3. (7 points) Suppose we have a population, for which the true conditional relationship between Y_i and X_i is

$$Y_i \mid X_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2 x_i), \quad x_i > 1. \tag{1}$$

A sample of n pairs of (x_i, y_i) measurements is collected from this population and the linear regression relationship that is fit using these data is

$$\hat{y}_i = \hat{\beta}_1^* x_i, \quad i = 1, \ldots, n \tag{2}$$

where the least squares estimate is found to be

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

The model fit to these data assumes all assumptions of linear regression hold.

(a) (2 points) Determine whether $\hat{\beta}_1^*$ is an unbiased estimator of β_1 in the true population relationship. Show all your work in arriving at your answer.

Solution:
To show that $\hat{\beta}_1^*$, we just need to show that $E(\hat{\beta}_1^*) = \beta_1$.
Putting $y_i = \beta_0 + \beta_1 x_i$ into the equation above, we have that

$$E(\hat{\beta}_1^*) = \frac{\sum_{i=1}^n x_i E(y_i)}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i E(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_{i=1}^n x_i^2}$$

$$= \beta_0 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_1 \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2} + \frac{\sum_{i=1}^n x_i E(\epsilon_i)}{\sum_{i=1}^n x_i^2}$$

$$= \beta_0 \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} + \beta_1$$

where, from the assumptions of linear regression, $E(\epsilon_i) = 0$. Therefore, the estimator is biased for the true population slope of β_1.

Rubric: if they make a mistake in their work in an early step, this will affect the remaining work in their answer. Judge all subsequent steps for correctness assuming that the mistake is correct (i.e. that the only reason their work is wrong on each subsequent step is due to an additional mistake that occurred in that step).

- *1 point - for using the conditional mean from the true population relationship*
- *1 point - for arriving at the expected value of the estimator.*

(b) (3 points) Derive the variance of $\hat{\beta}_1^*$, showing all your steps. Will the variance of $\hat{\beta}_1^*$ be larger, smaller, or equal to the variance of $\hat{\beta}_1^*$ if all assumptions of linear regression actually held in the population?

Solution: The variance of $\hat{\beta}_1^*$ is obtained as

$$\text{Var}(\hat{\beta}_1^* \mid X) = \text{Var}(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}) = \frac{1}{(\sum_{i=1}^n x_i^2)^2}\text{Var}(\sum_{i=1}^n x_i y_i)$$

$$= \frac{1}{(\sum_{i=1}^n x_i^2)^2}\Big[\sum_{i=1}^n x_i^2 \text{Var}(y_i) + \sum_{i \neq j} x_i x_j \text{Cov}(y_i, y_j)\Big]$$

.

With $\text{Var}(y_i) = \sigma^2 x_i$ and uncorrelated errors, this becomes

$$\text{Var}(\hat{\beta}_1^* \mid X) = \sigma^2 \frac{\sum_{i=1}^n x_i^3}{(\sum_{i=1}^n x_i^2)^2}$$

Because of the extra $x_i > 1$ term in the numerator, this variance is larger compared to the standard variance of $\hat{\beta}_1 = \frac{\sigma^2}{(\sum_{i=1}^n x_i^2)^2}$, if all assumptions of linear regression held in the population so that the errors have constant variance, i.e. $\epsilon \sim N(0, \sigma^2)$.

Rubric: if they make a mistake in their work in an early step, this will affect the remaining work in their answer. Judge all subsequent steps for correctness assuming that the mistake is correct (i.e. that the only reason their work is wrong on each subsequent step is due to an additional mistake that occurred in that step).

- *1 point - for correctly removing constants to get Var(Y) or equivalently Var(ϵ)*
- *1 point - for using true population error variance to get a variance expression (0.5 only if they simplified their constants incorrectly arriving at the wrong variance expression)*
- *1 point - for noting how variance compares to when assumptions hold (would be ideal to actually see this variance, but not necessary)*