ECOM-G314 Econometrics 1
Homework Assignment 2

This version of the assignment includes some additional links to the course material to help you find where a particular topic is being discussed. It should otherwise be identical to the original assignment.

Update 2024-11-25 11:55pm. The hint to problem 3(b) has been corrected. The dependent variable in R can be formed as `y <- log(s) - lag(log(f3),-3)`. The original hint incorrectly used the 3-period lag of the 1-period forward rate `f1`. Using the original hint should not be penalized in grading.

1. (Adapted from Verbeek, Exercise 4.1) This exercise uses data on 30 standard metropolitan statistical areas (SMSAs) in California for 1972 in the file `airq2.xlsx` containing the following variables: [40%]

   - *airq*: indicator for air quality (the lower the better)
   - *vala*: value added of compaies (in 1000 US$)
   - *rain*: amount of rain (in inches)
   - *coas*: dummy variable; 1 for SMSAs at the coast, 0 for others
   - *dens*: population density (per square mile)
   - *medi*: average income per head (in US$)

   (a) Estimate a linear regression model that explains *airq* from the other variables using ordinary least squares. Interpret the coefficient estimates.

   (b) Test the null hypothesis that average income does not affect the air quality. Test the joint hypothesis that none of the variables has an effect on air quality.

   (c) Perform a Breusch-Pagan test for heteroskedasticity related to all five explanatory variables.

   > **References**: code for "wages" (updated, week 4), slides "Generalised least squares estimator & heteroskedasticity" (p. 10), book chapter 4.4.2.

   (d) Perform a White test for heteroskedasticity. Comment upon the appropriateness of the White test in light of the number of observations and the degrees of freedom of the test.
   Note: some sources and R packages may call a different test a "White test". To keep the review process simple, please use the White test as defined in any of the following sources:

   - The textbook chapter 4.4.3, "The White Test"
   - Lecture slides on the generalized least squares estimator and heteroskedasticity (p. 10)
   - (implicitly defined) `wages.R` code under Week 4

References: code for "wages" (updated, week 4), slides "Generalised least squares estimator & heteroskedasticity" (p. 10), book chapter 4.4.3.

In a `formula` given to the `lm` function in R, note that `I(x^2)` will include covariate `x` squared in a regression, whereas `(x1+x2+x3)^2` includes covariates x_1 through x_3 *and* their crossings. See `?formula` and `?I` for details.

(e) Estimate the heteroskedasticity-consistent covariance matrix. Do the conclusions change compared to the case were the standard error are based on the assumption of homoskedasticity? Test the hypothesis that *rain* and *dens* jointly have no effect on air quality.

References: code for "wages" (updated, week 4), slides "Generalised least squares estimator & heteroskedasticity" (p. 8–9), book chapter 4.3.4.

(f) Assuming that we have multiplicative heteroskedasticity related to *coas* and *medi*, estimate the coefficients by running a regression of $\log(e_i^2)$ on these two variables. Are these variables related to heteroskedasticity (test the null hypothesis that their coefficients are jointly equal to zero)?

References: book chapters 4.4.1 and 4.3.5.

(g) Using the results from (f), estimate the model by feasible GLS. Compare your results with those obtained in (a). Redo the tests in (b).

References: slides "Generalised least squares estimator & heteroskedasticity" (p. 5–7), book chapter 4.3.5 (also 4.3.3, 4.3.2).

2. (Adapted from Verbeek, Exercise 4.2) Consider the example model for the demand ice cream covered in the lecture on Time series and autocorrelation. The file `icecream2.xlsx` holds the corresponding data. Extend the model by including lagged consumption (rather than lagged temperature). [15%]

The model is

$$cons_t = \beta_1 + \beta_2 income_t + \beta_3 price_t + \beta_4 temp_t + \varepsilon_t,$$

where $cons_t$ stands for ice cream consumption, $income_t$ for income, $price_t$ for price and $temp_t$ for temperature at time t. (See the lecture slides or the book chapter 4.8 for a more detailed description of the data.)

(a) Perform a test for first-order autocorrelation in this extended model.

References: code for "icecream" (week 4), slides "Time Series Data & Autocorrelation" (p. 11 in particular), book chapters 4.6 and 4.7.1.

(b) Compute the HAC covariance matrix estimator. Test the significance of each of the coefficients of the model using the robust standard errors, and compare the conclusions to those based on t-tests assuming homoskedasticity. Select the lag order by the automatic procedure provided by the `sandwich` package.

> **References**: code for "icecream" (week 4), slides "Time Series Data & Autocorrelation" (p. 10 in particular), book chapters 4.6 and 4.10.2.

3. Let s_t denote the log of spot exchange rate of the British pound against the euro (GBP/EUR) in month t, and f_t^1 is the corresponding log of the one-month forward exchange rate. There is no risk premium in the GBP/EUR exchange market if the conditional expectation of s_t conditional on the information available in period $t-1$ equals f_{t-1}^1. This can be tested by testing the hypothesis $H_0 : \beta = (\beta_1, \beta_2)' = 0$ in the following regression model: $s_t - f_{t-1}^1 = \beta_1 + \beta_2(s_{t-1} - f_{t-1}^1) + \varepsilon_t$. The regressor $s_t - f_t^1$ is called the one-month forward discount. Likewise, the absence of risk premia in the three-month market can be tested by running the regression $s_t - f_{t-3}^3 = \beta_1 + \beta_2(s_{t-3} - f_{t-3}^3) + \varepsilon_t$, where f_t^3 the log of the three-month forward exchange rate, and testing $H_0 : \beta = (\beta_1, \beta_2)' = 0$. The regressor $s_t - f_t^3$ is called the three-month forward discount. [20%]

   The file `forward2.txt` contains monthly observation of the GBP/EUR spot and 1- and 3-month forward rates from January 1979 to December 2001. You may use the R code below to load the data and form the forward discount time series if you like.

```
dataset <- read.table("forward2.txt", header=TRUE)
s  <- ts(dataset$EXEURBP, frequency = 12, start = c(1979, 1))
f1 <- ts(dataset$F1EURBP, frequency = 12, start = c(1979, 1))
f3 <- ts(dataset$F3EURBP, frequency = 12, start = c(1979, 1))

fd1 <- log(s) - log(f1)
fd3 <- log(s) - log(f3)
```

   (a) Estimate the model $s_t - f_{t-1}^1 = \beta_1 + \beta_2(s_{t-1} - f_{t-1}^1) + \varepsilon_t$, and test for the existence of risk premia in the one-month forward market by testing $H_0 : \beta = (\beta_1, \beta_2)' = 0$. Test for first-order autocorrelation by the Breusch-Godfrey test, and use the appropriate covariance matrix estimator in the test on β. [Hint: The dependent variable is formed as `y <- log(s) - lag(log(f1),-1)` in R.]

> **References**: code for "icecream" (week 4), slides "Time Series Data & Autocorrelation", book chapters 4.6 and 4.7.1.
>
> If your data has N (rows of) observations, you have $N-1$ available lags. This means both your dependent variable and the regressor should have $N-1$ available elements. Note that `dynlm` allows for a special lagging construct on the right hand side, namely `y ~ L(x,1)`, as in the ice cream example; see the help page for details.

(b) Estimate the model $s_t - f^3_{t-3} = \beta_1 + \beta_2(s_{t-3} - f^3_{t-3}) + \varepsilon_t$, and test for the existence of risk premia in the three-month forward market by testing $H_0 : \beta = (\beta_1, \beta_2)' = 0$. Because the model involves the three-month difference between the spot and forward rates, but it is estimated on monthly data, the error term, by construction, exhibits first- and second-order autocorrelation (see Verbeek, Section 4.11). Test for first and second-order autocorrelation by the Breusch-Godfrey test. Test H_0 using both the conventional covariance matrix estimator assuming homoskedasticity and the HAC covariance matrix estimator, and compare the results. [Hint: The dependent variable is formed as `y <- log(s) - lag(log(f3),-3)` in R.]

> **References**: code for "icecream" (week 4), slides "Time Series Data & Autocorrelation", book chapters 4.7.1 and 4.10.2.
>
> You can use the function `linearHypothesis` as in the example code. You can either specify the restriction matrices directly (as in the code), and this may be instructive. You can also implicitly use the syntax allowed by `linearHypothesis` from the previous homework; simply use the same names for the coefficients as you see under `summary(lm(...))` to set restrictions.

4. Consider the simple linear regression model [25%]

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Let $\beta_1 = \beta_2 = 1.0$. Assume that x_i is independently normally distributed with mean zero and variance $\sigma^2_X = 1$, and the error term ε_i is independently normally distributed with mean zero and variance $\sigma^2_\varepsilon = 1$.

(a) Consider three sample sizes, $N = 10$, $N = 50$ and $N = 100$. For each sample size, generate $S = 5000$ samples from the regression model, and for each generated sample, compute the OLS estimate of β_2 and the t-test statistic for $H_0 : \beta_2 = 1.0$ against $H_1 : \beta_2 \neq 1.0$.

Because 1.0 is the true value of β_2, H_0 should be rejected in $p\%$ of the replications in the t-test conducted at the $p\%$ level of significance (the nominal size of the test). Compute the rejection rate of the test, i.e., find the proportion of the replications where the absolute value of the t-test statistic exceeds the critical value. Consider two values of $p\%$, 10% and 5% (with critical values 1.64 and 1.96, respectively). How do the rejection rates vary with the sample size N? [Hint: If the model is estimated using the `lm()` function in R and the result is stored in `ols1`, the OLS estimate of β_2 is obtained as `coef(ols1)[2]` and the covariance matrix estimator of the OLS estimator as `vcov(ols1)`.]

(b) Repeat (a), but test (i) $H_0 : \beta_2 = 1.2$ against $H_1 : \beta_2 \neq 1.2$, and (ii) $H_0 : \beta_2 = 1.5$ against $H_1 : \beta_2 \neq 1.5$. The rejection rates can be interpreted as the power of the test against the null hypotheses that involve values of β_2 deviating from the true value 1.0. How does the power behave as a function of the sample size N, the nominal size of the test $p\%$, and the null hypothesis ($H_0 : \beta_2 = 1.2$ and $H_0 : \beta_2 = 1.5$)?

References: slides "Testing hypotheses under the normality assumption" (week 1, p. 10), book chapter 2.5.7.

The coding part is very similar to the previous homework.