# Question 1 of 21

L1.T2.88.2

Our linear regression produces a high coefficient of determination ($R^2$) but few significant t ratios. Which assumption is **most likely** violated?

- **A.** Homoscedasticity
- **B.** Multicollinearity
- **C.** Error term is normal with mean = 0 and constant variance = sigma ^2
- **D.** No autocorrelation between error terms

# Question 2 of 21

L1.T2.88.1

Which assumption underlies the multiple linear model but **NOT** the two-variable regression model?

- **A.** Homoscedasticity
- **B.** Multicollinearity
- **C.** Error term is normal with mean = 0 and constant variance = sigma ^2
- **D.** No autocorrelation between error terms

# Question 3 of 21

L1.T2.88.3

Our regression function is given by Y(t) = B1 + B2^2*X(t). Which assumption is violated?

- **A.** Homoscedasticity
- **B.** No autocorrelation between error terms
- **C.** No material specification error (specification bias)
- **D.** Model is linear

# Question 4 of 21

L1.T2.88.4

We observe the variance of the error term is an increasing function of the explanatory variable. Which assumption is violated?

- **A.** Homoscedasticity
- **B.** Multicollinearity
- **C.** No autocorrelation between error terms
- **D.** Model is linear

# Question 5 of 21

L1.T2.88.5

We reject the null hypothesis in a Durbin-Watson D test. Which assumption is violated?

- **A.** Homoscedasticity
- **B.** Multicollinearity
- **C.** No autocorrelation between error terms
- **D.** Model is linear

# Question 6 of 21

L1.T2.95.3

Which of the following is most likely in the case of high multicollinearity?

- **A.** Low F ratio and insignificant partial slope coefficients
- **B.** High F ratio and insignificant partial slope coefficients
- **C.** Low F ratio and significant partial slope coefficients
- **D.** High F ratio and significant partial slope coefficients

# Question 7 of 21

L1.T2.88.6

In a valid two-variable regression where the coefficient of determination is positive ($R^2 > 0$) and the slope is negative (in Y = b1 + b2*X, b2 < 0), which is true about the sum of the product of the residuals and the explanatory variables, i.e., SUM of {product of each e(i) and X(i)}?

- **A.** Less than zero
- **B.** Zero
- **C.** Greater than zero
- **D.** Need more information

# Question 8 of 21

P1.T2.218.2

Assume we have confirmed that all three of Stock & Watson's assumptions are true for our OLS linear regression model; i.e., the error term has a mean of zero conditional on the regressor; the [X(i),Y(i)] observations are i.i.d. random draws; and large outliers are unlikely. Our OLS regression model is: Y(i) = B(0) + B(1)*X(i) + u(i). Each of the following is true **EXCEPT** for:

- **A.** Whether the errors are homo- or heteroskedastic, the OLS estimators are unbiased, consistent, and asymptotically normal
- **B.** If the errors are heteroskedastic, we can compute heteroskedasticity-robust standard errors
- **C.** If it is true that, in addition to the three assumptions above, that the errors are homoskedastic, then our OLS estimator for B(1) is BLUE
- **D.** As heteroskedasticity is a special case of homoskedasticity, and given that homoskedasticity is more most prevalent, the safest practice is to employ homoskedasticity-robust standard errors

# Question 9 of 21

P1.T2.218.1

We want to regress hourly Earnings (the regressand) against years of Education (the regressor) based on the following OLS regression model: Earnings(i) = B(0) + B(1)*Education(i) + u(i), where u(i) is the error term. After we run the regression, which of the following statements **MOST NEARLY** demonstrates homoskedasticity?

- **A.** Education(i) is not a linear function of any other regressor
- **B.** Earnings(i) is independent of Education(i)
- **C.** The variance of the error, u(i), is independent of Education(i)
- **D.** The error term has a conditional mean of zero, E[u(i) | Education(i)] = 0

# Question 10 of 21

P1.T2.220.2

A multiple regression model, on a small sample of monthly returns for one year, has two regressors and is given by: Y(i) = 10.0 + 1.46*X(1,i) – 0.82*X(2,i) + u(i). The number of observations (n) is 12. The sum of squared residuals (SSR) is 106.0. The total sum of squares (TSS) is 166.0. What are, respectively, the standard error of the regression (SER) and the adjusted $R^2$?

- **A.** SER = 0.89 and Adjusted $R^2$ = -0.11
- **B.** SER = 2.25 and Adjusted $R^2$ = 0.64
- **C.** SER = 3.43 and Adjusted $R^2$ = 0.22
- **D.** SER = 11.87 and Adjusted $R^2$ = 0.64

# Question 11 of 21

P1.T2.220.1

Each of the following is true about the adjusted $R^2$ **EXCEPT** which is false?

- **A.** Adjusted $R^2$ = 1 - (SSR/TSS)*[(n-1)/(n-k-1)]
- **B.** Adding a regressor (independent variable) always causes the adjusted $R^2$ to decrease
- **C.** Adjusted $R^2$ is always less than $R^2$
- **D.** The adjusted $R^2$ can be negative

# Question 12 of 21

P1.T2.220.3

With respect to a linear regression with multiple regressors, each of the following is true **EXCEPT** which statement is false:

- **A.** Imperfect multicollinearity implies that we cannot estimate precisely ANY of the partial effects (slope coefficients)
- **B.** Imperfect multicollinearity means that two or more of the regressors are highly correlated
- **C.** The dummy variable trap is an example of perfect multicollinearity
- **D.** In contrast to perfect muticollinearity, imperfect multicollinearity it is not necessarily an error but likely just a feature of the OLS

# Question 13 of 21

P1.T2.218.3

You presented a regression model to your boss, the Chief Risk Officer (CRO). She is a certified FRM, so you know that she knows statistics, although she laments the decision to replace rigorous Gujarati with a softer, gentler Stock & Watson. She queries you on the dataset and your regression, and you admit to two realities: First, the error term is heteroskedastic. Second, there are many extreme outliers in the dataset. Your boss makes the following assertions:

I. "It is okay, for our purposes, that the error term is heteroskedastic: the slope (B1) estimator remains efficient and BLUE."

II. "Since we have many extreme outliers, the least absolute deviations (LAD) is a viable alternative to OLS, because its estimators may be more efficient (i.e., have smaller variances)"

Which of your boss' statements is (are) **TRUE**?

- **A.** Neither
- **B.** I. only
- **C.** II. only
- **D.** Both are true

# Question 14 of 21

P1.T2.222.1

You estimate the relationship between a security’s return and a market index under the assumption of homoskedasticity of the error terms. The regression output is as follows: Predicted[Return(i)] = 2.85% + 1.490*Index(i), and the standard error on the slope is 0.820. The homoskedasticity-only “overall” regression F-statistic for the hypothesis that the regression $R^2$ is zero is approximately?

- **A.** 1.35
- **B.** 1.82
- **C.** 3.30
- **D.** 10.90

# Question 15 of 21

P1.T2.222.2

You test the three-factor Fama-French model with a multiple OLS regression, which has three regressors: Return(i) = 1.2% + 0.38*HML + 1.23*SMB + 0.17*UMD and $R^2$ = 0.520, where HML is “high minus low” (book-to-market), SMB is “small minus big” (small capitalization), and UMD is “up minus down” (momentum). The number of observations, n, is 384. Then you perform a restricted regression which imposes the joint null hypothesis that the true coefficients on SMB and UMD are zero. The restricted OLS regression is given by Return(i) = 0.9% + 0.44*HML and $R^2$ = 0.490. Please note: as the unrestricted regression has three regressors and the restricted regression hypothesizes two of the coefficients are zero, we have unrestricted k = 3 and number of restriction (q) = 2. What is the homoskedasticity-only F-statistic?

- **A.** 1.7
- **B.** 3.4
- **C.** 11.9
- **D.** 23.3

P1.T2.20.19.2

Josh regressed house prices (as the response or dependent variable) against two explanatory variables: square footage (SQFEET) and the number of rooms in the house (ROOMS). The dependent variable, PRICE, is expressed in thousands of dollars ($000); e.g., the average PRICE is $728.283 because the average house price in the sample of 150 houses is $728,283. The units of SQFEET are unadjusted units; e.g., the average SQFEET in the sample is 1,893 ft^2. The variable ROOMS is equal to the sum of the number of bedrooms and bathrooms; because much of the sample is 2- and 3-bedroom houses with 2 baths, the average of ROOM is 4.35. Josh's regression results are displayed below.

House Price regressed against ft^2 (SQFEET) + ROOMS(#)
House Price in Thousands **($000)** of dollars

| Coefficient | Estimate | Std Error | t-stat | p value |
|---|---|---|---|---|
| **(Intercept)** | **-9.457** | **69.598** | **-0.136** | **$8.92 \times 10^{-1}$** |
| **SQFEET** | **0.370** | **0.027** | **13.522** | **$1.40 \times 10^{-27}$** |
| **ROOMS** | **8.784** | **8.612** | **1.020** | **$3.09 \times 10^{-1}$** |

Residual standard error: 211.1 on 147 degrees of freedom
Multiple R-squared: 0.5548, Adjusted R-squared: 0.5488
F-statistic: 91.61 on 2 and 147 DF, p-value: < 2.2e-16

Josh is concerned that the data might not be homoscedastic. He decides to conduct a White test for heteroskedasticity. In this test, he regresses the squared residuals against each of the explanatory variables and the cross-product of the explanatory variables (including the product of each variable with itself). The results of this regression are displayed below.

RESIDUAL^2 regressed against SQFEET + ROOMS + SQFEET^2 + ROOMS^2 + ROOMS*SQFEET
White's Test for Heteroskedasticity

| Coefficient | Estimate | Std Error | t-stat | p value |
|---|---|---|---|---|
| **(Intercept)** | **$1.59 \times 10^{5}$** | **$7.49 \times 10^{4}$** | **2.128** | **0.035** |
| **SQFEET** | **$-1.96 \times 10^{2}$** | **$6.94 \times 10^{1}$** | **-2.828** | **0.005** |
| **ROOMS** | **$-1.58 \times 10^{4}$** | **$1.49 \times 10^{4}$** | **-1.059** | **0.292** |
| **SQFEET^2** | **$6.02 \times 10^{-2}$** | **$1.74 \times 10^{-2}$** | **3.454** | **0.001** |

| | | | | |
|---|---|---|---|---|
| **ROOMS^2** | **$1.49 \times 10^2$** | **$1.27 \times 10^3$** | **0.118** | **0.906** |
| **SQFEET*ROOMS** | **$1.00 \times 10^1$** | **4.94** | **2.030** | **0.044** |

Residual standard error: 76920 on 144 degrees of freedom
Multiple R-squared: 0.3381, Adjusted R-squared: 0.3152, F-statistic: 14.71 on 5 and 144 DF, p-value: 1.201e-11

Is the data heteroskedastic?

- **A.** No, the data is probably homoskedastic because all coefficients are highly significant
- **B.** No, the data is probably homoskedastic because the F-statistic does not imply the rejection of the null hypothesis
- **C.** Yes, the data is probably heteroskedastic because the m-fold cross-validation failed
- **D.** Yes, the data is probably heteroskedastic because the residual variance has some dependence on SQFEET

# Question 17 of 21

P1.T2.20.19.3

Emily works for an insurance company and she has regressed medical costs (aka, the response or dependent variable) for a sample of patients against three independent variables: AGE, BMI, and CHARITY. The sample's average age is 38.5 years. Body mass index (BMI) is mass divided by height squared and the sample's average BMI is 22.24 kg/m^2. CHARITY is the dollar amount of charitable spending in the last year; the sample average is $511.66 donated to charity in the last year. Emily's regression results are displayed below.

Medical COST regressed against AGE + BMI + CHARITY($)
Simulated data

| Coefficient | Estimate | Std Error | t-stat | p value |
|---|---|---|---|---|
| **(Intercept)** | **-165.25** | **774.73** | **-0.21** | **$8.32 \times 10^{-1}$** |
| **AGE** | **63.67** | **29.66** | **2.15** | **$3.81 \times 10^{-2}$** |
| **BMI** | **102.46** | **13.61** | **7.53** | **$4.09 \times 10^{-9}$** |
| **CHARITY** | **-0.95** | **0.83** | **-1.14** | **$2.63 \times 10^{-1}$** |

Residual standard error: 325.1 on 39 degrees of freedom
Multiple R-squared: 0.6961, Adjusted R-squared: 0.6727, F-statistic: 29.77 on 3 and 39 DF, p-value: 3.514e-10

Emily wonders if the data exhibits multicollinearity. In order to test for multicollinearity, she conducts three additional regressions. She regresses each of the explanatory variables against the other two explanatory variables. Below are summarized the R-squared (R^2) values for each of those regressions:

Each response variable regressed against the others
Testing for multicollinearity

| Regression | Response | Explanatory | R-squared |
|---|---|---|---|
| **1** | **AGE** | **BMI + CHARITY** | **0.927** |
| **2** | **BMI** | **AGE + CHARITY** | **0.048** |
| **3** | **CHARITY** | **AGE + BMI** | **0.926** |

Note: According to GARP, the standard test of multicollinearity is the variance inflation factor (VIF)

Does Emily’s data contain multicollinearity?

- **A.** No, because none of the variance inflation factors (VIFs) are excessive
- **B.** No, because all estimates are significant and the Adjusted R-squared is above 0.50
- **C.** Yes, because two of the variance inflation factors (VIFs) are excessive
- **D.** Yes, because the R-squared of BMI is less than 5.0%

# Question 18 of 21

P1.T2.20.19.1

Jane manages a market-neutral equity fund for her investment management firm. The fund's market-neutral style implies (we will assume) that the fund's beta with respect to the market's excess return is zero. However, the fund does seek exposure to other factors. The size factor captures the excess return of small-capitalization stocks (SMB = "small minus big"). Jane tests her portfolio's exposure to the size factor by regressing the portfolio's excess return against the size factor returns. Her regression takes the form PORTFOLIO(i) = α + β1×SMB(i) + ε(i). The results of this single-variable (aka, simple) regression are displayed below.

| Market-neutral portfolio excess returns regressed against SMB (but HML is an ***omitted variable***) | | | | |
|---|---|---|---|---|
| Coefficient | Estimate | Std Error | t-stat | p value |
| **(Intercept)** | **0.0588** | **0.0053** | **$1.11 \times 10^{1}$** | **$4.74 \times 10^{-20}$** |
| **SMB** | **0.6771** | **0.1064** | **6.37** | **$3.86 \times 10^{-9}$** |

In this simple regression, we can observe that SMB's coefficient is 0.6771 and significant. Jane is concerned that this simple regression might suffer from omitted variable bias. Specifically, she thinks the value factor has been omitted. The value factor captures the excess returns of value stocks (HML = "high book-to-market minus low book-to-market")'. She confirms that the omitted variable, HML, is associated with her response variable. Further, the omitted variable, HML, is correlated to SMB. The correlation between HML and SMB is 0.30.

Further, it happens to be the case that the volatilities of SML and SMB are identical: σ(MHL) = σ(SMB) = 0.010. Jane runs a multivariate regression with both explanatory variables, SMB and HML; in this regression, HML's beta coefficient is 0.7240 such that the new term is β2×HML(i) = 0.7240×HML(i). Which of the following is nearest to the revised SMB coefficient; i.e., what is the revised β1?

- **A.** 0.677
- **B.** 0.230

- **C.** 0.460
- **D.** 1.253

# Question 19 of 21

P1.T2.20.20.1

Below are displayed 15 pairwise (X,Y) trials. The simple regression line based on all 15 observations is given by Y1 = 0.488 + 0.425*X. We consider the possibility that the 12th Trial, given by point (X = 2.50, Y = -3.00) might be an outlier. If this point is removed, then the regression based on the remaining 14 observations is given by Y2 = 0.761 + 0.574*X. These results are displayed, including selected summary statistics.

**All 15 observations: Y1 = 0.488 + 0.425 * X**
**14 observations (excludes 12th trial): Y2 = 0.761 + 0.574 * X**

| Trial | X | Y | Y1 | Y2 | $(Y1 - Y2)^2$ | (Y1 – Y) | $(Y1 - Y)^2$ |
|---|---|---|---|---|---|---|---|
| 1 | -3.00 | -0.89 | -0.79 | -0.96 | 0.030 | 0.103 | 0.011 |
| 2 | -2.50 | -1.36 | -0.57 | -0.67 | 0.010 | 0.782 | 0.612 |
| 3 | -2.00 | 0.05 | -0.36 | -0.39 | 0.001 | -0.416 | 0.173 |
| 4 | -1.50 | -0.12 | -0.15 | -0.10 | 0.002 | -0.032 | 0.001 |
| 5 | -1.00 | -0.99 | 0.06 | 0.19 | 0.015 | 1.050 | 1.102 |
| 6 | -0.50 | 0.69 | 0.28 | 0.47 | 0.039 | -0.411 | 0.169 |
| 7 | 0.00 | 1.22 | 0.49 | 0.76 | 0.074 | -0.731 | 0.535 |
| 8 | 0.50 | 1.48 | 0.70 | 1.05 | 0.120 | -0.777 | 0.604 |
| 9 | 1.00 | 1.38 | 0.91 | 1.33 | 0.177 | -0.468 | 0.219 |
| 10 | 1.50 | 2.75 | 1.13 | 1.62 | 0.245 | -1.628 | 2.651 |
| 11 | 2.00 | 2.13 | 1.34 | 1.91 | 0.324 | -0.792 | 0.627 |
| **12** | **2.50** | **-3.00** | **1.55** | **2.19** | **0.414** | **4.552** | **20.717** |
| 13 | 3.00 | 2.59 | 1.76 | 2.48 | 0.515 | -0.827 | 0.683 |
| 14 | 3.50 | 1.65 | 1.98 | 2.77 | 0.627 | 0.324 | 0.105 |
| 15 | 4.00 | 2.92 | 2.19 | 3.06 | 0.750 | -0.730 | 0.532 |
| **Average** | **0.500** | **0.701** | **0.701** | **1.047** | **0.223** | **0.000** | **s^2=1.916** |
| **Sum** | | | | | **3.341** | **0.000** | **28.740** |

According to Cook's distance, is the 12th Trial an outlier?

- **A.** No, because its Cook's distance is negative

- **B.** No, because its Cook's distance is 3.341/(2*1.916) = 0.872
- **C.** Yes, because its Cook's distance is +0.15 (as given by the slope change)
- **D.** Yes, because its Cook's distance is 1.916/(2*0.223) = +4.301

# Question 20 of 21

P1.T2.20.20.2

Patricia needs to specify a regression model, but she is only given nine (Y,x) pairwise observations, as displayed below. She employs m-fold cross-validation (CV) and selects three folds (aka, three blocks). Each of her three candidate regression models is “trained” on two of the folds, so that model can be “tested” on the remaining fold. The first model (M1 in light green) is a regression that is “trained” on the first six observations, and it is given M1: Y1 = 0.820 + 1.166*X. The second model (M2 in slightly darker green) is a regression that is “trained” on the last six observations, and it is given M2: Y2 = 0.089 + 1.117*X. The third model (M3 in darkest green) is a regression that is “trained” on the first three and last three observations, and it is given M3: Y3 = 1.1773 + 0.862*X.

| | | | Three models | | | Difference between Predicted Y1\|Y2\|Y3 & Observed Y | | | Residual sum of squares (RSS) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | M1 | M2 | M3 | | | |
| Obs # | Y | X | Y1 | Y2 | Y3 | Y-Y1 | Y-Y2 | Y-Y3 | $(Y-Y1)^2$ | $(Y-Y2)^2$ | $(Y-Y3)^2$ |
| 1 | 2.20 | 1.00 | **1.99** | 1.21 | **2.64** | 0.21 | 0.99 | (0.44) | 0.05 | **0.99** | 0.19 |
| 2 | 3.50 | 2.00 | **3.15** | 2.32 | **3.50** | 0.35 | 1.18 | 0.00 | 0.12 | **1.39** | 0.00 |
| 3 | 5.20 | 3.00 | **4.32** | 3.44 | **4.36** | 0.88 | 1.76 | 0.84 | 0.78 | **3.10** | 0.71 |
| 4 | 2.80 | 4.00 | **5.48** | **4.56** | 5.22 | (2.68) | (1.76) | (2.42) | 7.20 | 3.09 | **5.86** |
| 5 | 6.90 | 5.00 | **6.65** | **5.67** | 6.08 | 0.25 | 1.23 | 0.82 | 0.06 | 1.50 | **0.67** |
| 6 | 8.80 | 6.00 | **7.81** | **6.79** | 6.95 | 0.99 | 2.01 | 1.85 | 0.97 | 4.03 | **3.44** |
| 7 | 5.70 | 7.00 | 8.98 | **7.91** | **7.81** | (3.28) | (2.21) | (2.11) | **10.76** | 4.88 | 4.44 |
| 8 | 11.30 | 8.00 | 10.15 | **9.03** | **8.67** | 1.15 | 2.27 | 2.63 | **1.33** | 5.17 | 6.92 |
| 9 | 8.60 | 9.00 | 11.31 | **10.14** | **9.53** | (2.71) | (1.54) | (0.93) | **7.35** | 2.38 | 0.87 |
| | | **Intercept** | **0.820** | **0.089** | **1.773** | | | **Total RSS** | **28.62** | **26.53** | **23.09** |
| | | **Beta** | **1.166** | **1.117** | **0.862** | | | **CV RSS** | **19.44** | **5.47** | **9.97** |
| | | **R^2** | **0.721** | **0.509** | **0.767** | | | | **M1** | **M2** | **M3** |
| | | | **M1** | **M2** | **M3** | | | | | | |

For each model, the residual (i.e., the difference between the predicted and observed Y) is displayed. The final three columns display the squared residuals. If her criteria for model selection follows the principles of m-fold cross-validation then which of the three models should Patricia select?

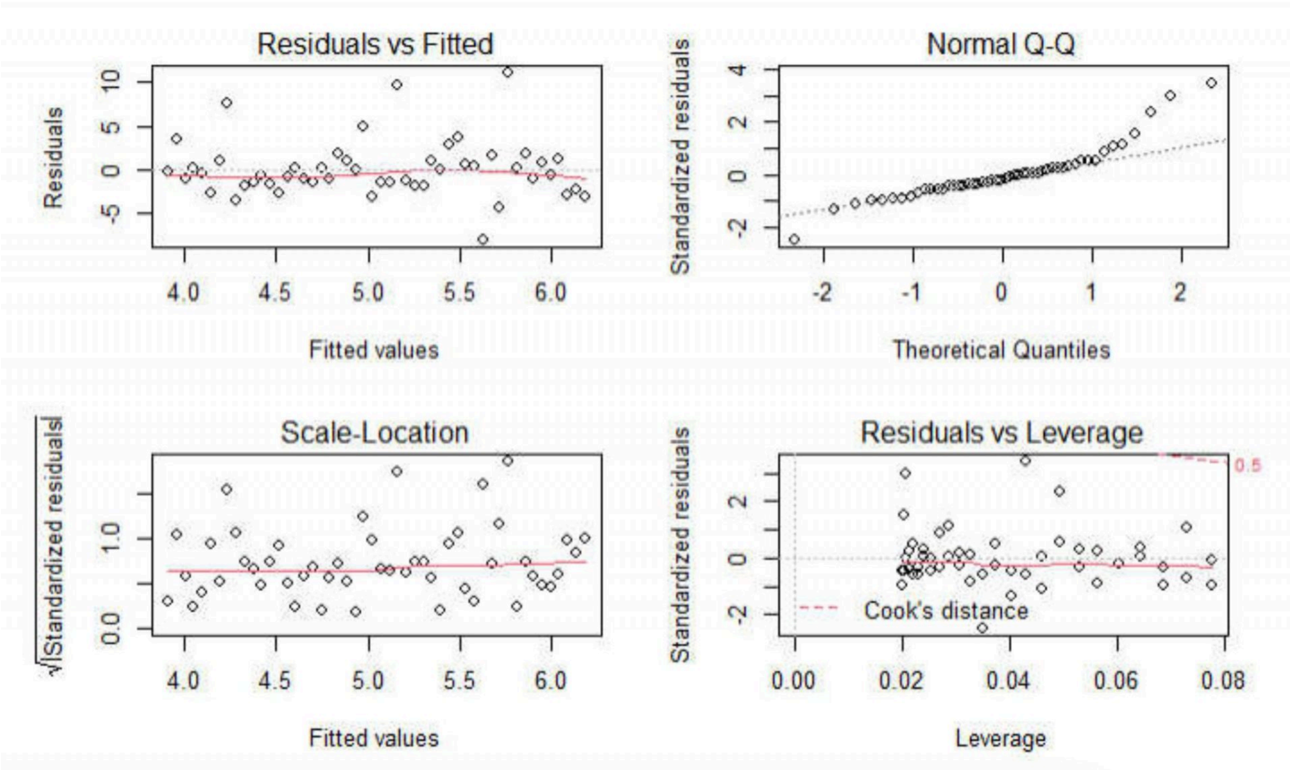
- **A.** She should select M1 because it has the highest CV RSS
- **B.** She should select M2 because it has the lowest CV RSS

- **C.** She should select M3 because it has the lowest total RSS
- **D.** She should select M3 because it has the highest coefficient of determination

# Question 21 of 21

P1.T2.20.20.3

Patrick generated a simple regression line for a sample of 50 pairwise observations. After generating the regression model, he ran R's built-in plot(model) function which produces a standard set of regression diagnostics. These four plots are displayed below.




About these diagnostic plots, which of the following statements is **TRUE**?

- **A.** There are many outliers
- **B.** The data is significantly heteroskedastic
- **C.** The residuals are a bit heavy-tailed (non-normal) on the right side

- **D.** The residuals reveal that the relationship between the explanatory and response variable is non-linear