Week 2 Practical Session

David Barron

Hilary Term 2018

Data analysis

Let's have a look at the Canadian occupational prestige data. This is a dataset that comes with the `car` package, so we can get access to it by using the `data(Prestige)` function.

```
library(car)
library(effects)
```

```
Loading required package: carData


Attaching package: 'carData'

The following objects are masked from 'package:car':

    Guyer, UN, Vocab

lattice theme set by effectsTheme()
See ?effectsTheme for details.
```

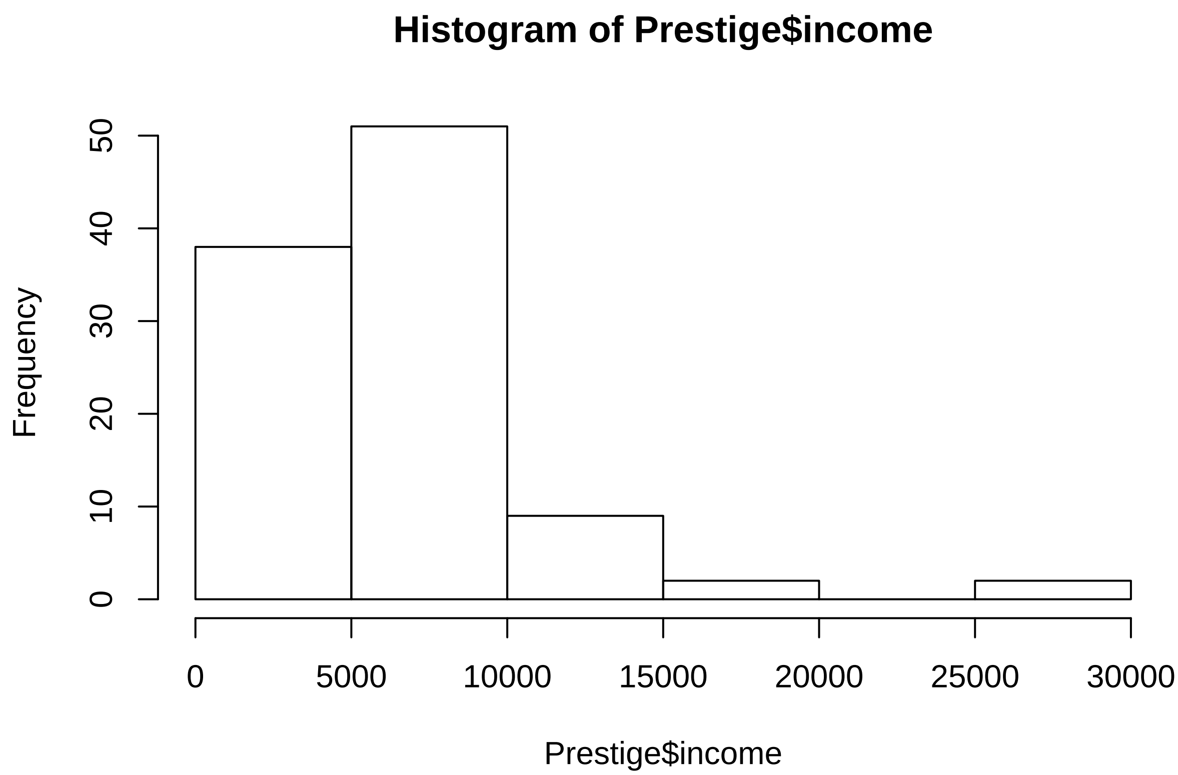
```
data(Prestige)
head(Prestige)
```

```
                    education income women prestige census type
gov.administrators      13.11  12351 11.16     68.8   1113 prof
general.managers        12.26  25879  4.02     69.1   1130 prof
accountants             12.77   9271 15.70     63.4   1171 prof
purchasing.officers     11.42   8865  9.11     56.8   1175 prof
chemists                14.62   8403 11.68     73.5   2111 prof
physicists              15.64  11030  5.13     77.6   2113 prof
```

```
str(Prestige)
```

```
'data.frame':   102 obs. of  6 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

```r
hist(Prestige$income)
```



Let's look at some data analysis. I will use the `update` function, which takes as its first argument an existing regression output and as its second, a modified formula. The `.` means all the variables in the original formula. It just saves a bit of typing; you can achieve the same result by using `lm` again if you prefer.

```r
l1 <- lm(prestige ~ income + education, data = Prestige)
summary(l1)
```

```

Call:
lm(formula = prestige ~ income + education, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.4040  -5.3308   0.0154   4.9803  17.6889 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.8477787  3.2189771  -2.127   0.0359 *  
income       0.0013612  0.0002242   6.071 2.36e-08 ***
education    4.1374444  0.3489120  11.858  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.81 on 99 degrees of freedom
Multiple R-squared:  0.798, Adjusted R-squared:  0.7939 
```

```
F-statistic: 195.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

```r
l2 <- update(l1, . ~ . + type)
summary(l2)
```

```

Call:
lm(formula = prestige ~ income + education + type, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.9529  -4.4486   0.1678   5.0566  18.6320 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.6229292  5.2275255  -0.119    0.905    
income       0.0010132  0.0002209   4.586 1.40e-05 ***
education    3.6731661  0.6405016   5.735 1.21e-07 ***
typeprof     6.0389707  3.8668551   1.562    0.122    
typewc      -2.7372307  2.5139324  -1.089    0.279    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.095 on 93 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8349,    Adjusted R-squared:  0.8278 
F-statistic: 117.5 on 4 and 93 DF,  p-value: < 2.2e-16
```

```r
anova(l1, l2)
```

```
Error in anova.lmlist(object, ...): models were not all fitted to the same size of dataset
```

```r
any(is.na(Prestige$type))
```

```
[1] TRUE
```

This is a common problem; there are some missing data in the `type` variable, so we can't compare the fit of these two regressions. The solution is to re-fit the first regression with the same data as was used for the second.

```r
l1a <- update(l1, subset = !is.na(type))
summary(l1a)
```

```

Call:
lm(formula = prestige ~ income + education, data = Prestige, 
    subset = !is.na(type))

Residuals:
     Min       1Q   Median       3Q      Max 
-16.9367  -4.8881   0.0116   4.9690  15.9280 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.6210352  3.1162309  -2.446   0.0163 *  
income       0.0012415  0.0002185   5.682 1.45e-07 ***
education    4.2921076  0.3360645  12.772  < 2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.45 on 95 degrees of freedom
Multiple R-squared:  0.814, Adjusted R-squared:  0.8101
F-statistic: 207.9 on 2 and 95 DF,  p-value: < 2.2e-16
```

```r
anova(l1a, l2)
```

```
Analysis of Variance Table

Model 1: prestige ~ income + education
Model 2: prestige ~ income + education + type
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     95 5272.4
2     93 4681.3  2    591.16 5.8721 0.003966 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This uses update again, but this time I've used the `subset` option to restrict the data to those rows that don't have an NA in the `type` variable. `!` is the *not* operator in R, while `is.na` is a function that returns `TRUE` for any row that is `NA`. Now we can see that the `type` variabe improves fit.

Stepwise regression

Does stepwise regression give us the same result? Yes!

```r
l.step <- step(lm(prestige ~ 1, data = Prestige, subset = !is.na(type)), scope = ~income +
    education + type + women, dir = "for")
```

```
Start:  AIC=557.4
prestige ~ 1

            Df Sum of Sq     RSS    AIC
+ education  1   21282.5  7064.4 423.23
+ type       2   19775.6  8571.3 444.18
+ income     1   14021.6 14325.3 492.51
<none>                   28346.9 557.40
+ women      1     343.9 28003.0 558.20

Step:  AIC=423.23
prestige ~ education

         Df Sum of Sq    RSS    AIC
+ income  1   1791.97 5272.4 396.56
+ type    2   1324.36 5740.0 406.89
+ women   1    763.48 6300.9 414.02
<none>                7064.4 423.23

Step:  AIC=396.56
prestige ~ education + income

        Df Sum of Sq    RSS    AIC
+ type   2    591.16 4681.3 388.90
<none>               5272.4 396.56
```

```
+ women  1     10.37 5262.1 398.36

Step:  AIC=388.9
prestige ~ education + income + type

        Df Sum of Sq    RSS    AIC
<none>                4681.3 388.90
+ women  1    2.2881 4679.0 390.86
```

Just to illustrate, we can also do this backwards:

```
cc.full <- lm(prestige ~ education + income + women + type, data = Prestige,
    subset = !is.na(type))
cc.back <- step(cc.full, dir = "back")
```

```
Start:  AIC=390.86
prestige ~ education + income + women + type

            Df Sum of Sq    RSS    AIC
- women      1      2.29 4681.3 388.90
<none>                   4679.0 390.86
- type       2    583.08 5262.1 398.36
- income     1    803.92 5482.9 404.39
- education  1   1635.49 6314.5 418.23

Step:  AIC=388.9
prestige ~ education + income + type

            Df Sum of Sq    RSS    AIC
<none>                   4681.3 388.90
- type       2    591.16 5272.4 396.56
- income     1   1058.77 5740.0 406.89
- education  1   1655.47 6336.7 416.58
```

Same result.

Diagnostics

Let's look at some diagnostics.

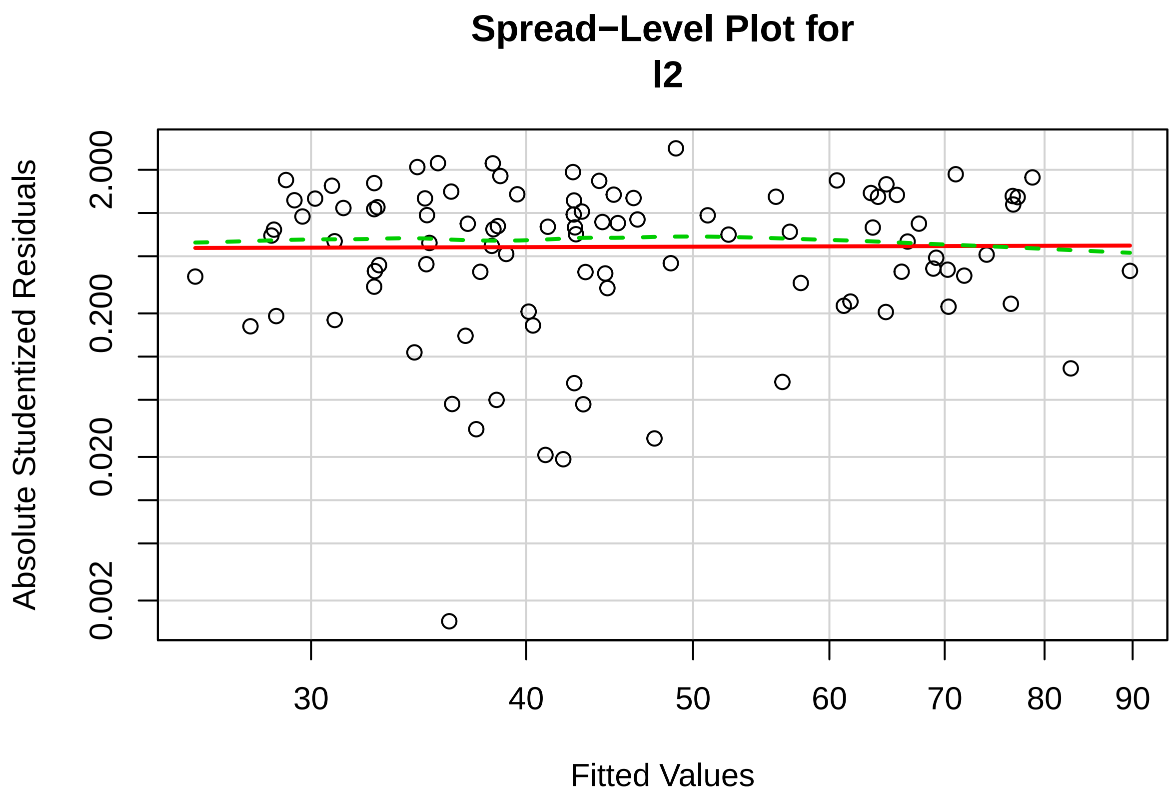
```
vif(l2)  # Looks OK
```

```
              GVIF Df GVIF^(1/(2*Df))
income    1.681325  1        1.296659
education 5.973932  1        2.444163
type      6.102131  2        1.571703
```

```
ncvTest(l2)  # This looks OK
```

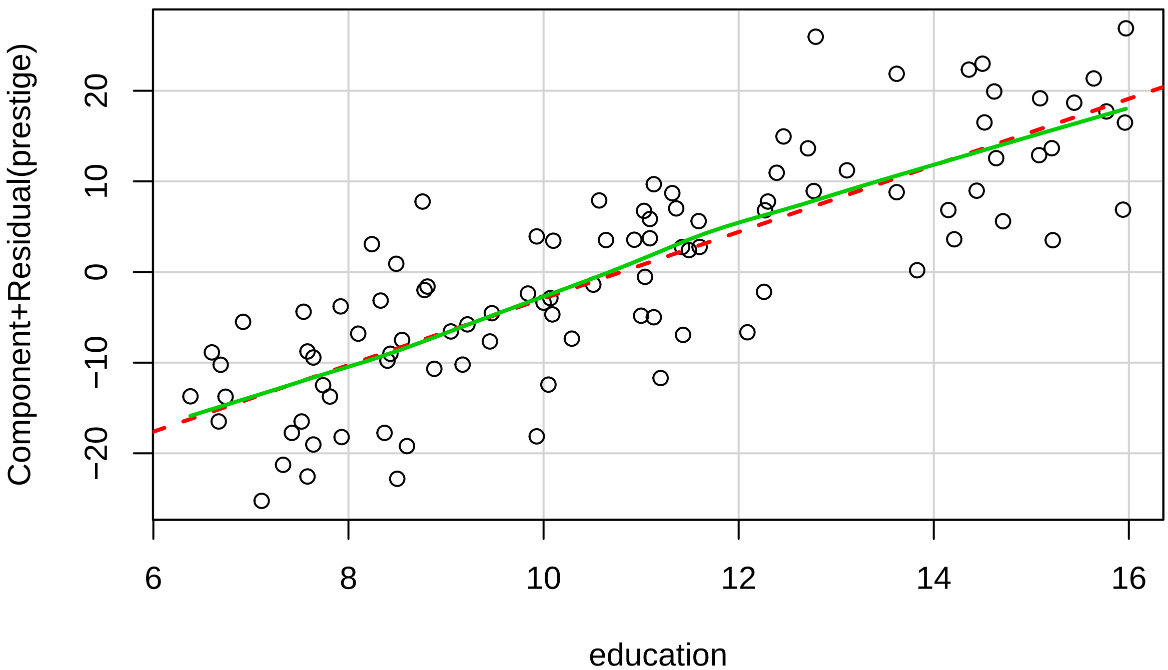
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.09830307    Df = 1     p = 0.7538756
```

```
spreadLevelPlot(l2)  # Also looks OK
```

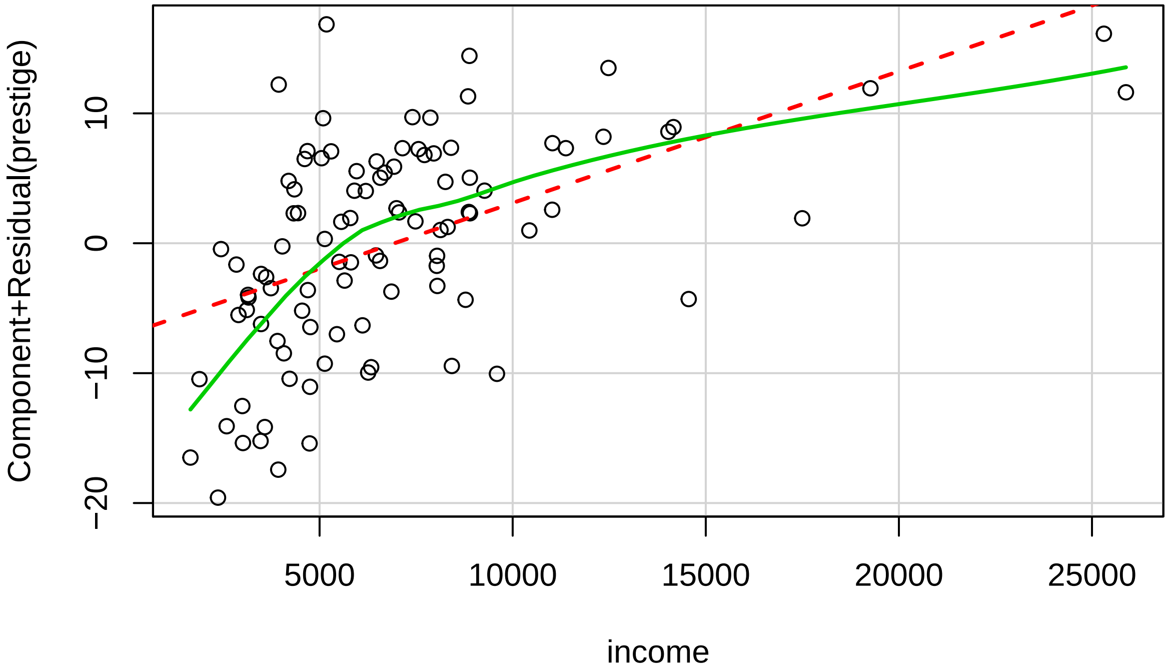



```
Suggested power transformation:  0.9695631
```

```r
crPlot(l2, "education")  # Looks OK
```




```
crPlot(l2, "income")
```



There is some evidence of an issue with `income`, which isn't surprising. Let's try a log transformation:

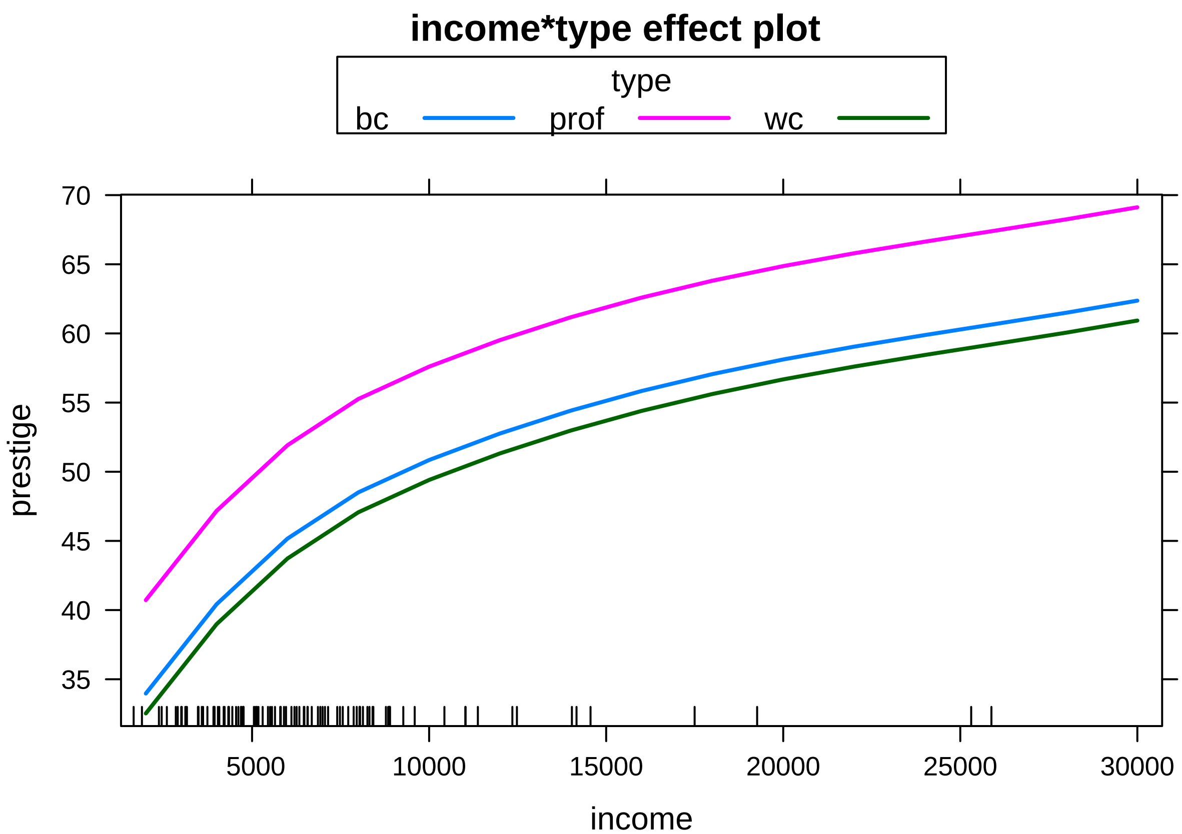
```
Prestige$log.income <- log(Prestige$income)

l3 <- update(l2, . ~ . - income + log(income))
summary(l3)
```

```
Call:
lm(formula = prestige ~ education + type + log(income), data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-13.511  -3.746   1.011   4.356  18.438

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -81.2019    13.7431  -5.909 5.63e-08 ***
education     3.2845     0.6081   5.401 5.06e-07 ***
typeprof      6.7509     3.6185   1.866   0.0652 .
typewc       -1.4394     2.3780  -0.605   0.5465
log(income)  10.4875     1.7167   6.109 2.31e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.637 on 93 degrees of freedom
  (4 observations deleted due to missingness)
```

```
Multiple R-squared:  0.8555,    Adjusted R-squared:  0.8493
F-statistic: 137.6 on 4 and 93 DF,  p-value: < 2.2e-16
```

You can see this is a better fit by looking at the R^2. It's a bit more tricky to work out the effect of income, though. For a unit increase in log income we get a 10.5 increase in prestige. Let's look at an effect plot:

```
plot(Effect(c("income", "type"), l3), multiline = TRUE)
```



Interaction effects

Let's try an interaction between income and type:

```
l4 <- update(l3, . ~ . + log(income):type)
summary(l4)
```

```
Call:
lm(formula = prestige ~ education + type + log(income) + type:log(income),
    data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-13.484  -4.453   1.122   4.123  18.737

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -118.4325    20.3728  -5.813 8.97e-08 ***
```

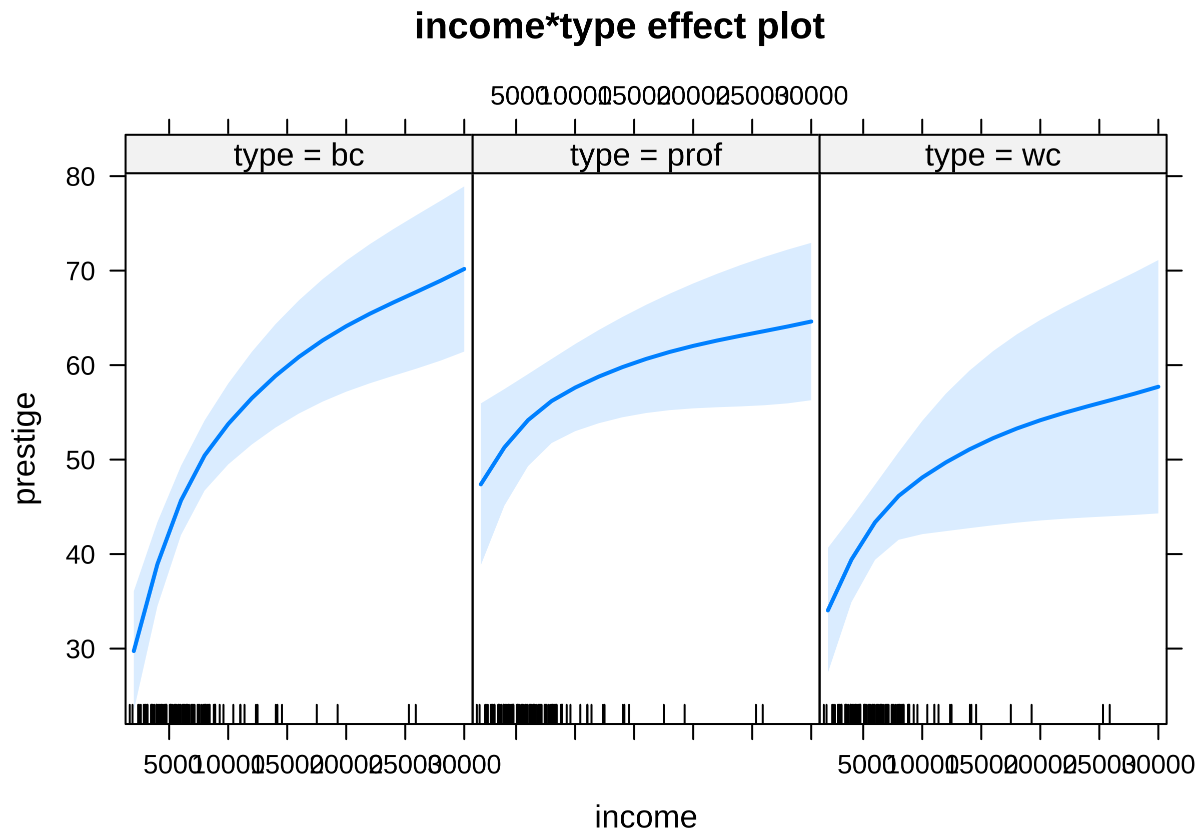
```
education               3.2107     0.5993   5.357 6.31e-07 ***
typeprof               82.7757    31.5059   2.627   0.0101 *
typewc                 51.3717    36.8521   1.394   0.1667
log(income)            14.9336     2.4928   5.991 4.12e-08 ***
typeprof:log(income)   -8.5690     3.5251  -2.431   0.0170 *
typewc:log(income)     -6.1925     4.3172  -1.434   0.1549
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.491 on 91 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.8647,    Adjusted R-squared:  0.8558
F-statistic: 96.96 on 6 and 91 DF,  p-value: < 2.2e-16
```

```
anova(l3, l4)
```

```
Analysis of Variance Table

Model 1: prestige ~ education + type + log(income)
Model 2: prestige ~ education + type + log(income) + type:log(income)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     93 4096.3
2     91 3834.2  2    262.13 3.1107 0.04934 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is just about statistically significant. Let's look at the effect plot again.

```
plot(Effect(c("income", "type"), l4), multi = TRUE)
```



We can see that the effect of income on prestige is greater for blue collar occupations than it is for the other two.

Homework

1. Load the data SLID in the `car` package.

2. Explore the data.
3. Perform a regression using `wage` as the outcome variable and all the other variables in the data as explanatory variables.
4. Test for normality of residuals. If necessary, transform data and perform a new regression.
5. Test for heterokedasticity. Is any action needed? If so, what?
6. Test for linearity of relationship between education and age and wages. Do either of these explanatory variable appear non-linear? If so, perform new regression as appropriate.
7. Consider an interaction between education and sex. Does including this improve the model? If so, display graphically the estimated relationship between education and wage separately for men and women.