Multiple Regression - Model Selection

Duncan's Occupational Prestige Data

```
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].

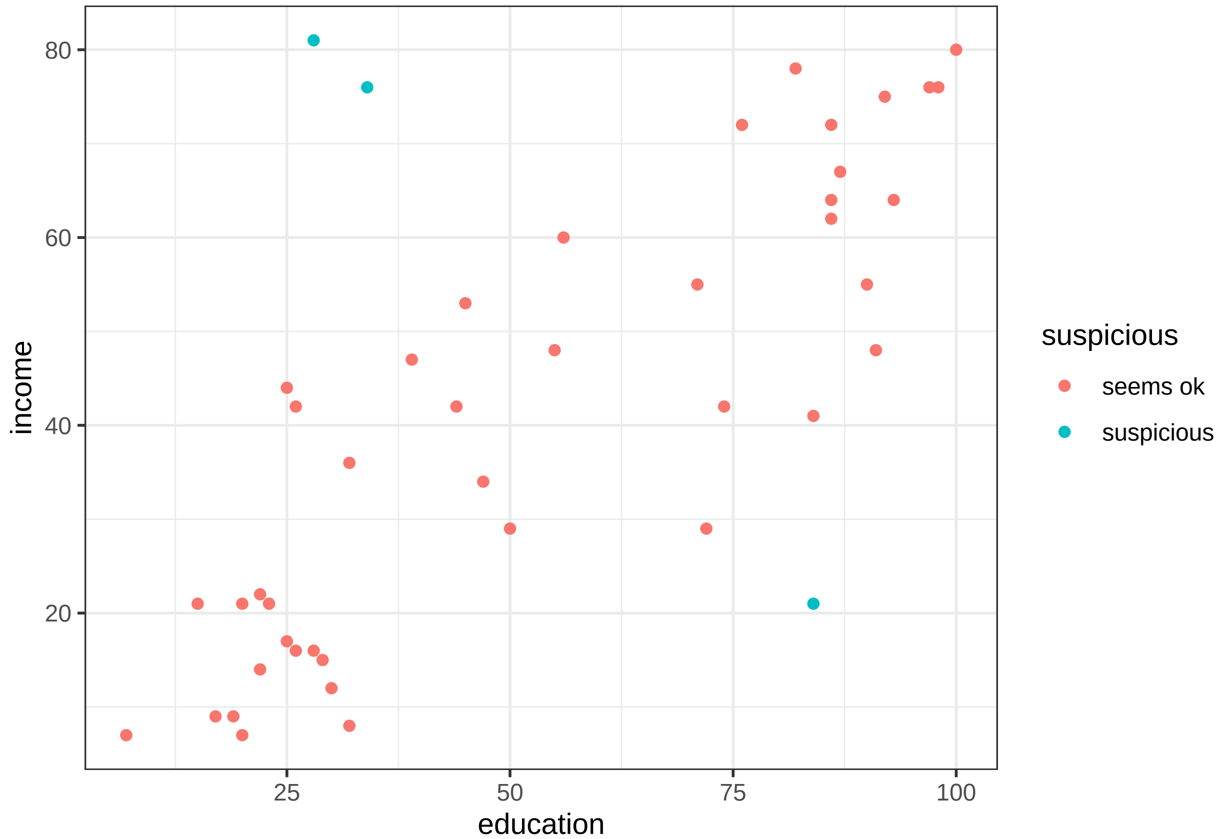
Where we left off:

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

```
Duncan[obs_to_investigate, ]
```

```
##             type income education prestige  occupation
## minister    prof     21        84       87    minister
## conductor     wc     76        34       38   conductor
## RR.engineer   bc     81        28       67 RR.engineer
```

```
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() +
  theme_bw()
```




```
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)
```

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

Schwarz's Bayesian Information Criterion (BIC)

- Used for model selection
- Takes a measure of lack of fit of a model (here, the residual sum of squares, SSRes) and adds a penalty for the number of terms in the model:

$$BIC = n \times \log\left(\frac{SSRes}{n}\right) + \log(n) \times (p+1)$$

- Subsets that produce smaller BIC values are better; within 2-3 points implies roughly similar performance

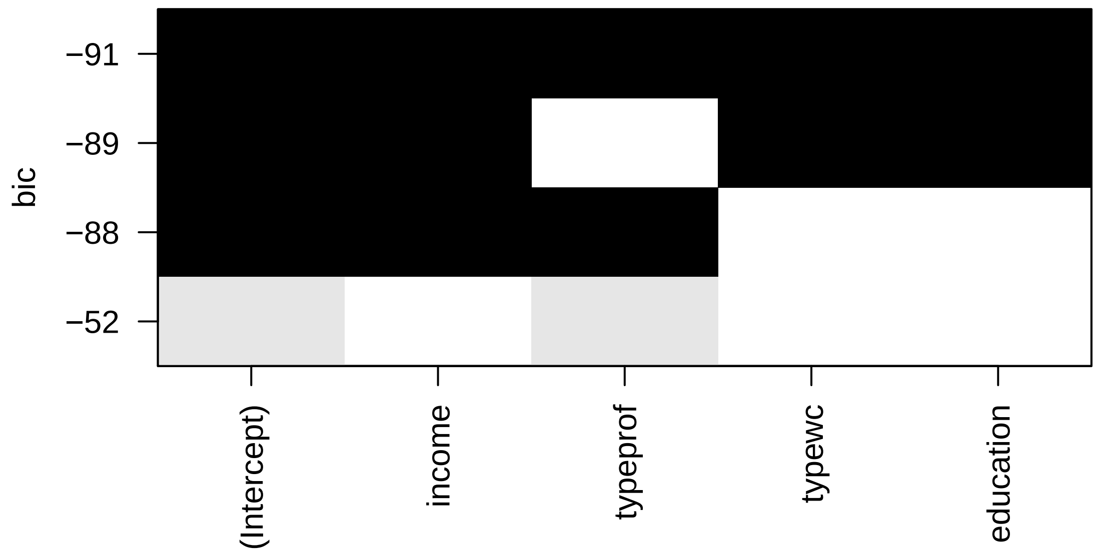
All subsets regression

- Involves fitting all possible subset models and identifying the ones with "best fit" as those that best satisfy some model-fitting criteria (here we are going to use BIC)
- Avoids problems with sequential variable selection techniques (i.e. forward selection, backward elimination, stepwise regression), which tend to select models with too many variables if the set contains unimportant ones

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

```
candidate_models1 <- regsubsets(prestige ~ income + type + education, data=Duncan)
plot(candidate_models1)
```




```r
summary(candidate_models1)
```

```
## Subset selection object
## Call: regsubsets.formula(prestige ~ income + type + education, data = Duncan)
## 4 Variables  (and intercept)
##           Forced in Forced out
## income        FALSE      FALSE
## typeprof      FALSE      FALSE
## typewc        FALSE      FALSE
## education     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          income typeprof typewc education
## 1  ( 1 ) " "    "*"      " "    " "
## 2  ( 1 ) "*"    "*"      " "    " "
## 3  ( 1 ) "*"    " "      "*"    "*"
## 4  ( 1 ) "*"    "*"      "*"    "*"
```

```r
str(summary(candidate_models1))
```

```
## List of 8
##  $ which : logi [1:4, 1:5] TRUE TRUE TRUE TRUE FALSE TRUE ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "1" "2" "3" "4"
```

```
##   .. ..$ : chr [1:5] "(Intercept)" "income" "typeprof" "typewc" ...
##  $ rsq    : num [1:4] 0.737 0.891 0.901 0.913
##  $ rss    : num [1:4] 11500 4743 4337 3798
##  $ adjr2 : num [1:4] 0.731 0.886 0.893 0.904
##  $ cp     : num [1:4] 80.12 10.95 8.67 5
##  $ bic    : num [1:4] -52.4 -88.5 -88.7 -90.9
##  $ outmat: chr [1:4, 1:4] " " "*" "*" "*" ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "1  ( 1 )" "2  ( 1 )" "3  ( 1 )" "4  ( 1 )"
##   .. ..$ : chr [1:4] "income" "typeprof" "typewc" "education"
##  $ obj    :List of 28
##   ..$ np       : int 5
##   ..$ nrbar    : int 10
##   ..$ d        : num [1:5] 45 10.8 14126.3 2.65 11891.84
##   ..$ rbar     : num [1:10] 0.4 52.556 0.133 41.867 47.963 ...
##   ..$ thetab   : num [1:5] 47.689 54.593 0.487 -6.5 0.598
##   ..$ first    : int 2
##   ..$ last     : int 5
##   ..$ vorder   : int [1:5] 1 3 5 4 2
##   ..$ tol      : num [1:5] 3.35e-09 3.65e-09 3.64e-07 2.16e-09 2.21e-07
##   ..$ rss      : num [1:5] 43688 11500 8156 8044 3798
##   ..$ bound    : num [1:5] 43688 11500 4743 4337 3798
##   ..$ nvmax    : int 5
##   ..$ ress     : num [1:5, 1] 43688 11500 4743 4337 3798
##   ..$ ir       : int 5
##   ..$ nbest    : int 1
##   ..$ lopt     : int [1:15, 1] 1 1 3 1 2 3 1 2 5 4 ...
##   ..$ il       : int 15
##   ..$ ier      : int 0
##   ..$ xnames   : chr [1:5] "(Intercept)" "income" "typeprof" "typewc" ...
##   ..$ method   : chr "exhaustive"
##   ..$ force.in : Named logi [1:5] TRUE FALSE FALSE FALSE FALSE
##   .. ..- attr(*, "names")= chr [1:5] "" "income" "typeprof" "typewc" ...
##   ..$ force.out: Named logi [1:5] FALSE FALSE FALSE FALSE FALSE
##   .. ..- attr(*, "names")= chr [1:5] "" "income" "typeprof" "typewc" ...
##   ..$ sserr    : num 3798
##   ..$ intercept: logi TRUE
##   ..$ lindep   : logi [1:5] FALSE FALSE FALSE FALSE FALSE
##   ..$ nullrss  : num 43688
##   ..$ nn       : int 45
##   ..$ call     : language regsubsets.formula(prestige ~ income + type + education, data = Duncan)
##   ..- attr(*, "class")= chr "regsubsets"
##  - attr(*, "class")= chr "summary.regsubsets"
```

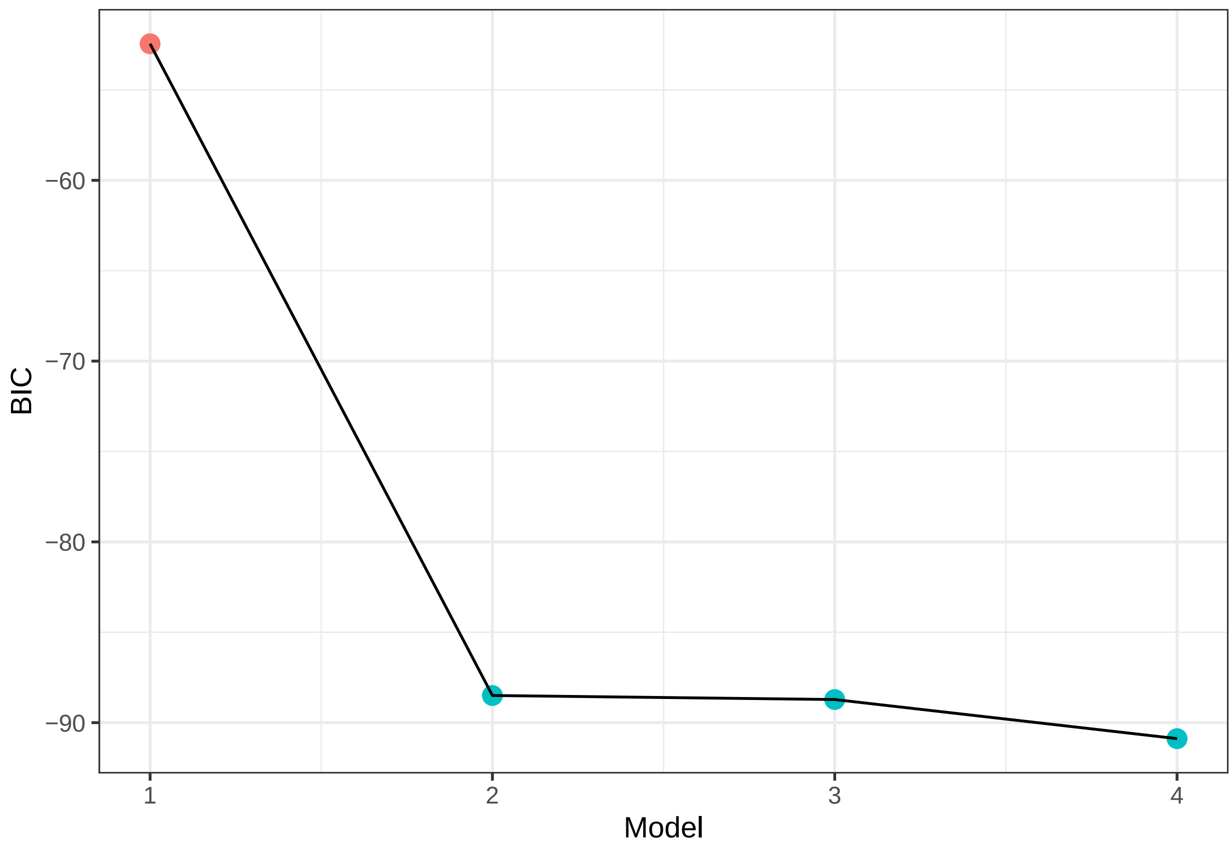
```r
summary(candidate_models1)$bic
```

```
## [1] -52.44958 -88.49874 -88.72119 -90.88381
```

```r
vis_bic1 <- data.frame(Model=1:4, BIC=summary(candidate_models1)$bic)

ggplot(data=vis_bic1, aes(x=Model, y=BIC)) +
  geom_point(aes(color=BIC < -88), size=3) +
  geom_line() +
```

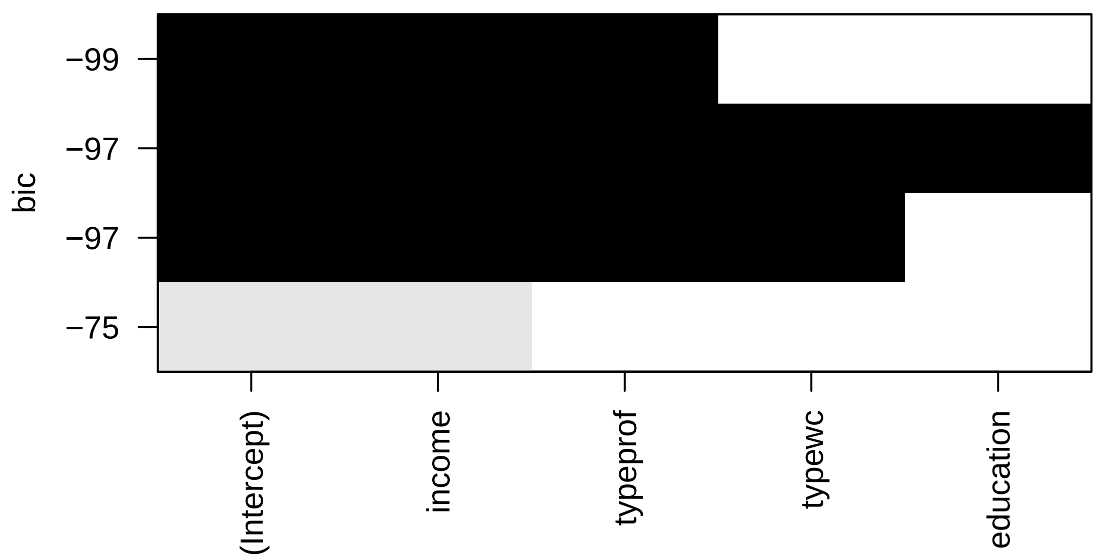
```r
  theme_bw() +
  theme(legend.position = "none")
```



Model 2, Model 3, and Model 4 have roughly similar performance.

- Model 2: income and typeprof (versus type_notprof) (BIC= -88.50)
- Model 3: income, typewc (versus type_notwc), education (BIC= -88.72)
- Model 4: income, typeprof, typewc, education (BIC= -90.88)

```r
candidate_models2 <- regsubsets(prestige ~ income +type + education, data=Duncan_minus_suspicious)
plot(candidate_models2)
```




```r
summary(candidate_models2)
```

```
## Subset selection object
## Call: regsubsets.formula(prestige ~ income + type + education, data = Duncan_minus_suspicious)
## 4 Variables  (and intercept)
##           Forced in Forced out
## income        FALSE      FALSE
## typeprof      FALSE      FALSE
## typewc        FALSE      FALSE
## education     FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##          income typeprof typewc education
## 1  ( 1 ) "*"    " "      " "    " "
## 2  ( 1 ) "*"    "*"      " "    " "
## 3  ( 1 ) "*"    "*"      "*"    " "
## 4  ( 1 ) "*"    "*"      "*"    "*"
```

```r
# str(summary(candidate_models2))

summary(candidate_models2)$bic
```

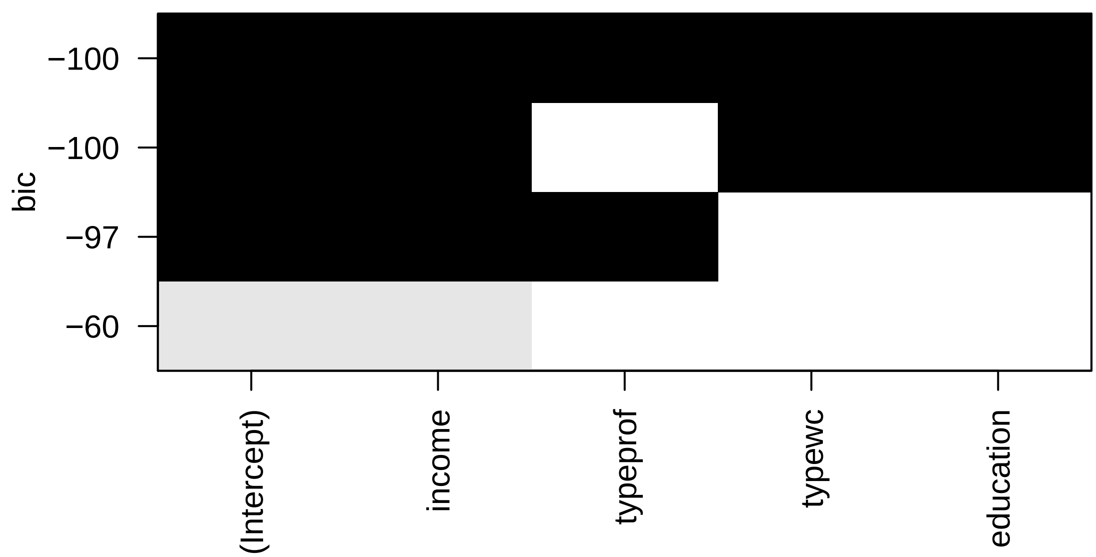
```
## [1] -74.73726 -99.45390 -97.28478 -97.28789
```

```r
vis_bic1 <- data.frame(Model=1:4, BIC=summary(candidate_models2)$bic)
```

Model 2, Model 3, and Model 4 have roughly similar performance.

- Model 2: income and typeprof (versus type_notprof) (BIC= -99.45)
- Model 3: income, typeprof, typewc (BIC=-97.28)
- Model 4: income, typeprof, typewc, education (-97.29)

```r
candidate_models3 <- regsubsets(prestige ~ income +type + education, data=Duncan_minus_minister)
plot(candidate_models3)
```




```r
summary(candidate_models3)
```

```
## Subset selection object
## Call: regsubsets.formula(prestige ~ income + type + education, data = Duncan_minus_minister)
## 4 Variables  (and intercept)
##            Forced in Forced out
## income         FALSE      FALSE
## typeprof       FALSE      FALSE
## typewc         FALSE      FALSE
## education      FALSE      FALSE
## 1 subsets of each size up to 4
```

```
## Selection Algorithm: exhaustive
##          income typeprof typewc education
## 1  ( 1 ) "*"    " "      " "    " "
## 2  ( 1 ) "*"    "*"      " "    " "
## 3  ( 1 ) "*"    " "      "*"    "*"
## 4  ( 1 ) "*"    "*"      "*"    "*"
```

```r
# str(summary(candidate_models3))

summary(candidate_models3)$bic
```

```
## [1]  -60.11700  -97.28833  -99.59875 -100.97645
```

Model 2, Model 3, and Model 4 have roughly similar performance.

- Model 2: income, typeprof (versus type_notprof) (BIC= -97.29)
- Model 3: income, typewc (versus type_notwc), education (BIC= -99.60)
- Model 4: income, typeprof, typewc, education (BIC= -100.98)

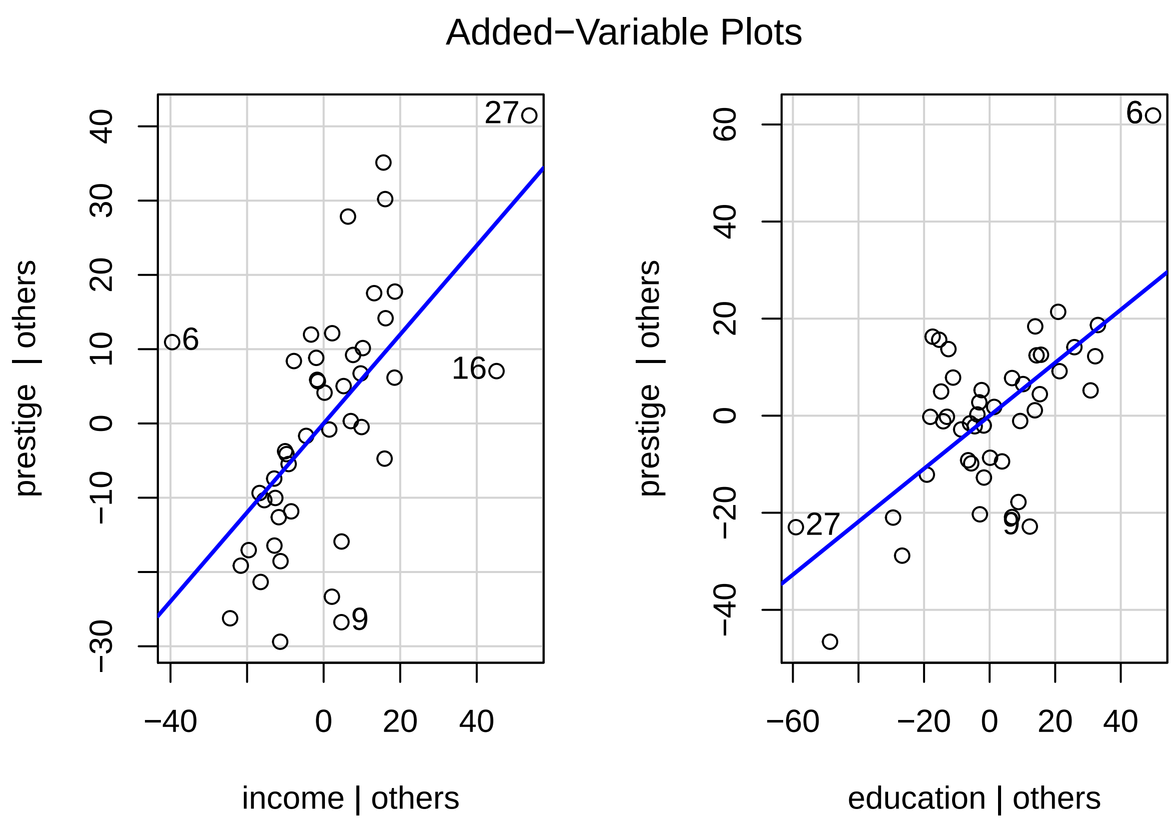
Added variable plots

- Also called partial regression plots
- Used to examine the effect of adding another explanatory variable to a model that already has one or more explanatory variables
- Strong linear relationship indicates that adding variable will likely be of value
- Can be used to identify influential points (6, 16, and 17 were already identified through other diagnostics)

Generating with avPlots (from R package car)

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

avPlots(fit_prestige)
```



Generating by hand

```
fit_prestige1 <- lm(prestige ~ education, data=Duncan)
fit_inc <- lm(education ~ income, data=Duncan)

fit_prestige2 <- lm(prestige ~ income, data=Duncan)
```

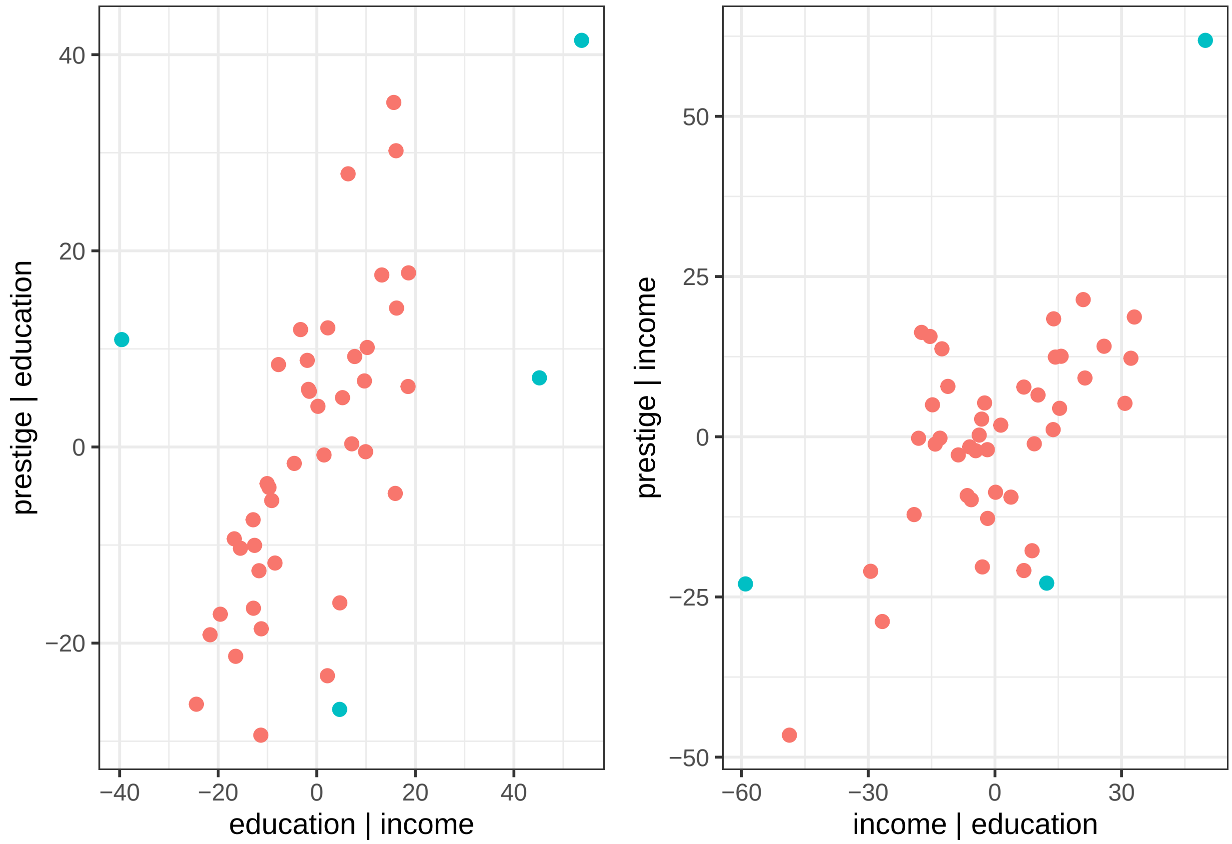
```r
fit_edu <- lm(income ~ education, data=Duncan)

Duncan <- Duncan %>%
  mutate(
    resid_edu = residuals(fit_inc),
    resid_inc = residuals(fit_edu),
    resid_prestige1 = residuals(fit_prestige1),
    resid_prestige2 = residuals(fit_prestige2),
    id = 1:nrow(Duncan)
  )

p1 <- ggplot(data=Duncan, aes(x=resid_inc, y=resid_prestige1)) +
        geom_point(aes(color=id %in% c(6, 9, 16, 27)), size=2) +
        theme_bw() +
        theme(legend.position="none") +
        ylab("prestige | education") +
        xlab("education | income")

p2 <- ggplot(data=Duncan, aes(x=resid_edu, y=resid_prestige2)) +
        geom_point(aes(color=id %in% c(6, 9, 27)), size=2) +
        theme_bw() +
        theme(legend.position="none") +
        ylab("prestige | income") +
        xlab("income | education")

grid.arrange(p1, p2, nrow=1)
```


prestige | education
education | income
prestige | income
income | education