1. Jae Park, CFA, is a manager of a hedge fund that bases its security selection on advanced quantitative analysis. For several open job positions with the fund, Park is looking to hire people with scientific and research backgrounds. Using multiple regression, she would like to evaluate the relationship between the expected salary of the candidates based on their years of experience (EXP), number of published research papers (PRP), and amount of grant funding received in their career (GF). The results of that regression are shown below, along with sample critical values. Park wishes to test the results at a 5% significance level (α = 0.05).

Exhibit 1 Selected Regression Output and ANOVA Data

	Coefficient	Standard Error	*t*-Statistic
Intercept	94.222	11.785	7.995
EXP	5.080	1.116	4.550
PRP	−0.820	1.873	−0.438
GF	0.212	0.136	1.552

ANOVA Data	*df*	Sum of Squares (SS)	Mean SS	*F*	Significance *F*
Regression (*k*)	3	30,430.34	10,143.40	18.643	0
Residual (*n* − *k* − 1)	22	11,969.66	544.08		
Total	25	42,400.00			

Observations	26
R^2	0.718
Standard error	23.325

Exhibit 2 Sample Values from *t*-Distribution Table

Significance - Two-tailed	0.100	0.050
Significance - One-tailed	0.050	0.025
df		
21	1.7207	2.0796
22	1.7171	2.0739
23	1.7139	2.0687
24	1.7109	2.0639
25	1.7081	2.0595

Park also notices that each candidate attended one of five universities. She is considering how to add a variable for university attended to the regression model and believes dummy variables are the best way to capture this.

Finally, Park suspects that her regression in its current form may violate regression assumptions. Her concern is that her model might have an artificially large R^2 and *t*-statistics that are understated.

Question 1 of 6

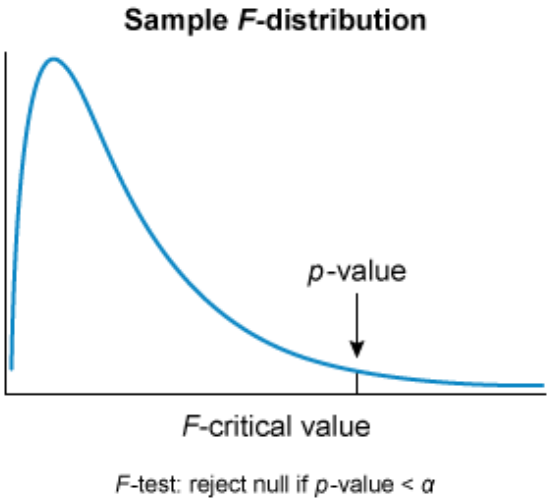
Based on the data in Exhibit 1, the regression is *most likely* a good predictor of projected salary since:

a. it has a high R^2.

b. the *F*-statistic has a low *p*-value.

c. most coefficients are statistically significant.

Explanation:



Hypothesis testing verifies whether a regression is a good predictor of the dependent variable. Testing can be performed on coefficients or the overall regression.

The ***F*-statistic** is used to test whether the **overall regression** (ie, the combination of all slope coefficients) is statistically significant. The null hypothesis (H_0) states that the model has no statistically significant coefficients, while the alternative hypothesis (H_a) states that at least one coefficient is significant.

The ANOVA table provides the *F*-statistic's *p*-value (labeled "significance *F*"). H_0 is rejected if the ***p*-value** is **less** than the **level of significance** (**α**). In this scenario, the ***p*-value** is **less** than **α (0 < 0.05)**, so $\boldsymbol{H_0}$ **is rejected** and the regression is **statistically significant**. In other words, the **regression** is a **good predictor** of the dependent variable.

(Choice A) R^2 gauges how closely the data fit the regression line. Adding independent variables will increase R^2, even if those variables are only slightly correlated with the dependent variable. Furthermore, R^2 does not measure statistical significance. Models may have multiple coefficients without statistical significance and still have a high R^2. The model's predictive power depends mainly on the statistical significance of the regression and the coefficients.

(Choice C) A *t*-test can determine if a slope coefficient is statistically significant. However, the model's overall fit cannot be determined from *t*-tests due to possible interactions among the independent variables. Individual coefficients can be statistically significant, but the overall regression is not.

Things to remember:
The *F*-statistic is used to test how well the regression explains the dependent variable. The null hypothesis H_0 states that none of the coefficients are statistically significant, and it can be rejected if the *p*-value of the *F*-statistic is less than α. Rejecting H_0 indicates that the model is a statistically significant predictor of the dependent variable.

LOS - Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

---

2. **Question 2 of 6**

According to Exhibit 1, the *most appropriate* interpretation of the coefficients is that a higher expected salary will result from:

a. more published papers and less grant funding.
b. more published papers and more grant funding.
c. fewer published papers and more grant funding.

Explanation:

Independent variable's (X_n) effect on dependent variable (Y)

Coefficient	Effect of X_n on Y
$b > 0$	↑ X_n results in ↑ Y
$b < 0$	↑ X_n results in ↓ Y

The **predicted value** of the **dependent variable** can be derived using the relationship expressed in the **regression equation**. The slope **coefficients** of the equation explain the change in the dependent variable, given a one-unit change to one of the independent variables.

The **sign** of the coefficient indicates the **direction** of the change in the dependent variable. If the sign of the coefficient is:

- **positive**, then the independent and dependent variables will move in the **same direction**.
- **negative**, then the independent and dependent variables will move in **opposite directions**.

In this scenario, the PRP and GF coefficients determine how changes to those variables affect the expected salary:

- The PRP coefficient is negative, so a higher PRP variable results in a lower projected salary **(Choice A)**.
- The GF coefficient is positive, so a higher GF variable results in a higher projected salary **(Choice B)**.

Thus, a **higher salary** would be expected from **fewer published papers** and **more grant funding**.

Things to remember:

The sign of a slope coefficient indicates how a change in an independent variable affects the dependent variable. If a coefficient has a positive sign, then the independent and dependent variables move in the same direction. If a coefficient has a negative sign, then the independent and dependent variables move in opposite directions.

LOS - Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests

---

3. **Question 3 of 6**

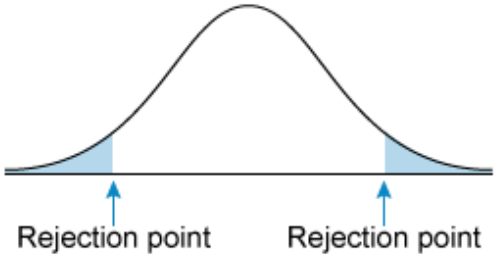
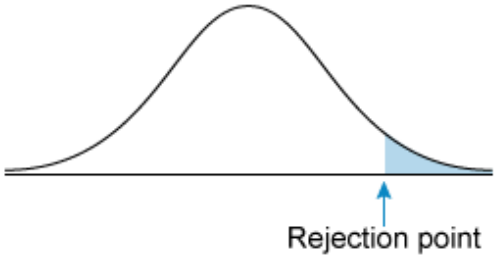
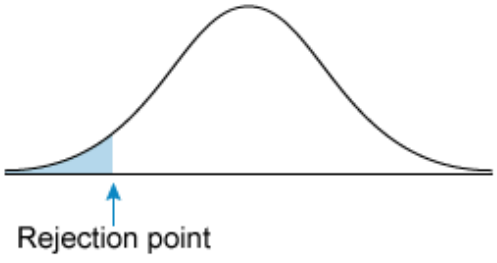
Based on Exhibits 1 and 2, which slope coefficient is *most likely* to be statistically significant?

a. GF

b. EXP

c. PRP

Explanation:

	Null hypothesis	Alternative hypothesis	Tails	Illustration
1	H_0: $\mu = \mu_0$	H_a: $\mu \neq \mu_0$	Two	
2	H_0: $\mu \leq \mu_0$	H_a: $\mu > \mu_0$	One	
3	H_0: $\mu \geq \mu_0$	H_a: $\mu < \mu_0$	One	

In a regression model, the slope coefficients describe how a change in an independent variable affects the dependent variable. A *t*-test is used to verify whether a coefficient meaningfully describes the relationship between an independent and dependent variable.

The test seeks to determine whether a coefficient is **statistically different from zero**. The null hypothesis (H_0) states that the coefficient equals zero, while the alternative hypothesis (H_a) states that the coefficient is not equal to zero. The test is also **two-tailed**, as shown in the image above.

The **test compares** the ***t*-statistic (*t*)** to the **critical value (t_c)**, and the **null** is **rejected** if the **absolute value of the *t*-statistic is greater than the critical value** (ie, reject H_0 if $|t| > t_c$). Rejection of H_0 means that the coefficient is statistically different from zero (ie, statistically significant).

The steps for the *t*-test on each of the coefficients are as follows:

Steps	Calculations
Solve for the residuals' degrees of freedom	$df = n - k - 1 = 26 - 3 - 1 = 22$
Solve for the critical *t*-value	Critical value for 2-tailed test, $df = 22$, and 5% significance = 2.0739.
Compare each *t*-statistic to the critical value; reject H0 if $\|t\| > t_c$	EXP: $\|4.550\| > 2.0739$; **reject** H_0 PRP: $\|-0.438\| < 2.0739$; **cannot reject** H_0 GF: $\|1.552\| < 2.0739$; **cannot reject** H_0

The **null** hypothesis is **rejected only for EXP**. Therefore, **EXP** is the **only statistically significant** coefficient; **PRP and GF** are **not statistically significant (Choices A and C)**.

Things to remember:
The *t*-test is used to assess whether a slope coefficient is statistically significant. A *t*-test assumes a null hypothesis where the coefficient equals zero. If the null hypothesis is rejected, then the coefficient is not equal to zero and is therefore statistically significant.

LOS - Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

---

4. **Question 4 of 6**

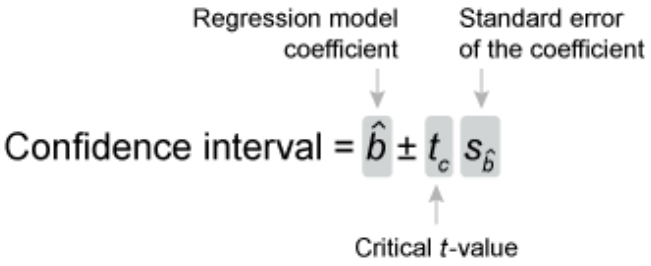
Based on Exhibits 1 and 2, the 95% confidence interval for the EXP coefficient has a lower bound *closest* to:

a. −4.36

b. 2.77

c. 3.16

Explanation:



In a linear regression, each of the variables' coefficients are estimates because the true value cannot be known with certainty. A **confidence interval** is a **range** of values that would be expected to **include** the **true value** of a **coefficient** for a given significance level. The upper and lower bounds of the interval are **calculated** using the regression coefficient and its standard error, along with critical value t_c, all of which assumes:

- $df = n - k - 1$, where df is the degrees of freedom, k is the number of independent variables, and n is the number of observations.
- The interval is **two-sided and symmetrical** since the formula specifies both upper and lower bounds.

The lower bound of the confidence interval is calculated as follows:

Steps	Calculations
Solve for degrees of freedom	$df = n - k - 1 = 26 - 3 - 1 = 22$
Solve for critical t-value	Critical value for 2-tailed test, $df = 22$, and 5% significance = 2.0739
Solve for the lower bound	Lower bound $= \hat{b} - t_c s_{\hat{b}}$ $= 5.080 - (2.0739 \times 1.116)$ ≈ 2.77

The lower bound is approximately equal to 2.77.

(Choice A) −4.36 results from incorrectly using the t-statistic in the ANOVA table instead of solving for the critical value.

(Choice C) 3.16 results from incorrectly using the critical *t*-value from a 5% significance level for a one-tailed test.

Things to remember:
A confidence interval is a range that would contain the true value of a dependent variable at a given significance level. Solving for the interval requires the regression coefficient and its standard error, along with the critical *t*-value at that significance level.

LOS - Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

---

5. **Question 5 of 6**

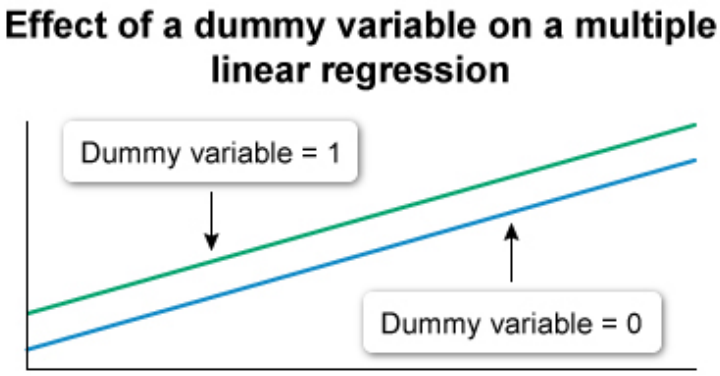
To account for the candidates' universities, the number of dummy variables that Park should add to the regression is *closest* to:

a. 1

b. 4

c. 5

Explanation:



Dummy variables are used to **incorporate qualitative independent variables** into a regression model. The value of these variables is binary: **equal to 1** if some attribute is **present**, or to **0** if an attribute is **absent**. Regressions may use **multiple dummy variables**, depending on the number of **categories** that the model is attempting to capture. If there are ***n*** **categories**, then the model requires ***n*− 1 dummy variables**.

For example, Park wants to use dummy variables to distinguish between universities attended by the candidates. Since there are five universities ($n = 5$), four dummy variables are needed ($n - 1 = 5 - 1 = 4$). Assume they are labeled W, X, Y, and Z. If a candidate attended:

- W University: W = 1, and X = Y = Z = 0.
- X University: X = 1, and W = Y = Z = 0.
- Y University: Y = 1, and W = X = Z = 0.
- Z University: Z = 1, and W = X = Y = 0.

Finally, if a candidate attended the fifth university, all four dummy variables equal 0, so there is no need for a separate variable to represent that fifth university (ie, the model requires only $n-1$ dummy variables). The variables are **mutually exclusive** (ie, there is no overlap) and **exhaustive** (ie, all possible real outcomes are described).

Any number of dummies less than four represents a misspecification of the regression, since this would not capture all possible combinations (ie, not exhaustive) **(Choice A)**.

Using more than four dummies would create instances that are not mutually exclusive due to overlap of the university attended. This would result in linear relationships among the independent variables and therefore violate regression assumptions **(Choice C)**.

Things to remember:
Regressions can have multiple dummy variables, but the proper number of dummy variables must be used. When distinguishing between n categories, the regression should have $n-1$ dummy variables.

LOS - Calculate and interpret a predicted value for the dependent variable, given an estimated regression model and assumed values for the independent variables

---

6. **Question 6 of 6**

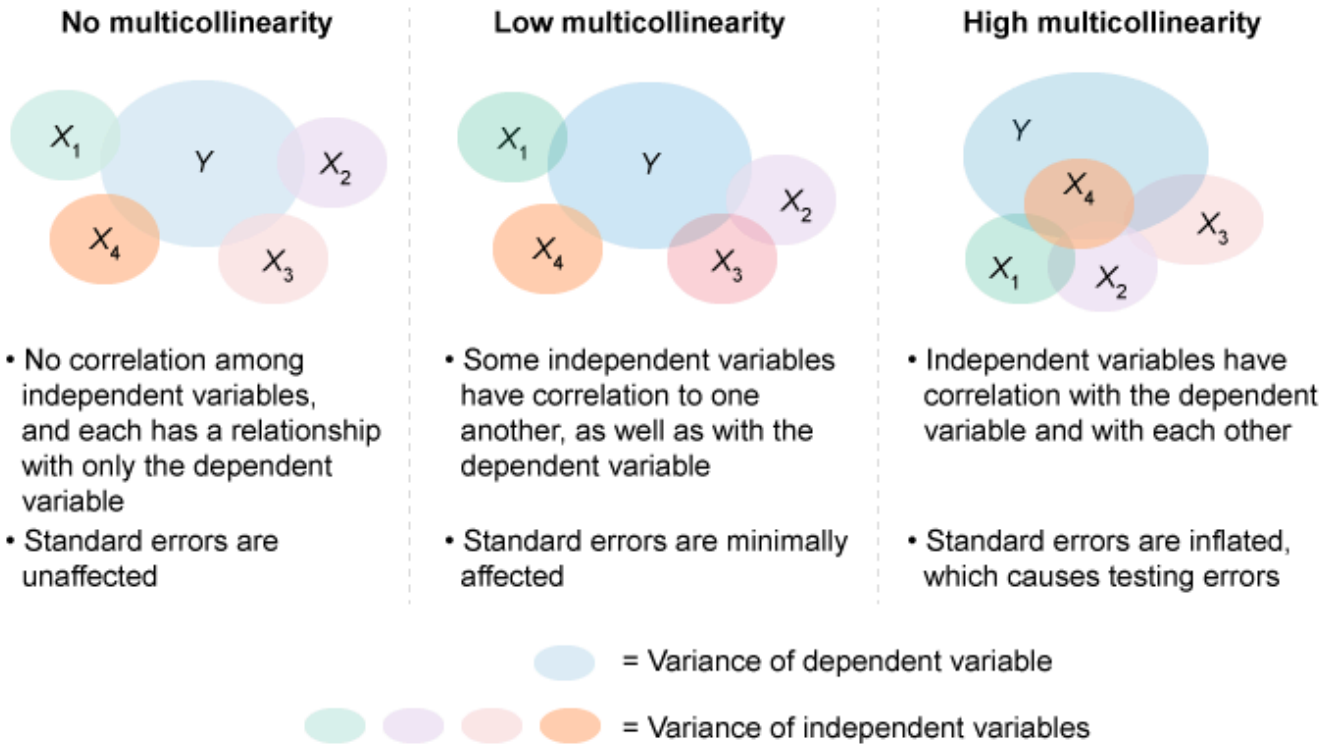
If Park's suspicions about her model are correct, then the model would *most likely* show signs of*:*

a. multicollinearity.

b. serial correlation.

c. heteroskedasticity.

Explanation:



A **key assumption** of a multiple **regression** equation is that there is **no linear relationship** between the independent variables. **Multicollinearity** refers to the condition where at least two **independent variables** are **highly correlated**.

Multicollinearity has several impacts on a regression. The estimates of the slope coefficients are not affected, but the standard errors for each coefficient become inflated. This results in understated *t*-statistics, which in turn leads to **coefficients** being incorrectly classified as **not statistically significant**. Furthermore, a regression with multicollinearity will have **inflated** R^2 and ***F*-statistic** values, and thus seem to be a "better fit" than it actually is.

(Choice B) A key assumption of multiple regression is that error terms are uncorrelated. Serial correlation occurs when regression errors are correlated across observations. This typically affects only the regression standard errors, not the R^2.

(Choice C) Another key assumption is that the variance of the error term is constant across observations. Heteroskedasticity refers to the variance of regression errors changing across

observations. This affects only a regression's standard errors.

Things to remember:
Several assumptions must hold for conclusions drawn from a regression to be valid. One of those assumptions is that there is no linear relationship between the independent variables. Multicollinearity occurs when at least two independent variables are highly correlated. This typically manifests in the form of high R^2 and F-statistic values without statistically significant coefficients.

LOS - Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

---

7. Lana Marek, CFA, a pricing analyst at a European auto manufacturer, is analyzing competitor pricing data. She is working with Micah Hould, CFA, a quantitative analyst at the same company, to create multiple regression models that help identify factors affecting vehicle pricing.

As part of her analysis, she has chosen three independent variables to regress against vehicle price: engine power output (*HP*), vehicle seating capacity (*SC*), and size of the car's interior (*CI*). Marek assumes that all three variables are positively related to price. To test this, she has compiled data for 30 different vehicle models from last year and developed a regression equation and statistics at a 5% level of significance.

Marek believes that there is another variable that affects vehicle price: whether the car belongs to a luxury vehicle brand. She adds a dummy variable and assigns a value of 1 for a luxury vehicle and a value of 0 for a nonluxury vehicle. Marek expects that there will be a positive relationship between the brand variable and price. The updated regression statistics and equation are as follows:

Exhibit 1 Regression Output and ANOVA Data

	Coefficient	Standard error	*t*-statistic	*p*-value (two-tailed)	*p*-value (one-tailed)
HP	0.113	0.013	8.531	0.000	0.000
SC	−1.596	1.270	−1.257	0.220	0.110
CI	11.990	3.677	3.261	0.003	0.002
LX	1.104	2.200	0.502	0.620	0.310
Intercept	−19.197	6.262	−3.066	0.005	0.003

ANOVA data	*df*	Sum of squares (SS)	Mean SS
Regression (k)	4	3164.10	791.03
Residual ($n - k - 1$)	25	770.20	30.81
Total ($n - 1$)	29	3934.30	

R^2	0.804
Adjusted R^2	0.773
Standard error	5.550
Observations (n)	30

Price = Dealership price, expressed in €1,000s
HP = Power output of the vehicle's engine, expressed in units of metric horsepower
SC = Seating capacity, expressed as number of seats in the vehicle
CI = Size of the car's interior, expressed in units of cubic meters
LX = Luxury vehicle dummy variable (1 = luxury, 0 = not luxury)

Marek notices that R^2 increased after she added the dummy variable to her analysis, but adjusted R^2 (or $\bar{R}^2$ decreased.

Finally, Marek is concerned about multicollinearity's effect on her model. She suspects the seating capacity and interior size variables are highly correlated, which may influence the regression results. She shares her concern with Hould, who responds as follows:

- Statement 1: "The most reliable way of detecting multicollinearity is to assess the magnitude of correlations between independent variables."
- Statement 2: "A key issue with multicollinearity is that *t*-tests on regression coefficients will result in instances of rejecting a true null hypothesis."
- Statement 3: "The most straightforward way to correct multicollinearity is to remove one or more correlated variables from the regression."

Question 1 of 6

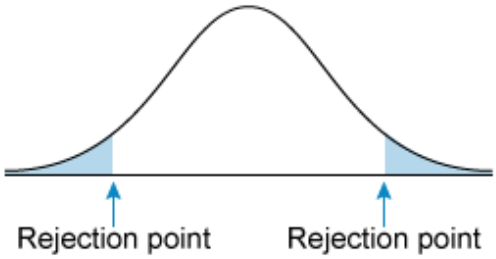
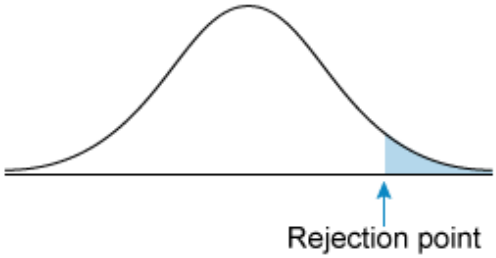
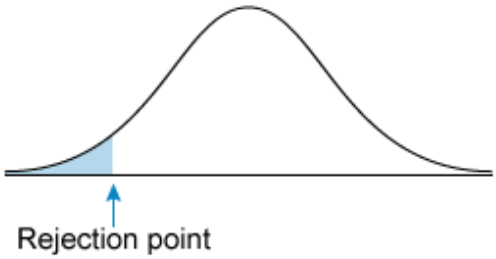
Based on Exhibit 1, which is the *most appropriate* alternative hypothesis and conclusion from Marek's assumption on the seating capacity variable (*SC*)?

a. H_a: $SC > 0$; reject the null.

b. H_a: $SC > 0$; fail to reject the null.

c. H_a: $SC \leq 0$; fail to reject the null.

Explanation:

	Null hypothesis	Alternative hypothesis	Tails	Illustration
1	H_0: $\mu = \mu_0$	H_a: $\mu \neq \mu_0$	Two	
2	H_0: $\mu \leq \mu_0$	H_a: $\mu > \mu_0$	One	
3	H_0: $\mu \geq \mu_0$	H_a: $\mu < \mu_0$	One	

Multiple linear regression uses hypothesis testing to assess whether a regression **coefficient explains the relationship** between each dependent variable and the independent variable; the null is tested by performing a *t*-test or checking the *p*-value of the coefficient:

- The null hypothesis (H_0) proposes that the coefficient is **not statistically different from zero**, and
- the alternative hypothesis (H_a) proposes that the coefficient is statistically significant.

Marek assumes that *SC* and price are **positively** correlated, in which case the *SC* **coefficient** would be **greater than zero**. This test is one-tailed: the H_0 is $SC \leq 0$, and the H_a is $SC > 0$.

Exhibit 1 shows that the *SC* variable has a one-tailed *p*-value of 0.110. The ***p*-value** is **greater** than the **level of significance** of 0.05, so Marek **cannot reject** H_0. Therefore, the sample data suggests that there is no meaningful relationship between seating capacity and price.

(Choice A) The alternative hypothesis is correct. However, the *p*-value is greater than the significance level, so the null cannot be rejected.

(Choice C) $SC \leq 0$ is the null hypothesis. Furthermore, the *p*-value is greater than the significance level, so the null hypothesis cannot be rejected.

Things to remember:
Hypothesis testing is used to verify whether the relationship between an independent and

dependent variable is statistically significant. The null, H_0, proposes that the condition to be tested is untrue, while the alternative, H_a, proposes that the tested condition is valid, and *p*-values are used to check whether the null hypothesis can be rejected.

LOS - Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests

---

8. **Question 2 of 6**

If Marek correctly calculated the F-statistic shown in Exhibit 1 and then compared it with a critical F-value of 2.975, the *most appropriate* conclusion is that for this regression model:

a. there are no statistically significant coefficients.
b. most of the coefficients are statistically significant.
c. there is at least one statistically significant coefficient.

Explanation:

***F*-statistic equation**

$$F = \frac{\frac{\text{Regression SS}}{k}}{\frac{\text{Residual SS}}{n-k-1}} = \frac{\text{MSR}}{\text{MSE}}$$

k = Number of slope coefficients
n = Number of observations
MSR = Mean regression sum of squares = $\frac{\text{Regression SS}}{k}$
MSE = Mean squared error = $\frac{\text{Residual SS}}{n-k-1}$

For a multiple regression, hypothesis testing can show whether the **regression model** (ie, all slope coefficients) **explains** changes in the **dependent** variable. A *t*-test can be performed on each independent variable to verify whether an **individual** independent variable is statistically significant.

However, a ***t*-test cannot** test coefficients as a **group** due to possible **interactions** among the **independent** variables (ie, multicollinearity). It is possible for **one** or more individual coefficients to be **statistically significant** while the **overall** regression **is not**.

To test the **overall regression**, an **F-test** is used instead, which allows for **testing slope coefficients jointly** by solving for the regression's F-statistic then comparing that with the critical F-value. In this case, the F-statistic can be solved from Exhibit 1 values, specifically *k*, *n*, and the sum of squares (SS).

Steps	Calculations
Solve degrees of freedom	Numerator $df = k = 4$ Denominator $df = n - k - 1 = 30 - 4 - 1 = 25$
Calculate F-statistic	$F = \frac{(3164.10 / 4)}{(770.20 / 25)} = 25.68$

The F-statistic can also be calculated using the mean SS values from Exhibit 1:

Steps	Calculations
Calculate F-statistic	$F = \frac{791.03}{30.81} = 25.68$

The null can be rejected since the calculated F-statistic of 25.68 is greater than the given critical F of 2.975. Therefore, the regression model has at least one statistically significant coefficient.

(Choice A) The calculation above shows that the null can be rejected. Therefore, there is at least one statistically significant coefficient in this regression.

(Choice B) The F-test determines whether all slope coefficients have no statistical significance, or if at least one coefficient is statistically significant. Beyond that, it does not make any determination on whether some, most, or all coefficients are significant.

Things to remember:
In a multiple regression model, it is possible for one or more individual coefficients to be statistically significant while the overall regression is not. To test the overall regression, an F-test is used, which allows for testing slope coefficients jointly by solving for the regression's F-statistic, then comparing that with the critical F-value.

LOS - Formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests

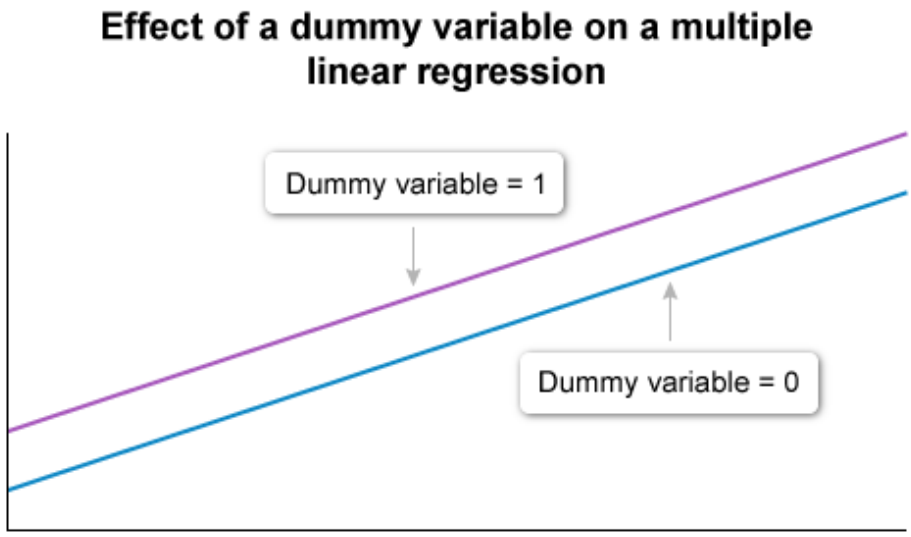
---

9. **Question 3 of 6**

Based on the *LX* dummy variable coefficient in Exhibit 1, Marek's *most appropriate* conclusion is that the model projects a luxury vehicle's price to be:

a. €1,104 higher, and Marek would reject the null hypothesis.
b. €1,104 lower, and Marek cannot reject the null hypothesis.
c. €1,104 higher, and Marek cannot reject the null hypothesis.

Explanation:



Dummy variables are used to **incorporate qualitative independent variables** into a regression model. These variables are binary: **equal to 1** if a specific attribute is **present** or to **0** if that attribute is **absent**.

The **regression coefficient** of a dummy variable represents the **incremental effect** that the **dummy variable** will have on the **dependent variable**. Once included in a regression, **hypothesis testing** for the dummy variable coefficient is performed the **same** way as for any other independent variable: by using a *t*-test or checking the *p*-value of the coefficient.

The values in Exhibit 1 show the *LX* coefficient equals approximately 1.104. Since the coefficient sign is positive, a value of 1 in the dummy variable will cause the dependent variable to increase by about 1.104 (ie, €1,104) **(Choice B)**.

Marek expects a **positive relationship** between the luxury variable and price. Therefore, the **null** hypothesis is $\boldsymbol{LX \leq 0}$ and the **alternative** hypothesis is $\boldsymbol{LX > 0}$. Exhibit 1 shows that the *p*-value for a one-tailed test equals 0.310, which is greater than the significance level of 5% or 0.05. Therefore, the null hypothesis cannot be rejected **(Choice A)**.

Things to remember:
Dummy variables are used to incorporate qualitative variables into multiple regressions. They have two possible values: 1 if the attribute being tested is present, or 0 if the attribute is

absent. The dummy variable coefficient represents the incremental effect that variable has on the dependent variable and can be subjected to hypothesis testing like other regression coefficients.

LOS - Calculate and interpret a predicted value for the dependent variable, given an estimated regression model and assumed values for the independent variables

---

10. **Question 4 of 6**

Based on Exhibit 1, the estimated price (in €) for a vehicle that has a 200 HP engine, seats 5 people, has an interior capacity of 2.5 cubic meters, and is not considered a luxury vehicle is *closest* to:

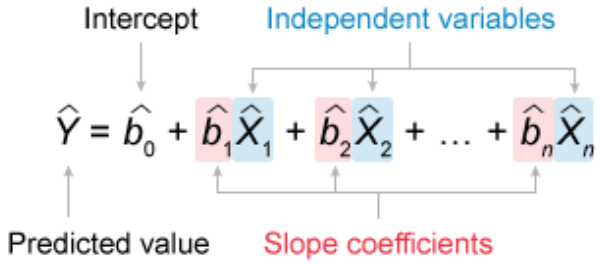
a. 25,400

b. 26,500

c. 59,300

Explanation:

Estimated multiple regression equation



A dependent variable's **predicted** (ie, **expected**) **value** is calculated by using the intercept, slope coefficients, and independent variables as expressed in the **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. The intercept is the estimate for the dependent variable if all the independent variables equal zero.

The independent variable values are given, and the dummy variable value for a car that is not a luxury vehicle is zero. The coefficients are shown in Exhibit 1. The regression's predicted value (or $\hat{Y}$) is:

Steps	Calculations
Express independent variable values in terms of regression variables	$HP = 200$ $SC = 5$ $CI = 2.5$ $LX = 0$
Calculate price given	$\hat{Y} = -19.197 + (1.104 \times 0) + (0.113 \times 200)$ $+ (-1.596 \times 5) + (11.990 \times 2.5)$ $= 25.398$

The dependent variable (in EUR thousands) is 25.398 or €25,398, closest to €25,400.

(Choice B) €26,500 results from incorrectly setting the dummy variable to 1 when solving for the predicted value.

(Choice C) €59,300 results from incorrectly applying independent variables to the wrong coefficients: in this case, the *SC* independent variable was applied to the *CI* coefficient, and vice versa.

Things to remember:
The predicted value of the dependent variable is estimated by using the multiple regression equation, which shows how the intercept and slope coefficients describe the relationship between the dependent and independent variables.

Note: A negative intercept does not mean that the model predicts negative and thus implausible dependent variables. Rather, the intercept term is used to make the regression "fit"; in other words, the combination of intercept and coefficients results in a line that forces the error term to have an expected value of zero. Stated differently, removing the intercept or forcing it to have a positive sign results in a line that is not a best fit and violates the assumption that the error term's expected value is zero.

LOS - Calculate and interpret a predicted value for the dependent variable, given an estimated regression model and assumed values for the independent variables

---

11. **Question 5 of 6**

The decline in $\bar{R}^2$ *most likely* suggests that:

a. R^2 is the more suitable measure of the regression's goodness of fit.
b. the regression model does not explain changes in the dependent variable.
c. it is caused by including an independent variable without a statistically significant relationship to the dependent variable.

Explanation:

R^2 and $\bar{R}^2$ equations

$$R^2 = \frac{\text{Regression } SS}{\text{Total } SS}$$

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2)$$

***R*²** gauges **how well** the **regression** equation **fits** the **data** (ie, how well changes to the independent variables explain changes to the dependent variable); a **higher** R^2 is interpreted as a **better fit**. However, for multiple regressions, the inclusion of **additional independent variables** tends to **increase** the ***R*²**, even if those variables are only **slightly correlated** with the dependent variable.

As a result, a statistic called **adjusted *R*²** (or $\bar{R}^2$) is typically used to measure a **multiple regression's goodness of fit**. Unlike R^2, $\bar{R}^2$ can **decrease** if independent variables added to a regression do **not** have a statistically significant connection to the dependent variable. $\bar{R}^2$ can even be negative if enough variables are added to a regression that do not help explain changes to the dependent variable.

(Choice A) $\bar{R}^2$ is the preferred measure when dealing with multiple regression, since R^2 can increase even in instances in which added variables do *not* have statistically significant slope coefficients.

(Choice B) The F-statistic, not R^2 or $\bar{R}^2$, is used to determine if the model as a whole explains changes in the dependent variable. The model may be statistically significant even if $\bar{R}^2$ declined.

Things to remember:
Adjusted R^2, or $\bar{R}^2$ measures how well a multiple regression explains fits the data. Unlike R^2, a regression model's $\bar{R}^2$ would decline if an independent variable is added to the regression that does not have a statistically significant connection to the dependent variable. Therefore, it is considered a better measure of the model's goodness of fit.

LOS - Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

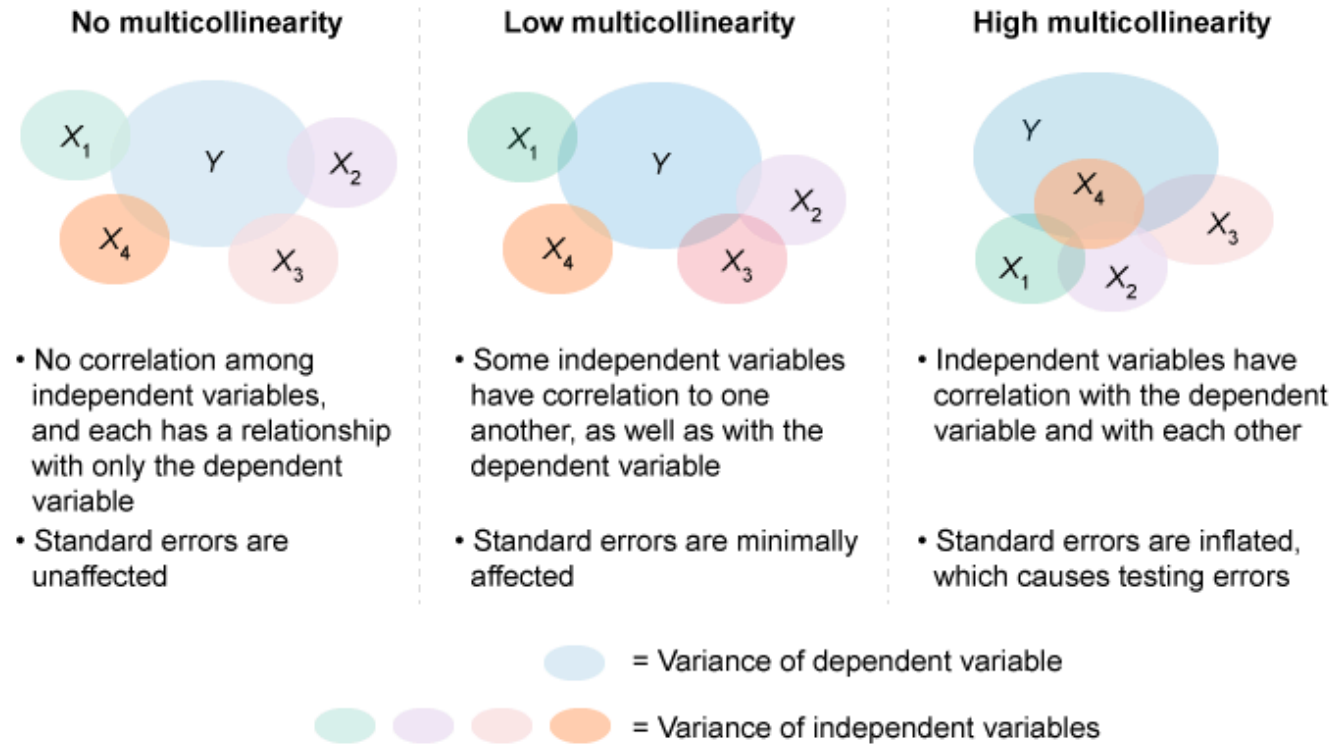
---

12. **Question 6 of 6**

Which of Hould's statements on multicollinearity is correct?

a. Statement 1
b. Statement 2
c. Statement 3

Explanation:



An important assumption of multiple regression is that there is no linear relationship among the independent variables. **Multicollinearity** occurs when **two or more independent variables** are **strongly correlated**. The easiest way to **correct** for this problem is to **remove** one or more correlated variables.

The magnitude of **correlations** between independent variables is **not** an indicator of **multicollinearity**. It is possible for regressions to have multicollinearity despite using variables with low correlations, and it is also possible to include correlated variables in a regression without doing so resulting in high multicollinearity **(Choice A)**.

A common sign of multicollinearity is that the overall **regression** may be statistically significant (ie, result in a significant F-statistic) and have **high** R^2, but individual **slope** coefficients do **not** appear to **be significant**. This is caused by overstated standard errors, which in turn leads to hypothesis tests in which null hypotheses that should be rejected are not.

Thus, multicollinearity leads to statistically significant coefficients being incorrectly classified as not statistically significant and increases the incidence of Type II errors **(Choice B)**.

Things to remember:
Multicollinearity occurs when independent variables are highly correlated. This results in a statistically significant regression while including one or more slope coefficients that are not statistically significant. This causes inflated standard errors, which in turn reduces the effectiveness of *t*-tests on the coefficients. The preferred fix for multicollinearity is to remove one or more correlated variables from the regression.

LOS - Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

---