Econometrics

Homework 3

Problem 1. Consider a simple linear regression model:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + U_i,\ i = 1, \dots, n; \\
\beta_0 &\neq 0; \\
E\left(U_i | X_1, \dots, X_n\right) &= 0.
\end{aligned}
$$

Define

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right) Y_i}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \\
\tilde{\beta}_1 &= \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \text{ and } \tilde{\beta}_0 = 0,
\end{aligned}
$$

where $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Define also

$$
\begin{aligned}
\hat{U}_i &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i, \\
\tilde{U}_i &= Y_i - \tilde{\beta}_1 X_i.
\end{aligned}
$$

For each of the following statements, indicate true or false and explain your answers.
(a) $\sum_{i=1}^{n} \hat{U}_i = 0$.
(b) $\sum_{i=1}^{n} \tilde{U}_i = 0$.
(c) $\sum_{i=1}^{n} U_i = 0$.
(d) $E\left(U_i X_i^4\right) = 0$.
(e) In this model, $\hat{\beta}_1$ is the OLS estimator, and therefore the Gauss-Markov Theorem implies that

$$
Var\left(\hat{\beta}_1 | X_1, \dots, X_n\right) \leq Var\left(\tilde{\beta}_1 | X_1, \dots, X_n\right).
$$

Assume that errors U_i's are homoskedastic and there is no serial correlation.

Problem 2. (Wooldridge 2.10) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS intercept and slope estimators, respectively, and let $\bar{U}$ be the sample average of the errors U_i, $i = 1, ..., n$.

1. Show that $\hat{\beta}_1$ can be written as $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i U_i$ where $w_i = d_i / SST_X$, $d_i = X_i - \bar{X}$ and $SST_X = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$.

2. Use part (i), along with $\sum_{i=1}^{n} w_i = 0$, to show that $\hat{\beta}_1$ and $\bar{U}$ are uncorrelated. Hint:You are being asked to show that $E\left[\left(\hat{\beta}_1 - \beta_1\right) \cdot \bar{U}\right] = 0$. Show first that

$$
\left(\hat{\beta}_1 - \beta_1\right) \bar{U} = \frac{1}{n} \left(\sum_{i=1}^{n} w_i U_i\right) \left(\sum_{i=1}^{n} U_i\right) = \frac{1}{n} \left(\sum_{i=1}^{n} w_i U_i^2 + \sum_{i=1}^{n} \sum_{j \neq i} w_i U_i U_j\right).
$$

3. Show that $\hat{\beta}_0$ can be written as $\hat{\beta}_0 = \beta_0 + \bar{U} - \left(\hat{\beta}_1 - \beta_1\right)\bar{X}$.

4. Use parts (ii) and (iii) to show that (conditional on X's) $Var\left(\hat{\beta}_0\right) = \sigma^2/n + \sigma^2\bar{X}^2/SST_X$. Hint: Show that

$$Var\left(\bar{U}\right) = \frac{1}{n^2}E\left(\sum_{i=1}^{n} U_i\right)^2 = \frac{1}{n^2}E\left(\sum_{i=1}^{n} U_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} U_i U_j\right).$$

5. Do the algebra to simplify the expression in part (iv) to

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma^2\left(n^{-1}\sum_{i=1}^{n} X_i^2\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}.$$

Hint: $SST_X/n = n^{-1}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$.

Problem 3. (Wooldridge Problem 2.7) Consider the saving function

$$sav = \beta_0 + \beta_1 inc + u, u = \sqrt{inc}\cdot e,$$

where e is a random variable with $E(e) = 0$ and $Var(e) = \sigma_e^2$. Assume that e is independent of *inc*.

1. Show that $E(u \mid inc) = 0$. (Hint:If e is independent of *inc*, then $E(e \mid inc) = E(e)$.)

2. Show that $Var(u \mid inc) = \sigma_e^2 inc$, so that the homoskedasticity Assumption is violated. In particular, the variance of *sav* increases with *inc*. (Hint:$Var(e \mid inc) = Var(e)$, if e and *inc* are independent.)

3. Provide a discussion that supports the assumption that the variance of savings increases with family income.

Problem 4. The econometrician obtained the following output from regressing the dependent variable "liver" against the independent variable "alcohol" and a constant, where "liver" is the number of liver disease deaths per 100,000 people in a country, and "alcohol" is consumption of alcohol in liters per capita in a country:

```
      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   22.62
       Model |  1554.38867     1  1554.38867           Prob > F      =  0.0001
    Residual |   1305.8181    19  68.7272685           R-squared     =  0.5435
-------------+------------------------------           Adj R-squared =  0.5194
       Total |  2860.20677    20  143.010338           Root MSE      =  8.2902
------------------------------------------------------------------------------
       liver |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   3.586388   .7541228       A       B            C           D
       _cons |   10.85482   2.802408    3.87   0.001     4.989313    16.72033
------------------------------------------------------------------------------
```

- Several entries in the output were replaced with letters. Find A - D. Show your work.

- Test at 5% significance level that the coefficient of "alcohol" is 5 (against the alternative that it is different from 5).
- Test the same hypothesis as in part (b) at 10% significance level.

Problem 5. (Wooldridge Problem 3.13)

1. Consider the simple regression model $Y_i = \beta_0 + \beta_1 X_i + U_i$ under the assumptions: for all $i = 1, \ldots, n$:

$$\begin{aligned} E\left(U_i | X_1, \ldots, X_n\right) &= 0, \\ E\left(U_i^2 | X_1, \ldots, X_n\right) &= \sigma^2, \\ E\left(U_i U_j | X_1, \ldots, X_n\right) &= 0 \text{ for } i \neq j. \end{aligned}$$

For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log(1 + x^2)$, define $Z_i = g(X_i)$. Define a slope estimator as

$$\tilde{\beta}_1 = \frac{\sum_{I=1}^{n} \left(Z_i - \bar{Z}\right) Y_i}{\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right) X_i}.$$

Show that $\tilde{\beta}_1$ is linear and unbiased. Remember, because $E(U_i \mid X_1, \ldots, X_n) = 0$, you can treat both X's and Z's as nonrandom in your derivation.

2. Show that (conditional on X's)

$$Var\left(\tilde{\beta}_1\right) = \frac{\sigma^2 \left(\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2\right)}{\left(\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right) X_i\right)^2}.$$

3. Show directly (without using the Gauss-Markov theorem) that, $Var\left(\hat{\beta}_1\right) \leq Var\left(\tilde{\beta}_1\right)$, where $\hat{\beta}_1$ is the OLS estimator. Hint: The Cauchy-Schwartz inequality implies that

$$\left(n^{-1} \sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)\left(X_i - \bar{X}\right)\right)^2 \leq \left(n^{-1} \sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2\right)\left(n^{-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2\right);$$

notice that we can drop $\bar{X}$ from the sample covariance.

Problem 6. Consider again the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + U_i, \; i = 1, \ldots, n;$$

with assumptions: (1) (X_i, Y_i), $i = 1, ..., n$ are independently and identically distributed (i.i.d.). (2) $E(U_i | X_i) = 0$, for $i = 1, ..., n$. (3) $E(U_i^2 | X_i) = \sigma^2$, for $i = 1, ..., n$, with some $\sigma > 0$. Define the estimator

$$\bar{\beta}_1 = \frac{\frac{\sum_{i=1}^{n} Y_i 1\{X_i \geq 0\}}{\sum_{i=1}^{n} 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^{n} Y_i 1\{X_i < 0\}}{\sum_{i=1}^{n} 1\{X_i < 0\}}}{\frac{\sum_{i=1}^{n} X_i 1\{X_i \geq 0\}}{\sum_{i=1}^{n} 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^{n} X_i 1\{X_i < 0\}}{\sum_{i=1}^{n} 1\{X_i < 0\}}}$$

where

$$1\{X_i \geq 0\} = \begin{cases} 1 & \text{if } X_i \geq 0 \\ 0 & \text{if } X_i < 0 \end{cases}$$

and

$$1\{X_i < 0\} = \begin{cases} 1 & \text{if } X_i < 0 \\ 0 & \text{if } X_i \geq 0. \end{cases}$$

In other words, $\bar{\beta}_1$ is the difference between the averaged Y's conditional on X being positive and the averaged Y's conditional on X being negative divided by the difference between the averaged X conditional on X being positive and the averaged X conditional on X being negative. Assume $\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} \neq \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}$.

1. Show that $\bar{\beta}_1$ is unbiased.
2. Is the conditional variance of $\bar{\beta}_1$ less than or equal to $\frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}$? Explain.

Problem 7. Suppose that a random variable X has a normal distribution with unknown mean μ. To simplify the analysis, we shall assume that σ^2 is known. Given a sample of observations, an estimator of μ is the sample mean, $\overline{X}$. When performing a (two-sided) test of the null hypothesis $H_0 : \mu = \mu_0$ at 5% significance level, it is usual to choose the upper and lower 2.5% tails of the normal distribution as the rejection regions, as shown in the first figure. s.d. is equal to $\sqrt{\sigma^2/n}$, the standard deviation of $\overline{X}$. The density function of $N\left(\mu_0, \sigma^2/n\right)$ is shown in the first figure. H_0 is rejected when $\left|\overline{X} - \mu_0\right|/\text{s.d.} > 1.96$. However, suppose that someone instead chooses the central 5% of the distribution as the rejection region, as in the second figure. Give a technical explanation, using appropriate statistical concepts, of why this is not a good idea.

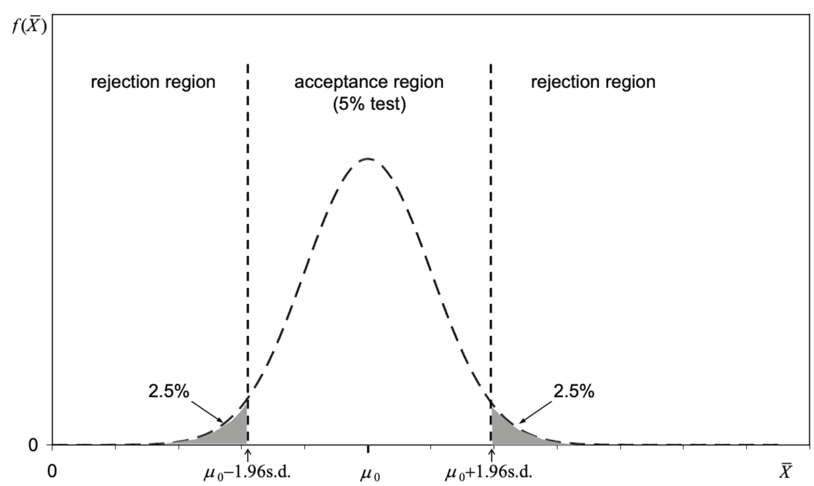

Figure 1: Conventional rejection regions.

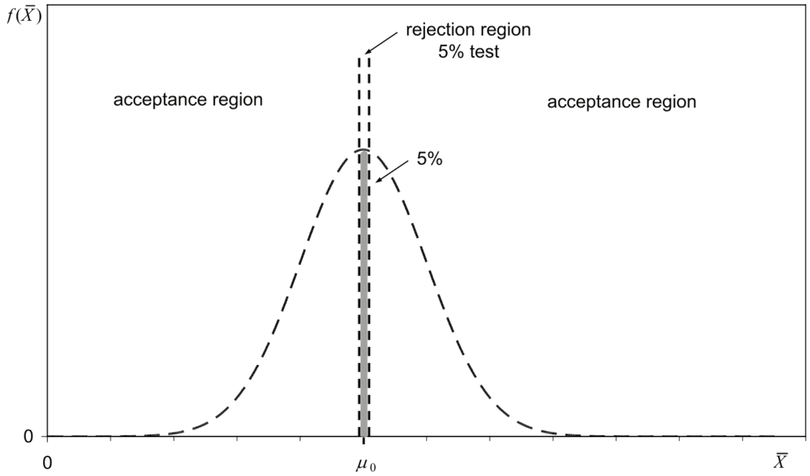

Figure 2: Central 5 per cent chosen as rejection region.