ECONOMETRICS I: PROBLEM SET 2

EXERCISE 1: FINITE PROPERTIES

a) Suppose that $E[e|X] = 0$ and $var[e|X] = \Omega$. Prove that:

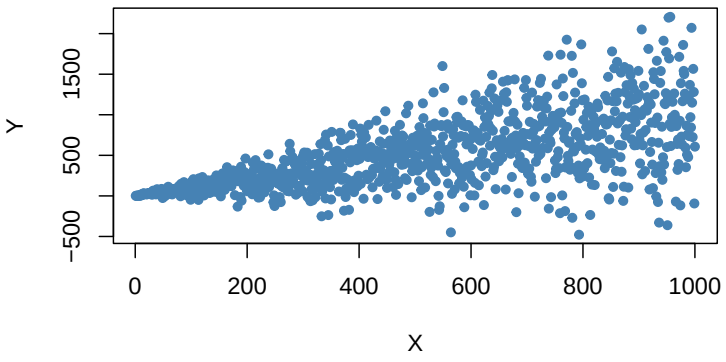
$$E[\hat{\beta}|X] = \beta;$$

$$var[\hat{\beta}|X] = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$$

b) The parameter β is esimtated by OLS $\hat{\beta} = (X'X)^{-1}X'Y$ and GLS $\tilde{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$. Let $\hat{e} = Y - X\hat{\beta}$ and $\tilde{e} = Y - X\tilde{\beta}$ denote the residuals. How would you compute $\hat{R}^2$ and $\tilde{R}^2$. If the error e_i is truly heteroskedastic will $\hat{R}^2$ or $\tilde{R}^2$ be smaller?

c) Follow the example in R below. Obtain the linear projection of Y on X using OLS. Find robust standard errors (hint: use the `lmtest` package) and show empirically that $HC1 < HC2 < HC3$.

```r
# Generate heteroskedastic data
X <- 1:1000
Y <- rnorm(n = 1000, mean = X, sd = 0.6 * X)
# Plot the data
plot(x = X, y = Y,
     pch = 19, col = "steelblue", cex = 0.8)
```



d) Use the dataset `cps09mar.txt` and follow the provided R code in BH p.91. Obtain the HC0 to HC3 errors.

- Modify h_{ii} so that it has a value close to 0.4 and obtain HC0 to HC3 again. What conclusions do you reach?

e) Obtain the clustered standard errors in matrix form using the dataset `DDK2011.txt`. Similar to what you did in d).

f) Load the dataset `PublicSchools` from the `sandwich` package. Fit a simple linear regression model using OLS where `Expenditure` is the dependent variable and `Income` is the independent variable.

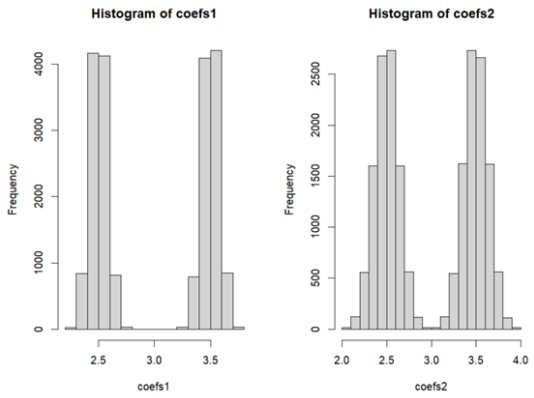
- Perform a regression diagnostic (you may use functions like `plot()`).
- Interpret the results of your diagnostics. Is there any state that disproportionately influences the model? How could you modify the model to handle these influences?

g) Load the dataset `Journals` from Stock and Watson (2007), which contains information on subscriptions to economics journals in U.S. libraries, using the `AER` package.

- Create a new `data frame` containing only the variables of interest: number of subscriptions (`subs`), subscription price (`price`), and calculated price per citation (`citeprice`).
- Transform the variables `subs` and `citeprice` using the natural logarithm for analysis.
- Fit a linear regression model using OLS where the logarithm of the number of subscriptions is explained by the logarithm of the price per citation.
- Perform a regression diagnostic.
- Interpret the results of the diagnostics.
- Report the correct standard errors for the estimated coefficients.

h) In the context of a national survey on satisfaction with public health services, a cluster sampling design was used. Researchers initially selected certain geographic areas (primary sampling units) and then randomly selected households within these areas for interviews. Discuss why it is important to use clustered standard errors in estimating the coefficients of a regression model analyzing the determinants of satisfaction with health services.

i) Based on the example below, run two simulations with $n = 500$ and 10,000 repetitions to reproduce, as closely as possible, the following plots:



Example: Here we create a matrix X with two vectors (regressors X_1 and X_2) with median 50 and 100 each, and $cov(X_1, X_2) = 2.5$

```
library(MASS)
X <-rmvnorm(n, c(50, 100), sigma = cbind(c(10, 2.5), c(2.5, 10)))
summary (X)
plot(X[,1], X[,2])
```

j) Explain how the variance changes according to the parameters n, σ^2, and ρ, following the definition of var$[\hat{\beta}|X]$ in BH p.121.

EXERCISE 2: MLE

a) Show that, under $e \sim N(0, \sigma^2)$ MLE estimators of β and σ^2 are equivalent to OLS.

b) Execute the following R code, which serves as an example to obtain robust standard errors for heteroscedasticity and the corresponding hypothesis t-tests.

Example:

```
#Environment:
library(AER)              # install.packages("AER")
data("CPSSWEducation")
attach(CPSSWEducation)    # ?CPSSWEducation

# Model
reg <- lm(earnings ~ education)
summary(reg)

# Plot observations and add the regression line
plot(education, earnings, ylim = c(0, 150))
abline(labor_model, col = "steelblue", lwd = 2)

# Compute homoskedastic-robust standard errors.
t <- linearHypothesis(reg, "X = 0")$'Pr(>F)'[2] < 0.05

# Compute heteroskedasticity-robust (HC1) standard errors
t.rob <- linearHypothesis(reg,
                          "X = 0", white.adjust = "hc1")$'Pr(>F)'[2] < 0.05

# Show both t-tests, where 1="true" meaning Ho is "true" at the 95% level.
round(cbind(t = mean(t), t.rob = mean(t.rob)), 3)

# Same fot the varcovar matrix
(vcov <- vcovHC(labor_model, type = "HC1"))

# Compute the square root of the diagonal elements in vcov
(robust_se <- sqrt(diag(vcov)))
```

```r
# We use `coeftest()` on our robust model and show the assumed-homoskedastic
# model:
coeftest(labor_model, vcov. = vcov)
summary(labor_model)
```

- Generate heteroscedastic data for two variables (X) and (Y) (as in 1.c), which satisfy the population equation $Y = \alpha + \beta X + e$, where $\beta = 1$.
- Create a scatterplot of these variables with a fitted line showing their relationship.
- Perform 10,000 regressions, each with new random samples drawn from the same distribution established in (1.c).
- For each $i = 1, \ldots, 10,000$, perform the "HC0" and "HC2" t-tests and store them in vectors `t` and `t.rob`.
- Compute the percentage of rejections of the null hypothesis $\beta = 1$.
- What is the conclusion of this result?

c) MLE is the technique that helps us determine the parameters of the distribution that best describe the given data. Imagine that we have a sample that was drawn from a normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 100$. The objective is to estimate these parameters with MLE.

The normal log-likelihood function is given by: $l = -\frac{1}{2}nln(2\pi) - \frac{1}{2}nln(\sigma^2) - \frac{1}{2\sigma^2}\sum(yi-\mu)^2$

*Note that minimizing a negative likelihood function is the same as maximizing the likelihood function.

```r
# We define:
X <- rnorm(n = 1000, mean = 5, sd = 10)
df <- data.frame(X)

# We program the log-likelihood function in R:
normal.lik1 <- function(theta,y){
  mu<-theta[1]
  sigma2<-theta[2]
  n<-nrow(y)
  logl<- -.5*n*log(2*pi)-.5*n*log(sigma2)-(1/(2*sigma2))*sum((y-mu)^2)
  return(-logl)
}

# Here theta is a vector containing the two parameters of interest
# (i.e. theta[1] is equal to mu). The remainder sets n, and the log-likelihood function.

# we use optim(starting values, log-likelihood, data) with starting values 0 and 1.
optim(c(0,1), normal.lik1, y=df)

#We can ask for the method-of-moments-mean directly:
mean(X)
var(X)
```

- Now, estimate the MLE parameters β and σ^2 with Y = 5 + 2X + e.