ASSIGNMENT 1

Instructions

The instructions for the assignment are as follows.

1. The assignment is due on 22 March (23:59).
2. Hand in your assignment in Canvas.
3. Although you are allowed to hand in handwritten answers for the first part, we highly encourage you to use R Markdown throughout the exercise. You can write both text and equations, and a lot of help on how to use it can be found online.[1] From the Markdown file, you can generate a PDF file by clicking on the "Knit" button, which then is nicely formatted and shows clear answers. Furthermore, we provide a template solution file, which you can then fill in accordingly. In case you get a package error on the university computers, please follow the following steps:
    1. Delete the folder `M:\R` (this folder should only contain R packages)
    2. Run the following code:

```
packageurl = "https://cran.r-project.org/src/contrib/Archive/caTools/caTools_1.14.tar.gz"
install.packages(packageurl, repos=NULL, type="source")
```

    If you use your own computer and run into this error, we recommend reinstalling R completely.
4. When answering the empirical exercise, the code used and the results (i.e. values of statistics) from the code need to be clearly stated and linked to your answers. In particular, first state the R code you use, then provide the values of the statistics you calculate below the code, and beneath that provide your answer that relates to the code. Again, we encourage you to use R Markdown, since it combines code with text in a readable format.
5. Using built-in R functions is not allowed. Build up everything from basic matrix algebra. For simple operations such as calculating a mean, and for operations related to distributions (e.g. taking random draws, getting the quantile), you can use the built-in R functions.
6. Work in groups of three.
7. State your full name and your student number on the front page of the assignment.

[1] E.g. http://www.stat.cmu.edu/~cshalizi/rmarkdown/ provides a great summary of all necessary commands.

Theoretical Exercises

1. Stated below are three estimators of $\mu = \mathbb{E}[y_i]$. For each estimator, show whether the estimator is unbiased and consistent.

   a) $\hat{\mu} = \frac{1}{n+1} \sum_{i=1}^{n} y_i$.
   b) $\hat{\mu} = \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} y_i$. Assume that n is even.
   c) $\hat{\mu} = \frac{0.1}{100} \sum_{i=1}^{100} y_i + \frac{0.9}{n-100} \sum_{i=101}^{n} y_i$. Assume that $n > 100$.

2. Consider the multiple linear regression model with k regressors. The sum of squared errors as a function of $\boldsymbol{\beta}$ can be expressed as $S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Show using the first-order conditions that the solution of the minimisation of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is equal to the OLS estimator $\hat{\boldsymbol{\beta}}$. Furthermore, show that this solution is unique.

3. Assume that the true linear regression model of interest is given by

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

   where i indexes observations, and there are n observations available so that $i = 1, \ldots, n$. $\boldsymbol{x}_i$ is a 2×1 column vector containing the explanatory variables x_{i1} and x_{i2} for individual i. $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ denote $n \times 1$ column vectors containing the n observations of each of these variables. $\boldsymbol{\beta}$ is the vector of coefficients. The unobservable ε_i is an independent and identically distributed error, has zero mean and a finite variance. $\boldsymbol{\beta}$ contains the true coefficients β_1 and β_2. Let $\hat{\beta}_1$ be the OLS estimator of β_1. Using the projection matrix for $\boldsymbol{x}_2$, show formally that if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are orthogonal to each other, $\hat{\beta}_1$ simplifies to $(\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{y}$, where $\boldsymbol{y}$ is a $n \times 1$ vector containing the observations y_i. Explain the implication of this result for the analysis of the 'partial' effect of $\boldsymbol{x}_1$ on y in this regression. Note: this question is a past exam question.

4. Assume that the true linear regression model of interest is given by

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

   where i indexes observations, and there are n observations available so that $i = 1, \ldots, n$. $\boldsymbol{x}_i$ is a $K \times 1$ column vector containing the constant term x_{i0}, and the explanatory variables x_{i1} and x_{i2}. $\boldsymbol{\beta}$ is the vector of coefficients. Let β_1 denote the coefficient of x_1, and $\hat{\beta}_1$ be the OLS estimator of β_1. Assume that the model satisfies all the assumptions of the standard linear regression model.

   a) *Using Monte Carlo simulation*, demonstrate the effect of multicollinearity (correlation between regressors) on the sampling distribution of the OLS estimator. In the simulation, consider two (or more) levels of correlation between the two independent variables. At a low and a high level of correlation, plot the two sampling distributions of the OLS estimator. Hint: Mimic the Monte Carlo simulation exercises you have studied in the lab. In the simulation, you need to draw random values for the two independent variables from a multivariate distribution so that the two independent variables are correlated. To do this, install and use the package `mvtnorm`. An example is provided in the template file.
   b) Based on the plots created in a), explain how multicollinearity affects (i) the unbiasedness property of the OLS estimator, (ii) the standard error of the OLS estimator, (iii) the t-statistic.

Empirical Exercise

For this exercise use the enclosed data in 'RData' format native to R. The data is based on a sample of young males, and it is collected in 1980 in the National Longitudinal Survey. It includes, among others, information on wage, education, work experience, race, region, and IQ score for 935 employees. The data is a subset of the data used by *Blackburn, M. and Neumark, D., 1992. Unobserved ability, efficiency wages, and interindustry wage differentials, Quarterly Journal of Economics, 107, 1421-1436.* An overview of the data is given in the table provided at the end.

The regression model of interest is:

$$wage_i = \beta_0 + \beta_1\, educ_i + \beta_2\, exper_i + \beta_3\, IQ_i + \beta_4\, age_i + \beta_5\, black_i + u_i$$

1. List the assumptions needed to consistently estimate the coefficients of this model using the OLS estimator. For each assumption, briefly explain why you need it.
2. Using the Frisch-Waugh-Lovell formula, compute *only* the OLS coefficient estimate of education. Interpret this estimate. Hint: Partition variables into two components, such that $\boldsymbol{X_1}$ contains only `educ` and $\boldsymbol{X_2}$ all other variables. Based on these components, construct the required projection matrices for the FWL formula.
3. Compute the estimate of the variance of the regression error ($\boldsymbol{s^2}$). State the value of this statistic.
4. Compute the estimate of the variance-covariance matrix of the OLS coefficient estimates. Present this matrix. What is the estimated standard error of $\hat{\boldsymbol{\beta}}_{\boldsymbol{3}}$? Explain what this estimate represents using the notion of the 'sampling distribution' of the OLS estimator.
5. It is claimed that innate ability, proxied by the IQ score, has a positive effect on wages. Conduct a formal hypothesis test to test this claim. When conducting the test, clearly state the null and the alternative hypotheses, the test statistic, and the value of the test statistic. Conduct the test at a significance level of 5 percent.
6. Using a formal hypothesis test, test the joint hypothesis that the effect of education is equal to 10 and the effect of experience is equal to 5. When conducting the test, clearly state the null and the alternative hypotheses, the test statistic, and the value of the test statistic. Conduct the test at a significance level of 5 percent.

```
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
wage            int    %9.0g                  monthly nominal earnings
hours           byte   %9.0g                  average weekly hours
IQ              int    %9.0g                  IQ score
KWW             byte   %9.0g                  knowledge of world work score
educ            byte   %9.0g                  years of education
exper           byte   %9.0g                  years of work experience
tenure          byte   %9.0g                  years with current employer
age             byte   %9.0g                  age in years
married         byte   %9.0g                  =1 if married
black           byte   %9.0g                  =1 if black
south           byte   %9.0g                  =1 if live in south
urban           byte   %9.0g                  =1 if live in SMSA
sibs            byte   %9.0g                  number of siblings
brthord         byte   %9.0g                  birth order
meduc           byte   %9.0g                  mother's education
feduc           byte   %9.0g                  father's education
```