# Case Study 2: Model choice: Chile

*Richard Wilkinson*

This case study is included to make you think about what questions we might want to ask, and how we might answer them.

The data are from a survey of 2700 people ahead of the 1988 Chilean national referendum to determine whether Augusto Pinochet should stay in power or not (Pinochet was the military dictator of Chile from 1973 to 1990. He killed at least 2000 political opponents, and tortured many more. He was considered an important ally to Margret Thatcher.).

```
library(car)
Chile[1:10,]
```

```
##    region population sex age education income statusquo vote
## 1       N     175000   M  65         P  35000    1.0082    Y
## 2       N     175000   M  29        PS   7500   -1.2962    N
## 3       N     175000   F  38         P  15000    1.2307    Y
## 4       N     175000   F  49         P  35000   -1.0316    N
## 5       N     175000   F  23         S  35000   -1.1050    N
## 6       N     175000   F  28         P   7500   -1.0469    N
## 7       N     175000   M  26        PS  35000   -0.7863    N
## 8       N     175000   F  24         S  15000   -1.1135    N
## 9       N     175000   F  41         P  15000   -1.0129    U
## 10      N     175000   M  41         P  15000   -1.2962    N
```

```
str(Chile)
```

```
## 'data.frame':    2700 obs. of  8 variables:
##  $ region    : Factor w/ 5 levels "C","M","N","S",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ population: int  175000 175000 175000 175000 175000 175000 175000 175000 175000 175000 ...
##  $ sex       : Factor w/ 2 levels "F","M": 2 2 1 1 1 1 2 1 1 2 ...
##  $ age       : int  65 29 38 49 23 28 26 24 41 41 ...
##  $ education : Factor w/ 3 levels "P","PS","S": 1 2 1 1 3 1 2 3 1 1 ...
##  $ income    : int  35000 7500 15000 35000 35000 7500 35000 15000 15000 15000 ...
##  $ statusquo : num  1.01 -1.3 1.23 -1.03 -1.1 ...
##  $ vote      : Factor w/ 4 levels "A","N","U","Y": 4 2 4 2 2 2 2 2 3 2 ...
```

**region** is a categorical variable with levels M = metropolitan, S = south, N = north, etc.

**population** is a continuous variable denoting the number of people in the city where the respondent lives

**education** is a categorical variable with levels P = Primary, S = secondary, PS = post-secondary

**income** is the respondents yearly salary in Pesos

**vote** is their reported voting intentions, with Y = yes (Pinochet should stay in power), N = no, A = abstain, U = undecided

The other variables have their obvious meaning. Note that **region, sex, education** and **vote** are categorical variables i.e., they take a discrete value, whereas **income, age** and **population** are continuous variables.

The data currently have some missing values, represented as 'NA' by R. Lets remove these from the data set. Note that we can't use x==NA as NA is a special character. We have to test using is.na(x) instead.

```
library(dplyr)
ChileData = filter(Chile, !is.na(vote), !is.na(education), !is.na(age), !is.na(income))
```

## Contingency tables

Suppose we want to see whether education level has an effect on voting intention

- How can we test for this?

```
votebyed <- table(ChileData$education, ChileData$vote)
votebyed
```

```
##
##        A   N   U   Y
##   P   49 262 285 410
##   PS  31 220  48 123
##   S   99 385 223 304
```

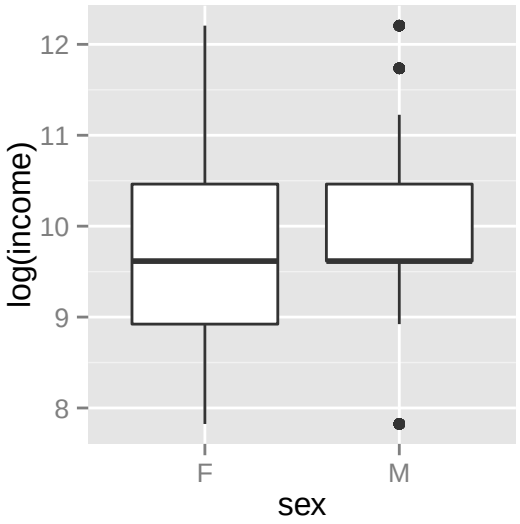
```
chisq.test(votebyed)
```

```
##
##  Pearson's Chi-squared test
##
## data:  votebyed
## X-squared = 135.3, df = 6, p-value < 2.2e-16
```

## One-way ANOVA

Lets ignore the voting part of the data. How would you test whether **income** was dependent upon **sex**?

```
library(ggplot2)
qplot(x=sex, y= log(income), data=ChileData, geom='boxplot')
```




```
fit1 <- lm(income~sex, data = ChileData)
fit0 <- lm(income ~ 1, data=ChileData)
anova(fit0, fit1)
```

```
## Analysis of Variance Table
##
## Model 1: income ~ 1
## Model 2: income ~ sex
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1   2438 3.84e+12
## 2   2437 3.84e+12  1  9.02e+09 5.73  0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is a *one-way analysis of variance (ANOVA)* model. Analysis of variance attempts to determine whether different groups have different mean responses, in this case, different average incomes between males and females. The idea is to look at the variability between groups (i.e. average salary of men vs salary of women), and the variability within a group (ie. salary variation amongst all males), and to see if the former is larger than one would expect if groups were the same.

This is a *one-way* ANOVA become there is just one discrete factor (sex).

## Two-way ANOVA

How would you test whether **income** was dependent upon **sex** after controlling for the effect of education?

```
fit2 <- lm(income~sex + education, data=ChileData)
fit3 <- lm(income~education, data=ChileData)
anova(fit3, fit2)
```

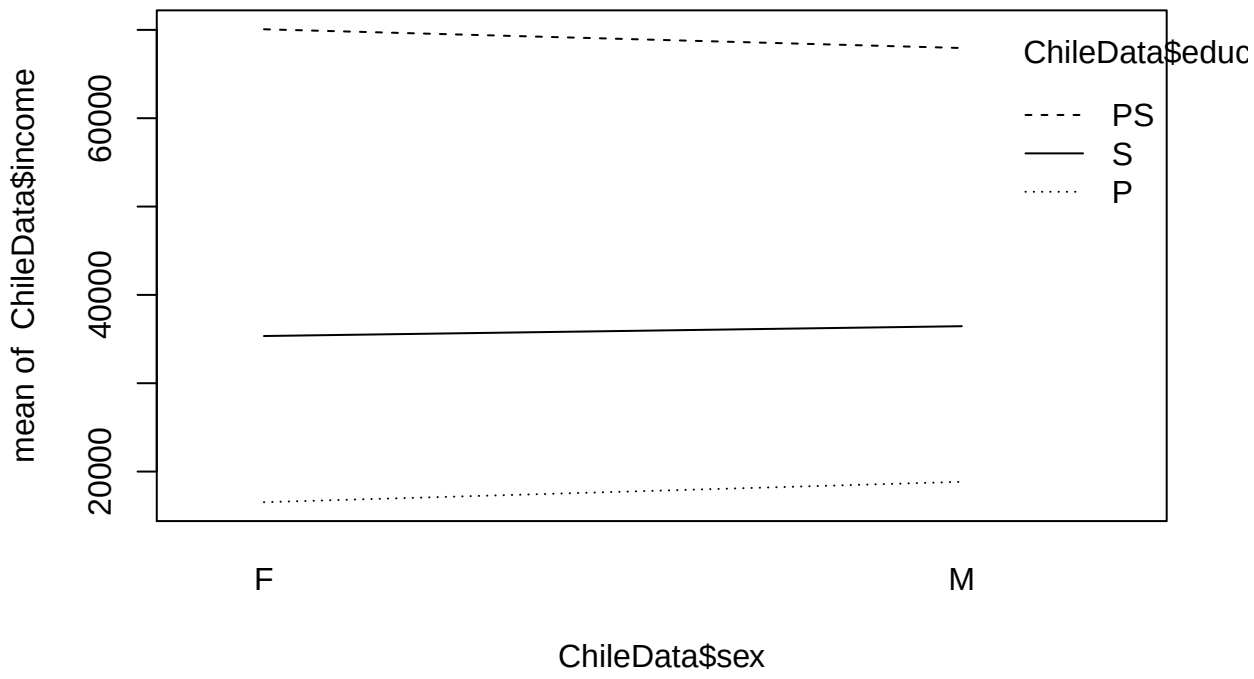
```
## Analysis of Variance Table
##
## Model 1: income ~ education
## Model 2: income ~ sex + education
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1   2436 3.06e+12
## 2   2435 3.06e+12  1  6.63e+08 0.53   0.47
```

This is a *two-way* ANOVA model as there are two discrete factors, sex and education.

**Two-way ANOVA with interaction**

Are the effects of sex and education additive or is there an interaction between them?

```
interaction.plot(ChileData$sex, ChileData$education, ChileData$income)
```




```
fit3 <- lm(income~sex *education, data=ChileData)
anova(fit3, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: income ~ sex * education
## Model 2: income ~ sex + education
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1   2433 3.05e+12
## 2   2435 3.06e+12 -2 -1.44e+09 0.57   0.56
```

### Simple linear regression

How would you test whether **income** was dependent upon **age**?

```
fit4 <- lm(income~age, data=ChileData)
summary(fit4)
```

```
##
## Call:
## lm(formula = income ~ age, data = ChileData)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -33155 -24022 -17714   1491 168478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37085.8     2247.6   16.50   <2e-16 ***
## age            -79.5       54.8   -1.45     0.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39700 on 2437 degrees of freedom
## Multiple R-squared:  0.000862,   Adjusted R-squared:  0.000452
## F-statistic:  2.1 on 1 and 2437 DF,  p-value: 0.147
```

## ANCOVA

How could you test whether **income** was dependent upon **education** after controlling for **age**?

```
fit5 <- lm(income~age + education, data=ChileData)
anova(fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: income ~ age
## Model 2: income ~ age + education
##   Res.Df      RSS Df Sum of Sq   F Pr(>F)
## 1   2437 3.84e+12
## 2   2435 3.02e+12  2  8.22e+11 332 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is an *analysis of covariance* model as there is a combination of continuous and discrete covariates.

## Summary

- The first type of analysis compared two categorical variables using **contingency tables**. The response variables was **vote** and the covariate was **education**.
- The second and third analyses had a continuous response variable **income**, but had categorical covariates (**sex** and **education**). We analysed this using a type of linear regression analysis called **Analysis of Variance (ANOVA)**.
- The fourth analysis had a continous resonse **income**, and one categorical covariate **education** and one continous covariate **age**. We analysed this using linear regression - a type of **Analysis of Covariance (ANCOVA)**.

| Explanatory variables | Response variable | Methods |
|---|---|---|
| Categorical | Categorical | Contingency |
| Categorical | Continuous | ANOVA |
| Continous | Continuous | Regression |
| Categorical and Continous | Continuous | ANCOVA |
| Continous | Categorical | Generalised regression (see G13MED) |

As we shall see, ANOVA and ANCOVA are really just special cases of regression.