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

Signature: ______________________

Printed name: ______________________

FINAL EXAMINATION VERSION B

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. Maximum total points are 200.

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 [2 pts each—44 pts total]

(1) A data set that follows a single individual firm, industry, country, or other geographic area over time is called
a. a cross-section.
b. a time-series.
c. a pooled data set.
d. a panel data set.

(2) Suppose we wish to fit the equation $y = \beta_1 + \beta_2 x$ to data by the method of *least squares*. This method minimizes which function of the data?
a. $f(\beta_1, \beta_2) = \sum(y_i - \beta_1 - \beta_2 x_i)$.
b. $f(\beta_1, \beta_2) = \sum(y_i^2 - (\beta_1 + \beta_2 x_i)^2)$.
c. $f(\beta_1, \beta_2) = \sum(y_i - \beta_1 - \beta_2 x_i)^2$.
d. $f(\beta_1, \beta_2) = \sum|y_i - \beta_1 - \beta_2 x_i|$.
e. $f(\beta_1, \beta_2) = \sum(\beta_1 + \beta_2 x_i)^2$.

(3) An estimator $\hat{\theta}$ of an unknown population parameter θ is said to be *consistent* if
a. $E(\hat{\theta}) = \theta$.
b. $E(\hat{\theta}) = 0$.
c. $\lim_{n\to\infty} E(\hat{\theta}) = 0$.
d. $\lim_{n\to\infty} E(\hat{\theta}) = \theta$.
e. $\lim_{n\to\infty} \text{Prob}(|\hat{\theta} - \theta| > \delta) = 0$, for all $\delta > 0$.

(4) Suppose the p-value for a test statistic is 0.037. If the size of the test is 5 percent, we
a. can reject the null hypothesis.
b. cannot reject the null hypothesis.
c. cannot compute the test statistic.
d. answer cannot be determined from the information given.

(5) Suppose the equation $y = \beta_1 + \beta_2 x$ is fitted to n observations on x and y by the method of least squares. Then the equation

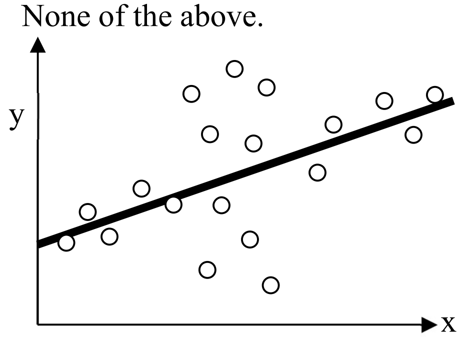
$$\sum_{i=1}^{n}(y_i - \overline{y})^2 - \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

holds

a. if there are enough observations.
b. always.
c. if the fit is good.
d. if the fit is poor.

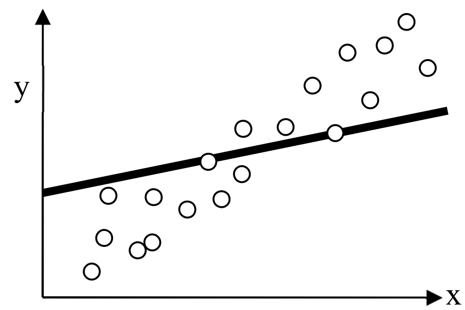
(6) In the graph below, the solid line is the true population regression line and the circles are observations in the sample. Which assumption appears to be violated in this sample?

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.



(7) 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.



(8) The LS predictor of $\hat{y}_{n+1}$ given x_{n+1} differs from the actual value y_{n+1} because

a. the LS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ differ from the true parameter values β_1 and β_2 .
b. the actual value y_{n+1} depends in part on a new error term ε_{n+1}.
c. both of the above.
d. They do not differ. The LS predictor equals exactly the actual value y_{n+1} by definition.

(9) Suppose we have the supply function

$$\ln(y) = 7 + 0.4 \ln(x),$$

where y denotes the quantity of corn supplied (in millions of tons) and x denotes the price of corn. Which of the following is correct?

a. If price increases by 1 dollar, then quantity supplied increases by 0.4 million tons.
b. If price increases by 1%, then quantity supplied increases by 0.4 million tons.
c. If price increases by 1 dollar then quantity supplied increases by 0.4%.
d. If price increases by 1%, then quantity supplied increases 0.4%.

(10) An *outlier* is an observation

a. whose y-value is quite different from the rest of the sample.
b. whose x-value is quite different from the rest of the sample.
c. that contains missing values for x and/or y.
d. that lies exactly on the true regression line.

(11) *Autocorrelation* 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.

(12) 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.

(13) Suppose we estimate the equation

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

If x_{i2} and x_{i3} are *closely but not 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.

(14) 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.

(15) 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 four 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.

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

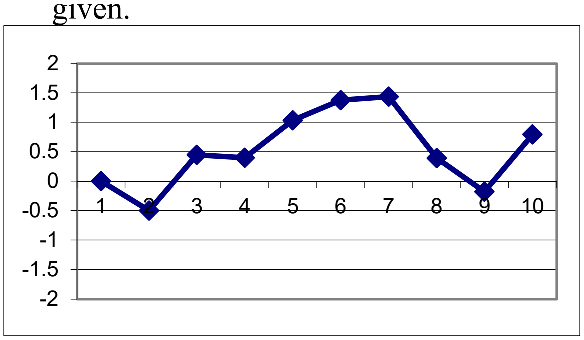
a. increase the degrees of freedom.
b. prevent omitted-variable bias.
c. improve the fit of the equation.
d. raise the sum of squared residuals.

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

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

(18) The time series u_t graphed below has mean zero. It appears to be

a. positively serially-correlated.
b. negatively serially-correlated.
c. serially uncorrelated.
d. Cannot be determined from information given.



(19) If the random process u_t has a unit root, then

a. $Var(u_t) = 1$.
b. the square root of $u_t = 1$.
c. $u_t = 1$.
d. u_t wanders away from its mean.
e. $E(u_t) = 1$.
f. All of the above.

(20) Suppose u_t is defined by

$$u_t = 3.1 + u_{t-1} + \varepsilon_t,$$

where ε_t is an independent identically-distributed process with mean zero and u_0 is nonrandom. Which is true?

a. The mean of u_t depends on t.
b. The series u_t does not tend to return to a trend line.
c. The variance of u_t depends on t.
d. All of the above.

(21) To identify and estimate any *integrated time-series process*, we must first compute

a. differences of the data from the sample mean.
b. cumulative sums of the data.
c. logarithms of the data.
d. first differences of the data.

(22) Given the random walk model with drift

$$y_t = 3 + y_{t-1} + \varepsilon,$$

and $y_T = 2.5$, the two-steps-ahead forecast of y_{T+2} is

a. 0.
b. 3.0.
c. 8.5.
d. 11.5.

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

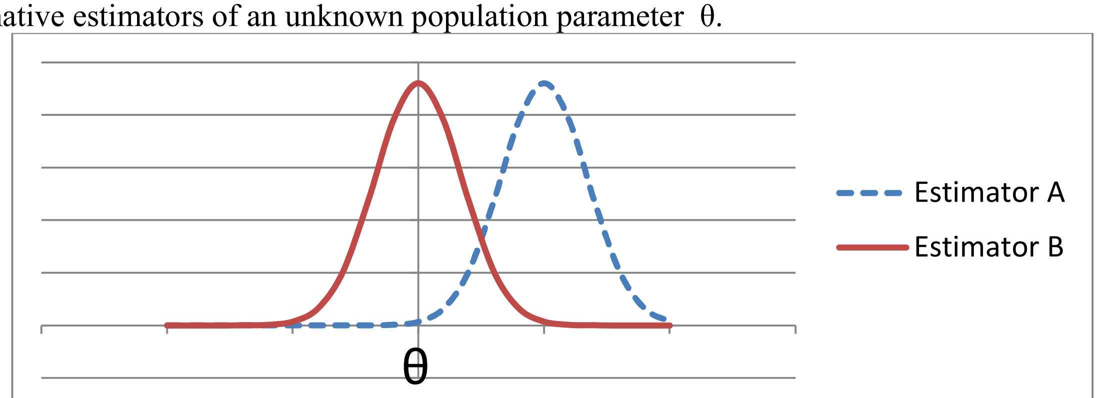
(1) [Summation operator: 6 pts] Consider the following equations. Write "TRUE" for equations that are necessarily true for any values of x_i and y_i. Write "FALSE" for equations that are not necessarily true. All sums run from i = 1, ..., n.

a. $\sum(\alpha\, x_i) = \alpha \sum x_i$	
b. $\sum(x_i - y_i)^2 = \sum x_i^2 - \sum y_i^2$	
c. $\sum(x_i - y_i) = \sum x_i - \sum y_i$	

(2) [Mean and variance of linear function: 4 pts] Suppose X is a random variable with mean E(X) = 3 and variance Var(X) = 2 . Suppose Y = 5 + 3 X .

a. Compute the mean of Y, that is, E(Y).	
b. Compute the variance of Y, that is, Var(Y).	

(3) [Properties of estimators: 4 pts] The graph below shows the density functions for two alternative estimators of an unknown population parameter θ.



a. Which estimator has greater **bias**? Answer "A," "B," or "EQUAL."	
b. Which estimator has greater **variance**? Answer "A," "B," or "EQUAL."	

(4) [Algebraic properties: 8 pts] Suppose the equation $y_i = \beta_1 + \beta_2 x_i, + \varepsilon_i$ is fitted by least squares. Which equations hold necessarily, regardless of the data? Write "TRUE" or "FALSE" in the boxes below.

a. $\sum x_i \hat{\varepsilon}_i = 0$	
b. $\sum (x_i - \bar{x})^2 = 0$	
c. $\sum x_i \, \hat{y}_i = 0$	
d. $\sum \hat{y}_i \, \hat{\varepsilon}_i = 0$	

(5) [Properties: 10 pts] Which assumptions are required for the least-squares estimators to be *unbiased* estimators? Write "REQUIRED" or "NOT REQUIRED" in the boxes below.

a. Variance of error term is exactly one: $Var(\varepsilon_i) = 1$.	
b. Conditional mean of error term is zero: $E(\varepsilon_i \vert x_i) = 0$.	
c. No autocorrelation: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.	
d. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$	
e. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.	

(6) [Properties: 10 pts] Which assumptions are required for the least-squares estimators to be *best linear unbiased estimators* (BLUE)? Write "REQUIRED" or "NOT REQUIRED" in the boxes below.

a. Variance of error term is zero: $Var(\varepsilon_i) = 0$.	
b. Error term is normally-distributed: $\varepsilon_i \sim N(0, \sigma^2)$	
c. Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$, a constant.	
d. Conditional mean of error term is zero: $E(\varepsilon_i \vert x_i) = 0$.	
e. No autocorrelation: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.	

(7) [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 $\mathrm{Var}(\varepsilon_i) = \sigma^2$.	

(8) [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."

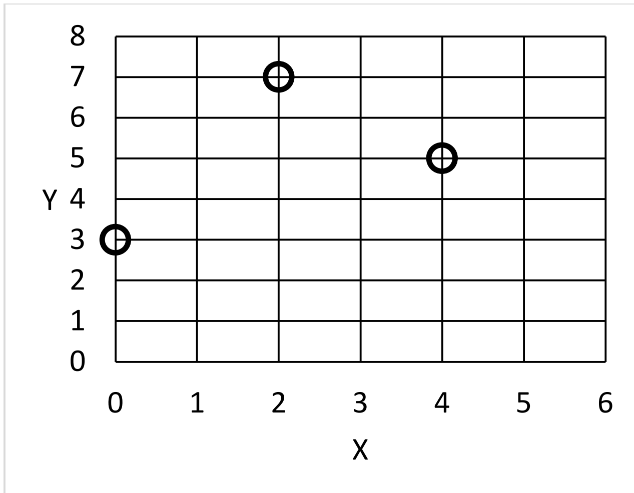
a. The values of the estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$…	
b. The standard errors of the estimated coefficients…	
c. The sum of squared residuals…	
d. The R^2 value…	
e. Theil's adjusted R^2 (also called "$\bar{R}^2$")...	

(9) [Stochastic processes: 10 pts] Let ε_t denote an independent identically-distributed (IID) process, and consider the stochastic processes defined by the equations below. Identify each process by filling in three numbers for ARIMA(p, d, q) .

a. $u_t = 3.2 + u_{t-1} + \varepsilon_t$.	ARIMA(, ,)
b. $u_t = 0.5\, u_{t-1} + 0.3\, u_{t-2} + \varepsilon_t$.	ARIMA(, ,)
c. $u_t = \varepsilon_t + 0.5\, \varepsilon_{t-1}$.	ARIMA(, ,)
d. $u_t = 0.3\, u_{t-1} + \varepsilon_t + 0.4\, \varepsilon_{t-1} - 0.2\, \varepsilon_{t-2}$.	ARIMA(, ,)
e. $u_t = u_{t-1} + \varepsilon_t + 0.3\, \varepsilon_{t-1}$.	ARIMA(, ,)

III. PROBLEMS: Please write your answers in the boxes on this question sheet. Show your work and circle your final answer.

(1) [Definition of least-squares: 15 pts] Suppose we have three observations on X and Y shown in the graph below.



It can be shown that the least-squares estimate for the intercept is $\beta_1 = 4$ and the least squares estimate for the slope is $\beta_2 = 0.5$.

a. Compute the three fitted values $\hat{y}_i$ of this least-squares estimated regression line.

b. Sketch the least-squares fitted line in the graph above.

c. Compute the sample means $(\bar{x}, \bar{y})$ and verify numerically that the fitted line passes through the sample means.

d. Compute the three residuals $\hat{\varepsilon}_i$ of this estimated least-squares regression line.

e. Compute the sum of squared residuals.

(2) [LS confidence intervals, tests: 21 pts] Suppose we estimate the effect of temperature on ice cream sales using a sample of $n=300$ days. Let y_i denote daily ice cream sales and x_i denote temperature. The model $y_i = \beta_1 + \beta_2 x_i$ is estimated with the following results. Numbers in parentheses are standard errors.

Ice cream sales	=	65.0 (20.0)	+	22.0 (5.0)	Temperature

a. [3 pts] Suppose the temperature is 80 degrees. According to these results, what are predicted sales?

b. [3 pts] Suppose the temperature increases by 5 degrees. By how much would ice cream sales increase? That is, what is the predicted change Δy when $\Delta x = 5$?

c. [6 pts] Compute a **90%** confidence interval for the **intercept,** β_1 .

d. [9 pts] Test the hypothesis that temperature has a **positive** effect on ice cream sales, against the null hypothesis that temperature has no effect (a **one-tailed test**) at **10%** significance. Give the value of the test statistic, the critical point(s) from a table, and your conclusion (whether you can reject null hypothesis).

Value of test statistic = ________________. Critical point(s) = __________________.

Can you reject null hypothesis? ______________________.

(3) [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? ______________________.

(4) [Forecasting, trends and seasonality: 8 pts] Suppose we have estimated the following model for sales, using 80 quarterly observations from the first quarter of 2005 to the fourth quarter of 2024.

$$sales_t = 4.5 + 0.2\ trend - 0.6\ q1_t - 0.2\ q2_t - 0.5\ q3_t + \varepsilon_t\ .$$

The regressor "trend" equals 1 in the first quarter of 2005, equals 2 in the second quarter of 2005, and so forth, and equals 80 in the fourth quarter of 2024. The regressors "q1," "q2," and "q3," are quarterly dummy variables for the first, second and third quarters respectively. The error term ε_t is an independent, identically-distributed process with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma^2$, constant.

a. [2 pts] If a dummy variable "q4" for the fourth quarter were also included, then what econometric problem would result?

b. Compute the forecast of sales in the first quarter of 2025.

c. Compute the forecast of sales in the second quarter of 2025.

(5) [Forecasting, AR model: 9 pts] Suppose we have estimated the following model.

$$unemp_t = 2.7 + 0.6\ unemp_{t-1} - 0.2\ unemp_{t-2} + \varepsilon_t\ .$$

where ε_t denotes an independent, identically-distributed process with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma^2$, constant. In our data set, $unemp_{T-1} = 5$ and $unemp_T = 4.5$. Compute the following forecast values.

a. Compute the forecast of $unemp_{T+1}$.

b. Compute the forecast of $unemp_{T+2}$.

c. Compute the limit of the forecast $unemp_{T+h}$ as h approaches infinity. [Hint: this is the unconditional mean of $unemp_t$.].

(6) [Forecasting, MA model: 9 pts] Suppose we have estimated the following model.

$$int_t = 3.8 + \varepsilon_t + 0.4\ \varepsilon_{t-1}\ .$$

where ε_t denotes an independent, identically-distributed process with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma^2$, constant. In our data set, $\hat{\varepsilon}_T = -1.5$. Compute the following forecast values.

a. Compute the forecast of int_{T+1}.

b. Compute the forecast of int_{T+2}.

c. Compute the limit of the forecast int_{T+h} as h approaches infinity. [Hint: this is the unconditional mean of int_t.]

IV. CRITICAL THINKING: [4 pts] To investigate the possible effect of libraries on violent crime, the following regression was estimated using data on U.S. states[1]:

$$y_i = \underset{(535)}{-1091.} + \underset{(1.30)}{11.85}\, x_i$$

where y_i denotes the number of violent crimes committed in state i in 2007 and x_i denotes the number of libraries in state i in the same year. Numbers in parentheses are standard errors of the coefficient estimates. The estimate for the slope coefficient is significantly greater than zero at 0.1 percent significance. So the number of violent crimes is clearly *positively correlated* with the number of libraries across states. Is this evidence that libraries *cause* violent crime? If yes, explain why. If no, explain why not and suggest a better way to estimate the regression equation using these state-level data.

[end of exam]

[1] These are actual least-squares estimates using data from the *Statistical Abstract of the United States*, 2010. n=50.