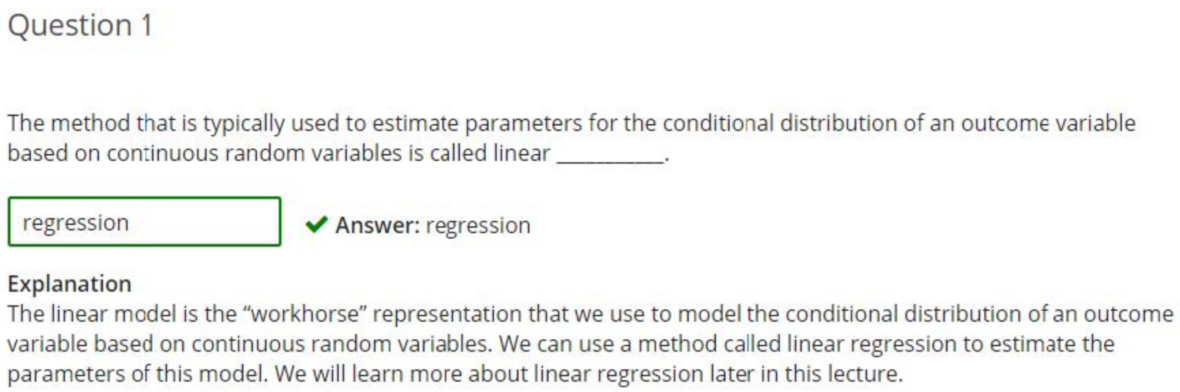
The Linear Model

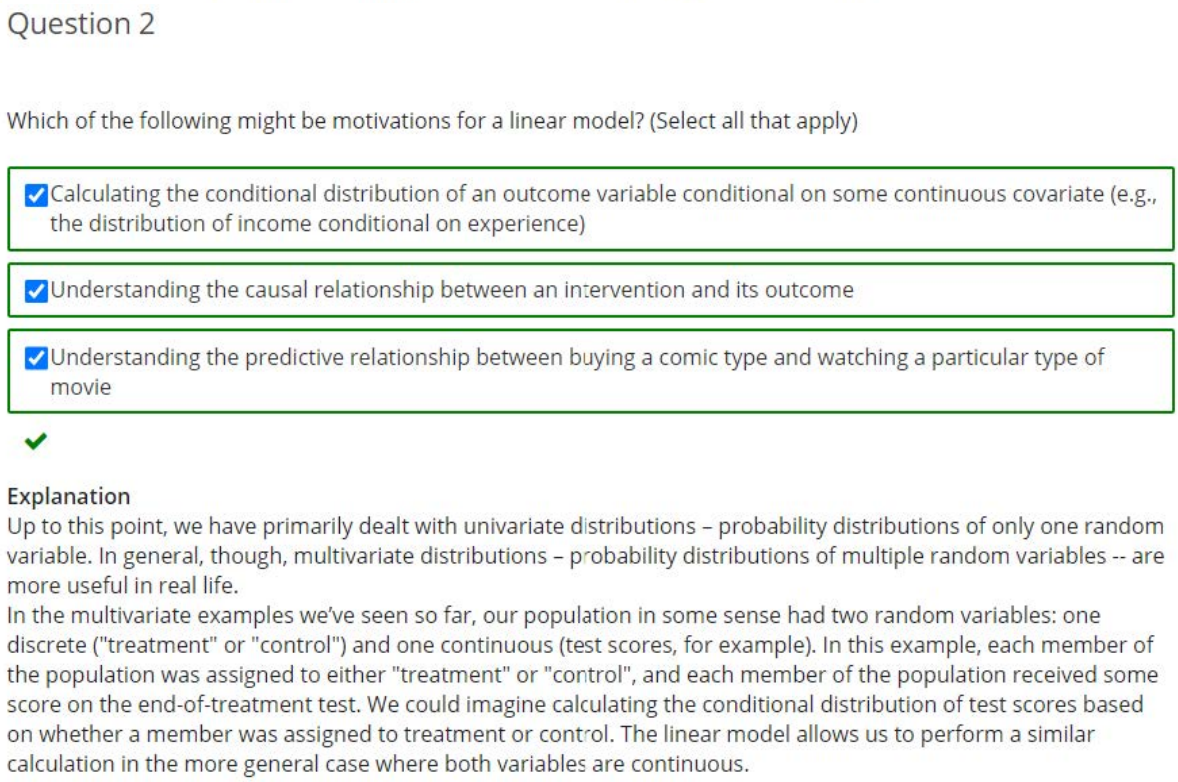


Question 1

The method that is typically used to estimate parameters for the conditional distribution of an outcome variable based on continuous random variables is called linear __________.

regression ✓ **Answer:** regression

Explanation
The linear model is the "workhorse" representation that we use to model the conditional distribution of an outcome variable based on continuous random variables. We can use a method called linear regression to estimate the parameters of this model. We will learn more about linear regression later in this lecture.



Question 2

Which of the following might be motivations for a linear model? (Select all that apply)

☑ Calculating the conditional distribution of an outcome variable conditional on some continuous covariate (e.g., the distribution of income conditional on experience)

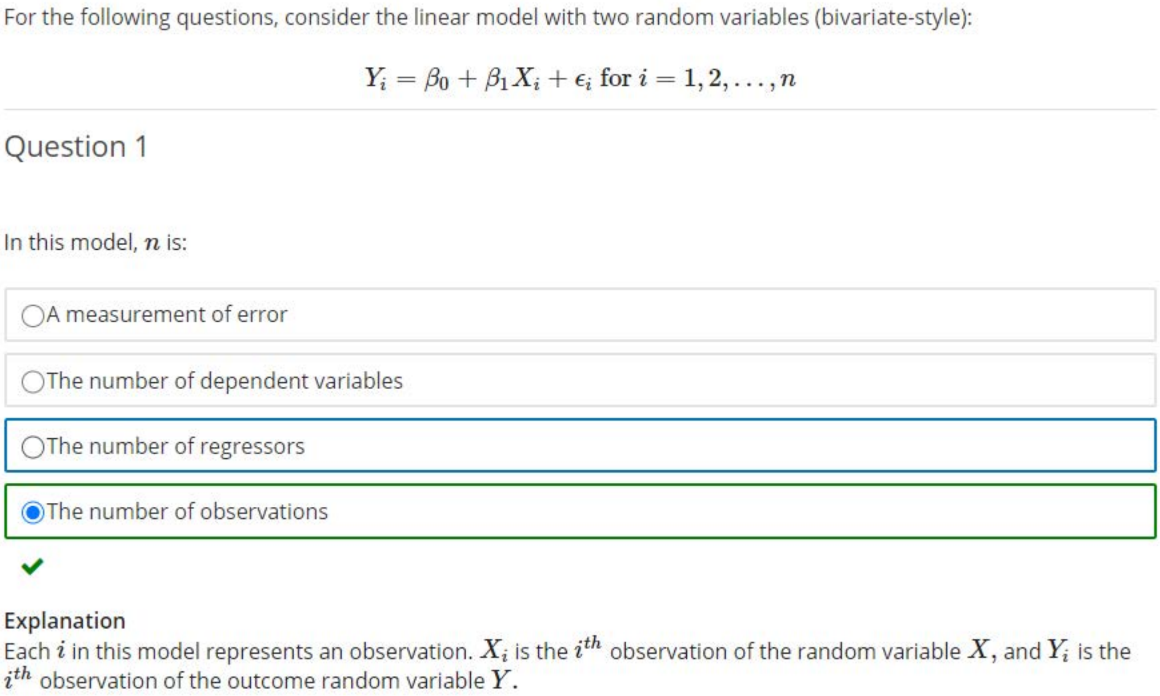
☑ Understanding the causal relationship between an intervention and its outcome

☑ Understanding the predictive relationship between buying a comic type and watching a particular type of movie

✓

Explanation
Up to this point, we have primarily dealt with univariate distributions – probability distributions of only one random variable. In general, though, multivariate distributions – probability distributions of multiple random variables -- are more useful in real life.
In the multivariate examples we've seen so far, our population in some sense had two random variables: one discrete ("treatment" or "control") and one continuous (test scores, for example). In this example, each member of the population was assigned to either "treatment" or "control", and each member of the population received some score on the end-of-treatment test. We could imagine calculating the conditional distribution of test scores based on whether a member was assigned to treatment or control. The linear model allows us to perform a similar calculation in the more general case where both variables are continuous.

The Linear Model



For the following questions, consider the linear model with two random variables (bivariate-style):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ for } i = 1, 2, \ldots, n$$

Question 1

In this model, n is:

○ A measurement of error

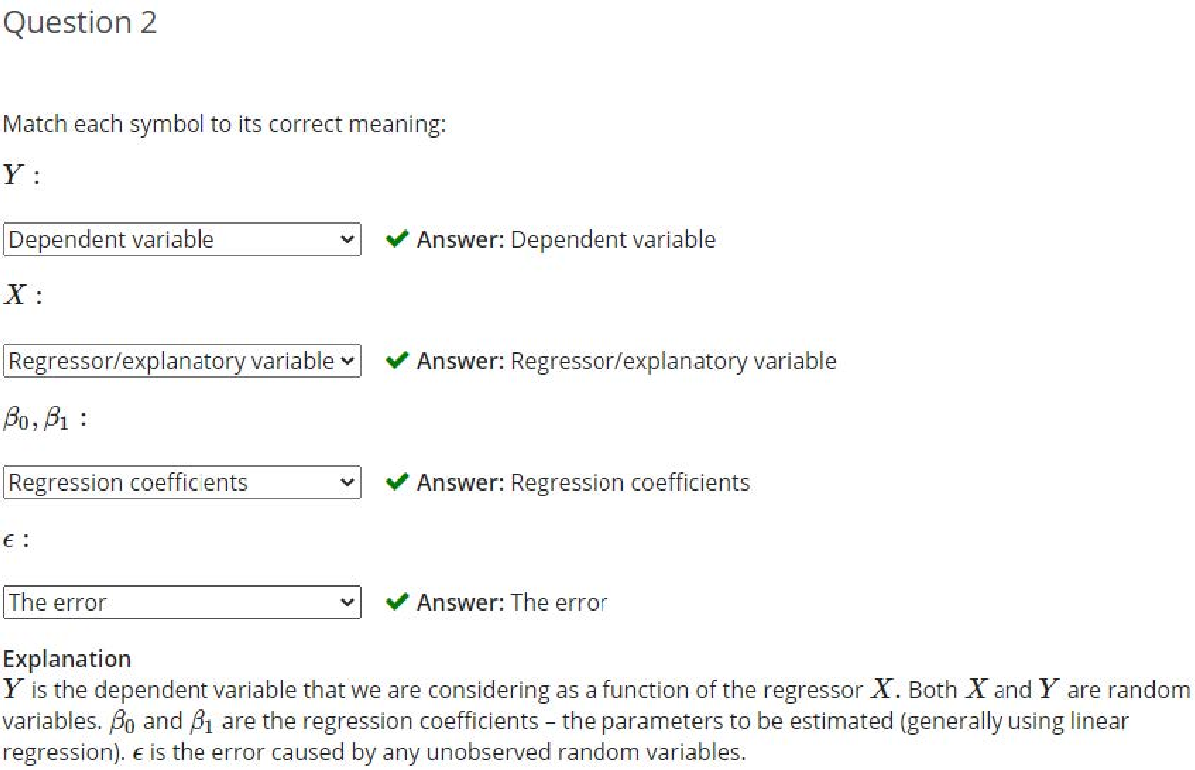
○ The number of dependent variables

○ The number of regressors

● The number of observations

✓

Explanation
Each i in this model represents an observation. X_i is the i^{th} observation of the random variable X, and Y_i is the i^{th} observation of the outcome random variable Y.



Question 2

Match each symbol to its correct meaning:

Y :

Dependent variable ✓ **Answer:** Dependent variable

X :

Regressor/explanatory variable ✓ **Answer:** Regressor/explanatory variable

β_0, β_1 :

Regression coefficients ✓ **Answer:** Regression coefficients

ϵ :

The error ✓ **Answer:** The error

Explanation
Y is the dependent variable that we are considering as a function of the regressor X. Both X and Y are random variables. β_0 and β_1 are the regression coefficients – the parameters to be estimated (generally using linear regression). ϵ is the error caused by any unobserved random variables.

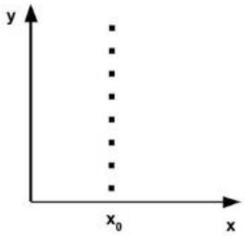
The Linear Model

Recall our linear model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon \text{ for } i = 1, 2, \ldots, n$$

Question 1

Which assumption of the linear model does the below scenario violate?



● Identification ✓

○ No serial correlation. $E[\epsilon_i \epsilon_j] = 0$ for $i \neq j$.

○ $E[\epsilon_i] = 0$

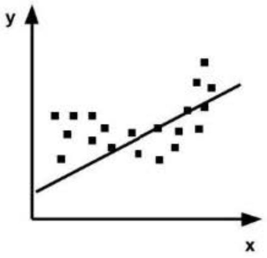
○ Homoskedasticity. $E[\epsilon_i^2] = \sigma^2$ for all i.

○ X_i, ϵ_i uncorrelated

Explanation
In the linear model, we assume that there is some variation in our regressor X. If X were always the exact same value regardless of the value of Y, then we would not be able to predict or learn anything about Y based on X.

Question 2

Which assumption of the linear model does the below scenario violate?



The Linear Model

○ Identification

● No serial correlation. $E\left[\epsilon_i \epsilon_j\right]=0$ for $i \neq j$.

○ $E\left[\epsilon_i\right]=0$

○ Homoskedasticity. $E\left[\epsilon_i^2\right]=\sigma^2$ for all i.

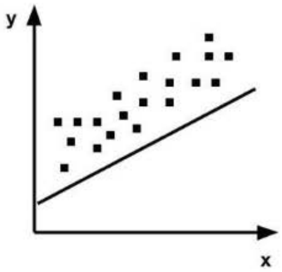
○ X_i, ϵ_i uncorrelated

✓

Explanation
In the linear model, we assume that there are no areas where errors are mostly positive or other areas where errors are mostly negative.

Question 3

Which assumption of the linear model does the below scenario violate?



○ Identification

○ No serial correlation. $E\left[\epsilon_i \epsilon_j\right]=0$ for $i \neq j$.

● $E\left[\epsilon_i\right]=0$

○ Homoskedasticity. $E\left[\epsilon_i^2\right]=\sigma^2$ for all i.

○ X_i, ϵ_i uncorrelated

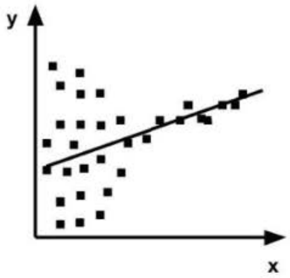
✓

Explanation
In the linear model, we assume that the expectation of the error is zero. A non-zero error mean indicates systematically under- or overpredicting β_0, so we just assume the error is zero in expectation.

The Linear Model

Question 4

Which assumption of the linear model does the below scenario violate?



● Homoskedasticity. $E\left[\epsilon_i^2\right] = \sigma^2$ for all i. ✓

○ Identification

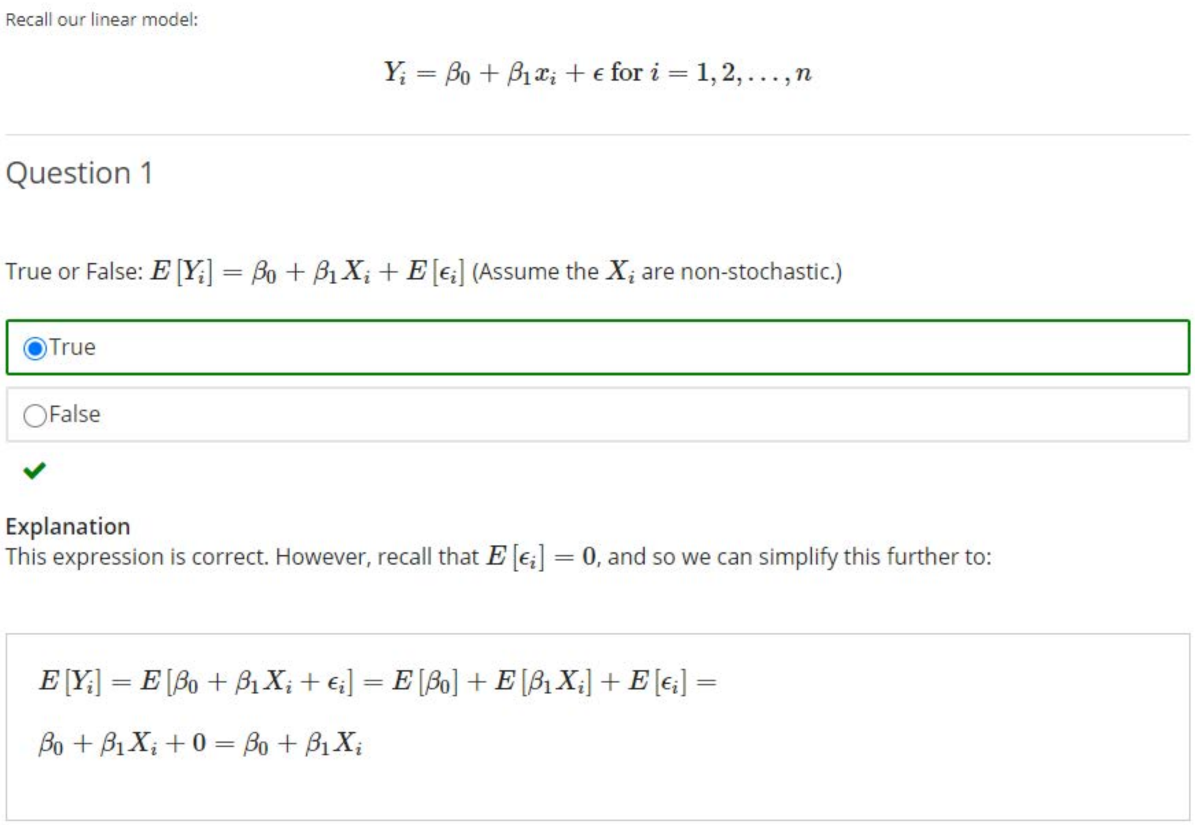
○ $E\left[\epsilon_i\right] = 0$

○ X_i, ϵ_i uncorrelated

○ No serial correlation. $E\left[\epsilon_i \epsilon_j\right] = 0$ for $i \neq j$.

Explanation
We assume homoskedasticity in the linear model, meaning the error variance should be consistent across all values of x. In the case above, the variance of the error is much higher at lower values of x.

The Linear Model



Recall our linear model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon \text{ for } i = 1, 2, \ldots, n$$

Question 1

True or False: $E\left[Y_i\right] = \beta_0 + \beta_1 X_i + E\left[\epsilon_i\right]$ (Assume the X_i are non-stochastic.)

● True

○ False

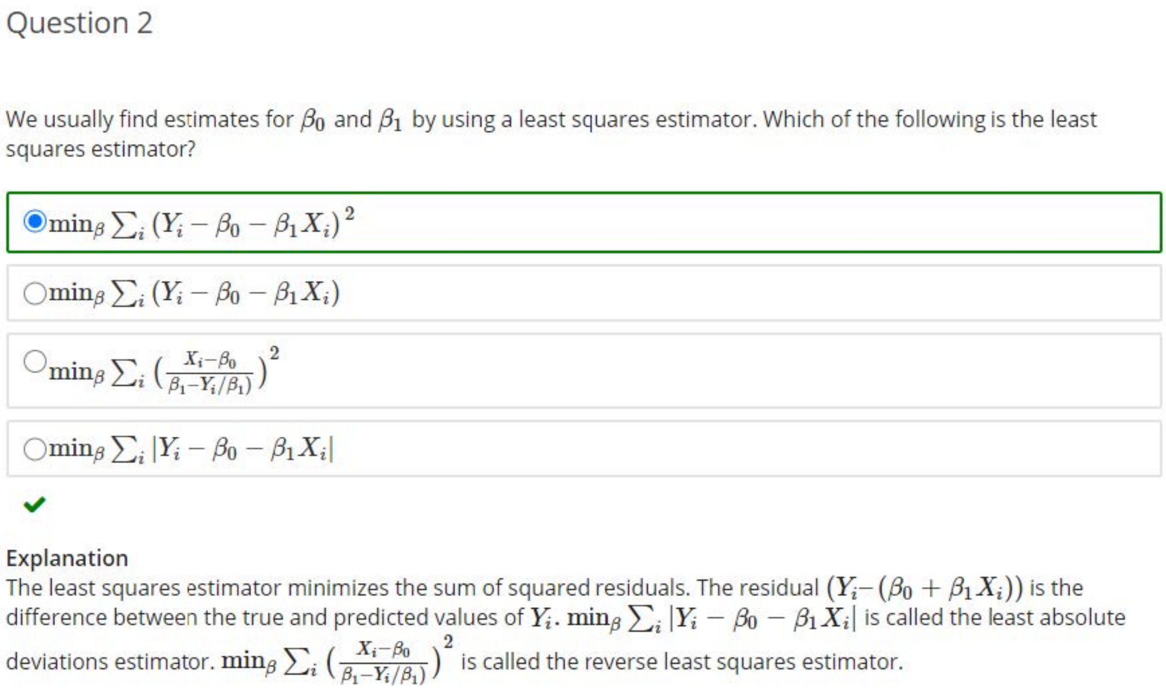
✓

Explanation

This expression is correct. However, recall that $E\left[\epsilon_i\right] = 0$, and so we can simplify this further to:

$$E\left[Y_i\right] = E\left[\beta_0 + \beta_1 X_i + \epsilon_i\right] = E\left[\beta_0\right] + E\left[\beta_1 X_i\right] + E\left[\epsilon_i\right] =$$

$$\beta_0 + \beta_1 X_i + 0 = \beta_0 + \beta_1 X_i$$



Question 2

We usually find estimates for β_0 and β_1 by using a least squares estimator. Which of the following is the least squares estimator?

● $\min_\beta \sum_i \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$

○ $\min_\beta \sum_i \left(Y_i - \beta_0 - \beta_1 X_i\right)$

○ $\min_\beta \sum_i \left(\frac{X_i - \beta_0}{\beta_1 - Y_i/\beta_1}\right)^2$

○ $\min_\beta \sum_i \left|Y_i - \beta_0 - \beta_1 X_i\right|$

✓

Explanation

The least squares estimator minimizes the sum of squared residuals. The residual $(Y_i - (\beta_0 + \beta_1 X_i))$ is the difference between the true and predicted values of Y_i. $\min_\beta \sum_i \left|Y_i - \beta_0 - \beta_1 X_i\right|$ is called the least absolute deviations estimator. $\min_\beta \sum_i \left(\frac{X_i - \beta_0}{\beta_1 - Y_i/\beta_1}\right)^2$ is called the reverse least squares estimator.

The Linear Model

Question 1

Why do we generally use OLS (the "ordinary least squares") estimator to estimate β_0 and β_1? (Select all reasons that hold under the Classical Linear Regression Model.)

☑ Provides the most efficient unbiased estimate of βs

☑ Assuming normality of errors, it is the maximum likelihood estimator

☐ Is the fastest estimator to calculate

✓

Explanation
We generally use OLS estimators because they hold several nice properties (providing the most efficient unbiased estimate of βs, maximum likelihood estimator assuming normality of errors, and provides estimates that are consistent and asymptotically normal (though this is not mentioned in the lecture)) that do not hold true of the other estimators we have seen. There may be times when the other estimators are useful (e.g. when you're worried about outliers having undue influence), but typically we choose to use OLS estimators because the properties are so good.

Question 1

Match each of these definitions with the correct terms:

$\hat{\beta}_0 + \hat{\beta}_1 X$

Regression line (fitted line) ✓ **Answer:** Regression line (fitted line)

$\hat{\beta}_0 + \hat{\beta}_1 X_i$

Fitted value Yi ✓ **Answer:** Fitted value Y_i

$Y_i - \hat{Y}_i$

Residual (ϵ^) ✓ **Answer:** Residual $(\hat{\epsilon})$

Explanation
The residual $(\hat{\epsilon}\ )$ is the deviation from an ordered pair (x, y) and the fitted regression line. The regression line is also known as the fitted line and is defined above as (b). The fitted value (Y_i) is the value of Y associated with a particular value X_i on the regression line.

Question 2

According to the properties of OLS, what is the value of $E\left[\hat{\beta}_0\right]$?

Note: type 'beta' for β. For a subscript (e.g. X_i), type "_" before the subscript (e.g. "X_i"). See here for more information on math formatting.

beta_0 ✓ **Answer:** beta_0

β_0

Explanation
As mentioned during our discussion of least squares estimators, one of the favourable properties of OLS is that the estimators are unbiased. This means that $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$.

The Linear Model

Question 1

If the error variance (σ^2) of our estimates is larger, we can estimate the linear relationship between Y and X more precisely.

○ True

● False

✓

Explanation

A higher error variance means that the variance of our $\hat{\beta}$s is higher, meaning that we should be less sure of our estimates. This means we should have less confidence in our ability to estimate the linear relationship precisely.

Question 2

Greater variance in the independent variable X (σ_x^2) means greater variance in our estimates $\hat{\beta}$.

○ True

● False ✓

Explanation

As the variance in X decreases, the variance in our estimates increases because we don't have a lot of variation in X to identify the effect we are interested in. Remember that the limit (when there is no variance in X), we cannot estimate the linear regression coefficients $\hat{\beta}$ at all.

Question 3

Which of the following is true about the relationship between $\bar{X}$ and estimates of β_0 and β_1? (Select all that apply)

☐ If $\bar{X} > 0,$ an overestimate of β_0 will likely lead to an overestimate of β_1.

☑ If $\bar{X} > 0,$ an underestimate of β_0 will likely lead to an overestimate of β_1.

☐ If $\bar{X} > 0,$ an underestimate of β_0 will likely lead to an underestimate of β_1.

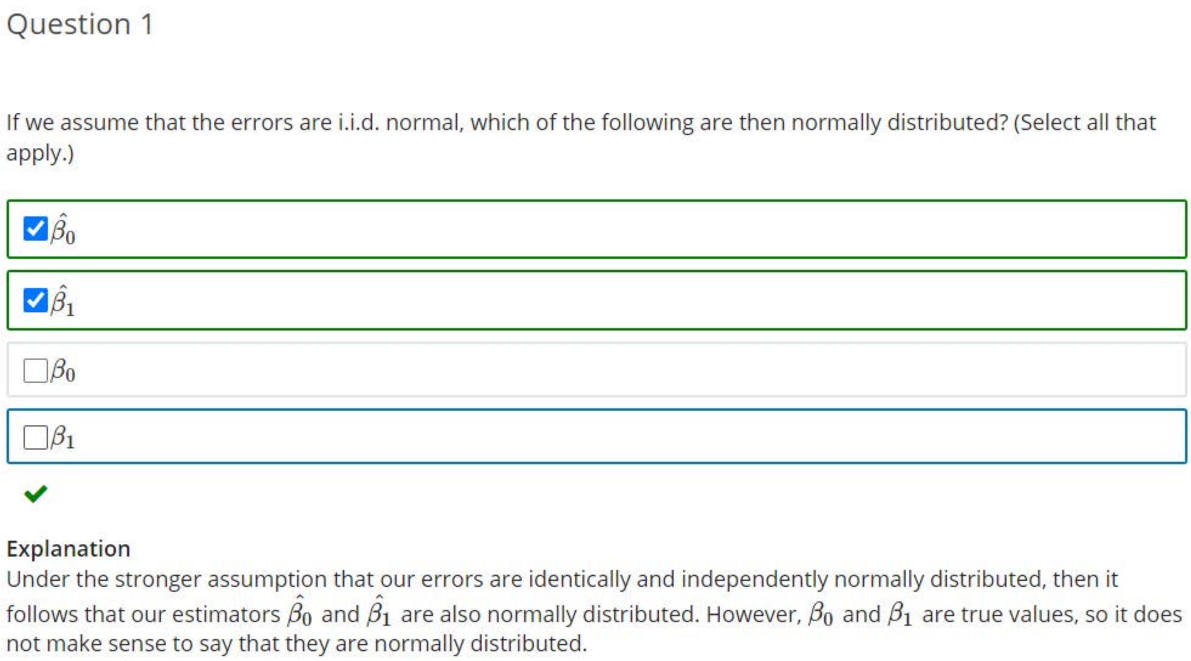
☑ If $\bar{X} > 0,$ an overestimate of β_0 will likely lead to an underestimate of β_1.

✓

Explanation

We know that $\hat{\beta}_0$ and $\hat{\beta}_1$ are related as follows: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$. Therefore, if we overestimate or underestimate the intercept β_0, then the slope β_1 will have to make up for it (by doing the opposite).

The Linear Model



Question 1

If we assume that the errors are i.i.d. normal, which of the following are then normally distributed? (Select all that apply.)

☑ $\hat{\beta_0}$

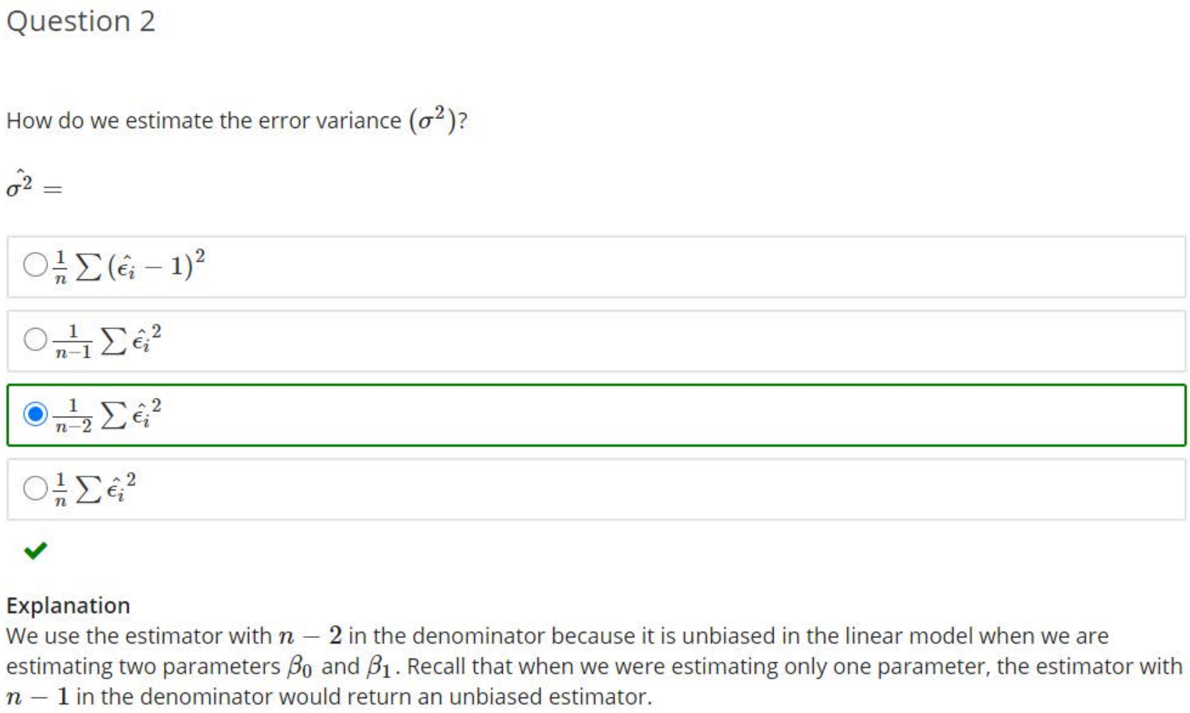
☑ $\hat{\beta_1}$

☐ β_0

☐ β_1

✓

Explanation
Under the stronger assumption that our errors are identically and independently normally distributed, then it follows that our estimators $\hat{\beta_0}$ and $\hat{\beta_1}$ are also normally distributed. However, β_0 and β_1 are true values, so it does not make sense to say that they are normally distributed.



Question 2

How do we estimate the error variance (σ^2)?

$\hat{\sigma^2} =$

○ $\frac{1}{n}\sum(\hat{\epsilon_i}-1)^2$

○ $\frac{1}{n-1}\sum\hat{\epsilon_i}^2$

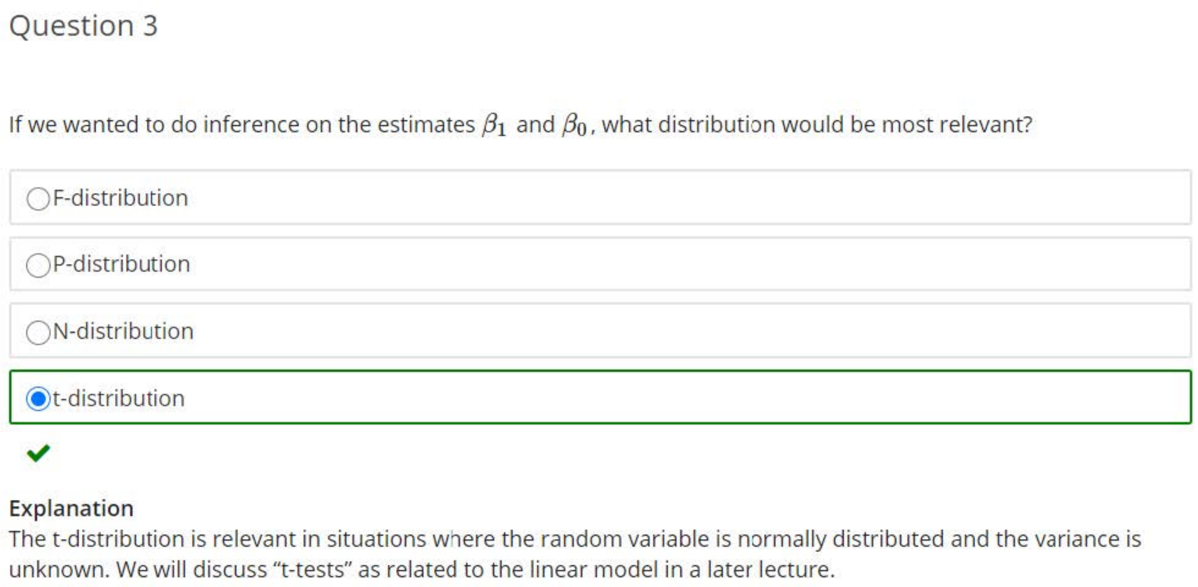
● $\frac{1}{n-2}\sum\hat{\epsilon_i}^2$

○ $\frac{1}{n}\sum\hat{\epsilon_i}^2$

✓

Explanation
We use the estimator with $n-2$ in the denominator because it is unbiased in the linear model when we are estimating two parameters β_0 and β_1. Recall that when we were estimating only one parameter, the estimator with $n-1$ in the denominator would return an unbiased estimator.

The Linear Model



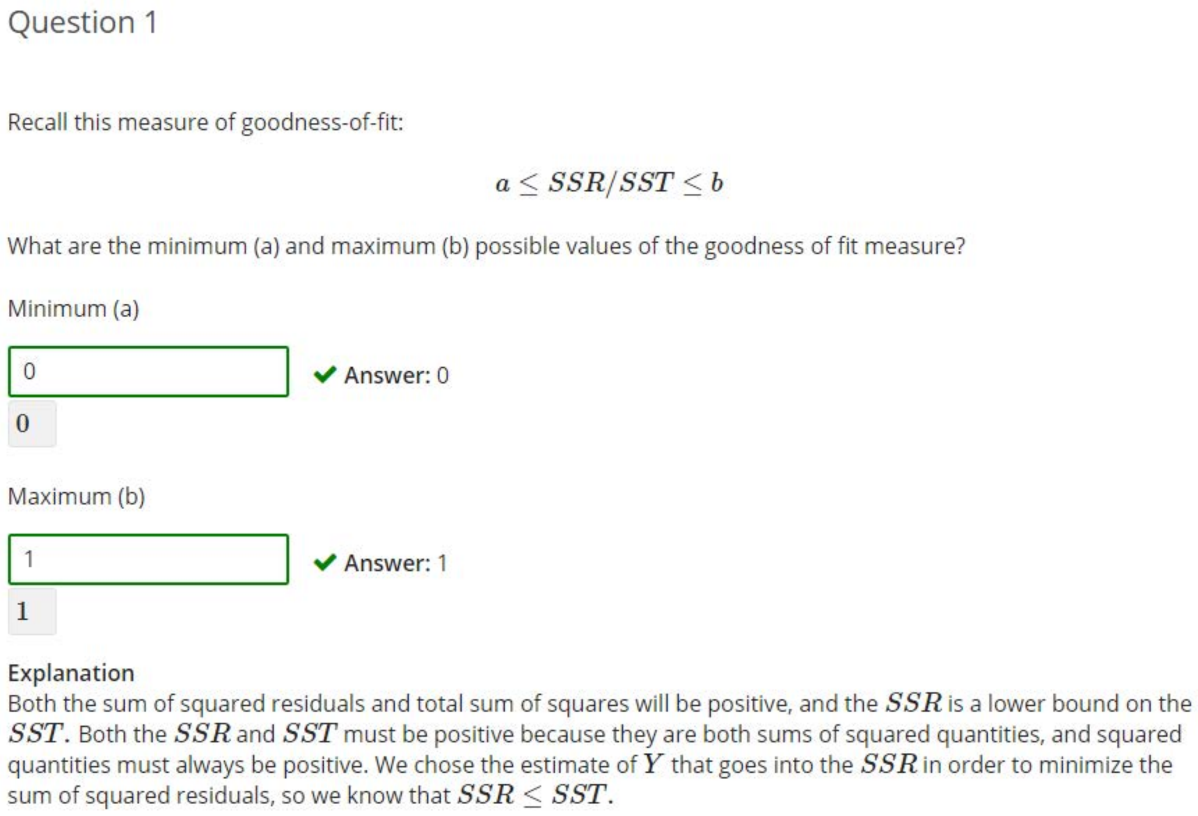
Question 3

If we wanted to do inference on the estimates β_1 and β_0, what distribution would be most relevant?

- ○ F-distribution
- ○ P-distribution
- ○ N-distribution
- ● t-distribution

✓

Explanation
The t-distribution is relevant in situations where the random variable is normally distributed and the variance is unknown. We will discuss "t-tests" as related to the linear model in a later lecture.



Question 1

Recall this measure of goodness-of-fit:

$$a \leq SSR/SST \leq b$$

What are the minimum (a) and maximum (b) possible values of the goodness of fit measure?

Minimum (a)

0 ✓ **Answer:** 0

0

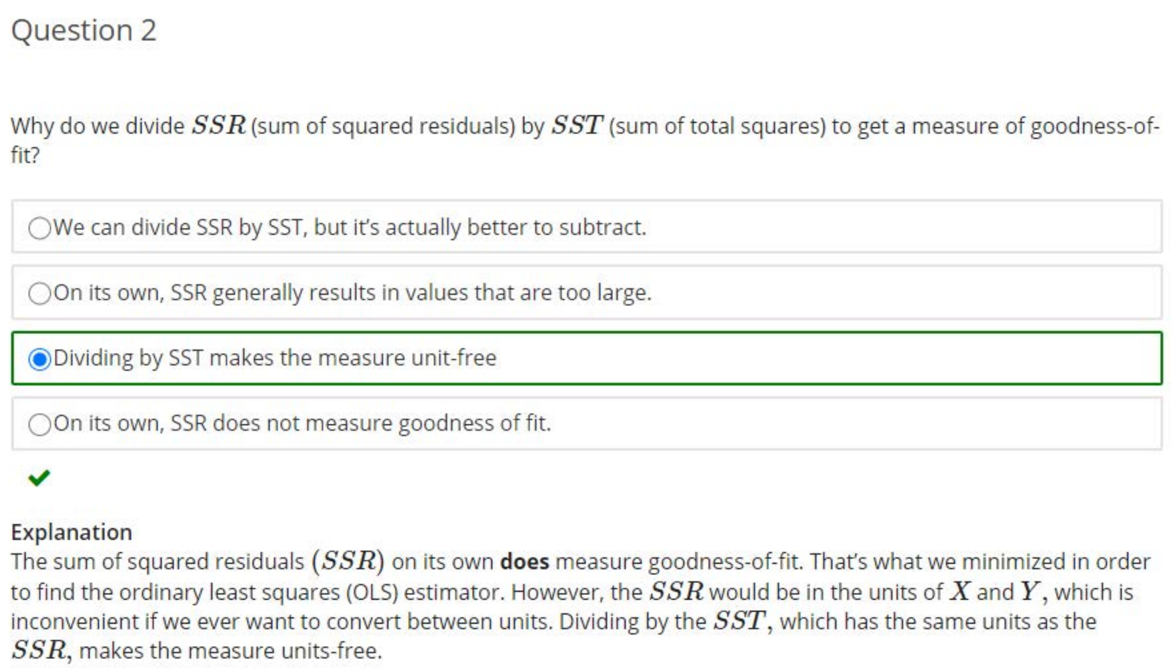
Maximum (b)

1 ✓ **Answer:** 1

1

Explanation
Both the sum of squared residuals and total sum of squares will be positive, and the SSR is a lower bound on the SST. Both the SSR and SST must be positive because they are both sums of squared quantities, and squared quantities must always be positive. We chose the estimate of Y that goes into the SSR in order to minimize the sum of squared residuals, so we know that $SSR \leq SST$.

The Linear Model



Question 2

Why do we divide SSR (sum of squared residuals) by SST (sum of total squares) to get a measure of goodness-of-fit?

○ We can divide SSR by SST, but it's actually better to subtract.

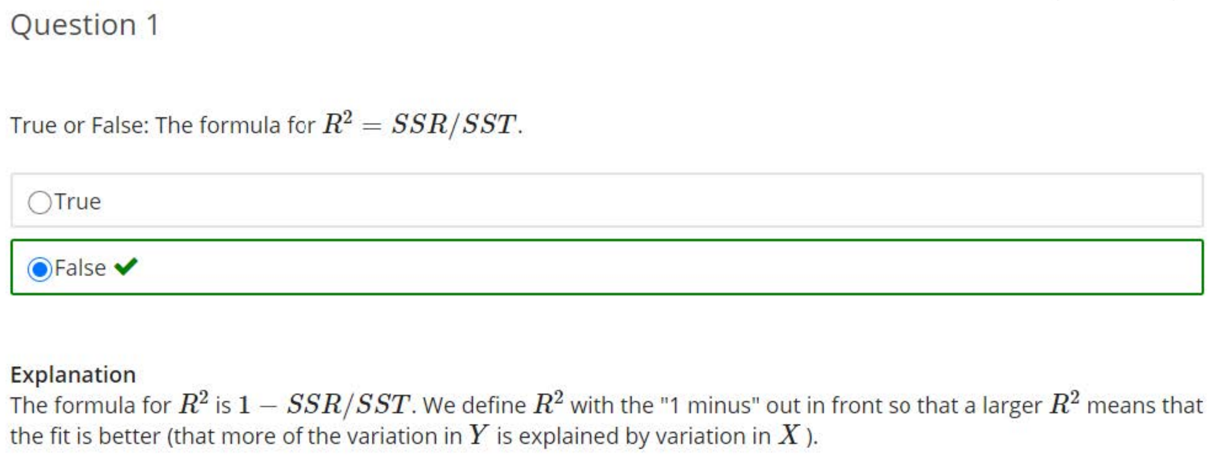
○ On its own, SSR generally results in values that are too large.

● Dividing by SST makes the measure unit-free

○ On its own, SSR does not measure goodness of fit.

✓

Explanation
The sum of squared residuals (SSR) on its own **does** measure goodness-of-fit. That's what we minimized in order to find the ordinary least squares (OLS) estimator. However, the SSR would be in the units of X and Y, which is inconvenient if we ever want to convert between units. Dividing by the SST, which has the same units as the SSR, makes the measure units-free.



Question 1

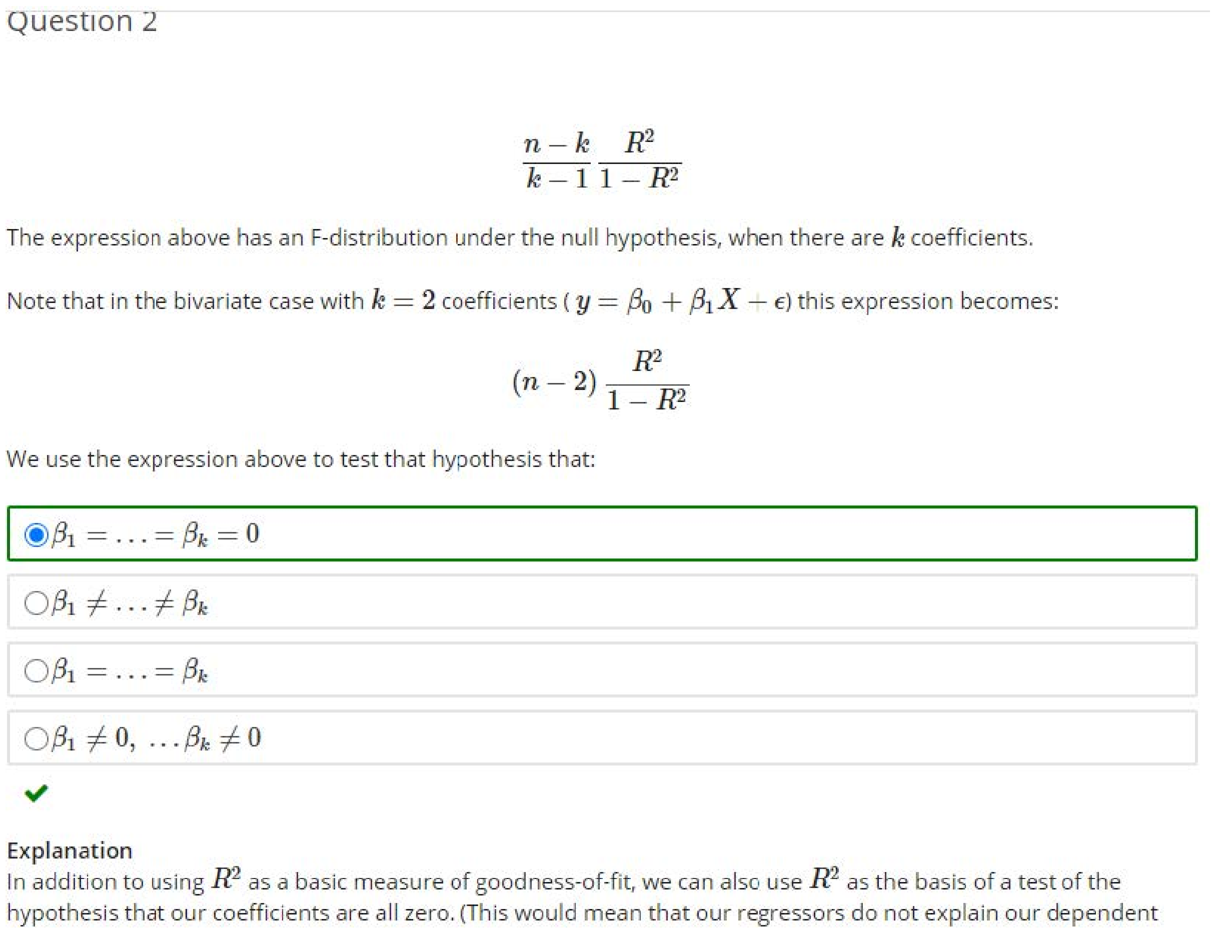
True or False: The formula for $R^2 = SSR/SST$.

○ True

● False ✓

Explanation
The formula for R^2 is $1 - SSR/SST$. We define R^2 with the "1 minus" out in front so that a larger R^2 means that the fit is better (that more of the variation in Y is explained by variation in X).

The Linear Model



Question 2

$$\frac{n-k}{k-1}\frac{R^2}{1-R^2}$$

The expression above has an F-distribution under the null hypothesis, when there are k coefficients.

Note that in the bivariate case with $k=2$ coefficients ($y=\beta_0+\beta_1 X+\epsilon$) this expression becomes:

$$(n-2)\frac{R^2}{1-R^2}$$

We use the expression above to test that hypothesis that:

● $\beta_1=\ldots=\beta_k=0$

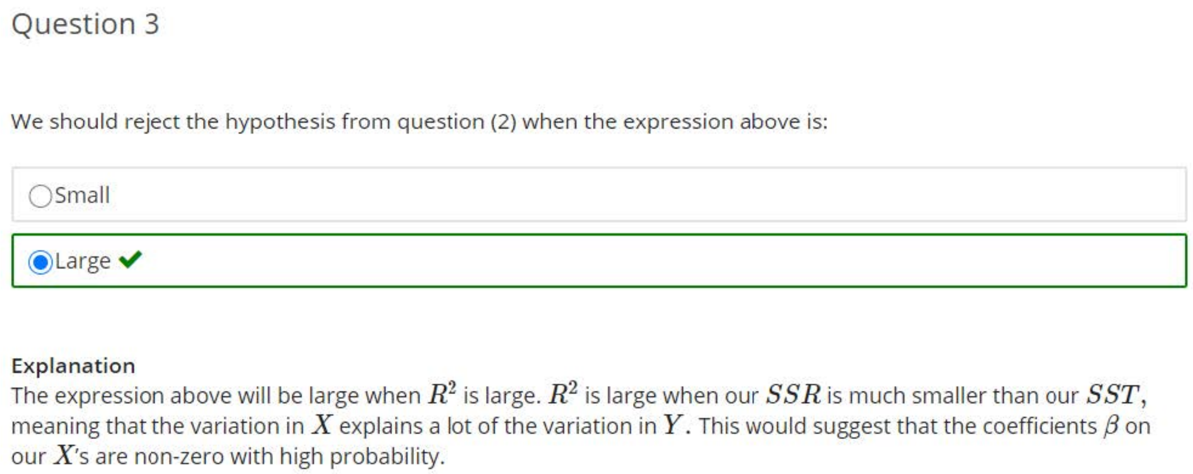
○ $\beta_1\neq\ldots\neq\beta_k$

○ $\beta_1=\ldots=\beta_k$

○ $\beta_1\neq 0,\ \ldots\beta_k\neq 0$

✓

Explanation
In addition to using R^2 as a basic measure of goodness-of-fit, we can also use R^2 as the basis of a test of the hypothesis that our coefficients are all zero. (This would mean that our regressors do not explain our dependent



Question 3

We should reject the hypothesis from question (2) when the expression above is:

○ Small

● Large ✓

Explanation
The expression above will be large when R^2 is large. R^2 is large when our SSR is much smaller than our SST, meaning that the variation in X explains a lot of the variation in Y. This would suggest that the coefficients β on our X's are non-zero with high probability.

The Linear Model

Question 1

Suppose our Stata or R output tells us the following: Prob $> \quad F = 0.0879$

We reject the null hypothesis that $\beta_1 = \ldots = \beta_k = 0$ under a 5% test.

○True

●False

✓

Explanation
This is telling us that we should not reject the null under a 5% test $(0.05 < 0.0879)$. However, we would reject the null hypothesis under a 10% test.

Question 1

When you run a regression in Stata or R, the output usually gives you t-tests for each coefficient. A t-test that says $Pr\,(> |t|) = 0.05$ in the row for β_1 would tell you:

○To reject the null hypothesis that $\beta_1 \neq 0$ for a two-sided test at the 0.05% level or above

○To reject the null hypothesis that $\beta_1 \neq 0$ for a two-sided test at the 5% level or above.

○To reject the null hypothesis that $\beta_1 = 0$ for a two-sided test at the 0.05% level or above

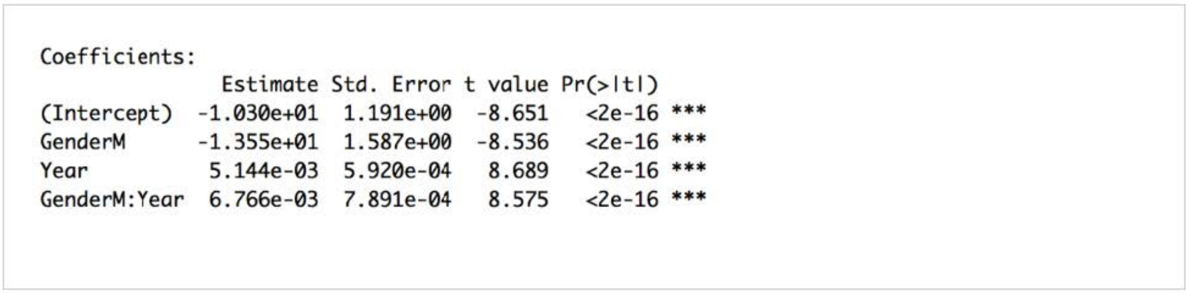
●To reject the null hypothesis that $\beta_1 = 0$ for a two-sided test at the 5% level or above.

✓

Explanation
T-tests given for each coefficient can be used to test whether or not to reject the null hypothesis that the given coefficient is equal to zero. You choose to reject the null hypothesis or not depending on what level test you choose to use. Common tests would be 1% or 5% tests. These percentage values are known as α.

Question 2

Suppose you get the following R output:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.030e+01  1.191e+00  -8.651   <2e-16 ***
GenderM      -1.355e+01  1.587e+00  -8.536   <2e-16 ***
Year          5.144e-03  5.920e-04   8.689   <2e-16 ***
GenderM:Year  6.766e-03  7.891e-04   8.575   <2e-16 ***
```

The Linear Model

Which of the following values corresponds to $\hat{\beta}_0$?

○ 7.891e-04

○ 6.766e-03

● -1.030e+01

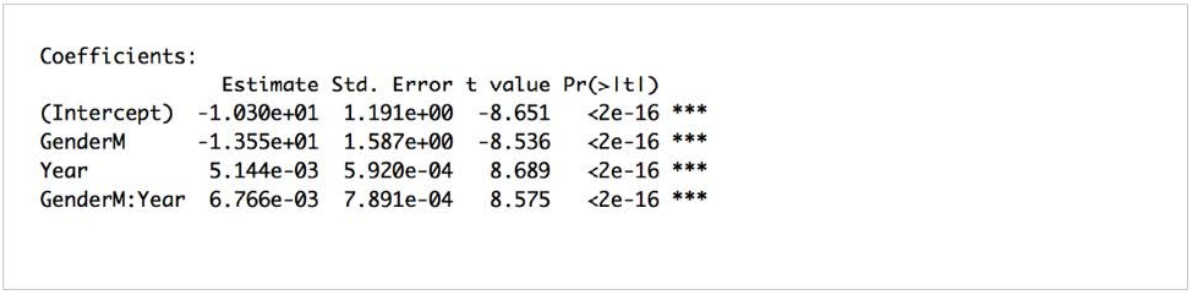
○ -8.651

✓

Explanation

The "Estimate" value for "(Intercept)" is the estimate of the y-intercept, $\hat{\beta}_0$.

Question 3



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.030e+01  1.191e+00  -8.651   <2e-16 ***
GenderM      -1.355e+01  1.587e+00  -8.536   <2e-16 ***
Year          5.144e-03  5.920e-04   8.689   <2e-16 ***
GenderM:Year  6.766e-03  7.891e-04   8.575   <2e-16 ***
```

Which of the following values corresponds to the standard error for $\hat{\beta}_1$, where β_1 is the estimate for coefficient on the Year random variable?

○ 1.191e+00

○ 8.689

○ 2e-16

● 5.920e-4

✓

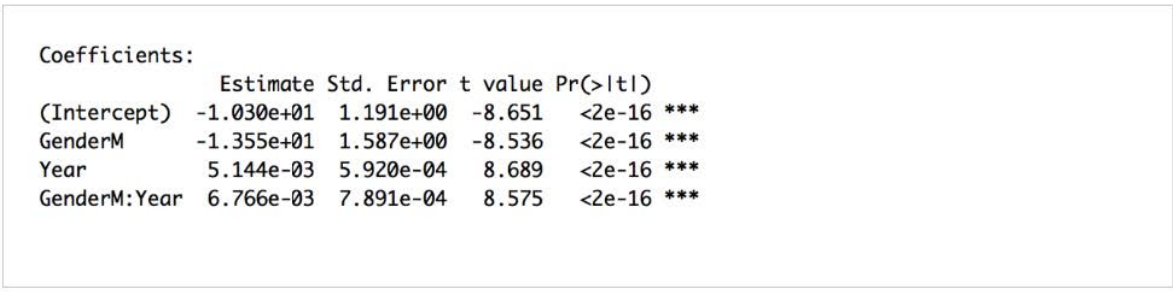
Explanation

The regression output gives you a standard error for the estimator for each coefficient. Standard error is a measurement of the accuracy of predictions made with the estimator and is generally abbreviated to "Std. Err." in regression outputs.

The Linear Model

Question 4

Consider the R output again:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.030e+01  1.191e+00  -8.651   <2e-16 ***
GenderM      -1.355e+01  1.587e+00  -8.536   <2e-16 ***
Year          5.144e-03  5.920e-04   8.689   <2e-16 ***
GenderM:Year  6.766e-03  7.891e-04   8.575   <2e-16 ***
```

Suppose our dependent variable is the number of students interested in computer science courses.

Given what we currently know, how do we interpret the value 5.144e-03?

○ One additional student becomes interested in computer science every 5.144e-03 years.

○ One additional year is associated with a decrease of 5.144e-03 students interested in computer science courses.

○ Student interest in computer science has been increasing steadily over time.

● One additional year is associated with an increase of 5.144e-03 students interested in computer science courses.

✓

Explanation
The estimate of a coefficient on a regressor variable is the change in dependent variable associated with one unit of change in the regressor. This doesn't necessarily mean it's a causal relationship, though.

Question 1

Which of the following is true about dummy variables? (Select all that apply)

☑ They only take on one of two values: 0 or 1. ✓

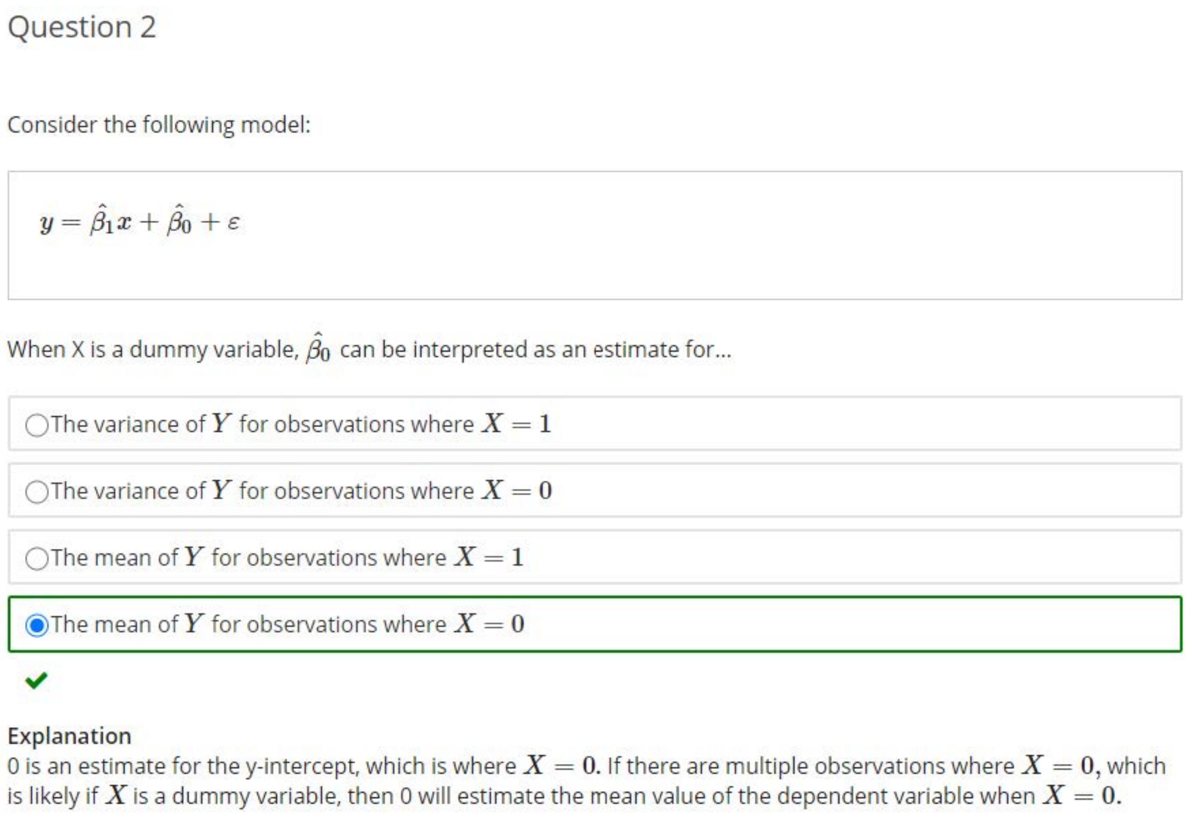
☐ They violate the basic assumptions of the linear model.

☑ A dummy variable contains useful information when representing a characteristic that is true of some, but not all, observations. ✓

☐ Dummy variables can only be used for characteristics that are randomly assigned, like in RCTs.

Explanation
Dummy variables are variables that take on only one of two values, 0 or 1. They do not violate any assumptions of linear models and can be used in the linear regressions that we have looked at. Dummy variables are useful in RCTs; often, members of the treatment group are assigned a "1" and members of the control group are assigned a "0". However, dummy variables can also be used in any case to separate members of a population who fulfil one characteristic from the members of the population that do not fulfil that characteristic.

The Linear Model



Question 2

Consider the following model:

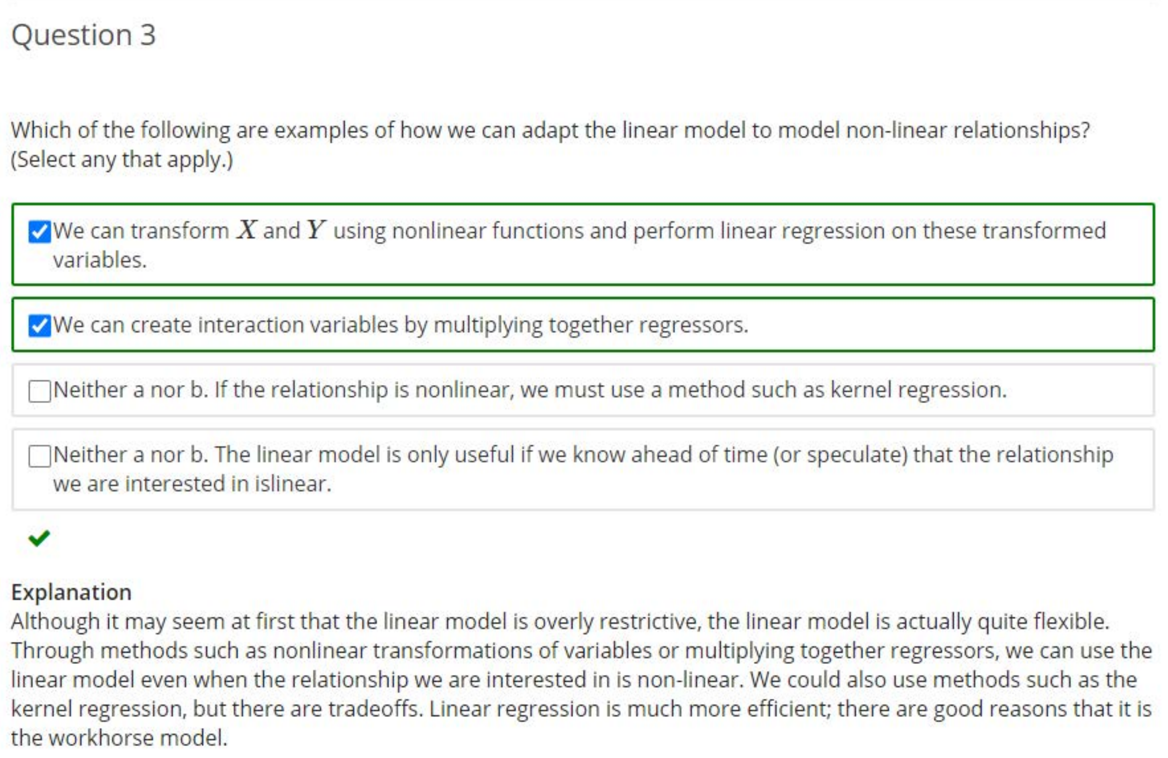
$$y = \hat{\beta}_1 x + \hat{\beta}_0 + \varepsilon$$

When X is a dummy variable, $\hat{\beta}_0$ can be interpreted as an estimate for...

- ○ The variance of Y for observations where $X = 1$
- ○ The variance of Y for observations where $X = 0$
- ○ The mean of Y for observations where $X = 1$
- ● The mean of Y for observations where $X = 0$

✓

Explanation
0 is an estimate for the y-intercept, which is where $X = 0$. If there are multiple observations where $X = 0$, which is likely if X is a dummy variable, then 0 will estimate the mean value of the dependent variable when $X = 0$.



Question 3

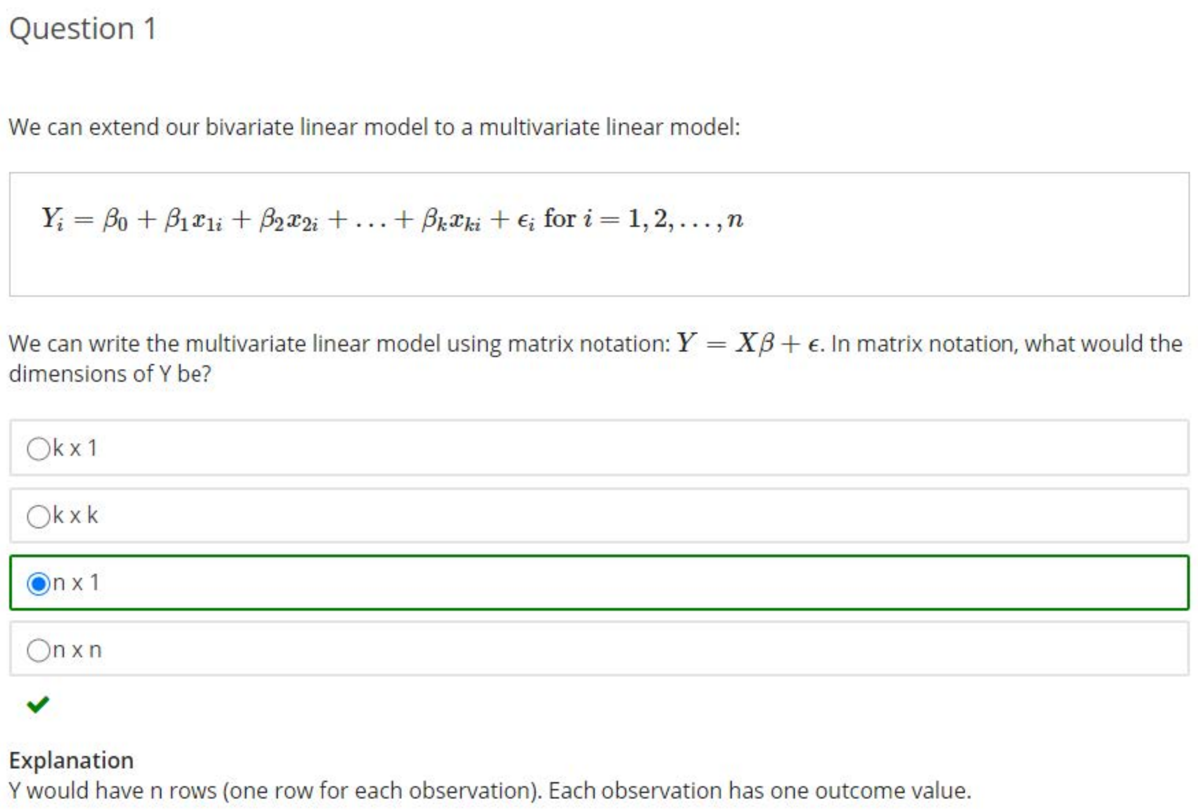
Which of the following are examples of how we can adapt the linear model to model non-linear relationships? (Select any that apply.)

- ☑ We can transform X and Y using nonlinear functions and perform linear regression on these transformed variables.
- ☑ We can create interaction variables by multiplying together regressors.
- ☐ Neither a nor b. If the relationship is nonlinear, we must use a method such as kernel regression.
- ☐ Neither a nor b. The linear model is only useful if we know ahead of time (or speculate) that the relationship we are interested in islinear.

✓

Explanation
Although it may seem at first that the linear model is overly restrictive, the linear model is actually quite flexible. Through methods such as nonlinear transformations of variables or multiplying together regressors, we can use the linear model even when the relationship we are interested in is non-linear. We could also use methods such as the kernel regression, but there are tradeoffs. Linear regression is much more efficient; there are good reasons that it is the workhorse model.

The Linear Model



Question 1

We can extend our bivariate linear model to a multivariate linear model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i \text{ for } i = 1, 2, \ldots, n$$

We can write the multivariate linear model using matrix notation: $Y = X\beta + \epsilon$. In matrix notation, what would the dimensions of Y be?

○ k x 1

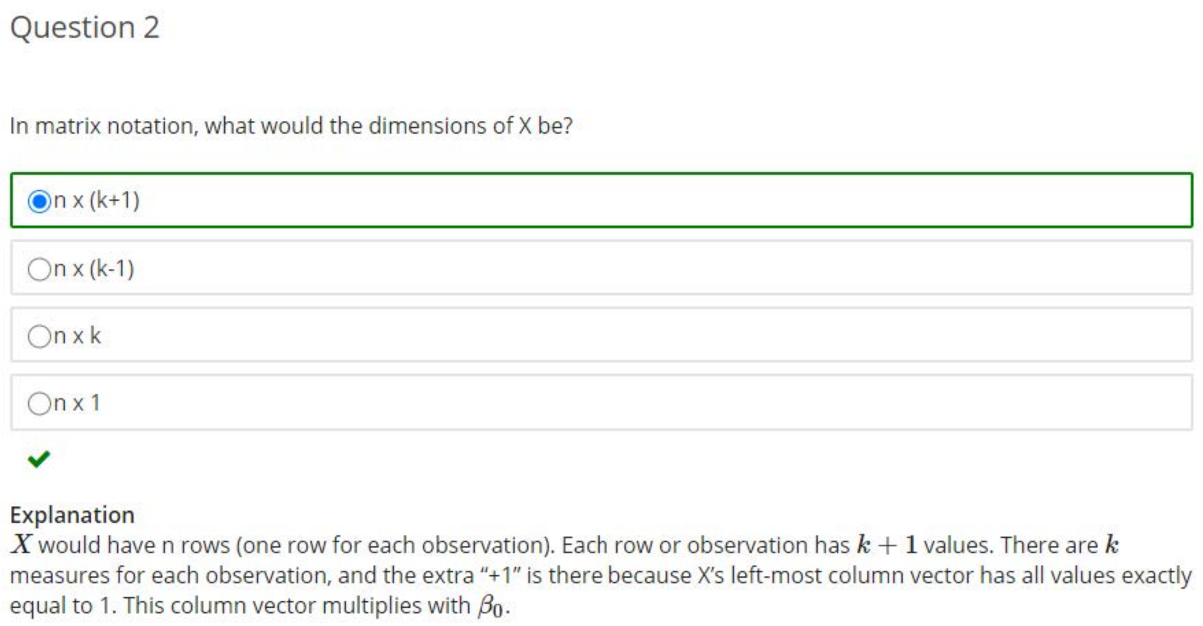
○ k x k

● n x 1

○ n x n

✓

Explanation
Y would have n rows (one row for each observation). Each observation has one outcome value.



Question 2

In matrix notation, what would the dimensions of X be?

● n x (k+1)

○ n x (k-1)

○ n x k

○ n x 1

✓

Explanation
X would have n rows (one row for each observation). Each row or observation has $k + 1$ values. There are k measures for each observation, and the extra "+1" is there because X's left-most column vector has all values exactly equal to 1. This column vector multiplies with β_0.