Solutions to test exercises

Michał Bojanowski
ICM UW
mbojan@icm.edu.pl

Zbigniew Karpiński
IFiS PAN
zkarpinski@ifispan.waw.pl

June 2, 2012

Exercise text is in serif font, answers are in sans-serif font, `R code is in typewriter font`. Some of the commentaries are also as **R** comments (`# in typewriter with preceding '#'`).

1 Exercise 1

In analyzing data from face-to-face surveys it is often overlooked that the interviewer might influence the responses of the respondent.

The file `pgssint.rda` contains a data frame `pgssint` which is an excerpt from PGSS data containing the following variables:

`pgssyear` PGSS edition year

`female` whether the respondent is a female.

`intfemale` whether the interviewer is a female.

`q9age` age of respondent.

`q7d` whether the respondent agrees or not (1=Strongly agree, 2, 3, 4=Strongy disagree) with the statement that *it is a role of a man to earn money and a woman to take care of the household.*

Special values of the variables have been already recoded to missing data (`NA`).

Using this dataset answer the question to what extent the responses to the question `q7d` depend on the gender of the respondent and the gender of the interviewer. In particular:

Create a cross-tabulation of the responses to `q7d`, respondent's gender, and interviewer's gender.

```
> # load data
> load("../pgssint.rda")
> # create the frequency table
> x <- with(pgssint, table(q7d, intfemale, female))
> x

, , female = FALSE

   intfemale
q7d FALSE TRUE
  1   430  368
  2   625  654
  3   202  306
  4    31   46

, , female = TRUE

   intfemale
q7d FALSE TRUE
  1   420  433
  2   650  754
  3   320  511
  4    64   92
```

Compute the percentage distribution of responses to `q7d` *given* the gender of the respondent *and* the interviewer.

In other words, the percentages have to sum-up to 100 in *every* combination of respondent's *and* interviewer's gender

```
> p <- prop.table(x, c(2,3)) * 100
> p

, , female = FALSE

   intfemale
q7d     FALSE      TRUE
  1 33.385093 26.783115
  2 48.524845 47.598253
  3 15.683230 22.270742
  4  2.406832  3.347889

, , female = TRUE

   intfemale
q7d     FALSE      TRUE
  1 28.885832 24.189944
  2 44.704264 42.122905
  3 22.008253 28.547486
  4  4.401651  5.139665

> # verify that columns sum-up to 100
> apply(p, c(2,3), sum)

         female
intfemale FALSE TRUE
    FALSE   100  100
    TRUE    100  100
```

Demonstrate graphically (with e.g. a bar chart) how the responses to `q7d` depend on the gender of respondent and interviewer. For simplification use the total percentage of answers "strongly agree" and "agree".

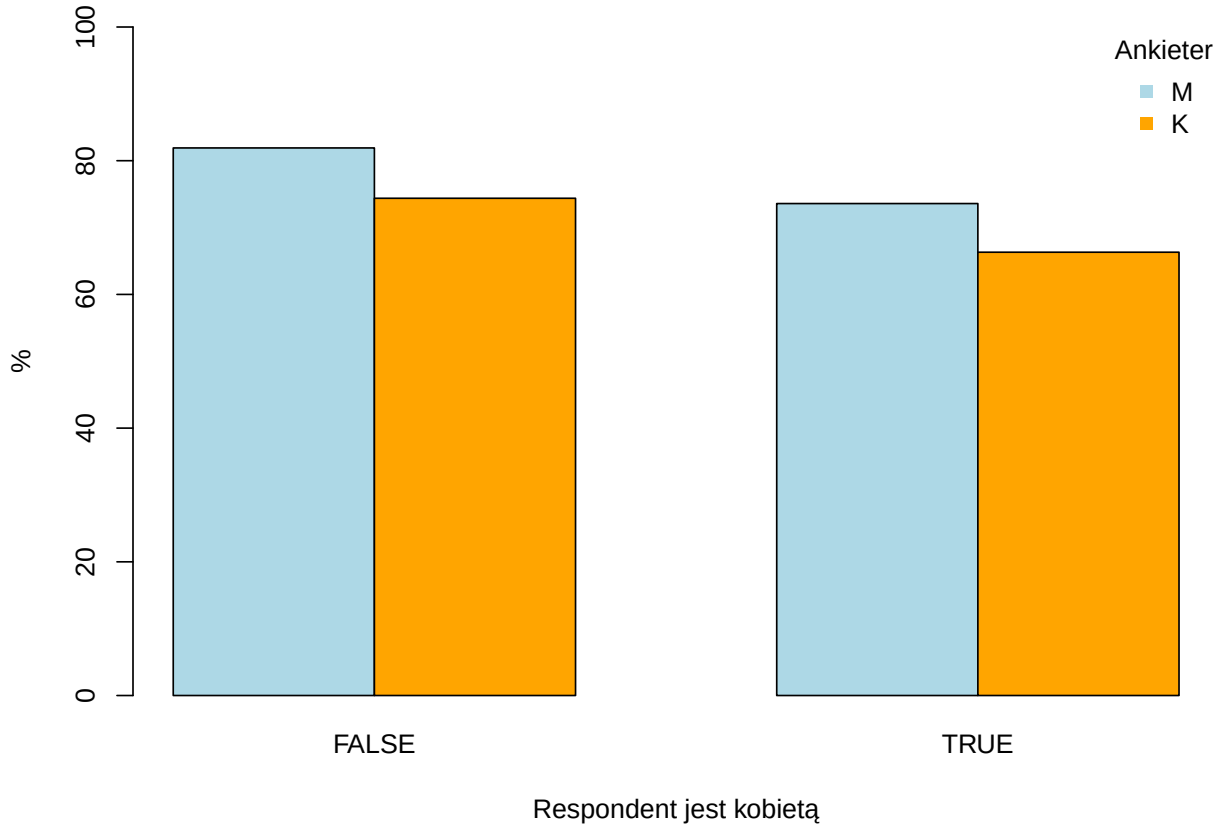
Computing the percentage of "strongly agree" and "agree" can be done in several ways, here we use `apply`:

From the table of percentages 'p' directly. We would like to take sums of the first two rows in every column of this three dimensional table, i.e., for every combination of "intfemale" (second dimesion) and "female" (third dimension) take first two entries ("strongly agree" and "agree") and sum them up:

```
> tab1 <- apply(p, c(2,3), function(x) sum(x[1:2]))
> tab1

         female
intfemale    FALSE     TRUE
    FALSE 81.90994 73.59010
    TRUE  74.38137 66.31285
```

Now we make a barchart from `tab1`:

```
> k <- c("lightblue", "orange")
> barplot(tab1, beside=TRUE, xlab="Respondent jest kobietą", col=k, ylab="%",
+         ylim=c(0,100))
> legend("topright", col=k, title="Ankieter", legend=c("M", "K"), bty="n",
+        pch=15)
```



What can we conclude about the interviewers influencing respondents' answers to that particular survey question? Write your answer.

> Men agree with the given statement more often than women. At the same time, respondents agree more often if the interviewer is a man rather than a woman given the gender of the respondent. Consequently, there is a interviewer's effect as people responded differently depending on the gender of the interviewer.

2 Exercise 2

Using the data `pgssint` (the same as in Exercise 1) investigate how the answers to question `q7d` vary between different cohorts and over time. In particular:

Create a variable "year of birth" based on age and year of study and categorize it into intervals using breakpoints at 1940, 1960, and 1980. These will be the cohorts.

```
> yb <- with(pgssint, pgssyear - q9age)
> cohort <- cut( yb, c(-Inf, 1940, 1960, 1980, Inf), dig.lab=4)
> table(cohort)

cohort
(-Inf,1940] (1940,1960] (1960,1980] (1980, Inf]
       1422        2263        2015         468
```

Compute the (conditional) percentage of answers "strongly agree" and "agree" given the cohorts and year of study.

> This can be done in several ways, here we show two ways:
>
> 1. Similarly to Exercise 1, create a table of percentages and extract sums of selected entries with `apply`:

```
> tab <- with(pgssint, table(q7d, cohort, pgssyear))
> ptab <- prop.table(tab, c(2,3)) * 100
> # result
```

```
> r1 <- apply(ptab, c(2,3), function(x) sum(x[1:2]))
> r1

             pgssyear
cohort            1997     1999     2002     2005     2008
  (-Inf,1940] 88.20225 89.86486 88.69258 90.04739 80.00000
  (1940,1960] 76.90583 79.08654 75.50111 77.46479 71.33028
  (1960,1980] 67.76860 69.04110 64.61916 62.61905 59.41645
  (1980, Inf]      NaN 50.00000 53.06122 55.36723 56.31068
```

2. Take advantage of the fact that taking a mean of a binary (or logical, aka dummy) variable is equal to the proportion of 1s.

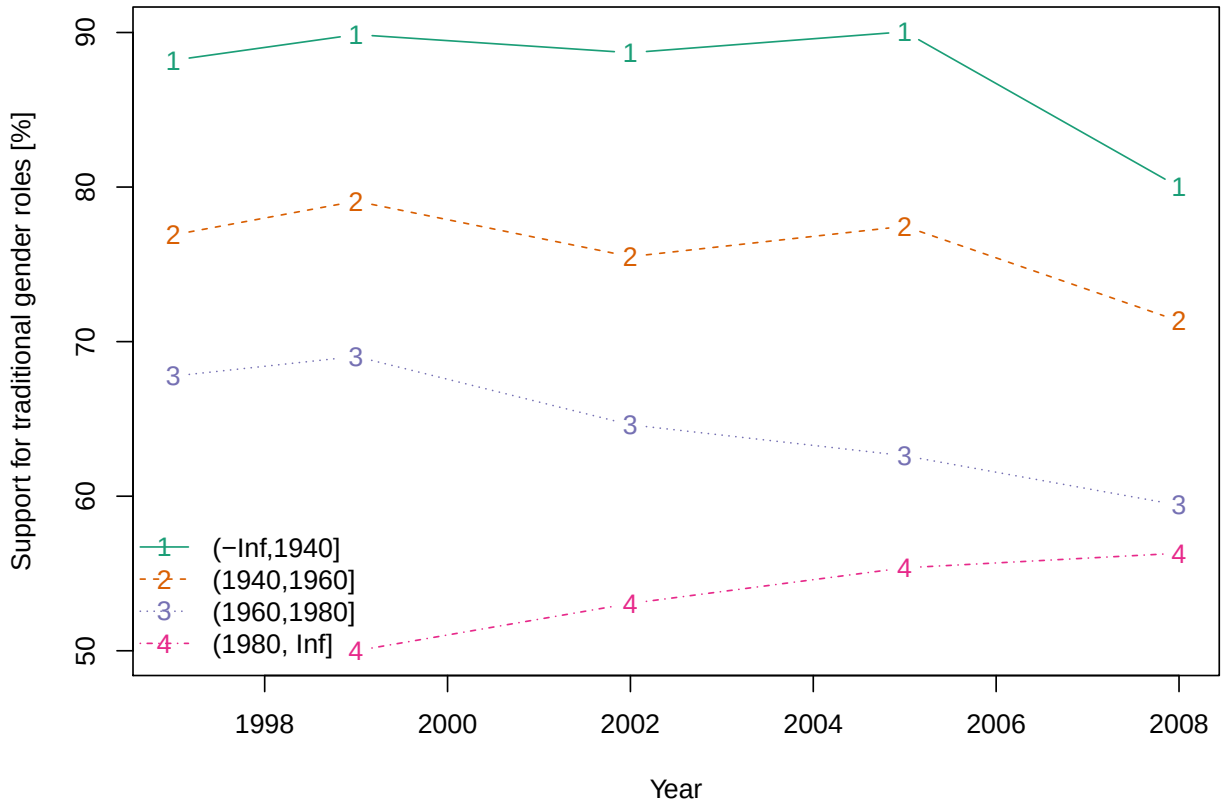
```
> # logical (binary) variable if 'q7d' is equal to 1 or 2
> y <- with(pgssint, q7d %in% c(1, 2))
> # as some of the FALSEs in 'y' correspond to missing data in 'q7d', re-create
> # them in 'y'
> y[is.na(pgssint$q7d)] <- NA
> # result: means of 'y' in subgroups defined by 'pgssyear' and 'cohort'
> r2 <- with(pgssint, tapply(y, list(cohort, pgssyear), mean, na.rm=TRUE)) * 100
> r2

                1997     1999     2002     2005     2008
(-Inf,1940] 88.20225 89.86486 88.69258 90.04739 80.00000
(1940,1960] 76.90583 79.08654 75.50111 77.46479 71.33028
(1960,1980] 67.76860 69.04110 64.61916 62.61905 59.41645
(1980, Inf]       NA 50.00000 53.06122 55.36723 56.31068
```

Both ways give identical numeric results

How did opinions of the cohorts evolve during the period covered by PGSS?

The older the cohort, the more often people express support for traditional gender roles in a family (more often agree with the statement in Q7D). People in all the cohorts but the youngest tend to support the traditional gender roles *less often* over time. The youngest cohort support them *more often* over time. This can be seen in the figure below (creating the figure was not part of the exercise).

```
> library(RColorBrewer)
> # take 4 colors from "Dark2" palette available in package "RColorBrewer"
> k <- brewer.pal(4, "Dark2")
> matplot( c(1997, 1999, 2002, 2005, 2008), t(r1), type="b",
+         ylab="Support for traditional gender roles [%]",
+         xlab="Year", col=k)
> legend("bottomleft", lty=1:4, col=k, legend=levels(cohort), bty="n",
+        pch=as.character(1:4))
```



3 Exercise 3

`status.csv` is a dataset containing 690 observations on 5 variables:

`status` an estimate of subject's social status on a 10-point scale, with larger values indicating higher status

`earnings` subject's net monthly income in PLN

`degree` subject's degree of education

`prestige` prestige category of subject's occupation

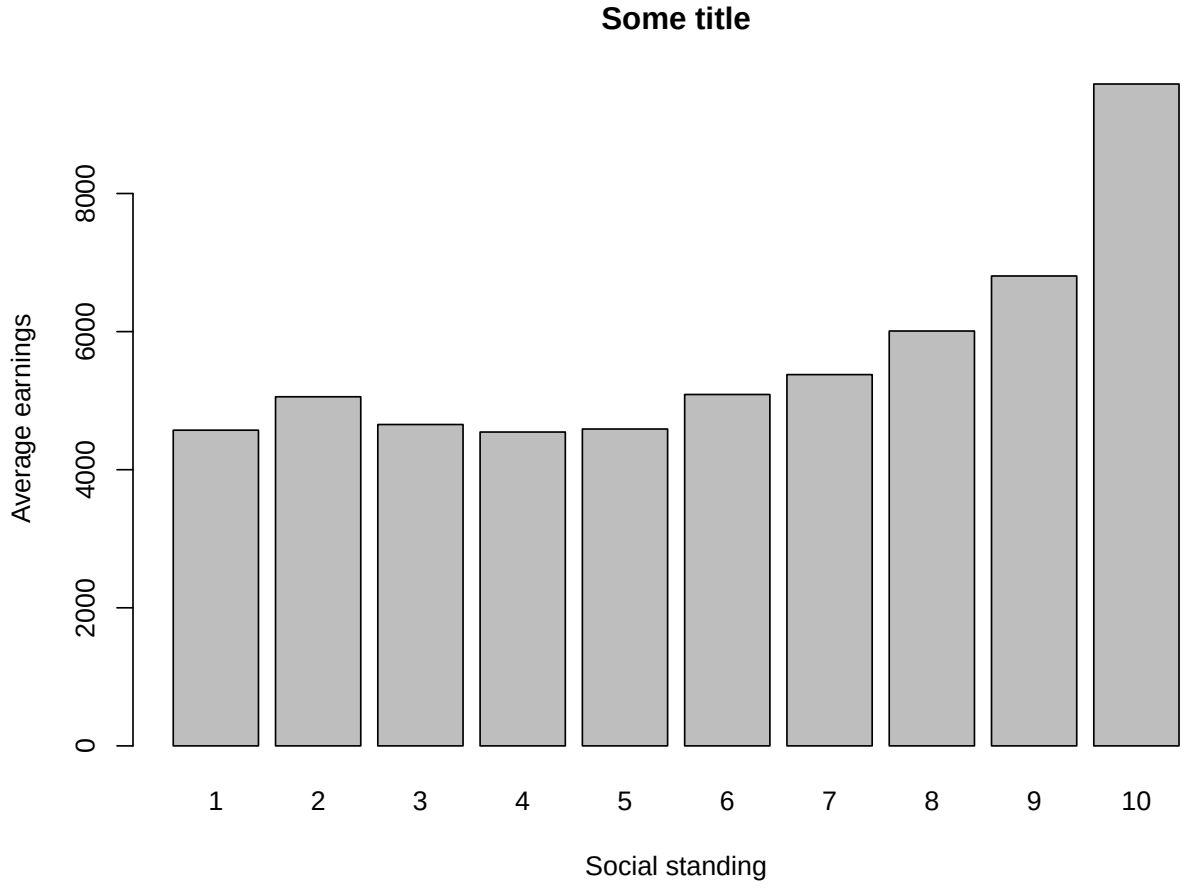
`gender` subject's gender.

The data are written in a text file with variables names in the first row and columns separated by semi-colons (;).

Using the `status` data:

Create a barplot showing mean values of income by social status. Add appropriate labels and titles to the plot and save as a PDF.

```
> status <- read.table( "../status.csv", head=TRUE, sep=";" )
> # Average earnings by status
> tab <- with( status, tapply( earnings, status, mean ) )
```



Code creating the figure and saving it to PDF. You could have also saved the PDF using the menus/buttons available in RStudio.

```
> pdf( file="Status_barplot.pdf" )
> # Creating the barplot
> barplot( tab, main="Some title", xlab="Social standing",
+         ylab="Average earnings" )
> dev.off()
```

Estimate a regression model that has status as a dependent variable (DV), and earnings (in thousands of PLN), degree of education, occupational prestige, and gender as independent variables (IVs). Is the DV significantly related to all the IVs?

```
> mod1 <- lm( status ~ I(earnings/1000) + degree + prestige + gender, data=status )
> summary(mod1)

Call:
lm(formula = status ~ I(earnings/1000) + degree + prestige +
    gender, data = status)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0411 -1.0781 -0.1577  1.0156  3.8064

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.81533    0.19793  34.433  < 2e-16 ***
I(earnings/1000)  0.07560    0.01840   4.109 4.45e-05 ***
degreelow        -0.80794    0.15709  -5.143 3.53e-07 ***
degreemedium     -0.28554    0.15757  -1.812   0.0704 .
prestigelow      -2.58520    0.13469 -19.194  < 2e-16 ***
```

```
prestigemedium   -1.12427    0.18710  -6.009 3.04e-09 ***
gendermale        0.02689    0.11144   0.241   0.8094
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.461 on 683 degrees of freedom
Multiple R-squared: 0.403,        Adjusted R-squared: 0.3978
F-statistic: 76.86 on 6 and 683 DF,  p-value: < 2.2e-16
```

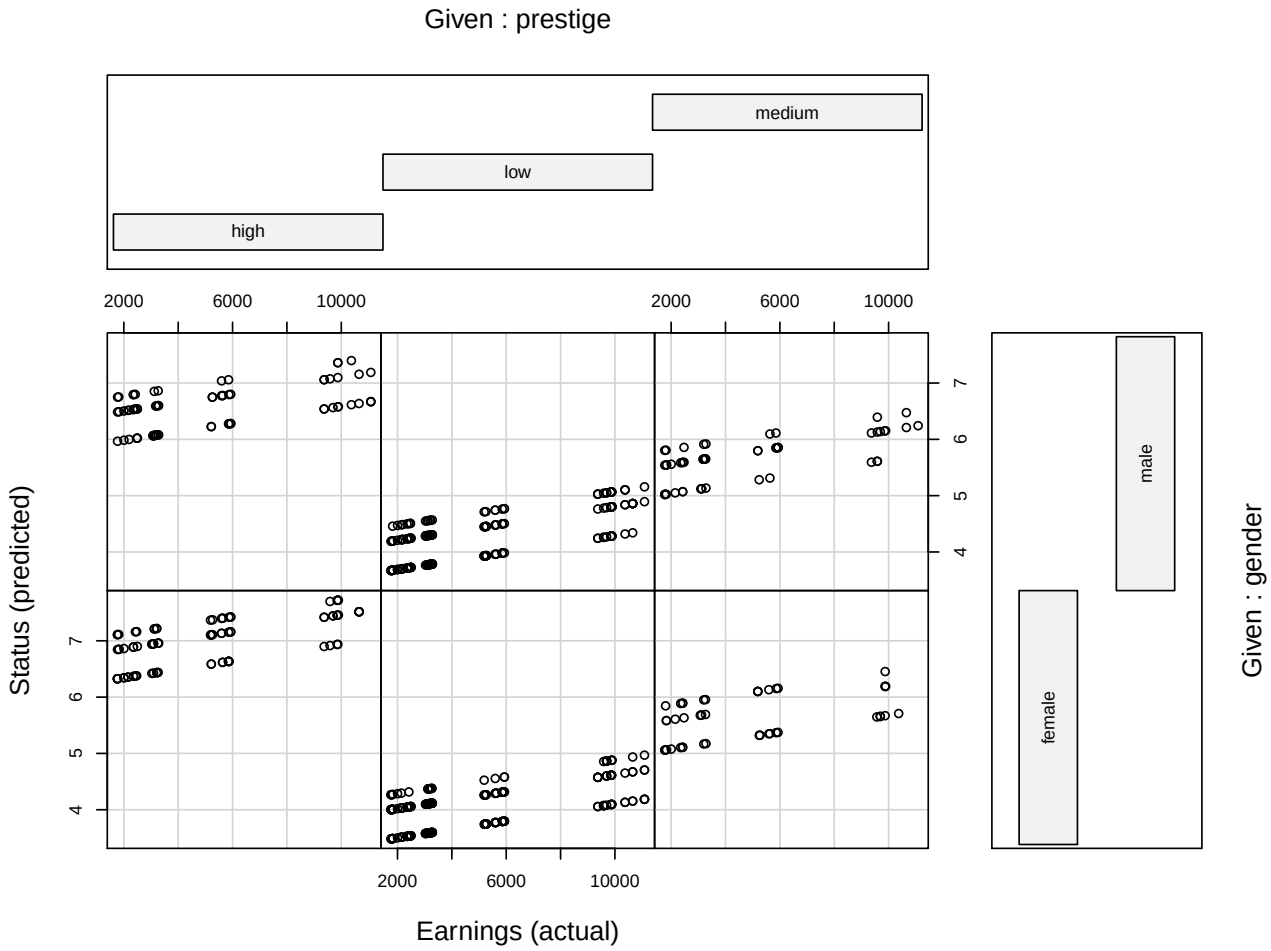
No, the DV is not related significantly to gender, and its relationship to the dummay variable 'medium degree of education' is significant at a level 0.1 only.

Update the model by adding an interaction effect between occupational prestige and gender. Does that improve the model's fit? Does the social standing of men and women "respond" differently to changes in prestige?

```
> mod2 <- update( mod1, .~. + prestige:gender )
> summary(mod2)

Call:
lm(formula = status ~ I(earnings/1000) + degree + prestige +
    gender + prestige:gender, data = status)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1029 -1.0438 -0.1086  0.9471  3.7199

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.97595    0.21579  32.327  < 2e-16 ***
I(earnings/1000)           0.07537    0.01837   4.103 4.57e-05 ***
degreelow                 -0.78327    0.15731  -4.979 8.10e-07 ***
degreemedium              -0.26407    0.15824  -1.669   0.0956 .
prestigelow               -2.84225    0.18480 -15.380  < 2e-16 ***
prestigemedium            -1.26622    0.26808  -4.723 2.82e-06 ***
gendermale                -0.35933    0.22979  -1.564   0.1183
prestigelow:gendermale     0.54672    0.26991   2.026   0.0432 *
prestigemedium:gendermale  0.32068    0.37596   0.853   0.3940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.459 on 681 degrees of freedom
Multiple R-squared: 0.4067,        Adjusted R-squared: 0.3997
F-statistic: 58.34 on 8 and 681 DF,  p-value: < 2.2e-16
```

Adding the interaction term does not result in a improvement of the model's fit. Consequently, there is not enough evidence for us to conclude that that women's social standing responds differently to changes in prestige than men's social standing.

Create a coplot showing a predicted relationship between status and earnings in categories of occupation and gender. Add titles and labels to the plot and save it as a PDF.




```
> pdf( file="status_coplot.pdf" )
> coplot( predict(mod2) ~ earnings|prestige*gender, data=status,
+         xlab="Earnings (actual)", ylab="Status (predicted)" )
> dev.off()
```

4 Exercise 4

Using PGSS data available in the file JustEarn (plain text file with columns separated with tabs):

Read the data into R. Select a subset containing complete cases only.

```
> pgss <- read.table( "../pgss1999in.tab", head=TRUE, sep="\t" )
> pgss <- pgss[ complete.cases(pgss), ]
```

For each subject in the reduced data set, compute the mean perceived earnings and the mean of just earnings. Is there a positive relationship between the two?

```
> ## Means of the actual earnings
> actl <- apply( subset(pgss, select=in5a:in5j), 1, mean )
> # Alternatively
> actl <- rowMeans( subset(pgss, select=in5a:in5j) )
> head(actl, 20) # first 20 values
```

```
    5     7    10    11    13    23    26    34    39    40    50
 9830  9330  6160  8980  9550  4590  7850  3070  3480  2455  1690
   51    53    56    57    58    61    62    64    67
 1560  3010 16650  1630  9770  3770  2320  3830 16220
```

```
> ## Mean of the just earnings
> just <- apply( subset(pgss, select=in6a:in6j), 1, mean )
> # Alternatively
> just <- rowMeans( subset(pgss, select=in6a:in6j) )
> head(just) # first 20 values
```

```
   5    7   10   11   13   23
5250 6350 5170 4930 5100 4700
```

Estimate a regression model that has the log of mean of just earnings as a DV, with the log of the mean of perceived earnings as an IV.

```
> mod3 <- lm( log(just) ~ log(actl) )
> summary(mod3)

Call:
lm(formula = log(just) ~ log(actl))

Residuals:
     Min       1Q   Median       3Q      Max
-1.93435 -0.27527 -0.01597  0.29672  3.06330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.20828    0.33300   9.635   <2e-16 ***
log(actl)    0.60861    0.03751  16.226   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5586 on 364 degrees of freedom
Multiple R-squared: 0.4197,        Adjusted R-squared: 0.4181
F-statistic: 263.3 on 1 and 364 DF,  p-value: < 2.2e-16
```

Interpret the values of the regression coefficients.

The estimate of the slope is 0.609 which means that increasing the IV by 1 unit translates into an increase in the value of the DV by 0.609 units on a logarithmic scale.

How strong is the relationship between the DV and IV?

R-squared for the model is 0.42 meaning that 42 per cent of the total variation in the DV can be attributed to the IV. The correlation coeficient for the two variables equals 0.65.