1. Ursula Beijer is the portfolio manager of Scheldt Equities Fund. The fund holds stocks from emerging markets, and Beijer is concerned about the risk of government debt default. She asks Roland Drenth, a quantitative analyst at the firm, to build a multiple regression model based on macroeconomic factors to predict whether or not a country will default. Drenth explains that the objective of the model is to predict a discrete variable that takes the value of:

    - 1 when the model predicts a debt default or
    - 0 when a default is not expected.

    Beijer also wants to investigate how the returns of the Scheldt fund are influenced by distinct market factors. She asks Drenth to build a regression model of the fund's excess returns against the three factors in the Fama-French model. Drenth presents the following multiple regression equation:

    $$SEF_i = b_0 + b_1 MKTRF_i + b_2 SMB_i + b_3 HML_i + \varepsilon_i, \text{ where:}$$

    - SEF = Excess return of the Scheldt Equities Fund
    - MKTRF = Market excess return over the risk-free rate
    - SMB = Small minus big (size effect)
    - HML = High minus low (value premium)

    Beijer asks Drenth how to interpret the regression variables and estimated coefficients. Drenth makes the following statements:

    Statement 1:The intercept b_0 is the fund's expected excess return if the error term ε_i is zero.
    Statement 2:The slope b_1 describes how SEF is affected by a unit change in MKTRF while holding SMB and HML constant.
    Statement 3:MKTRF, SMB, and HML constitute the stochastic part of the model.

    Drenth then runs the regression with 60 monthly return observations, stated in whole percentages (ie, 1 = 1%), and arrives at the following equation with the estimated coefficients:

    $$SEF = 0.06 + 0.51MKTRF - 0.72SMB + 0.30HML$$

    Beijer asks for a numeric example, and Drenth estimates Scheldt's excess return in a hypothetical month, given the following assumptions:

    - MKTRF = 1.00%,
    - SMB = 1.00%, and
    - HML = 0.50%.

    Beijer wants to use the regression model to predict the fund's performance. However, Drenth advises Beijer that several additional steps must be taken before deeming the model acceptable for making predictions and performing further analysis.

Question 1 of 4

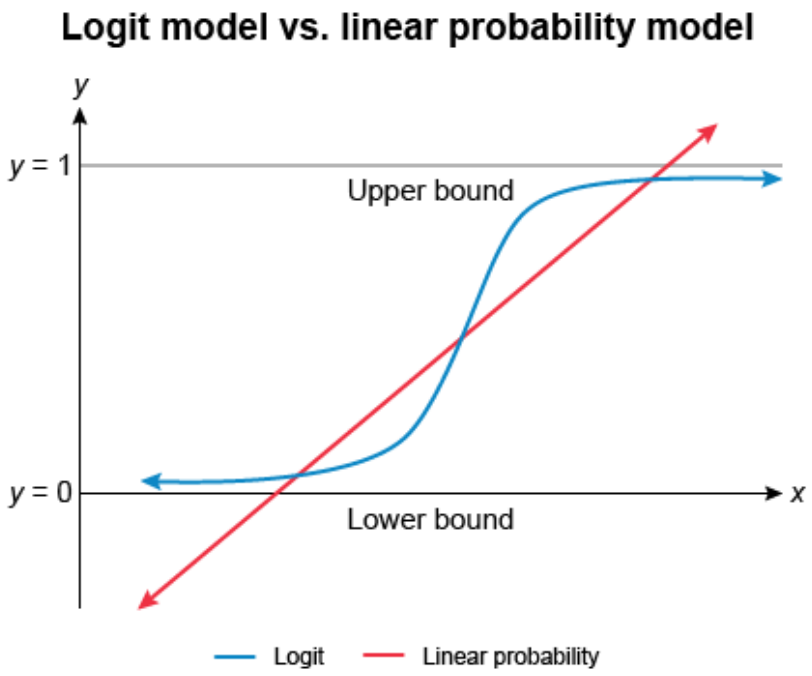
To estimate whether or not a country will default, Drenth should *most appropriately* use a:

a. logistic regression model.

b. restricted regression model.

c. regression model with a dummy independent variable.

Explanation:



Multiple regression models are used to explain and predict the value of the **dependent** variable (Y) based on changes in a set of **independent** (or explanatory) variables (X_1,X_2,…,X_k). Multiple regressions can address different types of **investment problems**, including:

- Analyzing and explaining the relationship between variables (eg, explaining returns based on volatility)
- Forecasting a variable based on changes in other variables (eg, predicting credit ratings based on financials)
- Testing existing financial theories (eg, testing the Fama-French model)

In this situation, Beijer wants to use a set of macroeconomic factors (ie, independent variables) to forecast whether or not a country will default on its debt. Drenth proposes using a model to predict a discrete variable known as a **qualitative dependent variable**, which takes the values of 1 (default) or 0 (no default).

Regressing a discrete (ie, not continuous) variable using a traditional regression model would likely result in coefficients that may generate invalid results (ie, predicted values of less than 0 or greater than 1). A **logistic regression** (logit) model represents the dependent variable as a natural logarithm of probability ratios, **confining results** to a **range between 0** and **1**.

Therefore, a logistic regression is the **appropriate approach** to forecast whether a default will occur.

(Choice B) A restricted model excludes independent variables from the full (unrestricted) model. This is not the type of model Drenth suggests.

(Choice C) A dummy variable is a qualitative independent variable used in regression models. Drenth's proposed model requires a discrete *dependent*, not independent, variable.

Things to remember:
Multiple regressions can explain the relationship between variables, forecast a variable based on changes in other variables, and test existing financial theories. A logistic regression (logit) model should be used when the dependent variable is discrete (ie, not continuous).

LOS - Describe the types of investment problems addressed by multiple linear regression and the regression process

---

2. **Question 2 of 4**

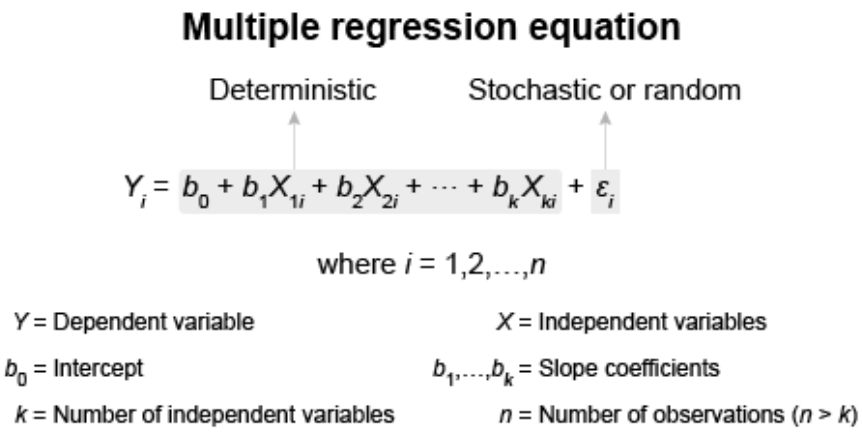
Which of the following statements about variables and coefficients is *correct*?

a. Statement 1

b. Statement 2

c. Statement 3

Explanation:



The **slope coefficient** ($b_1,b_2,\ldots,b_k$) in a multiple regression is known as a **partial** slope coefficient since it describes how the dependent variable, *Y*, is affected by changes in one independent variable while holding the other independent variables constant. If the slope associated with a given independent variable ($X_1,X_2,\ldots,X_k$) is positive (negative), an increase in that independent variable will increase (decrease) *Y*. In this situation, for example, the **slope b_1** describes how **SEF** is **affected** by a unit **change** in **MKTRF** while holding **SMB** and **HML constant**.

The intercept is the value of *Y* when all independent variables equal zero, that is, it is the expected variation in the estimate that is unrelated to the independent variables. The Scheldt fund is expected to have an excess return of b_0 (ie, the intercept) if MKTRF, SMB, and HML are zero and there is no random error **(Choice A)**.

The intercept and slope coefficients form the deterministic part of the regression model, producing the same output for the same set of independent variables. The uncertainty is added to the model by the stochastic (ie, random) part of the model, which is the error term ε **(Choice C)**.

Things to remember:
A partial slope coefficient describes how *Y* is affected by changes in one independent variable while holding the other independent variables constant. The intercept is the value of *Y* when all independent variables equal zero and there is no random error; it is the portion of *Y*'s value unrelated to the independent variables. The slope coefficients and the intercept form the deterministic part of the regression, while the error term is the model's stochastic part.

LOS - Formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients

---

3. **Question 3 of 4**

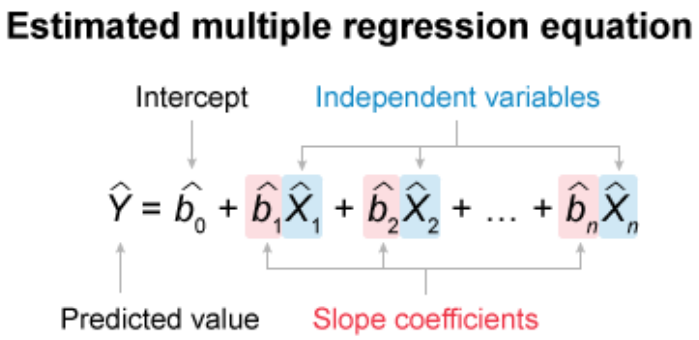
Based on the estimated regression equation and on Drenth's assumptions, Scheldt's estimated excess return is *closest* to:

a. −0.06%

b. 0.00%

c. 0.15%

Explanation:



The **value** of a **dependent variable** is predicted by using the **intercept**, slope **coefficients**, and **independent variables** expressed in the **estimated** multiple **regression** equation. Each coefficient estimates how much the dependent variable is expected to change given a one-unit change in the corresponding independent variable and holding all other independent variables constant. The intercept is the estimate for the dependent variable if the independent variables each equal zero.

In this scenario, the variables are all excess returns, so their units are percentage points. Based on the estimated regression equation, an increase of 1 percentage point in the independent variable:

- MKTRF increases SEF by 0.51.
- SMB decreases SEF by 0.72.
- HML increases SEF by 0.30.

These changes, however, assume that the other variables remain constant. To calculate the excess return in a given month, Drenth must add the intercept to the product of each independent variable and respective slope coefficient. Therefore, in a hypothetical month in which MKTRF = 1.00%, SMB = 1.00%, and HML = 0.50%:

$$SEF = 0.06 + 0.51MKTRF - 0.72SMB + 0.30HML$$

$$SEF = 0.06 + (0.51 \times 1) - (0.72 \times 1) + (0.30 \times 0.50) = 0$$

Therefore, the **estimated excess return** in the month is **0.00%**.

(Choice A) −0.06% results from failing to include the intercept (0.06) in the calculation.

(Choice C) 0.15% results from using a change of 1% in HML, instead of 0.50%.

Things to remember:
In an estimated multiple regression equation, the value of the dependent variable is predicted by adding the intercept to the product of each independent variable and respective slope coefficient.

LOS - Formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients

---

4. **Question 4 of 4**

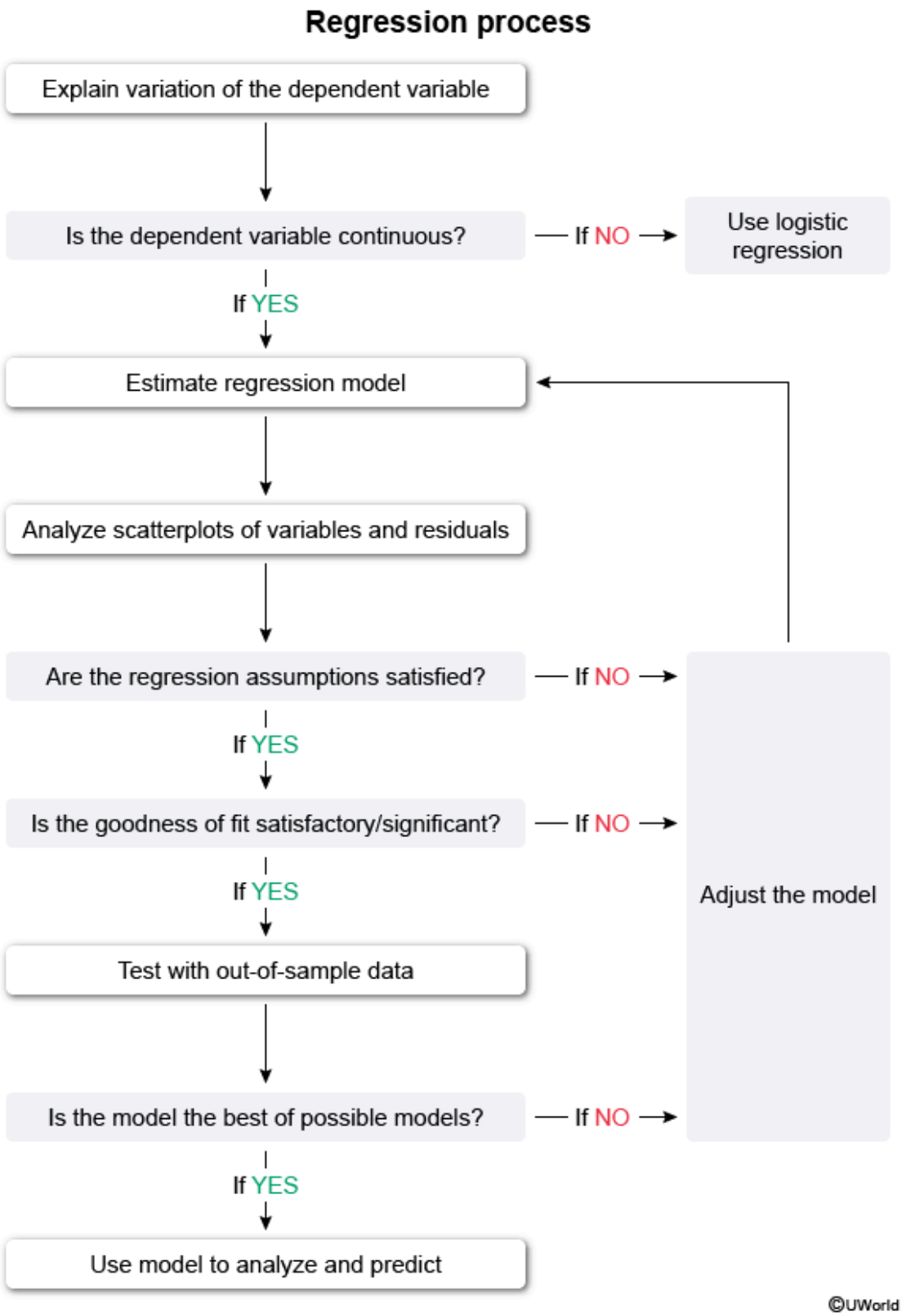
After estimating the regression model, Drenth's *most appropriate* next step is to:

a. analyze the residuals.

b. measure the goodness of fit.

c. test the model's performance with out-of-sample data.

Explanation:



A multiple regression model may be used to explain the relationship between variables, forecast a variable based on changes in other variables, and test existing financial theories. However, defining the appropriate dependent variable, the significant independent variables, and the form and goal of the model requires **several decisions** from the analyst.

Before estimating the regression model, the analyst should analyze the variables to define the correct model type: a logistic regression should be used when the dependent variable is discrete (ie, not continuous). After defining the correct model type, the regression equation can be estimated. Subsequently, although analysts may use different regression processes, these steps must be taken in the following order:

1. **Assess** whether the regression assumptions are **satisfied** by **analyzing**scatterplots, which are graphs that display the relationship between two variables (eg, Y and X_1, X_1 and X_2) or between variables and outputs (eg, residual plots).

2. Investigate how well the independent variables explain Y, using measures of goodness of fit. Also, test the significance of the variables and of the overall model.

3. Validate the performance of the model using new data (ie, out-of-sample data) to fine-tune the model and test its ability to predict.

Therefore, **before assessing** the **fit**, **significance**, and **performance** of the model, it is **essential** to **check** whether the **regression assumptions** are satisfied. If any assumption is violated, the model must be adjusted and the regression equation must be recalculated **(Choices B and C)**.

Things to remember:
Appropriately defining a regression model demands several decisions from the analyst. After analyzing the variables and estimating the equation, the analyst should assess whether the regression assumptions are violated, test the model's fit and significance, and validate the model's performance.

LOS - Describe the types of investment problems addressed by multiple linear regression and the regression process

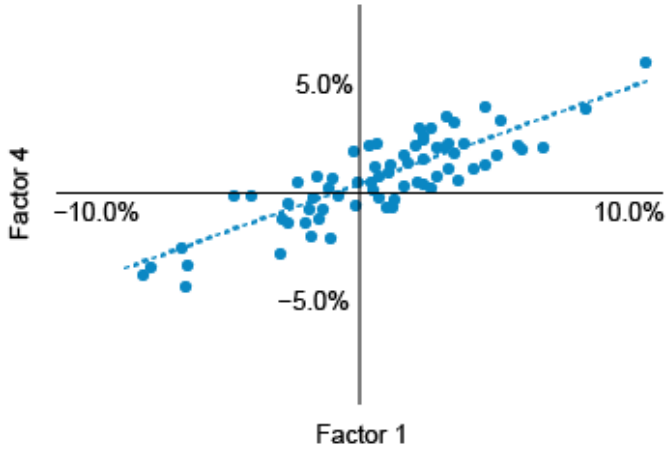
---

5. Henrik Lippert is a junior quantitative analyst at Regnitz. Lippert is running a multiple regression to predict revenue growth for an energy stock he covers and presents his findings to Tanja Vogl, his supervisor at Regnitz. Vogl analyzes the residual plot and notices a pattern: the variance of residuals changes across observations and is correlated with one of the independent variables. She tells Lippert that he needs to adjust the model since it appears to violate a multiple linear regression assumption.

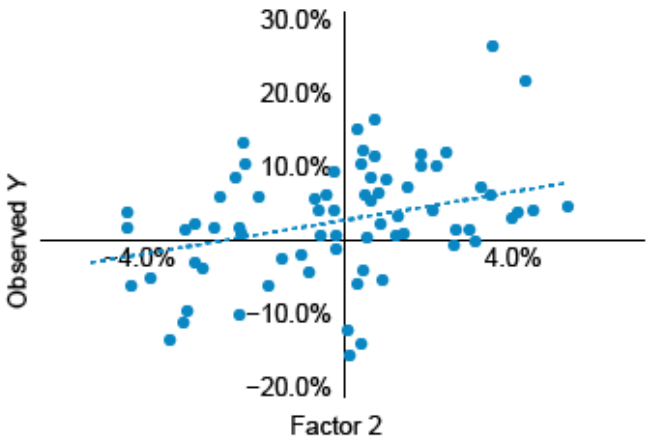
Lippert uses 72 observations to recalculate the regression of the dependent variable against four factors, and he plots the three graphs presented in Exhibit 1:

Exhibit 1 Selected Scatterplots for Multiple Regression Analysis

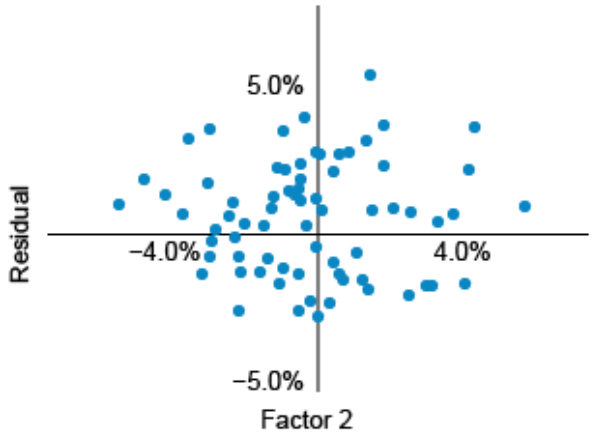
Graph 1: Factor 4 against Factor 1



Graph 2: Observed *Y* against Factor 2



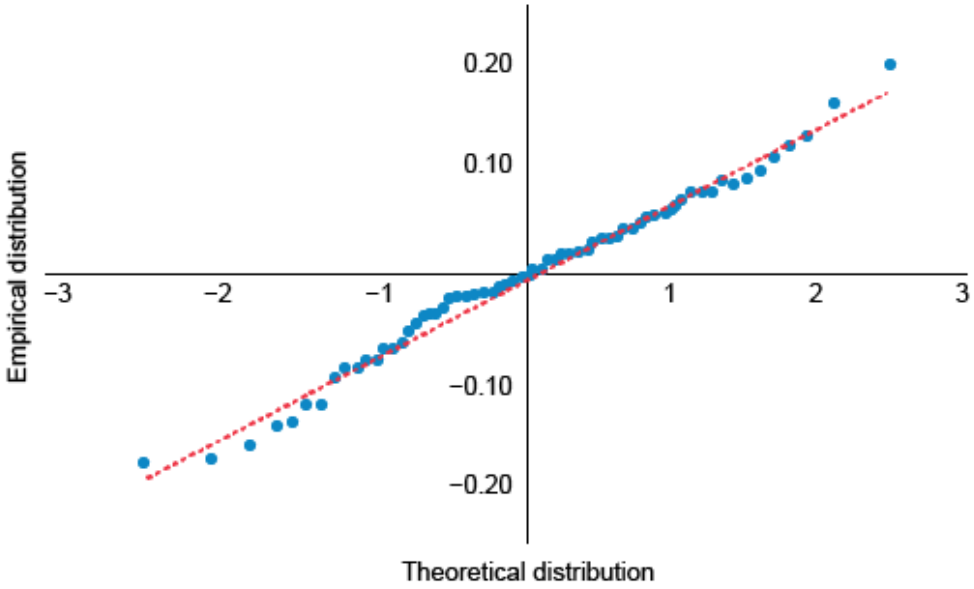
Graph 3: Regression residuals against Factor 3



Vogl analyzes the graphs and says that Graph 1 suggests a potential violation of another multiple linear regression assumption. She adds that scatterplots can also be used to identify the violation of other assumptions, such as linearity.

Finally, Vogl asks Lippert to create a Q–Q plot of regression residuals; he presents the graph in Exhibit 2:

Exhibit 2 Q–Q Plot of Regression Residuals



Question 1 of 4

Which of the following violations is *most likely* indicated by the changes in the residuals' variances across observations?

a. Nonlinearity
b. Autocorrelation
c. Heteroskedasticity

Explanation:

Multiple linear regression assumptions

Assumption	Description	Violation
Linearity	Dependent and independent variables have linear relationship	Nonlinearity
Homoskedasticity	Variance of residuals constant across all observations	Heteroskedasticity
Independence of errors	Observations are independent of each other; errors (ie, residuals) uncorrelated across all observations	Serial correlation or autocorrelation
Normality	Residuals normally distributed, with expected value of zero	Non-normality
Independence of independent variables	Independent variables are not random; no exact linear relation between independent variables	Multicollinearity

A multiple linear regression evaluates the relationship between a dependent variable and a set of independent variables, presenting that relationship as a linear equation. A multiple linear regression **has five key assumptions**, shown in the table above. If one or more of the assumptions is (are) violated:

- The estimated coefficients may be inconsistent and invalid estimations of the true coefficients.
- The standard errors (of the coefficients) and the mean square error (MSE) may be incorrectly estimated.
- *F*-tests and *t*-statistics may be unreliable, leading to mistakes in hypothesis testing.

In this instance, Vogl notes that the residuals' variance is not constant and is correlated with one of the independent variables. This indicates a violation of **homoskedasticity**, which is the assumption that **observations** have a **similar dispersion** and residuals have **constant variance**. The **violation** of homoskedasticity is called heteroskedasticity and may cause an **underestimation** of the **standard errors** of the regression coefficients.

(Choice A) Nonlinearity, a violation of the linearity assumption, occurs when the dependent variable and one or more of the independent variables do not have a linear relationship. This is not the pattern identified by Vogl.

(Choice B) Autocorrelation, a violation of the independence of errors assumption, is often identified by plotting the residuals against time (ie, observation order), and then determining whether neighboring residuals have similar signs and magnitudes. This is not the pattern noticed by Vogl.

Things to remember:
A multiple linear regression has five key assumptions: linearity (between dependent and independent variables), homoskedasticity (constant variance of residuals), independence of errors (and observations), normality (of residuals), and independence of independent variables (ie, not random and with no linear relation). Heteroskedasticity is a violation of homoskedasticity that may cause an underestimation of the standard errors of the regression coefficients.

LOS - Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

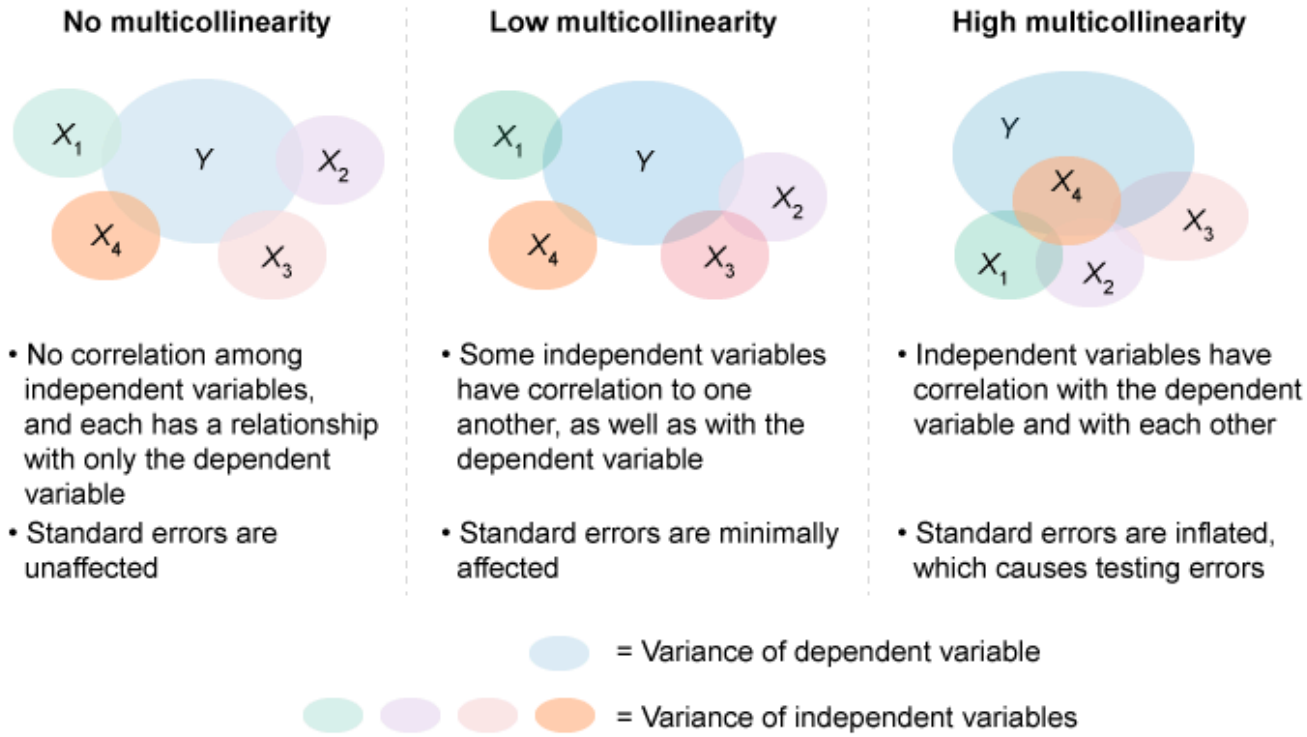
---

6. **Question 2 of 4**

Graph 1 in Exhibit 1 indicates a possible violation of the assumption of independence of:

a. errors.
b. observations.
c. independent variables.

Explanation:



A key assumption of a multiple regression equation is the **independence of independent variables**. This means that independent variables:

- are **not random** and
- do not have an **exact linear relation**.

Multicollinearity occurs when two or more **independent variables** are **strongly correlated**. Multicollinearity does not affect the slope coefficients of the regression, but the standard errors for each coefficient become inflated, making it difficult to interpret the significance of independent variables. A typical method to detect this and other violations of regression assumptions is to graph pairwise scatterplots for all (dependent and independent) variables in the regression.

The presence of patterns in the scatterplots often indicates the violation of one or more regression assumptions. Graph 1 shows that Factor 1 and Factor 4, two of the model's **independent variables**, are **strongly correlated**. This means that the variables may have an **approximate linear relationship** and are **not independent**, although additional tests may be necessary.

(Choices A and B) If the observations of the multiple regression are independent of (ie, uncorrelated with) each other, then the errors (ie, residuals) will also be uncorrelated across all observations. However, identifying the independence of errors or observations requires plotting residuals against time, which is not done in Graph 1.

Things to remember:
An assumption of a multiple regression equation is the independence of independent variables, which means that independent variables are not random and do not have an exact linear relation. A typical method to detect the violation of this assumption is to use pairwise scatterplots to identify strong correlations between independent variables.

LOS - Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

---

7. **Question 3 of 4**

Based on Exhibit 1, which of the following graphs is *most appropriately* used to identify a violation of the linearity assumption?

a. Graph 1

b. Graph 2

c. Graph 3

Explanation:

Examples of diagnostic plots

Vertical axis	Horizontal axis	Regression assumption tested
Y	X_i	Linearity
X_i	X_k	Independence of independent variables
Residuals	Y or X_i	Homoskedasticity and independence of errors
Standardized residuals	Normal distribution	Normality

Y = Dependent variable

X_i, X_k = Independent variables

Residual = Observed Y− Predicted Y

Diagnostic plots are widely used to detect **potential violations** of the multiple regression assumptions. A **scatterplot matrix**, also known as a pairs plot, shows pairwise scatterplots between the dependent variable (Y) and each independent variable (X_i) and between each pair of independent variables. These graphs can be plotted before the regression is calculated since they are based on observed values of Y and X_i. In this scenario, Graph 1 and Graph 2 are examples of pairs plots.

Graph 2 plots the observed results of *Y* against Factor 2 (ie, ***Y*againstX_2**) and can be used to identify a potential **violation** of the **linearity assumption**. The scatterplot shows a positive relationship between both variables and is also useful to identify extreme values and outliers, like the two points near the top of the first quadrant (ie, upper right) in Graph 2.

Graph 1 plots Factor 4 against Factor 1 (ie, X_4 against X_1). This plot shows the degree of correlation between two independent variables and is used to test the independence of such variables **(Choice A)**.

Graph 3, a residual plot, shows the residuals on the vertical axis and an independent variable (Factor 3 or X_3) on the horizontal axis. Graph 3 is apparently randomly scattered, which is desirable since the presence of patterns could indicate a violation of assumptions such as homoskedasticity and independence of errors **(Choice C)**.

Things to remember:
A scatterplot matrix shows pairwise scatterplots between regression variables and can be used to detect a violation of the assumption of either linearity or the independence of independent variables. A residual plot may indicate a violation of the homoskedasticity or independence of errors assumptions.

LOS - Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

---

8. **Question 4 of 4**

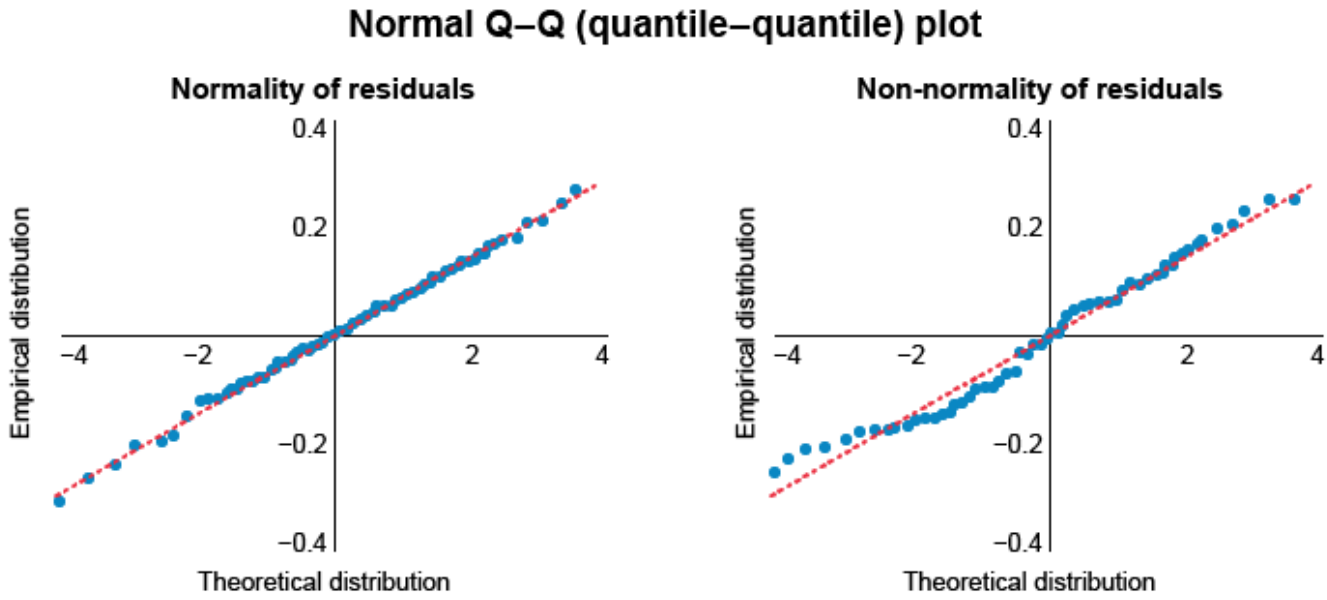
Based on Exhibit 2, which of the following multiple linear regression assumptions is *most likely* violated?

a. Linearity

b. Normality

c. Homoskedasticity

Explanation:



One of the key assumptions of multiple linear regressions is the **normal distribution** of the **residuals** (ie, the **normality assumption**). The violation of this assumption compromises some uses of the regression model (eg, prediction intervals), but it does not affect the estimated coefficients (ie, slopes and intercept). As the number of observations increases, the normality assumption becomes less relevant, according to the central limit theorem.

The **normal Q–Q plot** compares the distribution of a variable with a normal distribution. In a multiple linear regression, the Q–Q plot is a type of diagnostic plot that compares:

- the **empirical distribution** (ie, standardized residuals ordered from smallest to largest) with
- the **theoretical distribution** (ie, a theoretical standard normal distribution).

The Q–Q plot also includes a line representing a linear relationship between the empirical and theoretical distributions. **Normally distributed residuals** should be plotted **along** this **line**, as shown on the left graph in the image. If the residuals are not normally distributed, the points should not be aligned, as shown at right. In this instance, the residuals in the **Q–Q plot** are **not aligned**, indicating a potential **violation** of the **normality** assumption.

(Choice A) The violation of linearity is typically detected by using scatterplots of the dependent variable against each of the independent variables, rather than using the Q–Q plot.

(Choice C) The violation of homoskedasticity is typically detected using residual plots, not the Q–Q plot.

Things to remember:
In a multiple regression, the Q–Q plot compares the distribution of residuals with a theoretical standard normal distribution. If the residuals are not normally distributed, the points in the graph will not be aligned.

LOS - Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

---