Tutorial 4: Generalised Linear Models III

We need the **AER** and **car** packages.

Exercise 1

Consider a Gaussian linear regression model $Y = X\beta+\epsilon$. The OLS estimator of β is $\widehat{\beta} = (X'X)^{-1}X'Y$. The purpose of this exercise is to investigate how $\widehat{\beta}$ is computed. The formula $(X^tX)^{-1}Y$ is correct, but it turns out to be inefficient even for moderate values of p (the number of explanatory variables). The command `lm` does not use this formula for computing $\widehat{\beta}$, which is the minimiser of of $(Y-X\beta)^t(Y-X\beta)$.

Generate some artificial data by:

```
> x <- 1:20
> y <- x + rnorm(20)
```

Fit a polynomial in `x` for predicting `y`; that is, fit the model

$$Y = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \epsilon.$$

Compute $\widehat{\beta}$ in two ways, firstly by `lm()` and secondly by direct computation of $\widehat{\beta} = (X^tX)^{-1}X^tY$. At what degree of polynomial does the direct computation present difficulties? What about `lm()`?

`lm()` uses a different minimisation procedure which is more efficient than the formula. In fact, the formula is hardly ever the best way to proceed, particularly when assumptions about linear independence of the columns of X cannot be made.

Exercise 2

The data set `CPS1988` is found in the **AER** package. Read the description and activate it.

```
> library("AER")
> data("CPS1988")
```

To get a summary of the data, type

```
> summary(CPS1988)
```

1. Regress the logarithm of the wage on all available regressors plus experience squared.

   ```
   > cps_lm <- lm(log(wage)~experience + I(experience^2)+education +
   ethnicity, data=CPS1988)
   > summary(cps_lm)
   ```

What is the return on `education` for this model? (Answer: 8.57%)

2. We now test the relevance of `ethnicity` using the function `anova()`. Create a model without `ethnicity`

```
> cps_noeth <- lm(log(wage)~experience + I(experience^2)+education,
data = CPS1988)
> anova(cps_noeth,cps_lm)
```

What are your conclusions? Interpret the results of the command:

```
> anova(cps_lm)
```

3. Now try creating the model without `ethnicity` using the `update` command:

```
> cps_noeth2 <- update(cps_lm, formula = .~.-ethnicity)
```

and verify that this gives the same result.

4. Now use the package **lmtest** to get a Wald test.

```
> install.packages("lmtest")
> library("lmtest")
> waldtest(cps_lm,.~.-ethnicity)
```

What are your conclusions?

Exercise 3

Consider the data set `PSID1982` in the **AER** package.

1. Regress the logarithm of the wage on all regressor variables plus experience squared.
2. Perhaps $\log Y$ is not the best way of transforming the data. Box and Cox considered transformations of the form $\frac{Y^\gamma - 1}{\gamma}$. Look it up under 'Power transform' in wikipedia; the $\log Y$ transformation can be considered as a Box-Cox transform with $\gamma = 0$.

   The function `boxcox()` in **MASS** tests for an optimal value of γ. Use this for deciding on the best transformation. Is the logarithm a good choice of transformation? If not, suggest a better one.
3. Does gender interact with education and / or experience?
4. Consider the possible models including and excluding interactions. Which model seems best to you? Consider R^2 adjusted, AIC and BIC criteria for deciding between models with different numbers of parameters.

Exercise 4: Partially Linear Models

Consider the data set `CPS1988` from the **AER** package. This is the 'current population survey' from 1988 collected by the US Census Bureau. It has 28155 observations on males aged 18 to 70 with positive annual incom greater than $ 50 in 1992 who are no self-employed or working without pay. Wages are deflated by the deflator of personal consumption expenditure for 1992. The variables are: `wage`, `education`, `experience`, `ethnicity`.

1. Fit a model

$$\log(\texttt{wage}) = \beta_1 + g(\texttt{experience}) + \beta_2\texttt{education} + \beta_3\texttt{ethnicity} + \epsilon$$

   where g is an unknown function, estimated using *splines*. Here, we don't know the function; we try to build up the best function, so that the effect of `experience` can be removed; we're interested in the other variables. Try:

```
> library("splines")
> cps_plm <- lm(log(wage)~bs(experience,df=5)+education+ethnicity,data=CPS1988)
```

2. For choice of spline, try to find the best degree of freedom, using the AIC score. This may be done as follows:

```
> cps_bs<-lapply(3:10,function(i) lm(log(wage)~bs(experience,df=i)
+education+ethnicity,data=CPS1988))
> structure(sapply(cps_bs,AIC,k=log(nrow(CPS1988))),.Names=3:10)
```

   What do you conclude?

3. The following plots the quadratic function in `experience` as a dashed line and the cubic spline as a full line.

```
> cps<-data.frame(experience=-2:60, education=with(CPS1988,
mean(education[ethnicity=="cauc"])),ethnicity="cauc")
> cps$yhat1<-predict(cps_lm,newdata=cps)
> cps$yhat2<-predict(cps_plm,newdata=cps)
> plot(log(wage)~jitter(experience,factor=3),pch=19,
col=rgb(0.5,0.5,0.5,alpha=0.02),data=CPS1988)
> lines(yhat1~experience,data=cps,lty=2)
> lines(yhat2~experience,data=cps)
> legend("topleft",c("quadratic","spline"),lty=c(2,1),bty="n")
```

   Type in the commands and interpret the output.

4. What happens if we use the BIC instead of the AIC (the BIC has greater penalisation. On large data sets, it may give a smaller model).

Exercise 5

We consider the `CPS1988` data set indicating wages against experience, education and ethnicity.

1. Recall the basic model (without interactions):

```
> cps_lm<-lm(log(wage)~experience+I(experience^2)+education+ethnicity,data=CPS1988)
```

2. Consider a model where the categorical variables (education and ethnicity) only affect the intercept.

```
> cps_int <- lm(log(wage)~experience + I(experience^2)+
education*ethnicity, data=CPS1988)
> coeftest(cps_int)
```

   Interpret the results.

3. Now consider separate regressions for each level:

```
> cps_sep <- lm(log(wage)~ethnicity/(experience + I(experience^2)
+ education)-1, data = CPS1988)
```

   Interpret the syntax. This model specifies that the terms within the parentheses are nested within `ethnicity`, hence the intercept is not needed. It is best replaced by two separate intercepts, one for each level of ethnicity. The R^2 value is computed differently in the summary. Type

```
> ?summary.lm
```

   The estimated coefficients for the two groups defined by the levels of `ethnicity` are found as follows:

```
> cps_sep_cf <- matrix(coef(cps_sep),nrow=2)
> rownames(cps_sep_cf)<-levels(CPS1988$ethnicity)
> colnames(cps_sep_cf)<-names(coef(cps_lm))[1:4]
> cps_sep_cf
```

4. For any regression containing an unordered factor, R by default uses the first level of the factor as the reference category. In `CPS1988`, ''`cauc`'' is the reference category for `ethnicity`. The reference category can be changed using the `relevel()` command. Try the following and interpret the syntax and output:

```
> CPS1988$region <- relevel(CPS1988$region, ref="south")
> cps_region <- lm(log(wage)~ethnicity+education+experience+I(experience^2)
+region,data=CPS1988)
```