Duncan Data Analysis

J. Fox and S. Weisberg

2018-04-24

This R Markdown file reproduces the data analysis described in Section 1.5 of *An R Companion to Applied Linear Regression, 3rd Edition*. See the text for further explanation.

# Preliminaries

As a first step, we load the **car** package, and examine the first few rows of the Duncan data:

library("car") # load car and carData packages

## Loading required package: carData

head(Duncan, n=10)

## type income education prestige  
## accountant prof 62 86 82  
## pilot prof 72 76 83  
## architect prof 75 92 90  
## author prof 55 90 76  
## chemist prof 64 86 90  
## minister prof 21 84 87  
## professor prof 64 93 93  
## dentist prof 80 100 90  
## reporter wc 67 87 52  
## engineer prof 72 86 88

dim(Duncan)

## [1] 45 4

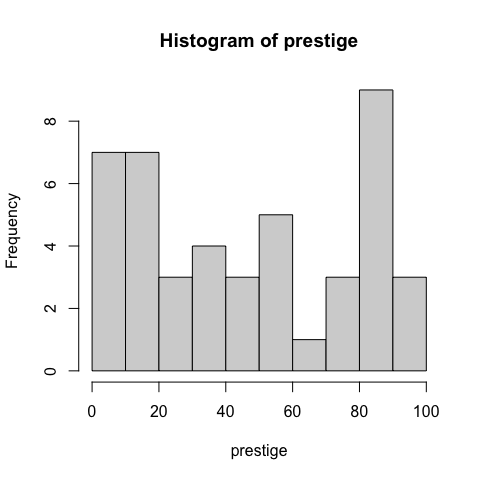
Obtain summary statistics for the variables in Duncan:

summary(Duncan)

## type income education prestige   
## bc :21 Min. : 7.00 Min. : 7.00 Min. : 3.00   
## prof:18 1st Qu.:21.00 1st Qu.: 26.00 1st Qu.:16.00   
## wc : 6 Median :42.00 Median : 45.00 Median :41.00   
## Mean :41.87 Mean : 52.56 Mean :47.69   
## 3rd Qu.:64.00 3rd Qu.: 84.00 3rd Qu.:81.00   
## Max. :81.00 Max. :100.00 Max. :97.00

As a first graph, we view a histogram of the variable prestige:

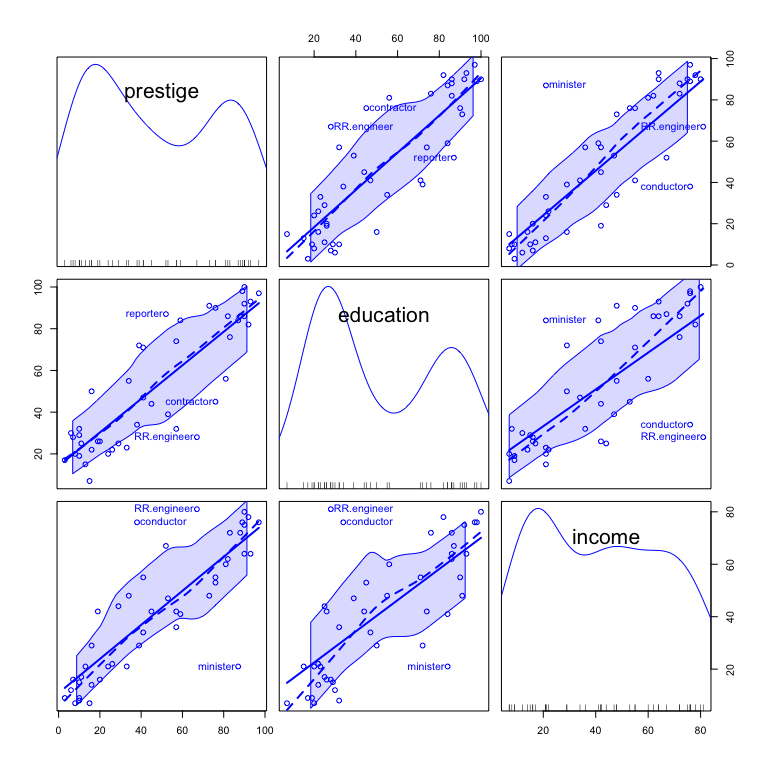
with(Duncan, hist(prestige))



## 1.5.1 Examining the Data

The scatterplotMatrix() function in the **car** package produces scatterplots for all paris of variables. A few relatively remote points are marked by case names, in this instance by occupation.

scatterplotMatrix( ~ prestige + education + income,   
 id=list(n=3), data=Duncan)



## 1.5.2 Regression Analysis

We use thelm() function to fit a linear regression model to the data:

(duncan.model <- lm(prestige ~ education + income, data=Duncan))

##   
## Call:  
## lm(formula = prestige ~ education + income, data = Duncan)  
##   
## Coefficients:  
## (Intercept) education income   
## -6.0647 0.5458 0.5987

summary(duncan.model)

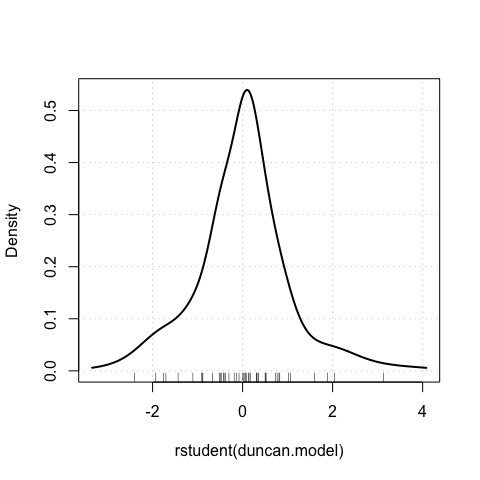
##   
## Call:  
## lm(formula = prestige ~ education + income, data = Duncan)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -29.538 -6.417 0.655 6.605 34.641   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.06466 4.27194 -1.420 0.163   
## education 0.54583 0.09825 5.555 1.73e-06 \*\*\*  
## income 0.59873 0.11967 5.003 1.05e-05 \*\*\*  
## ---  
## 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

The brief() and S() functions, both in the **car** package, provide alternative summaries of a regression fit.

## 1.5.3 Regression Diagnostics

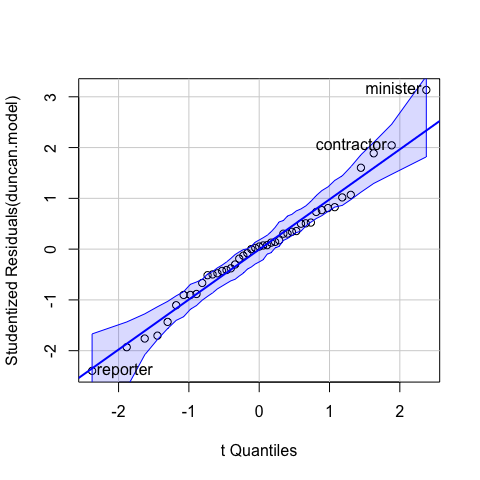
The rstudent() function returns studentized residuals, and the densityPlot() function fits an adaptive kernel density estimator to the distribution of the studentized residuals:

densityPlot(rstudent(duncan.model))



A qqPlot() can be used as a check for nonnormal errors, comparing the studentized residuals to a t-distribution:

qqPlot(duncan.model, id=list(n=3))



## minister reporter contractor   
## 6 9 17

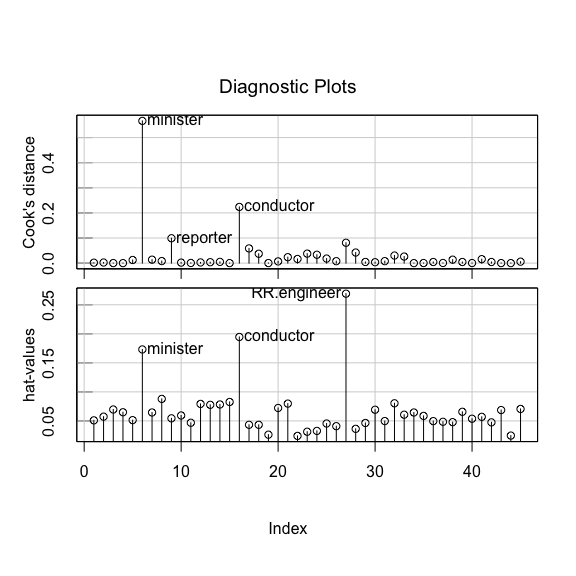
This next function tests for outliers in the regression:

outlierTest(duncan.model)

## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## minister 3.134519 0.0031772 0.14297

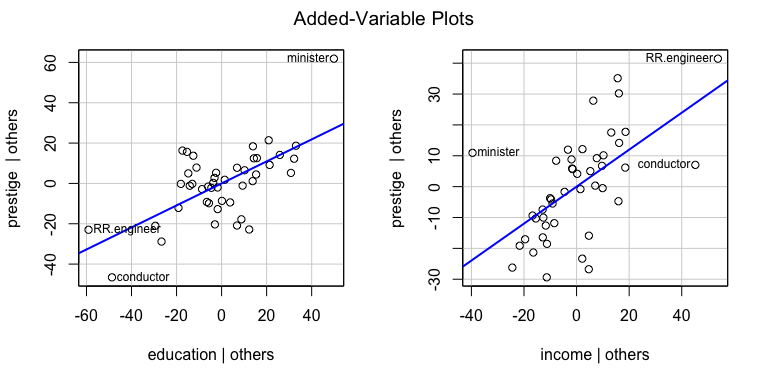
This graph displays influence measures in index plots:

influenceIndexPlot(duncan.model, vars=c("Cook", "hat"),   
 id=list(n=3))



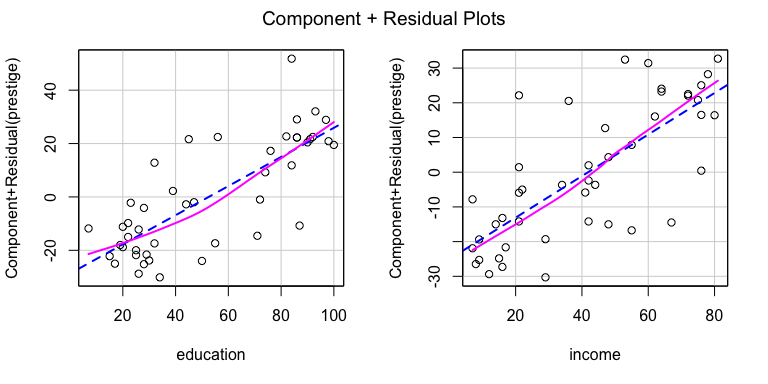
Added-variable plots for Duncan’s regression, looking for influential cases:

avPlots(duncan.model,   
 id=list(cex=0.75, n=3, method="mahal"))



Component-plus-residual plots for the regression, checking for nonlinearity:

crPlots(duncan.model, smooth=list(span=0.7))



Tests for non-constant error variance:

ncvTest(duncan.model)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.3810967, Df = 1, p = 0.53702

ncvTest(duncan.model, var.formula= ~ income + education)

## Non-constant Variance Score Test   
## Variance formula: ~ income + education   
## Chisquare = 0.6976023, Df = 2, p = 0.70553

Removing the cases "minister" and “`conductor’”:

whichNames(c("minister", "conductor"), Duncan)

## minister conductor   
## 6 16

duncan.model.2 <- update(duncan.model, subset=-c(6, 16))  
summary(duncan.model.2)

##   
## Call:  
## lm(formula = prestige ~ education + income, data = Duncan, subset = -c(6,   
## 16))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -28.612 -5.898 1.937 5.616 21.551   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.40899 3.65263 -1.755 0.0870 .   
## education 0.33224 0.09875 3.364 0.0017 \*\*   
## income 0.86740 0.12198 7.111 1.31e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.42 on 40 degrees of freedom  
## Multiple R-squared: 0.876, Adjusted R-squared: 0.8698   
## F-statistic: 141.3 on 2 and 40 DF, p-value: < 2.2e-16

Comparing the regressions with and without these two cases:

compareCoefs(duncan.model, duncan.model.2)

## Calls:  
## 1: lm(formula = prestige ~ education + income, data = Duncan)  
## 2: lm(formula = prestige ~ education + income, data = Duncan, subset = -c(6,  
## 16))  
##   
## Model 1 Model 2  
## (Intercept) -6.06 -6.41  
## SE 4.27 3.65  
##   
## education 0.5458 0.3322  
## SE 0.0983 0.0987  
##   
## income 0.599 0.867  
## SE 0.120 0.122  
##