STAT 170 – Regression and Time Series
Drake University, Fall 2024
William M. Boal

Signature: ______________________________

Printed name: ______________________________

EXAMINATION 3 VERSION A
"Multiple Regression With Cross-Section Data"
November 12, 2024

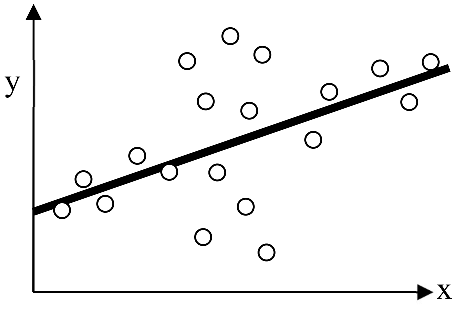
INSTRUCTIONS: This exam is closed-book, closed-notes. You may use a calculator on this exam, but not a graphing calculator, a calculator with alphabetical keys, nor a mobile phone. Point values for each question are noted in brackets. Tables of the t-distribution, the F-distribution, and the chi-square distribution are attached.

NOTATION: In this exam, $\hat{\beta}_j$ denotes the least-squares coefficient estimators of the equation $y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i$. The least-squares fitted value is denoted $\hat{y}_i$. The least-squares residual is denoted $\hat{\varepsilon}_i$. The sample size is denoted *n*. The true or population value of the variance of the unobserved error term ε_i is denoted σ^2. The (unbiased) least-squares estimator of σ^2 is denoted $\hat{\sigma}^2$. The sample mean of y is denoted $\bar{y}$. The natural logarithm is denoted *ln(.)*.

I. MULTIPLE CHOICE: Circle the one best answer to each question. Use margins for scratch work [1 pts each—12 pts total]

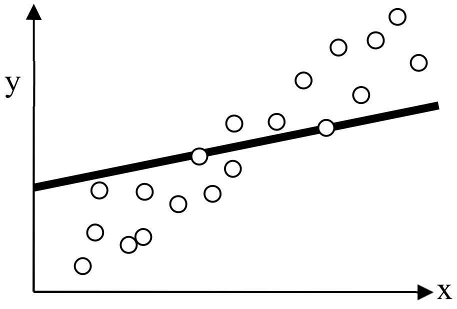
(1) In the graph below, the solid line is the true population regression line and the circles are observations in the sample. Which assumption appears to be violated in this sample?

a. $E(\varepsilon_i|x_i) = 0$.
b. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.
c. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j)=0$ for $i \neq j$.
d. All of the above.
e. None of the above.



(2) In the graph below, the solid line is the true population regression line and the circles are observations in the sample. Which assumption appears to be violated in this sample?

a. $E(\varepsilon_i|x_i) = 0$.
b. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.
c. No autocorrelation: $Cov(\varepsilon_i,\varepsilon_j)=0$ for $i \neq j$.
d. All of the above.
e. None of the above.



(3) Let $\hat{y}_i$ denote the least-squares fitted values and let $\hat{\varepsilon}_i$ denote the least-squares residuals. Which of the following must necessarily hold?

a. $\sum(\hat{y}_i - \bar{y})^2 = \sum(y_i - \bar{y})^2 + \sum \hat{\varepsilon}_i^2$.
b. $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum \hat{\varepsilon}_i^2$.
c. $\sum \hat{\varepsilon}_i^2 = \sum(y_i - \bar{y})^2 + \sum(\hat{y}_i - \bar{y})^2$.
d. $\sum \hat{\varepsilon}_i^2 = \sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2$.

(4) *Heteroskedasticity* means that

a. some x variables are perfectly correlated with each other.
b. random error terms of different observations are correlated.
c. random error terms of different observations have different variances.
d. random error terms are correlated with one or more of the x variables.

(5) Suppose we estimate the equation

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4},$$

and we want to test the null joint hypothesis that $\beta_2 = \beta_3 = \beta_4 = 0$. We should reject the null hypothesis at 5% significance if

a. THE F statistic is less than its 5% critical point.
b. THE F statistic is greater than its 5% critical point.
c. THE F statistic is either less than its lower 2.5 % critical point or greater than its upper 2.5 % critical point.
d. *all* the t-statistics for β_2, β_3, and β_4 are greater in absolute value than their 5% critical points.
e. *any* of the t-statistics for β_2, β_3, or β_4 are greater in absolute value than their 5% critical points.

(6) If another regressor is added to a multiple-regression equation and the equation is re-estimated on the same observations, the ordinary R^2 value

a. will necessarily increase.
b. will necessarily decrease.
c. may increase or decrease.
d. will necessarily remain the same because the dependent variable has not changed.

(7) Suppose we estimate the equation

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} .$$

If x_{i2} and x_{i3} are *perfectly* correlated, then the least-squares estimators of their coefficients

a. will have large standard errors.
b. will be zero.
c. cannot be computed.
d. will be biased.
e. will be inconsistent.

(8) An *interaction term* means

a. a correlation between a regressor x and the error term ε_i .
b. the product of two regressors x_2 and x_3.
c. a correlation between a regressor x and the dependent variable y .
d. the covariance between the least-squares intercept estimator and the least-squares slope estimator.

(9) Suppose we wish to estimate the relationship between the size of a firm and the number of patents it receives using data on 1000 business firms. Moreover, we want to allow the intercept to be different by industry (manufacturing, hospitality, retail, etc.). If we have five industries in our data, we need

a. one dummy variable.
b. two dummy variables.
c. three dummy variables.
d. four dummy variables.
e. five dummy variables.

(10) A model relating hourly earnings to educational background and gender was estimated on a random sample of workers with the following results.

$$ln(E) = 0.51 + 0.12\,S - 0.17\,D,$$

where E denotes the worker's hourly earnings, S denotes the worker's years of schooling, and D is a dummy variable that equals one for female workers and zero for male workers. According to this model, female workers with the same amount of schooling as male workers earn about

a. $0.17 more per hour.
b. $0.17 less per hour.
c. 17 percent more per hour.
d. 17 percent less per hour.

(11) If the purpose of our regression is *prediction*, then we should include additional regressors if they

a. increase the degrees of freedom.
b. prevent omitted-variable bias.
c. improve the fit of the equation.
d. raise the sum of squared residuals.

(12) If the random error term is heteroskedastic then the least squares estimators of the coefficients will

a. be biased.
b. be inconsistent.
c. be impossible to compute.
d. have incorrect standard errors.

II. SHORT ANSWER: Please write your answers in the boxes on this question sheet. Use margins for scratch work.

(1) [Algebraic properties: 6 pts] Suppose we estimate the equation $y_i = \beta_1 + \beta_2\, x_{i2}, + \beta_3\, x_{i3} + \varepsilon_i$ by ordinary least squares. Which equations below hold necessarily, regardless of the data or the model? Write "TRUE" or "FALSE" in the boxes below.

a. $\sum \hat{y}_i \hat{\varepsilon}_i = 0$

b. $\sum \hat{\varepsilon}_i = 0$

c. $\sum x_{i2}\, x_{i3} = 0$

(2) [Properties: 6 pts] Which of the following conditions cause the least-squares estimators for the slope coefficients (the $\hat{\beta}$s) to be biased and inconsistent? Write "YES" or "NO."

a. The error term (ε_i) is heteroskedastic.

b. The error term (ε) is autocorrelated.

c. The error term (ε_i) is correlated with a regressor.

(3) [Variance of LS estimators: 8 pts] Suppose we estimate the equation $y_i = \beta_1 + \beta_2 x_{i2}, + \beta_3 x_{i3} + \varepsilon_i$. Assume that the Gauss-Markov assumptions hold. Answer TRUE or FALSE below: The variance of the least-squares slope estimator $\hat{\beta}_2$ is smaller, and thus the true value of β_2 is estimated more precisely,

a. the larger the sample size.

b. the closer the correlation between x_{i2} and x_{i3} .

c. greater the variation of x_{i2} around the sample mean $\bar{x}_2$.

d. the larger the variance of the error term $\mathrm{Var}(\varepsilon_i) = \sigma^2$.

(4) [Adding regressors: 10 pts] Suppose we first estimate the equation $y_i = \beta_1 + \beta_2 x_{i2}$ by least squares, and then estimate $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3}$. What are the consequences of adding the regressor x_{i3} ? In each box below, write one of the following:

- “must increase, ”
- “must decrease, ”
- “can either increase or decrease, ”
- “must remain constant.”

a. The standard errors of the estimated coefficients…

b. The sum of squared residuals…

c. The R^2 value…

d. Theil’s adjusted R^2 (also called “$\bar{R}^2$”)...

e. The values of the estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$…

III. PROBLEMS: Write your answers in the boxes on this question sheet.

(1) [Analysis of variance table, R^2, F-test: 20 pts] A regression program computed the following analysis-of-variance (ANOVA) table:

	Degrees of freedom ("DF")	Sums of squares ("SS")	Mean squares ("MS")
Regression (or "Model" or "Explained")	4	272	68.0
Residual (or "Error")	60	240	4.0
Total	64	512	8.00

a. What is the sample size?

b. How many β coefficients were estimated, including the intercept?

c. What is the unbiased estimate of the variance of the error term?

d. Compute the value of R^2 (sometimes called the "coefficient of determination") to at least three decimal places.

e. Compute the value of Theil's adjusted R^2 (sometimes called "$\overline{R}^2$") to at least three decimal places.

f. [10 pts] Test the joint null hypothesis that all the coefficients except the intercept are zero (against the alternative hypothesis that at least one of these coefficients is not zero) at 5% significance. Give the value of the test statistic, its degrees of freedom, the critical point, and your conclusion (whether you can reject the null hypothesis).

Degrees of freedom in numerator = _____ Degrees of freedom in denominator = _____

Value of F statistic = _________________ Critical point = ____________________

Reject null hypothesis? ______________________.

(2) [Dummy variables and structural change: 20 pts] Suppose we wish to estimate the effect of income on spending for travel, using a sample of **120** households.

spending = spending for travel by household.
income = income of household.
retired = 1 if all members of household are retired.
= 0 if at least one member is still working.

The following four equations were estimated, with the sums of squared residuals (SSR) as shown.

[1] $spending_i = 56 + 0.2\ income$ SSR=372

[2] $spending = 46 + 0.3\ income + 8.0\ retired$ SSR=351

[3] $spending = 52 + 0.5\ income - 0.1\ (retired \times income)$ SSR=350

[4] $spending = 33 + 0.4\ income + 9.0\ retired - 0.1\ (retired \times income)$ SSR=232

First, consider equation [4].

a. According to equation [4], what is the intercept for non-retired households?

b. According to equation [4], what is the intercept for retired households?

c. According to equation [4], what is the slope for retired households?

Second, test the null hypothesis that all households have the same intercept and slope, against the alternative hypothesis that the intercept and slope for retired households are *both* different from the intercept and slope for nonretired households, at 5% significance. Assume the Gauss-Markov assumptions are satisfied and the error term is normally-distributed.

d. Which equation, [1], [2], [3], or [4], is the *restricted* equation, representing the null hypothesis?

e. Which equation, [1], [2], [3], or [4], is the *unrestricted* equation, representing the alternative hypothesis?

f. [10 pts] Give the value of the test statistic, its degrees of freedom, the critical point, and your conclusion (whether you can reject the null hypothesis).

Degrees of freedom in numerator = _____ Degrees of freedom in denominator = _____

Value of F statistic = __________________ Critical point = _____________________

Reject null hypothesis? _______________________.

(3) [Heteroskedasticity: 12 pts] We have estimated the following equation by ordinary least squares, using total data for **60** countries:

$$\textit{Total electricity consumption} \;=\; \beta_1 + \beta_2 \textit{ Total national income}$$

We believe that all the Gauss-Markov assumptions are satisfied, except that we fear that the error term (ε) might be heteroskedastic, with variance related to country population.

a. If the error term (ε) is heteroskedastic, are the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ unbiased? (Answer *yes* or *no*.)

b. If the error term (ε) is heteroskedastic, are usual standard errors for the least squares estimators valid? (Answer *yes* or *no*.)

c. Given that the dependent variable is a total, is the variance of the error term (ε) more likely to be *positively* or *negatively* related to the country's population?

To test for heteroskedasticity, we save the least-squares residuals from the above equation and estimate the following auxiliary regression by least squares:

$$\hat{\varepsilon}_i^2 = \alpha_1 + \alpha_2\, population_i + \nu_i$$

where "population" is the country's population and ν_i is a new error term. The R^2 value from this auxiliary regression is **0.05.**

d. Compute the value of the Breusch-Pagan test statistic.

e. Find the critical point in the appropriate table at 5% significance.

f. Can you reject the null hypothesis of no heteroskedasticity at 5% significance?

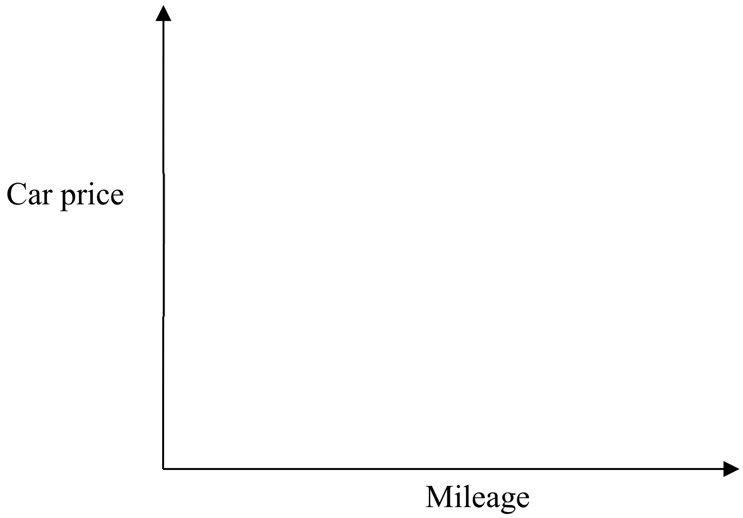
IV. CRITICAL THINKING: [6 pts] Suppose you want to estimate the effect of car mileage on the price of cars, holding all other car features constant (*ceteris paribus*). Using a dataset on n=50 different car models, you plan to estimate the following equation:

$$\textit{car price} = \beta_1 + \beta_2 \, \textit{mileage} + \varepsilon_i$$

where the true value of β_2 is expected to be positive because car buyers value good mileage, all else equal. Now suppose car buyers also value horsepower, and mileage is **negatively** correlated with horsepower.

a. If horsepower is omitted from the regression equation, as shown above, will the least-squares estimator of β_2 be biased *down* (closer to zero), biased *up* (too large), or unbiased? Explain why.

b. Draw a graph showing the true *ceteris-paribus* relationship between the *car price* and *mileage* as a solid line, the likely pattern of observations as dots or circles, and the least squares line as a dotted line.



[end of exam]