Homework 3

Problem 1. Suppose we observe a random sample $\{(Y_i, D_i)\}_{i=1}^n$, where Y_i is the dependent variable and D_i is a binary independent variable: for all $i = 1, 2, ..., n$, $D_i = 1$ or $D_i = 0$. Suppose we regress Y_i on D_i with an intercept. Show: the LS estimate of the slope is equal to the difference between the sample averages of the dependent variable of the two groups, observations with $D_i = 1$ and observations with $D_i = 0$. Hint: The sample average of Y of observations with $D_i = 1$ can be written as $\frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i}$. What is the sample average of Y of observations with $D_i = 0$? Also note: $D_i = D_i^2$.

Problem 2. Suppose that assumptions of the Classical Linear Regression model hold, i.e.

$$\begin{aligned} \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},\ \boldsymbol{\beta} \in \mathbb{R}^k \\ \mathbb{E}(\boldsymbol{e}|\boldsymbol{X}) = 0, \\ \operatorname{rank}(\boldsymbol{X}) = k, \end{aligned}$$

however,

$$\mathbb{E}(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}) = \boldsymbol{\Omega},$$

where $\boldsymbol{\Omega}$ is an $n \times n$, positive definite and symmetric matrix, but different from $\sigma^2 \boldsymbol{I}_n$.

1. Derive the conditional variance (given $\boldsymbol{X}$) of the LS estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.
2. Derive the conditional variance (given $\boldsymbol{X}$) of the Generalized LS estimator $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}$.
3. Without relying on the Gauss-Markov Theorem, show that

   $$\operatorname{Var}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}) - \operatorname{Var}(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}) \geq 0$$

   (in the positive semidefinite sense). Hint: Show

   $$\left(\operatorname{Var}(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X})\right)^{-1} - \left(\operatorname{Var}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X})\right)^{-1} \geq 0$$

   by showing that the expression on the left-hand side depends on a symmetric and idempotent matrix of the form $\boldsymbol{I}_n - \boldsymbol{H}(\boldsymbol{H}'\boldsymbol{H})^{-1}\boldsymbol{H}'$ for some $n \times k$ matrix $\boldsymbol{H}$ of rank k.

Problem 3. Consider the GLS estimator $\widetilde{\boldsymbol{\beta}}$ defined in the previous question.

1. Show that $\widetilde{\boldsymbol{\beta}}$ satisfies $\widetilde{\boldsymbol{e}}'\boldsymbol{\Omega}^{-1}\boldsymbol{X} = 0$, where $\widetilde{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$.
2. Using the result in (i), show that the generalized squared distance function $S(\boldsymbol{b}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})'\boldsymbol{\Omega}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$ can be written as

   $$S(\boldsymbol{b}) = \widetilde{\boldsymbol{e}}'\boldsymbol{\Omega}^{-1}\widetilde{\boldsymbol{e}} + (\widetilde{\boldsymbol{\beta}} - \boldsymbol{b})'\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{b}).$$

3. Using the result in (ii), show that $\widetilde{\boldsymbol{\beta}}$ minimizes $S(\boldsymbol{b})$.

Problem 4. Use FWL Theorem to show that in a simple (one-regressor) regression model,

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

the LS estimate for β_1 is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.$$

Then assume (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 $\sigma > 0$. Show that

$$\text{Var}\left(\widehat{\beta}_1|X_1, ..., X_n\right) = \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}.$$

Problem 5. 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 $\text{Var}\left(\bar{\beta}_1|X_1, ..., X_n\right)$ less than or equal to $\frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$ (the variance of the LS estimator)? Explain.

Problem 6. 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.

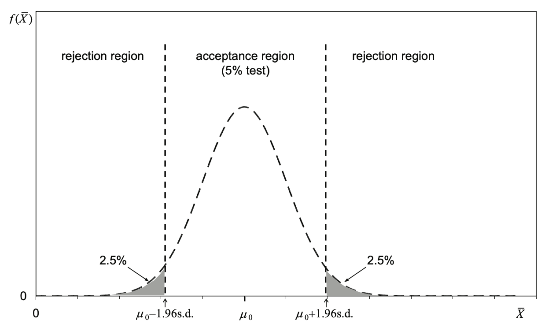

Figure 1: Conventional rejection regions.

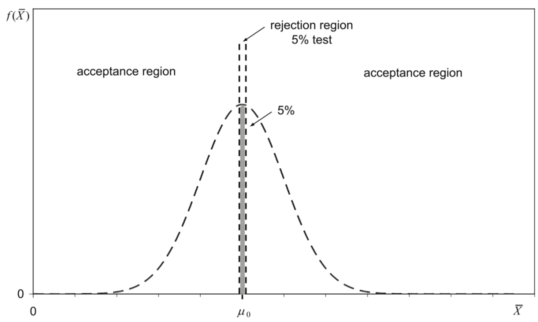

Figure 2: Central 5 per cent chosen as rejection region.

Problem 7. Consider the following model:

$$Y_i = \beta + U_i,$$

where U_i are iid $N(0,1)$ random variables, $i = 1, \dots, n$.

1. Find the LS estimator of β and its mean, variance, and distribution.
2. Suppose that a data set of 100 observation produced OLS estimate $\widehat{\beta} = 0.167$.
   (a) Construct 90% and 95% symmetric two-sided confidence intervals for β.
   (b) Construct a 95% one-sided confidence interval of the form $[A, +\infty)$ for β. In other words, find a random variable A such that $\Pr(\beta \in [A, +\infty)) = 1 - \alpha$, where $\alpha \in (0, 0.5)$ is a known constant chosen by the econometrician.
   (c) Construct a 95% one-sided confidence interval of the form $(-\infty, A]$ for β.

Problem 8. Consider the following regression model:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}, \\ \mathbb{E}(\boldsymbol{e}|\boldsymbol{X}_1, \boldsymbol{X}_2) &= 0, \\ \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}_1, \boldsymbol{X}_2\right) &= \sigma_e^2 \boldsymbol{I}_n. \end{aligned}$$

Let $\widetilde{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}$ be the LS estimator for $\boldsymbol{\beta}_1$ which omits $\boldsymbol{X}_2$ from the regression.

1. Find $\mathbb{E}(\tilde{\boldsymbol{\beta}}_1|\boldsymbol{X}_1)$.
2. Define
$$\boldsymbol{V} = \boldsymbol{X}_2\boldsymbol{\beta}_2 - \mathbb{E}\left(\boldsymbol{X}_2\boldsymbol{\beta}_2|\boldsymbol{X}_1\right).$$
   Find $\mathbb{E}\left(\boldsymbol{e}\boldsymbol{V}'|\boldsymbol{X}_1\right)$.
3. Find $\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}_1\right)$.
4. Assume that
$$\mathbb{E}\left(\boldsymbol{V}\boldsymbol{V}'|\boldsymbol{X}_1\right) = \sigma_v^2 I_n,$$
   and find $\mathrm{Var}(\tilde{\boldsymbol{\beta}}_1|\boldsymbol{X}_1)$.
5. Let $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y}$ be the OLS estimator for $\boldsymbol{\beta}_1$ from a regression of $\boldsymbol{Y}$ against $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, where $\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$. Compare $\mathrm{Var}(\tilde{\boldsymbol{\beta}}_1|\boldsymbol{X}_1)$ derived in part (iv) with $\mathrm{Var}(\hat{\boldsymbol{\beta}}_1|\boldsymbol{X}_1, \boldsymbol{X}_2)$. Can you say which of the two variances is bigger (in the positive semi-definite sense)? Explain your answer.