Econometrics

Homework 3

Problem 1. Consider a simple linear regression model:

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

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

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

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

Solution. (a) True. $\hat{U}_i$'s are the fitted residuals from a regression with an intercept, and $\sum_i \hat{U}_i = 0$ is the normal equation obtained from the OLS first-order conditions for the intercept.
(b) False (in general). $\tilde{U}_i$'s are the fitted residuals from a regression without an intercept, and therefore the OLS first-order condition corresponding to the intercept does not have to hold.
(c) False (in general). $EU_i = 0$, however, a sample average of (finitely many) U_i's does not have to be zero.

(d) True. By the law of iterated expectation (LIE),

$$\begin{aligned} E\left(U_i X_i^4\right) &= E\left(E\left(U_i X_i^4 | X_i\right)\right) \\ &= E\left(X_i^4 E\left(U_i | X_i\right)\right) \\ &= E\left(X_i^4 \cdot 0\right) \\ &= 0. \end{aligned}$$

(e) False. Since $\tilde{\beta}_1$ is a biased estimator, the Gauss-Markov Theorem does not apply in this case. In fact, one can show that $\tilde{\beta}_1$ has a smaller conditional variance given $X_1, \ldots, X_n$.

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

Solution. (1) Omitted. (2) Because (conditional on the X's) $Cov(\hat{\beta}_1, \overline{U}) = E\left[(\hat{\beta}_1 - \beta_1)\overline{U}\right]$, we show that the latter is zero. But, from part (i), $E\left[(\hat{\beta}_1 - \beta_1)\overline{U}\right] = E\left[\left(\sum_{i=1}^n w_i U_i\right)\overline{U}\right]$. Because the U_i are pairwise uncorrelated (they are independent) $E\left(U_i \overline{U}\right) = E\left(U_i^2/n\right) = \sigma^2/n$, (because $E(U_i U_h) = 0, i \neq h$). Therefore,

$$\sum_{i=1}^n w_i E(U_i \overline{U}) = \sum_{i=1}^n w_i(\sigma^2/n) = (\sigma^2/n)\sum_{i=1}^n w_i = 0.$$

(3) The formula for the OLS intercept is $\hat{\beta}_0 = \overline{Y} - \hat{\beta}\overline{X}$ and, plugging in $\overline{Y} = \beta_0 + \beta_1\overline{X} + \overline{U}$ gives

$$\hat{\beta}_0 = (\beta_0 + \beta_1\overline{X} + \overline{U}) - \hat{\beta}_1\overline{X} = \beta_0 + \overline{U} - (\hat{\beta}_1 - \beta_1)\overline{X}.$$

(4) Because $\hat{\beta}_1$and $\overline{U}$ are uncorrelated, and $E\left(\hat{\beta}_0\right) = \beta_0$,

$$\begin{aligned} Var(\hat{\beta}_0) &= Var(\overline{U}) + Var(\hat{\beta}_1)\overline{X}^2 \\ &= \sigma^2/n + (\sigma^2/SST_X)\overline{X}^2 \\ &= \sigma^2/n + \sigma^2\overline{X}^2/SST_{X,} \end{aligned}$$

which is what we want to show. (5) Using the hint and substitution gives

$$\begin{aligned} Var(\hat{\beta}_0) &= \sigma^2\left[(SST_X/n) + \overline{X}^2\right]/SST_X \\ &= \sigma^2\left[\left(n^{-1}\sum_{i=1}^{n} X_i^2 - \overline{X}^2\right) + \overline{X}^2\right]/SST_X \\ &= \sigma^2(n^{-1}\sum_{i=1}^{n} X_i^2)/SST_X. \end{aligned}$$

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.

Solution. (1) When we condition on *inc* in computing an expectation, $\sqrt{inc}$ becomes a constant. So $E(u|inc) = E(\sqrt{inc}{\cdot}e|inc) = \sqrt{inc}E(e|inc) = \sqrt{inc}{\cdot}0$ because $E(e|inc) = E(e) = 0$. (2) Again, when we condition on inc in computing a variance, $\sqrt{inc}$ becomes a constant. So $Var(u|inc) = Var(\sqrt{inc}{\cdot}e|inc) = (\sqrt{inc})^2 Var(e|inc) = \sigma_e^2\text{inc}$ because $Var(e|inc) = \sigma_e^2$. (3) Families with low incomes do not have much discretion about spending; typically, a low-income family must spend on food, clothing, housing, and other necessities. Higher income people have more discretion, and some might choose more consumption while others more saving. This discretion suggests wider variability in saving among higher income families.

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.

Solution. To find A, compute

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\widehat{Var}\left(\hat{\beta}_1\right)}} = \frac{3.586 - 0}{0.754} = 4.756.$$

B is the p-value for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. Since the degree of freedom is 21-2=19, and since the largest critical value in the t-table is

$$t_{19,1-0.0005} = 3.883$$

which is smaller than $T = 4.756$, we conclude the p-value (B) is smaller than $0.0005 \times 2 = 0.001$. C and D are the lower and upper bounds for 95% confidence interval for β_1. First

$$t_{19,1-0.05/2} = 2.093.$$

Next,

$$\begin{aligned}
[C, D] &= \left[\hat{\beta}_1 - t_{n-2,1-0.05/2} \times (standard\, error), \hat{\beta}_1 + t_{n-2,1-0.05/2} \times (standard\, error)\right] \\
&= [3.586 - 2.093 \times 0.754, 3.586 + 2.093 \times 0.754] \\
&= [2.008, 5.164].
\end{aligned}$$

The 95% confidence interval for β_1 is $[2.008, 5.164]$, which includes 5. Therefore $H_0 : \beta_1 = 5$ cannot be rejected in favor of $H_1 : \beta_1 \neq 5$ at 5% significance level.
The 90% confidence interval for β_1 is

$$\begin{aligned}
&\left[\hat{\beta}_1 - t_{n-2,1-0.10/2} \times (standard\, error), \hat{\beta}_1 + t_{n-2,1-0.10/2} \times (standard\, error)\right] \\
&= [3.586 - 1.729 \times 0.754, 3.586 + 1.729 \times 0.754] \\
&= [2.282, 4.890].
\end{aligned}$$

Since 5 is outside of the confidence interval, we reject $H_0 : \beta_1 = 5$ in favor of $H_1 : \beta_1 \neq 5$ 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 $i = 1, \ldots, n$:

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

For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log\left(1+x^2\right)$, 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 $E\left(U_i \mid X_1,\ldots,X_n\right) = 0$, 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.

Solution. (i) For notational simplicity, define $S_{ZX} = \sum_{i=1}^{n}(Z_i - \overline{Z})X_i$. Then we can write $\widetilde{\beta}_1$ as

$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})Y_i}{S_{ZX}}.$$

This is clearly a linear function of the Y's: take the weights to be $w_i = (Z_i - \overline{Z})/S_{ZX}$. To show unbiasedness, as usual we plug $Y_i = \beta_0 + \beta_1 X_i + U_i$ into this equation, and simplify:

$$\begin{aligned} \widetilde{\beta}_1 &= \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})\left(\beta_0 + \beta_1 X_i + U_i\right)}{S_{ZX}} \\ &= \frac{\beta_0\sum_{i=1}^{n}(Z_i - \overline{Z}) + \beta_1 S_{ZX} + \sum_{i=1}^{n}(Z_i - \overline{Z})U_i}{S_{ZX}} \\ &= \beta_1 + \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})U_i}{S_{ZX}}, \end{aligned}$$

where we use the fact that $\sum_{i=1}^{n}(Z_i - \overline{Z}) = 0$ always. Now S_{ZX} is a function of the Z's and X's and the expected value of each U_i is zero conditional on all Z's and X's in the sample. Therefore, conditional on these values,

$$E(\widetilde{\beta}_1) = \beta_1 + E\left(\frac{\sum_{i=1}^{n}(Z_i - \overline{Z})U_i}{S_{ZX}}\right) = \beta_1.$$

because $E(U_i) = 0$ for all i.

(ii) Again conditional on the Z's and X's in the sample,

$$\begin{aligned} Var(\widetilde{\beta}_1) &= \frac{Var\left[\sum_{i=1}^{n}(Z_i - \overline{Z})U_i\right]}{S_{ZX}^2} \\ &= \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})^2 Var(U_i)}{S_{ZX}^2} \\ &= \sigma^2 \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})^2}{S_{ZX}^2} \end{aligned}$$

because of the homoskedasticity assumption $[Var(U_i) = \sigma^2$ for all $i]$. Given the definition of S_{zx}, this is what we wanted to show.

(iii) We know that $Var(\hat{\beta}_1) = \sigma^2 / \left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]$. Now we can rearrange the inequality in the hint, drop $\overline{X}$ from the sample covariance, and cancel n^{-1} everywhere, to get $\sum_{i=1}^{n}(Z_i - \overline{Z})^2 / S_{ZX}^2 \geq 1/\left(\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right]\right)$. When we multiply through by σ^2 we get $Var(\widetilde{\beta}_1) > Var(\hat{\beta}_1)$, which is what we wanted to show.

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) $E\left(U_i | X_i\right) = 0$, for $i = 1, ..., n$. (3) $E\left(U_i^2 | X_i\right) = \sigma^2$, 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.

Solution. (i) As we have done in class we should: (1) substitute $Y_i = \beta_0 + \beta_1 X_i + U_i$ and then (2) use the properties of expectations to simplify.

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

$$= E\left[\frac{\frac{\sum_{i=1}^n (\beta_0 + X_i\beta_1 + U_i) 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n (\beta_0 + X_i\beta_1 + U_i) 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}{\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}\right]$$

rearranging

$$= E\Bigg[\frac{\Big(\beta_0 \frac{\sum_{i=1}^n 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} + \beta_1 \frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} + \frac{\sum_{i=1}^n U_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}}\Big) - \Big(\beta_0 \frac{\sum_{i=1}^n 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}} + \beta_1 \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}} + \frac{\sum_{i=1}^n U_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}{\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}$$

simplifying

$$= \beta_1 + E\left[\frac{\frac{\sum_{i=1}^n U_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n U_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}{\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}\right]$$

using iterated expectations

$$= \beta_1 + E\left[E\left[\frac{\frac{\sum_{i=1}^n U_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n U_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}{\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}} | X_1, ..., X_n\right]\right]$$

using the linearity of $E[\cdot | X_1, ..., X_n]$ we have

$$= \beta_1 + E\left[\frac{\frac{\sum_{i=1}^n E[U_i | X_1, ..., X_n] 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n E[U_i | X_1, ..., X_n] 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}{\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}\right]$$

$E[U_i | X_1, ..., X_n] = 0$ by assumption, so

$$= \beta_1$$

(ii) The previous part showed $\bar{\beta}_1$ is unbiased. It is also linear because it is equal $\sum_{i=1}^n \bar{c}_i Y_i$ with

$$\bar{c}_i = \frac{\frac{1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}{\frac{\sum_{i=1}^n X_i 1\{X_i \geq 0\}}{\sum_{i=1}^n 1\{X_i \geq 0\}} - \frac{\sum_{i=1}^n X_i 1\{X_i < 0\}}{\sum_{i=1}^n 1\{X_i < 0\}}}.$$

Therefore, by the Gauss-Markov theorem, $Var(\bar{\beta}_1) > Var(\hat{\beta}_1)$.

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.

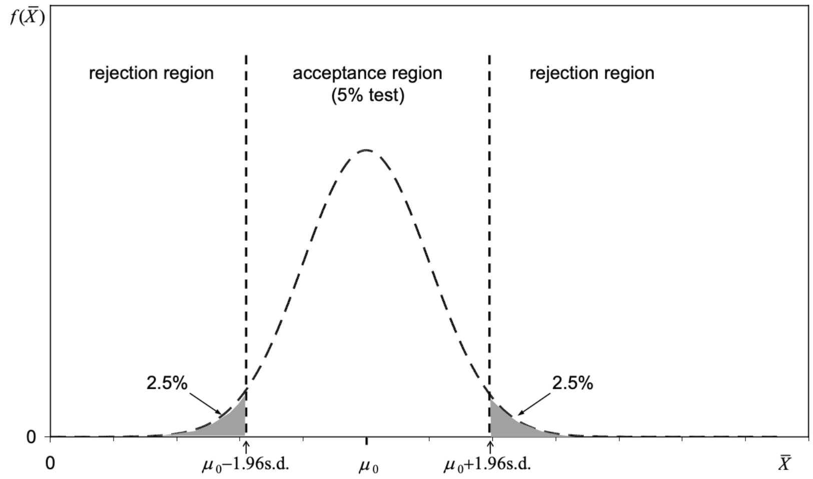

Figure 1: Conventional rejection regions.

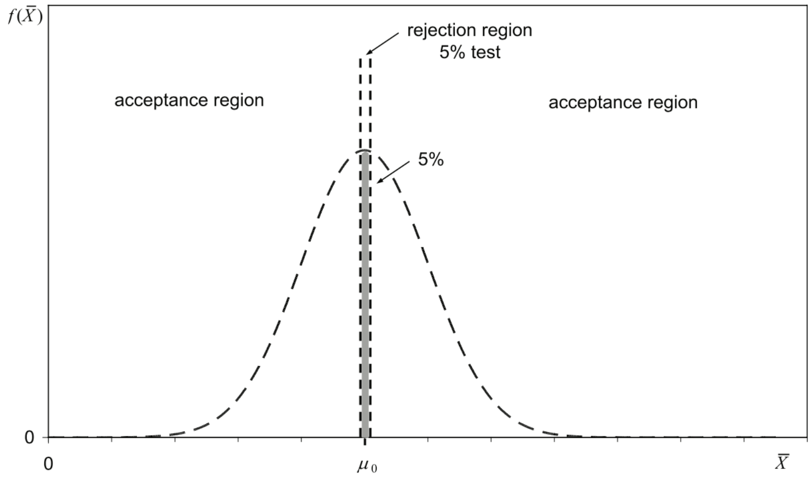

Figure 2: Central 5 per cent chosen as rejection region.

Solution. The following discussion assumes that you are performing a 5 per cent significance test, but it applies to any significance level. If the null hypothesis is true, it does not matter how you define the 5 per cent rejection region. By construction, the risk of making a Type I error will be 5 per cent. Issues relating to Type II errors are irrelevant when the null hypothesis is true.

The reason that the central part of the conditional distribution is not used as a rejection region is that it leads to problems when the null hypothesis is false. The probability of not rejecting H_0 when it is false will be lower. To use the obvious technical term, the power of the test will be lower. The figure opposite shows the power functions for the test using the conventional upper and lower 2.5 per cent tails and the test using the central region. The horizontal axis is the difference between the true value and the hypothetical value μ_0 in terms of standard deviations. The vertical axis is the power of the test. The first figure has been drawn for the case where the true value is greater than the hypothetical value. The second figure is for the case where the true value is lower than the hypothetical value. It is the same, but reflected horizontally.

The greater the difference between the true value and the hypothetical mean, the more likely it is that the sample mean will lie in a tail of the distribution conditional on H_0 being true, and so the more likely it is that the null hypothesis will be rejected by the conventional test. The figures show that the power of the test approaches 1 asymptotically. However, if the central region of the distribution is used as the rejection region, the probability of the sample mean lying in it will diminish as the difference between the true and hypothetical values increases, and the power of the test approaches zero asymptotically. This is an extreme example of a very bad test procedure.

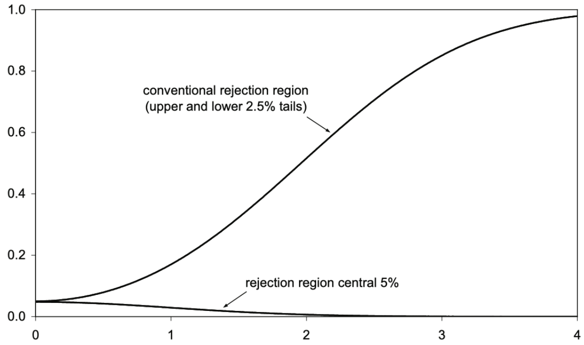

Figure 3: Power functions of a conventional 5 per cent test and one using the central region (true value $> \mu_0$).

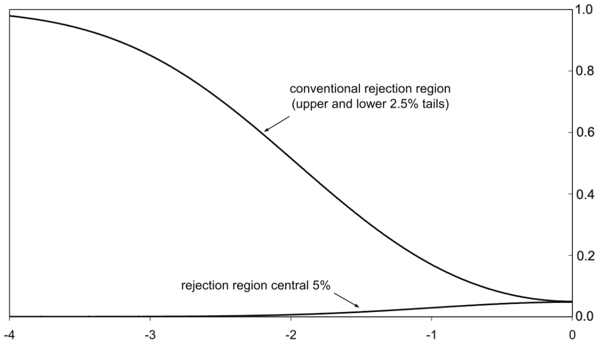

Figure 4: Power functions of a conventional 5 per cent test and one using the central region (true value $< \mu_0$).