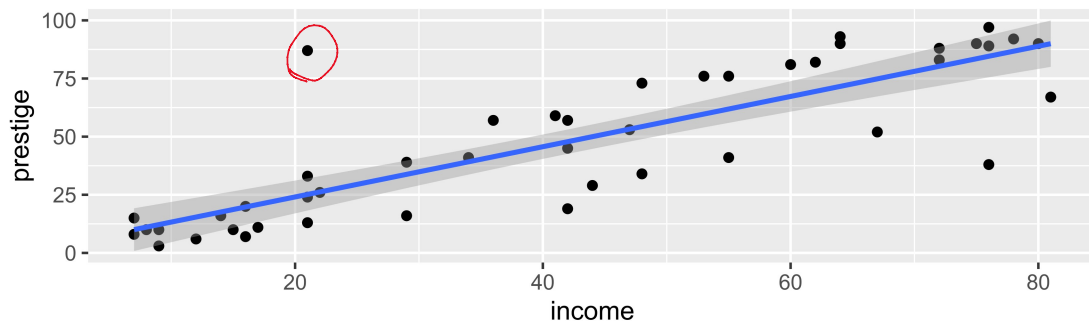
yes, there are potential outliers (in the Y direction).

A first model - income only explanatory variable

```
lm_fit_1 <- lm(prestige ~ income, data = Duncan)
summary(lm_fit_1)
```

```
##
## Call:
## lm(formula = prestige ~ income, data = Duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -46.566  -9.421   0.257   9.167  61.855
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.4566     5.1901   0.473   0.638
## income         1.0804     0.1074  10.062 7.14e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.4 on 43 degrees of freedom
## Multiple R-squared:  0.7019, Adjusted R-squared:  0.695
## F-statistic: 101.3 on 1 and 43 DF, p-value: 7.144e-13
```

```
ggplot(data = Duncan, mapping = aes(x = income, y = prestige)) +
  geom_point() +
  geom_smooth(method = "lm")
```



What is the equation of the estimated line?

$$\hat{\mu}(Y | X = \text{income}) = 2.457 + 1.08 (\text{income})$$

What is the interpretation of the coefficient estimate for income?

For a one percent increase in income, we expect a 1.08 percent increase in mean prestige in a population similar to the one in the survey (for US occupations in the 1950s).

Second Model: income and education as explanatory variables

```
lm_fit_2 <- lm(prestige ~ income + education, data = Duncan)
summary(lm_fit_2)
```

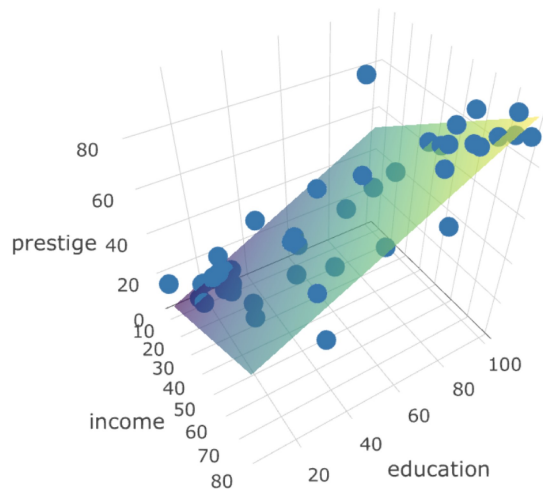
```
##
## Call:
## lm(formula = prestige ~ income + education, data = Duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.538  -6.417   0.655   6.605  34.641
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.06466    4.27194  -1.420   0.163
## income         0.59873    0.11967   5.003 1.05e-05 ***
## education     0.54583    0.09825   5.555 1.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.37 on 42 degrees of freedom
## Multiple R-squared:  0.8282, Adjusted R-squared:  0.82
## F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

What's the estimated equation for the mean from this model?

$$\hat{\mu}(Y | X_1 = \text{income}, X_2 = \text{education}) = -6.065 + 0.599 \text{ income} + 0.546 \text{ edu.}$$

This can be visualized as a plane

Plotly code suppressed because it's awful.

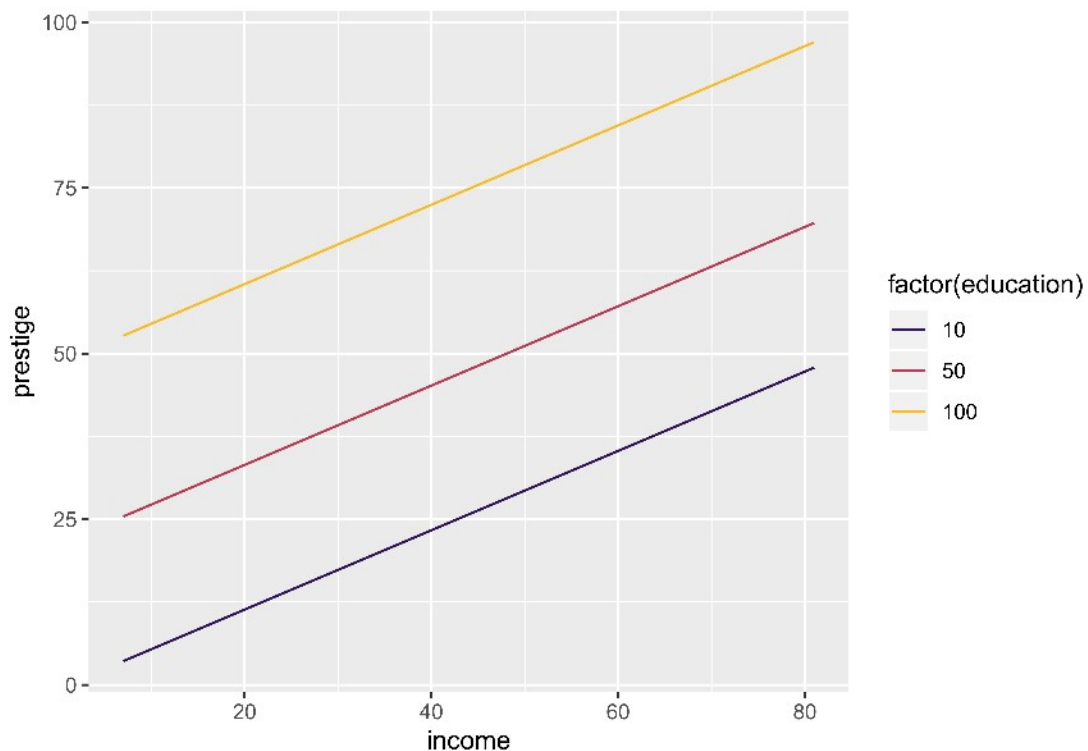


What is the interpretation of the coefficient estimate for income?

$$\hat{\beta}_{\text{inc}} = 0.599.$$

For a 1 percent increase in income, we expect a 0.599 percent increase in prestige for the same education (while education is held constant), for a population of 45 professions in the US in the 1950s.

Here is a plot showing the estimated relationship between income and prestige, for three different values of education:

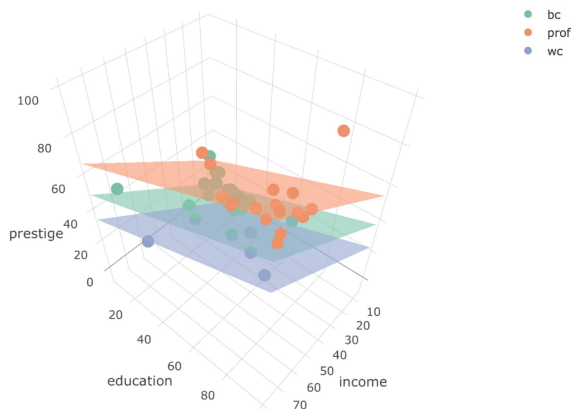


Third Model: All 3 explanatory variables!

```
lm_fit_3 <- lm(prestige ~ income + education + type, data = Duncan)
summary(lm_fit_3)

##
## Call:
## lm(formula = prestige ~ income + education + type, data = Duncan)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.890  -5.740  -1.754   5.442  28.972
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.18503     3.71377  -0.050  0.96051
## income         0.59755     0.08936   6.687 5.12e-08 ***
## education     0.34532     0.11361   3.040  0.00416 **
## typeprof     16.65751     6.99301   2.382  0.02206 *
## typewc      -14.66113     6.10877  -2.400  0.02114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.744 on 40 degrees of freedom
## Multiple R-squared:  0.9131, Adjusted R-squared:  0.9044
## F-statistic: 105 on 4 and 40 DF, p-value: < 2.2e-16
```

Plotly code suppressed because it's awful.



What is the estimated equation for the mean from this model fit?

$$\hat{\mu}(Y|X_1 = \text{income}, X_2 = \text{Edu.}, X_3 = \text{prof}) = -0.185 + 0.598(\text{income}) + 0.345(\text{education}) + 16.658(\text{prof}) - 14.661(\text{wc})$$

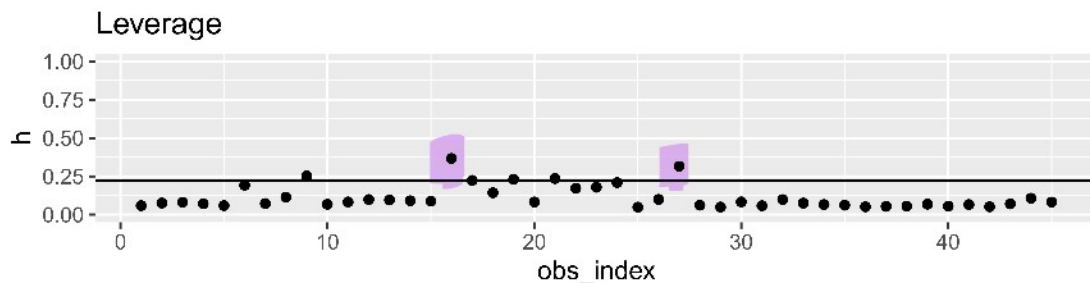
What is the interpretation of the estimated coefficient for income?

For a 1 percent increase in income we expect a 0.598 percent increase in prestige, while holding education constant and maintaining

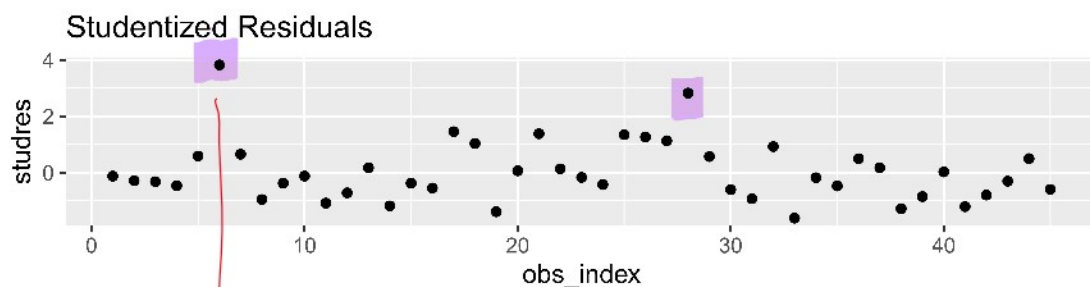
exp. -- ...
while holding education constant and maintaining
the same type of profession. ... in a pop ...

Diagnostic Plots

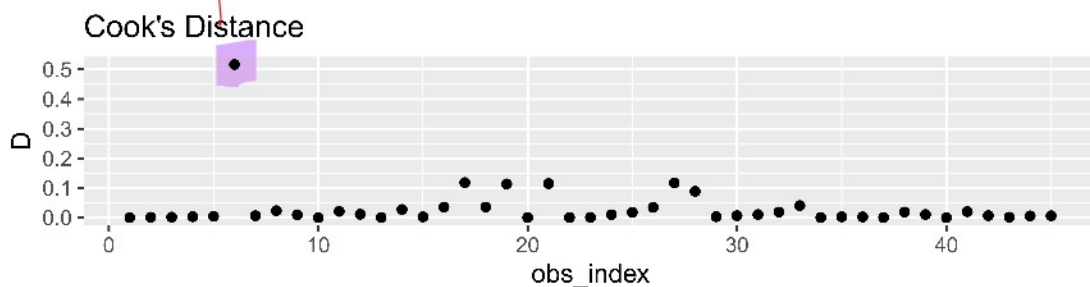
```
Duncan <- Duncan %>%  
  mutate(  
    obs_index = row_number(),  
    h = hatvalues(lm_fit_3),  
    studres = rstudent(lm_fit_3),  
    D = cooks.distance(lm_fit_3)  
  )  
ggplot(data = Duncan, mapping = aes(x = obs_index, y = h)) +  
  geom_point() +  
  geom_hline(yintercept = 2 * 5 / nrow(Duncan)) +  
  ylim(0, 1) +  
  ggtitle("Leverage")
```



```
ggplot(data = Duncan, mapping = aes(x = obs_index, y = studres)) +  
  geom_point() +  
  ggtitle("Studentized Residuals")
```



```
ggplot(data = Duncan, mapping = aes(x = obs_index, y = D)) +  
  geom_point() +  
  ggtitle("Cook's Distance")
```



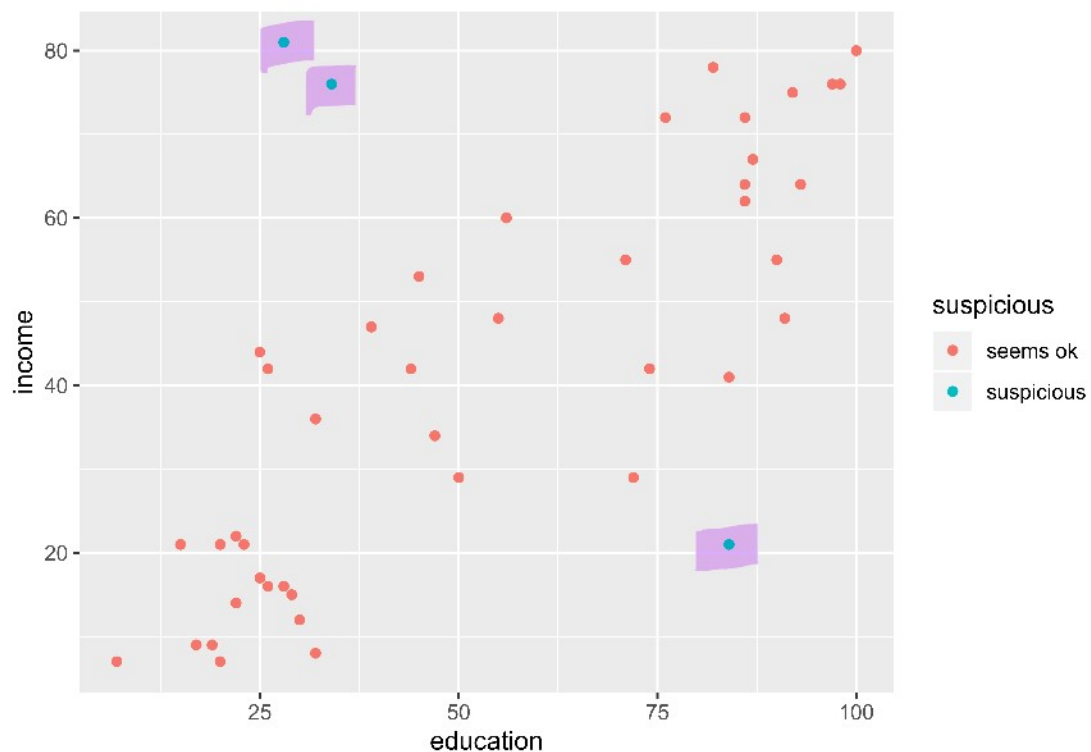


```
obs_to_investigate <- c(6, 16, 27)
```

```
Duncan[obs_to_investigate, ]
```

```
##      type income education prestige occupation obs_index      h      studres      D
## 6  prof     21      84      87    minister         6 0.1912053  3.8293960 0.51680533
## 16  wc      76      34      38   conductor        16 0.3663519 -0.5505711 0.03567303
## 27  bc      81      28      67 RR.engineer        27 0.3146829  1.1339763 0.11725367
```

```
Duncan <- Duncan %>%
  mutate(
    suspicious = ifelse(row_number() %in% obs_to_investigate, "suspicious", "seems ok")
  )
ggplot(data = Duncan, mapping = aes(x = education, y = income, color = suspicious)) +
  geom_point()
```



```
Duncan_minus_suspicious <- Duncan[-obs_to_investigate, ]
lm_fit_without_suspicious <- lm(prestige ~ income + education + type, data = Duncan_minus_suspicious)
summary(lm_fit_without_suspicious)
```

```
##
## Call:
## lm(formula = prestige ~ income + education + type, data = Duncan_minus_suspicious)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-18.0415	-5.3802	-0.6189	5.0992	23.2906

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.1053	3.2745	-0.338	0.7376
income	0.7733	0.1171	6.607	9.53e-08 ***
education	0.2180	0.1174	1.857	0.0714 .
typeprof	15.2512	6.4123	2.378	0.0227 *
typewc	-12.3622	5.9478	-2.078	0.0447 *

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.432 on 37 degrees of freedom
## Multiple R-squared:  0.9368, Adjusted R-squared:  0.93
## F-statistic: 137.1 on 4 and 37 DF,  p-value: < 2.2e-16
```

```
Duncan_minus_minister <- Duncan[-6, ]
lm_fit_without_minister <- lm(prestige ~ income + education + type, data = Duncan_minus_minister)
summary(lm_fit_without_minister)
```

```
##
## Call:
## lm(formula = prestige ~ income + education + type, data = Duncan_minus_minister)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-17.0521	-6.4105	-0.7819	4.6552	23.5212

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.62984	3.22841	-0.505	0.61651
income	0.71813	0.08332	8.619	1.44e-10 ***
education	0.28924	0.09917	2.917	0.00584 **
typeprof	13.43111	6.09592	2.203	0.03355 *
typewc	-15.87744	5.28357	-3.005	0.00462 **

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.413 on 39 degrees of freedom
## Multiple R-squared:  0.9344, Adjusted R-squared:  0.9277
## F-statistic: 139 on 4 and 39 DF,  p-value: < 2.2e-16
```

What do we say?