ECONOMETRICS I: PROBLEM SET 2

EXERCISE 1: FINITE PROPERTIES

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

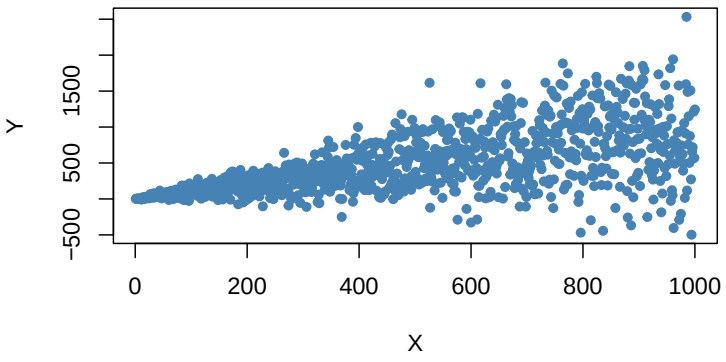
$$E[\hat{\beta}|X] = \beta;$$

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

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.

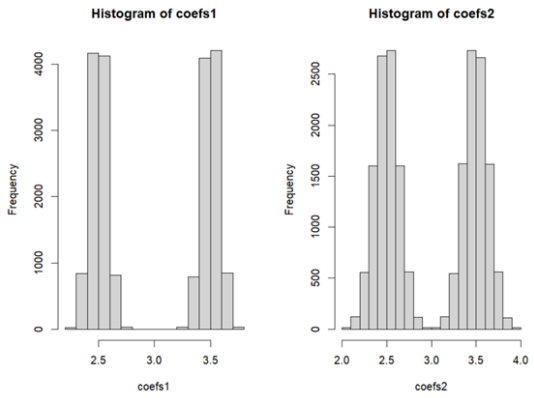
- Perform a regression diagnostic (you may use functions like `plot()`).
- Interpret the results of your diagnostics. Is there any state that disproportionately influences the model? How could you modify the model to handle these influences?

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.

k) Show that $E[\hat{V}_{\hat{\beta}}^{HC0}|X]$ is biased under homoscedasticity. Additionally, demonstrate that $HC2$ is unbiased under homoscedasticity.

l) Take the linear homoskedastic CEF $Y^* = X'\beta + e$; $E[e|X] = 0$; $E[e^2|X] = \sigma^2$. Suppose that Y* is measured with error. Instead, we observe $Y = Y^* + u$ where u is a measurement error, e and u are independent and $E[u|X] = 0$; $E[u^2|X] = \sigma_u^2(X)$. Describe the effect if any of this measurement error on variance calculation for $\hat{\beta}$.