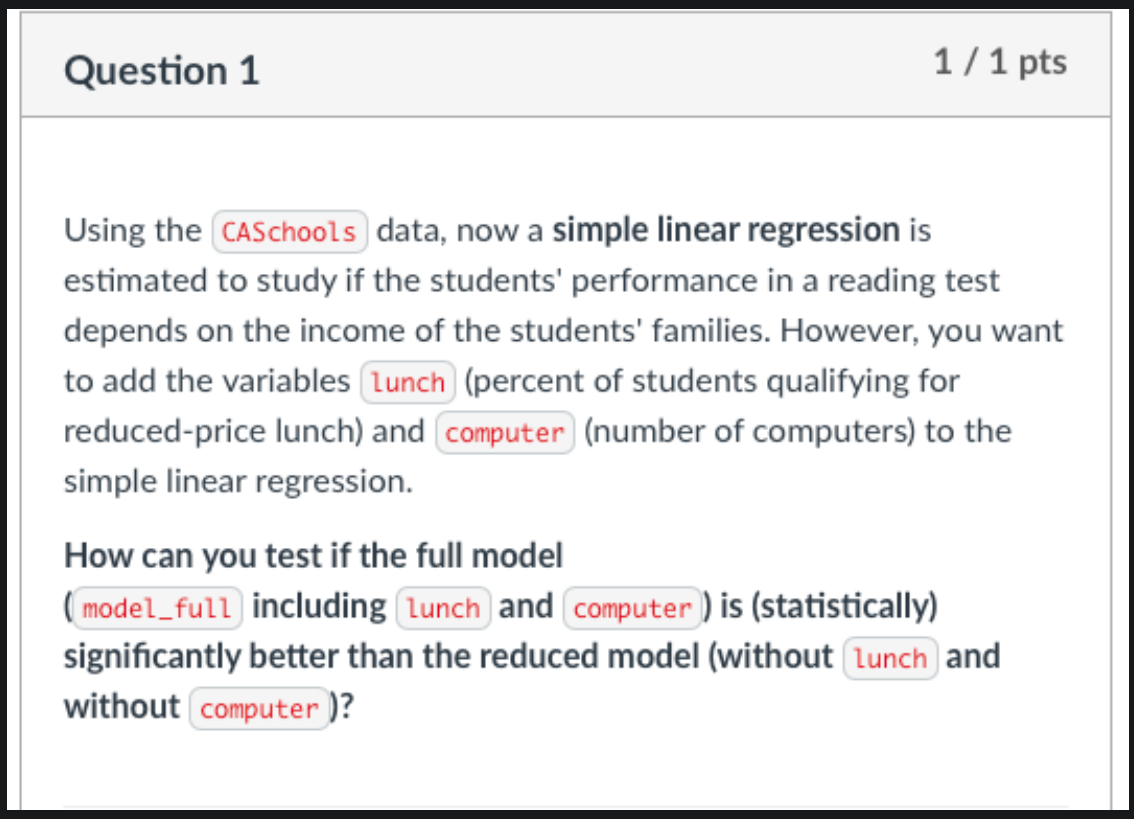
Midterm

- question 1

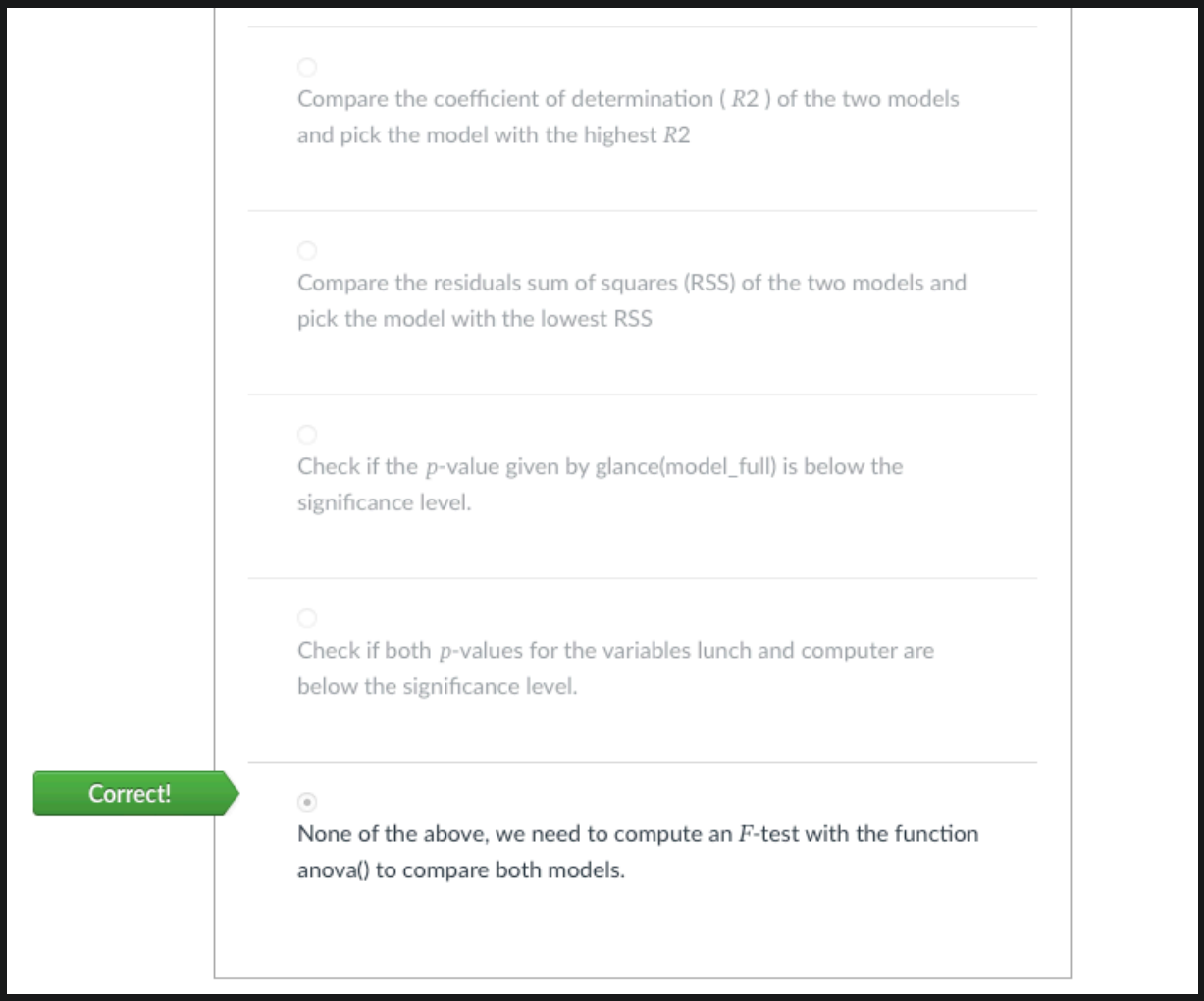


Question 1

1 / 1 pts

Using the `CASchools` data, now a **simple linear regression** is estimated to study if the students' performance in a reading test depends on the income of the students' families. However, you want to add the variables `lunch` (percent of students qualifying for reduced-price lunch) and `computer` (number of computers) to the simple linear regression.

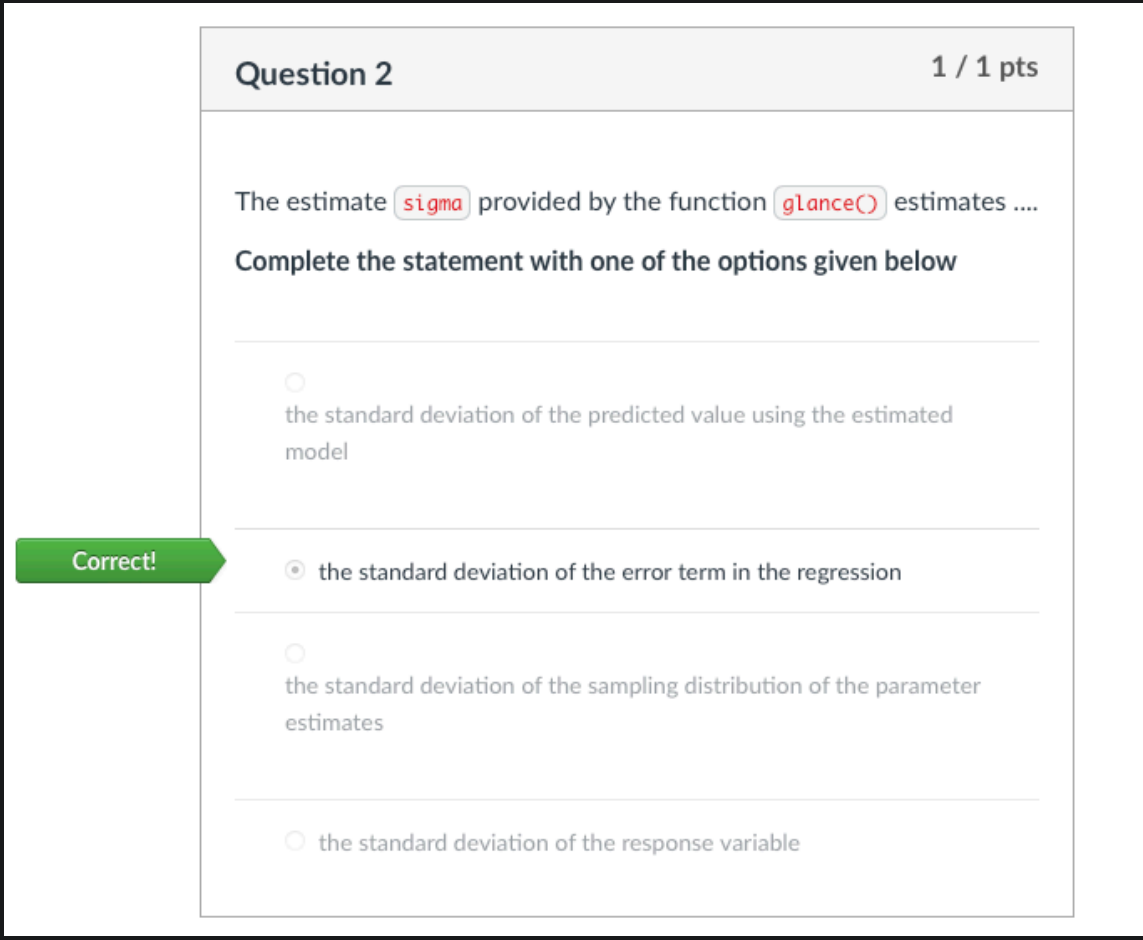
How can you test if the full model (`model_full` including `lunch` and `computer`) is (statistically) significantly better than the reduced model (without `lunch` and without `computer`)?



1. False - R^2 always increases as you add more predictors, cannot be used to compare models of different sizes
2. False - RSS is a measure of the model's unexplained variance, and while a lower RSS indicates a better fit, it doesn't consider the complexity of the model
    - also, we didn't really do this in class
3. False - `glance` is telling us if the `model_full` is significant or not (is our `model_full` better than the null, or is all the additional predictors irrelevant)
    - however, this does not tell us how it compares to the smaller model
4. False - again, checking the `p-value` does tell us if the newly added predictors in `model_full` is significant or not, but not how it compares to the smaller model

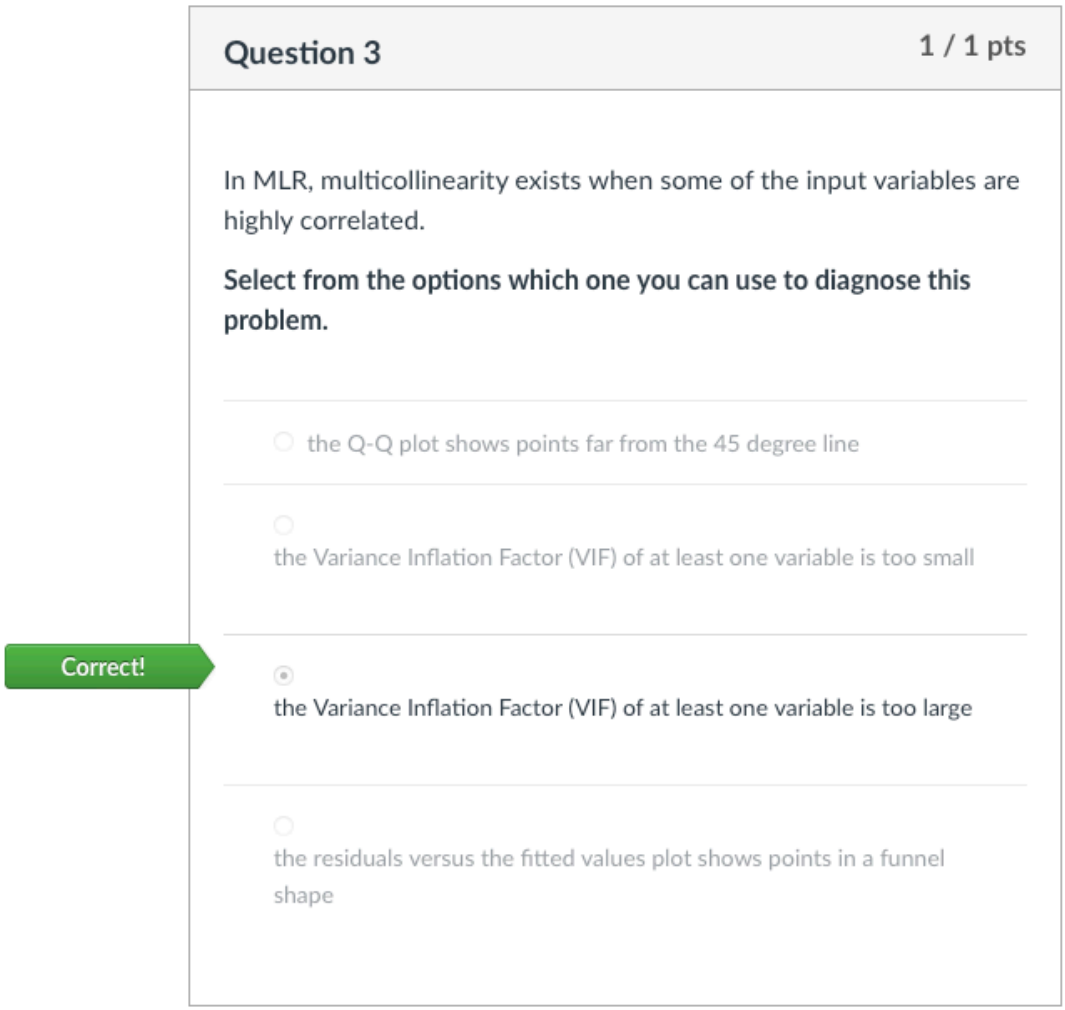
5. True - this is what we've been doing in class, the F-test doesn't just tell us if the newly predictors are significant or not, but rather "Do the additional variables improve the explanatory power of the model significantly, beyond what could be expected by chance?"
    - it's possible for individual coefficients to be statistically significant, yet the overall improvement in model fit might not justify the added complexity (e.g., due to overfitting)
    - so F-test is taking all of this into account for us, and there is a direct comparison being made between `model_full` and the smaller model

- question 2



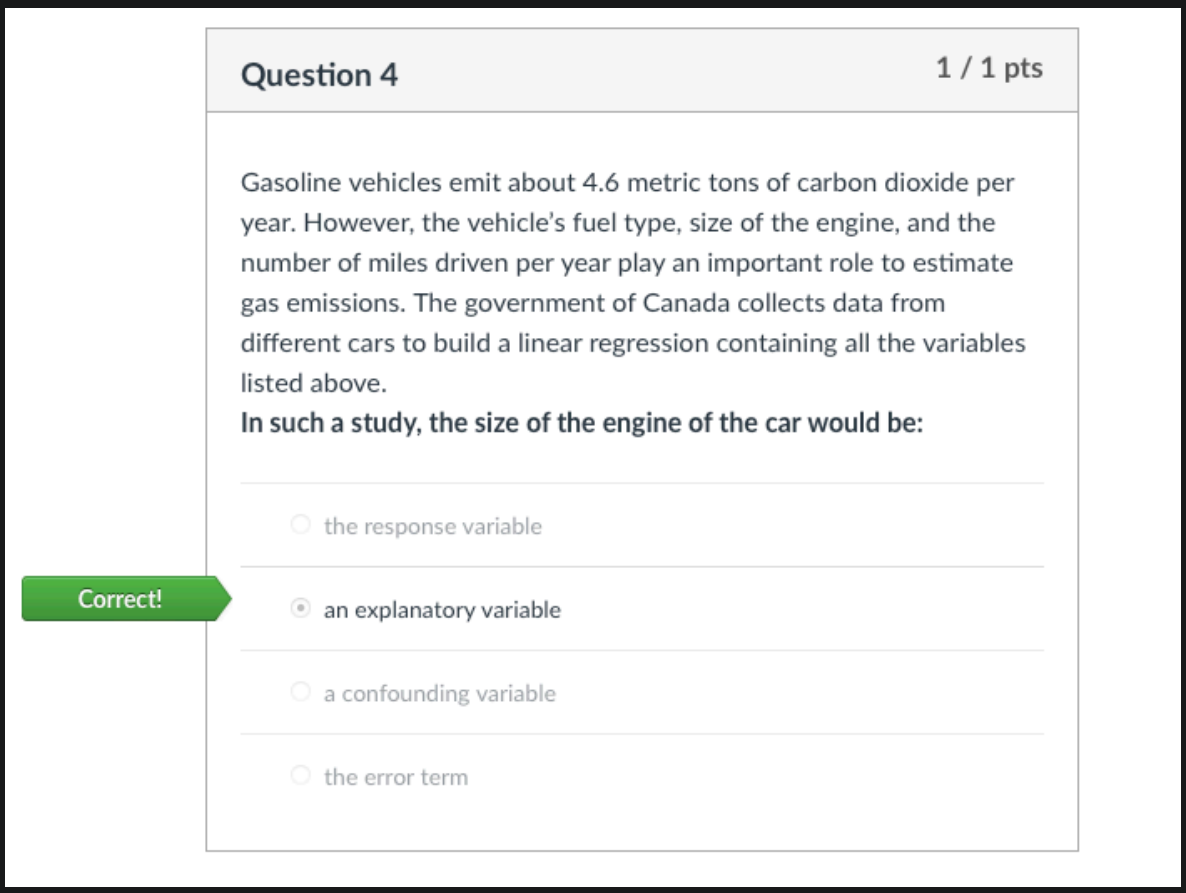
    - this was just a fun fact from the notes

- question 3



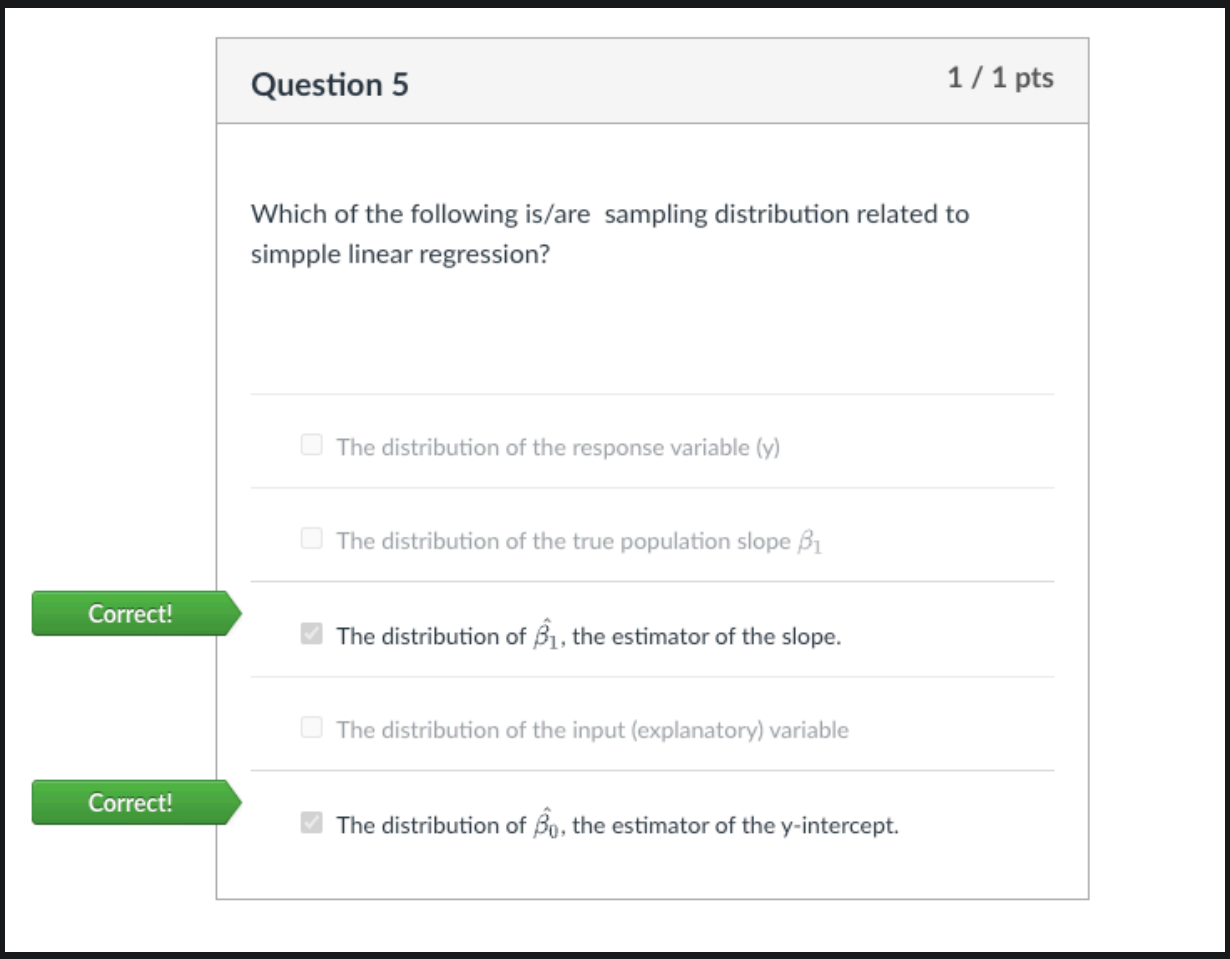
1. False - this is talking about the Normality of the data
2. False - VIF being small is good
3. True - VIF being large is indication that there's multicollinearity
4. False - this is talking about constant of variance

- question 4



- not much to say, we're trying to predict gas emission so that would be the response variable
- we're using engine size to predict gas emission so engine size is an explanatory variable

- question 5



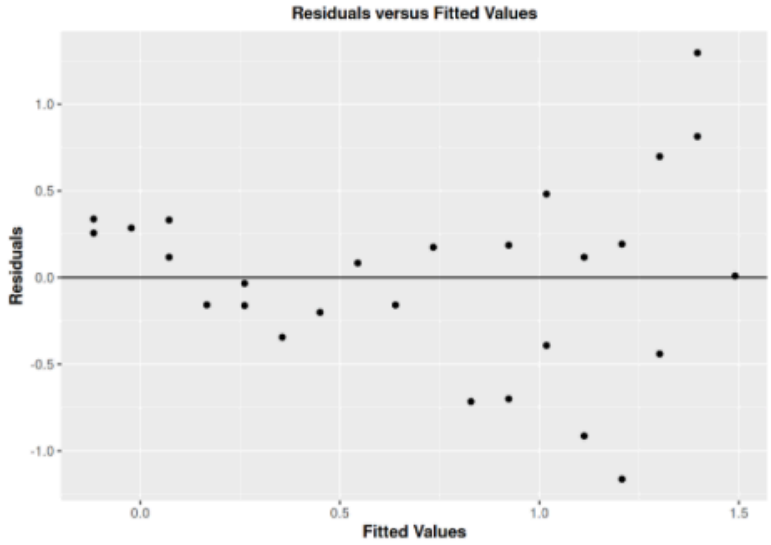
- when we're talking about sampling distributions, we're talking about the estimators themselves i.e $\hat{\beta}_i$
- distribution of y is simply the range of values that the response variable can take
- β_1 is the true population slope which is unknown but assumes to be constant
  - another way to think about is that it does not have a sampling distribution because it is not a statistic that varies from sample to sample

- question 6

Question 6

1 / 1 pts

The birth weight of a newborn baby is associated with the length of the pregnancy (gestation period) and the smoking habits of the mother. Using a data set called `baby_data` the estimated linear regression lm_`baby_weight` you obtained the following plot of the residuals versus the fitted values.



Which assumption of linear regression do you believe it has been violated?

- all input variables are independent
- there are no confounding factors

Correct!

- equal variance of the error terms

- we see a funnel shape here, which mean that it violates the assumption of equal variance

- question 7

Question 7

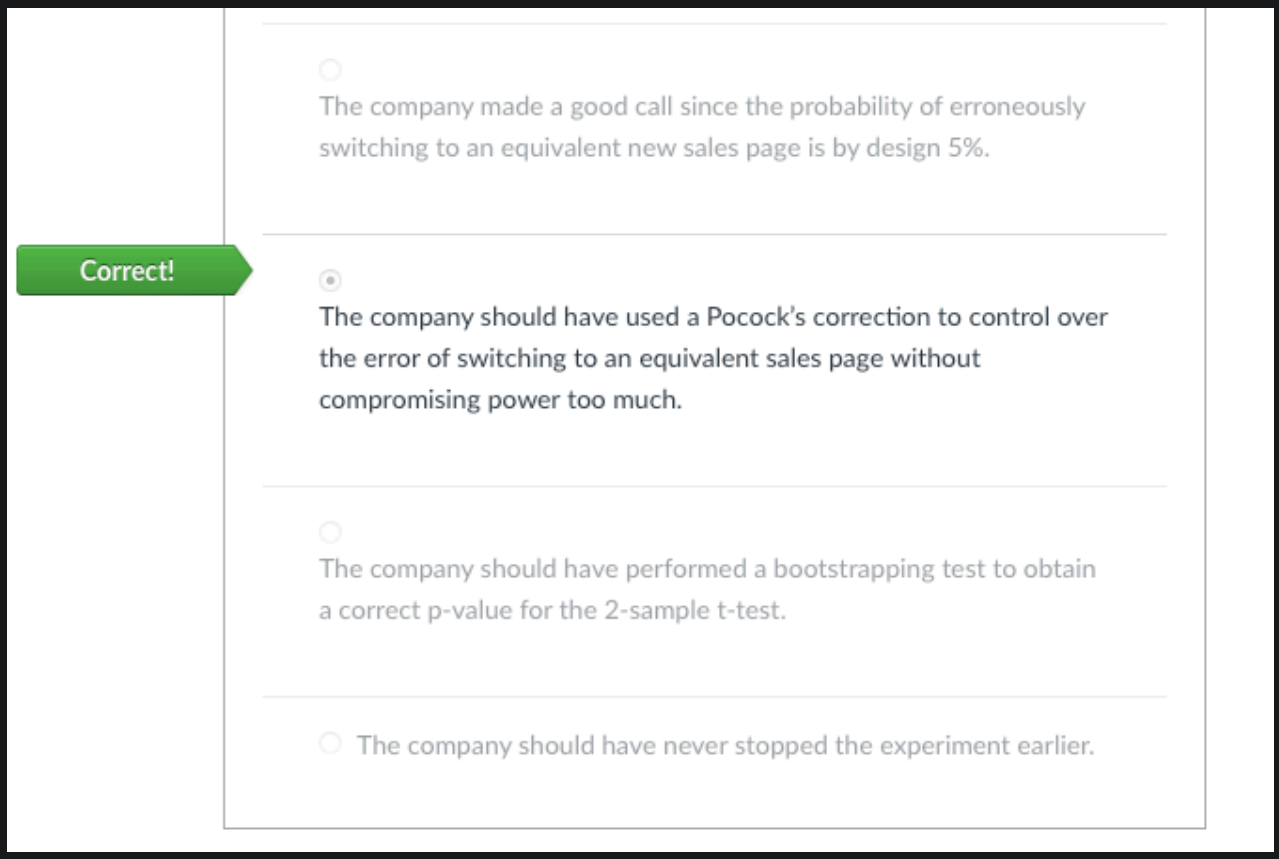
1 / 1 pts

A successful media company released a new version of one of its most popular applications, HomeDec3, which allows users to construct and decorate their own house. The members of the marketing department designed an A/B experiment to decide if the sales page should be changed:

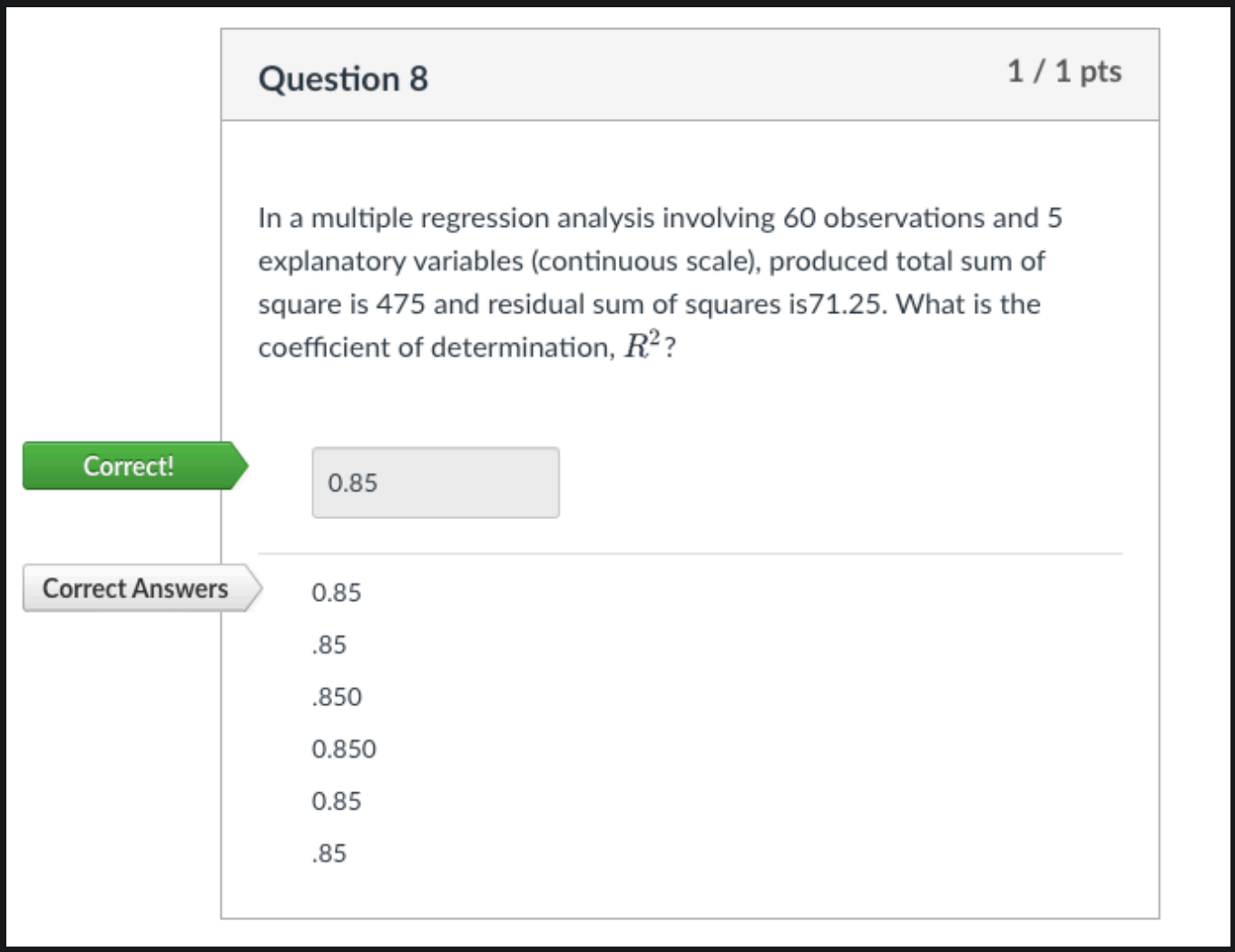
- **Variation A**: (control) the sales page offers 20 percent off a future purchase for anyone who buys HomeDec3.
- **Variation B**: (new) the sales page offers a 10-day free trial of the new version HomeDec3.

They have planned to randomly allocate customers to each page over the course of 3 months. However, after only one month onto the experiment, they believed they have sufficient data to make a decision since the p-value from a 2-sample t-test based on the partial data collected is 0.01.

Using a significance level of 5%, the company stopped the experiment earlier and changed the sales page to increase revenue. Which of the following observations is correct?



    - from class, we know that when we do early stopping, we need to do some kind of correction to prevent the inflation of type I error
    - here, the only option that mentions any adjustment is 2 which uses Pocock
- question 8



- the math is

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{71.25}{475} = 0.85$$

- note that the 60 observations and 5 explanatory variables thing is a red-herring - they're asking for R^2 and not adjusted R^2

- question 9

Question 9

1 / 1 pts

The function `glance()` provides the statistic and *p*-value of an *F*-test that compares two nested models for a given data. Complete the code below to compute the same values for this test.
IMPORTANT: do **NOT** use spaces between variables and symbols.
Complete **ONLY** the missing part in the code below.
`anova(.........., lm(y~.,data=dat))`

- important to know that `glance()` is doing an ANOVA test between the full model vs the null model
  - null model is `lm(y ~ 1, data = dat)`
- so what `glance` is really doing is `anova(null_model, full_model)` which gives us

```
1  anova(lm(y~1,data=dat), lm(y~.,data=dat))
```

- question 10

Question 10

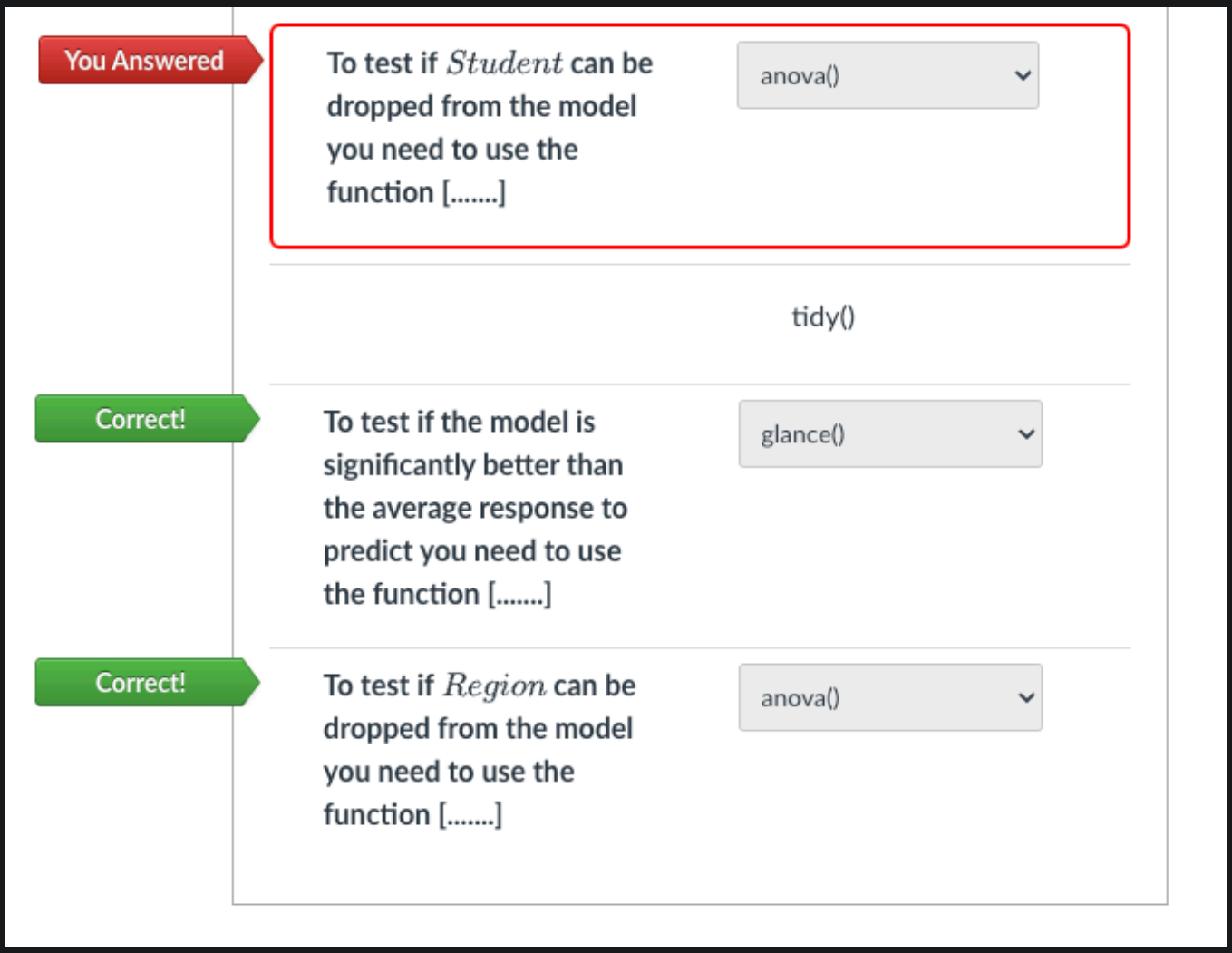
0.67 / 1 pts

The `Credit` data set records values of credit card debts for 400 individuals as well as the following variables that can be used to explain the variation observed in credit card debts:

- `Balance`: (numerical) credit card debt in dollars for each individual
- `Age`: (numerical) age of the individual
- `Cards`: (numerical) number of credit cards the individual has
- `Education`: (numerical) years of education of the individual
- `Income`: (numerical) income of the individual in thousands of dollars
- `Limit`: (numerical) credit limit
- `Rating`: (numerical) credit rating
- `Student`: (factor w/ 2 levels "No","Yes") to indicate whether the individual is a student
- `Region`: (factor w/ 3 levels "East","South", "West") to indicate the location where the individual lives

An additive model is estimated to understand the relation of these variables with `Balance`.

Which functions do you need to complete the statements below? Match



- (note: I'm pretty salty about this question as ANOVA can still be used for the first question but whatever)
- I'm pretty sure the choice between tidy and anova here is whether or not there are multiple variables involved
  - since `region` has 3 levels, it requires 2 dummy variables, while `student` only require 1
- for 2) `glance` will give us the statistics needed to compare against the null model (predicting using the average response)

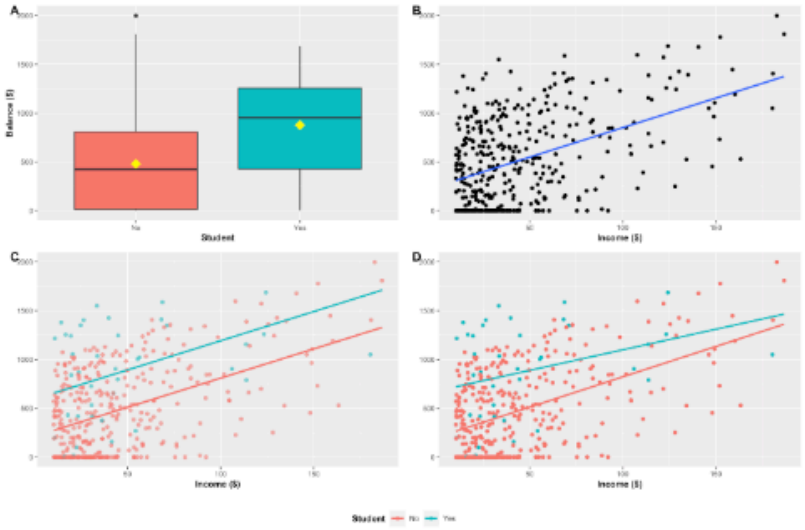
- question 11

Question 11

1.5 / 1.5 pts

Different plots were generated using the data set `Credit`.

Match the plots in the figure to the equation, code or statements given below.



Correct!	Assumption: the difference between the mean balance for students and non-students is the same for any income value.	C
Correct!	Balance = b0 + b1 *Income + e (with subindex omitted for simplicity)	B
Correct!	The model contains 4 different parameters	D

Other Incorrect Match Options:

- A

1. it's basically saying that the slope is the same between non-students and students
    - so we pick C as the slope looks the same there with differing intercept

    - so we pick C as the slope looks the same there with differing intercept
2. this is just an SLR (so only 1 predictor and 1 line) - so we pick B
3. model having 4 parameters mean that there is a different intercept AND different slope between student and non-student
    - the graph in D reflects this pretty well

- question 12

> In an experiment on study habits and the relation to final exam grades, data were collected on the scores on the final examination and the estimated hours spent revising for each of the 35 students on a course.
>
> Some of the students reported that most of the time they spent revising was in the presence of some form of distraction, such as a TV or radio. The remaining students studied most of the time with no such distractions. It is of interest to model how the final test score, Y depends on the amount of hours spent revising (X say). A model of the form
>
> $$Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 XW + \varepsilon$$
>
> was fitted, where ε is a Normally distributed error and the variable W is defined as
>
> $$W = \begin{cases} 0 & \text{if a student who revised mostly without distraction} \\ 1 & \text{if a student who revised mostly with a distraction} \end{cases}$$
>
> The following estimates and standard errors were obtained:
>
> | Parameter | Estimate | Standard error |
> |---|---|---|
> | β_0 | 35.260 | 3.250 |
> | β_1 | 1.520 | 0.346 |
> | β_2 | –0.216 | 0.430 |
> | β_3 | –0.473 | 0.032 |
>
> (a) [2 marks] Based on the model fitted above, for a student who revised mostly without distraction, by how much would you predict their grade to increase for each

> Without distraction, by how much would you predict their grade to increase for each additional hour of studying? Provide approximate 95% confidence interval for your estimate.(hint: $Z_{0.975} = 1.96$)
>
> (b) [2 mark] Based on the model fitted above, for a student who revised mostly with a distraction, by how much would you predict their grade to increase for each additional hour of studying?
>
> (c) [2 mark] Say there is another categorical explanatory variable H that indicate whether a student mostly participated online office hours or mostly participated in-person office hours. How many parameters (βs) are there in the largest linear model with interactions you can fit here considering all these explanatory variables.
>
> (d) [2 marks] After the preliminary analysis, the researcher decided on the following estimated model to predict the final exam score:
>
> $$\hat{Y} = 32.95 + 1.74X - 0.39XW$$
>
> Using this model predict the final exam grade for a student who spent 18 hours revising, mostly in the presence of a distraction.

- a) it's asking about the slope here of students studying without distraction, which is just β_1, so we can construct a CI

$$\begin{aligned} CI &= \hat{b}_1 \pm q_{0.975} se(\hat{b}_1) \\ &= [1.520 - 1.96 \times 0.345, 1.520 + 1.96 \times 0.345] \\ &= [0.8438, 2.1962] \end{aligned}$$

- b) this is asking about the slope of students studying with distraction, which is $\beta_1 + \beta_3$ as there's an interaction term (different slopes)
  - so I will expect their grades to increase by $1.520 - 0.473 = 1.047$ points
- c) this is actually kind of a hard question - requires some STAT 306 knowledge
  - one way to think of it is to think of all possible levels or "states" a student can take on
    - `(distraction, in person)`, `(distraction, online)`, `(no distraction, in person)`, `(no distraction, online)`
  - base line: has 2 terms - just β_0 and β_1

base line: has 2 terms - just β_0 and β_1

- first level: need to add 2 terms to the base line model, so we can have a different slope and different intercept to them
- same for every additional level, so we have $2 + 3(2) = 8$ terms

- full model looks like

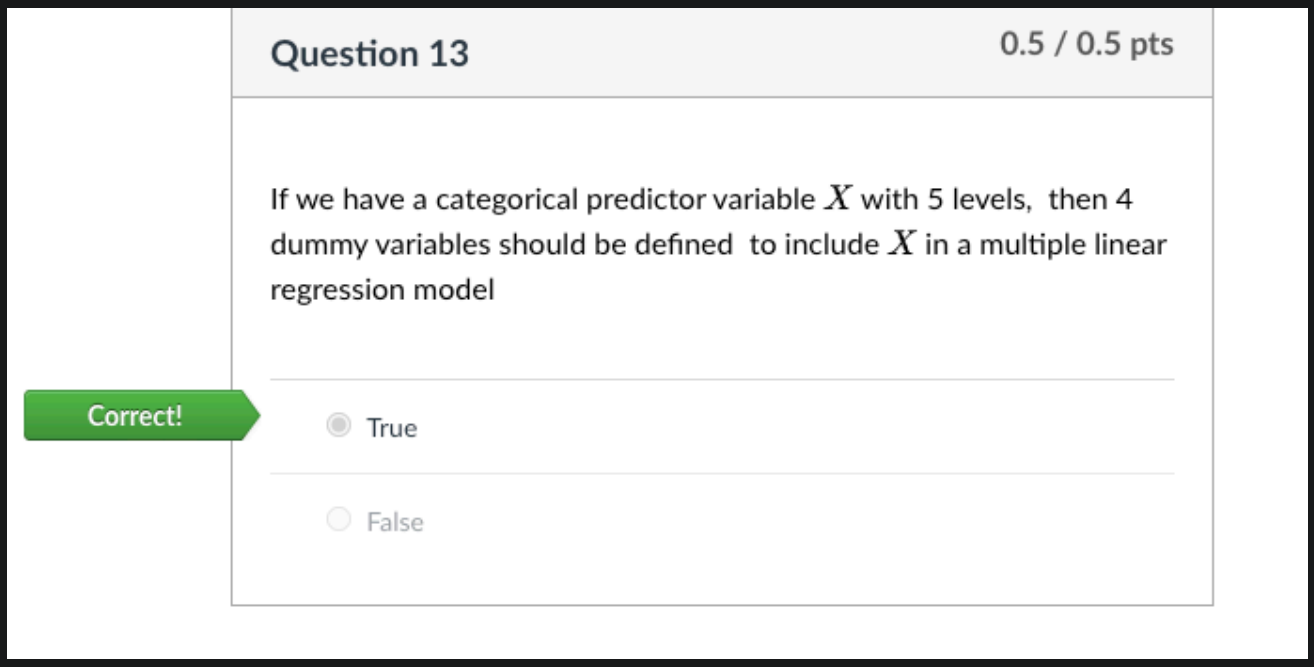
$$Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 XW + \beta_4 H + \beta_5 XH + \beta_6 WH + \beta_7 XWH$$

> 1. $\beta_4 H$: This is the main effect of the new variable H. It represents the difference in the outcome (test score) that is associated with in-person office hours compared to online office hours, ignoring all other factors. It's the added benefit or detriment to the score that is solely attributed to the type of office hours, assuming no other interaction.
> 2. $\beta_5 XH$: This term captures the interaction between the number of hours spent studying (X) and whether those office hours were online or in-person (H). It answers the question: Does the effectiveness of each hour of study differ depending on the type of office hours a student attends? For example, maybe studying for one additional hour is more beneficial for those who attend in-person office hours compared to those who attend online.
> 3. $\beta_6 WH$: This interaction term explores whether the presence of a distraction (W) has a different impact on the test score depending on whether a student attends in-person or online office hours. Perhaps the distraction is less detrimental for students who have the benefit of in-person guidance.
> 4. $\beta_7 XWH$: The three-way interaction term looks at the combined effect of study hours, distraction, and type of office hours. This can indicate a more nuanced dynamic, such as: Does the distraction impact study effectiveness differently for an hour of studying when comparing those who attend in-person vs. online office hours?

- TODO: general formula for this case?
  - I think it's $2 \times \text{number of levels}$

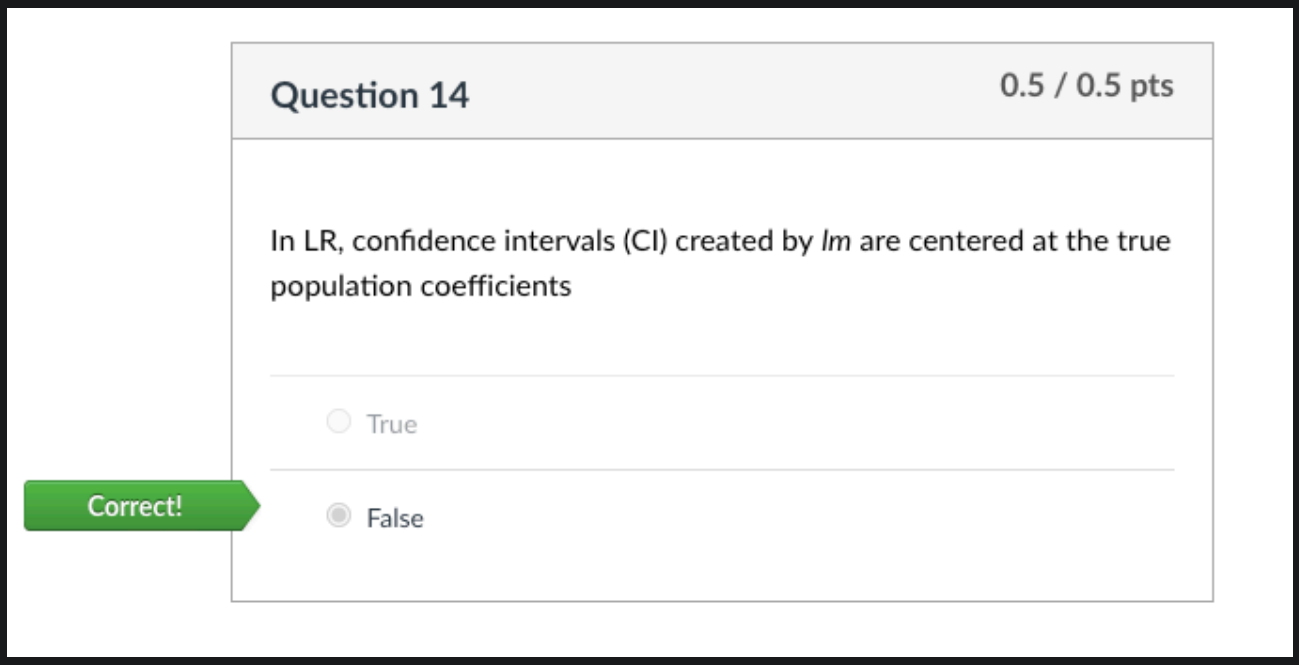
- d) you just have to do the math, $32.95 + 1.74(18) - 0.39(18)(1) = 57.25$

- question 13



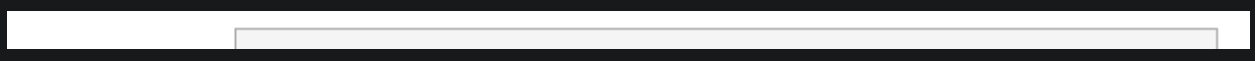
- if there are k levels, we need $k-1$ dummy variables

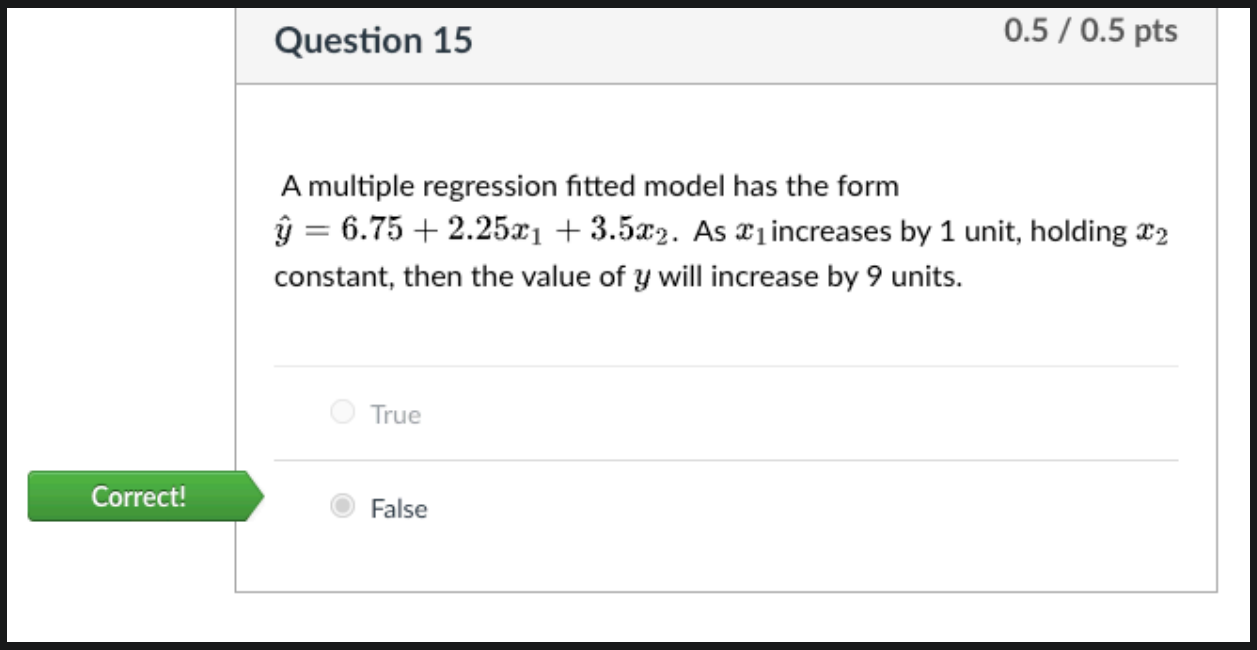
- question 14



- confidence intervals are centered around the sample statistics (i.e the $\hat{b}_1$ we found via our sample)

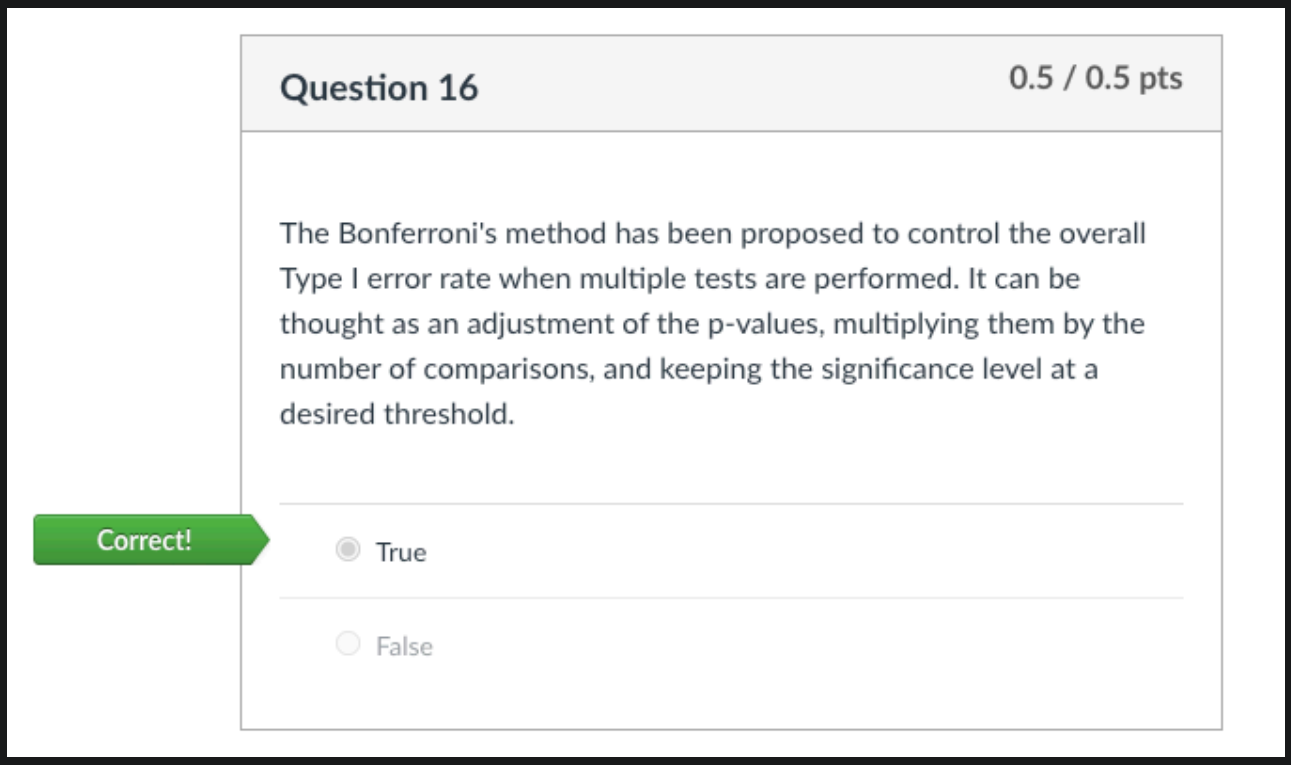
- question 15





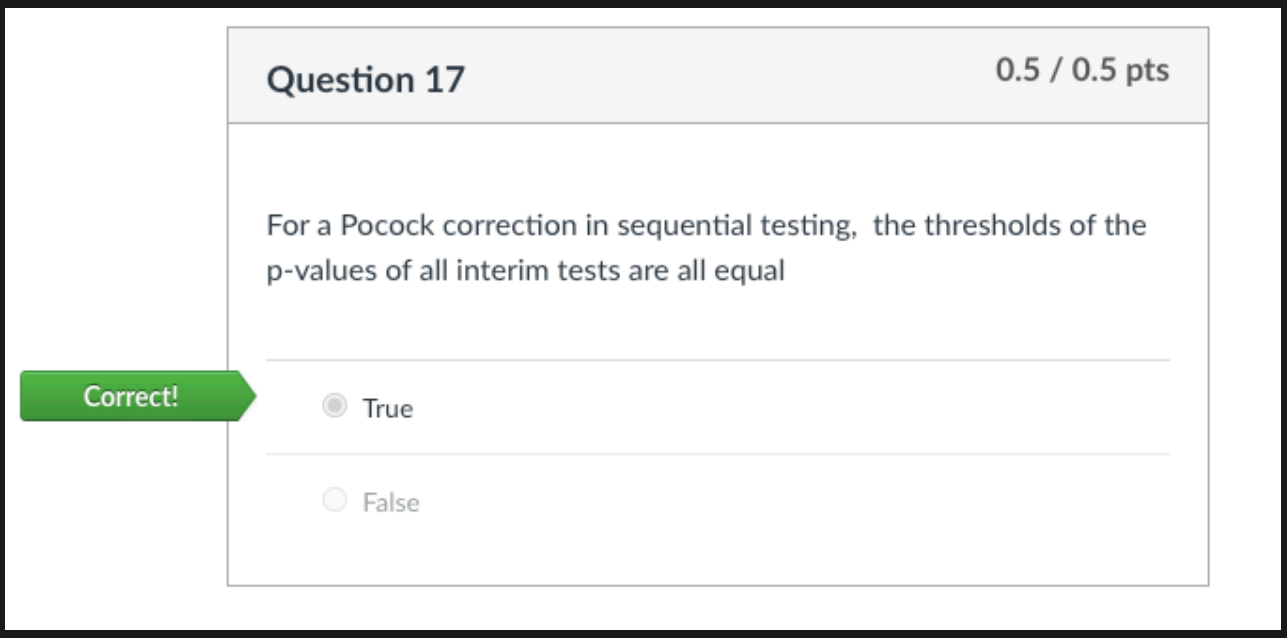
- holding x_2 constant, the value of y will increase by 2.25
  - (just try out 2 random number of x_1 that's 1 apart)

- question 16



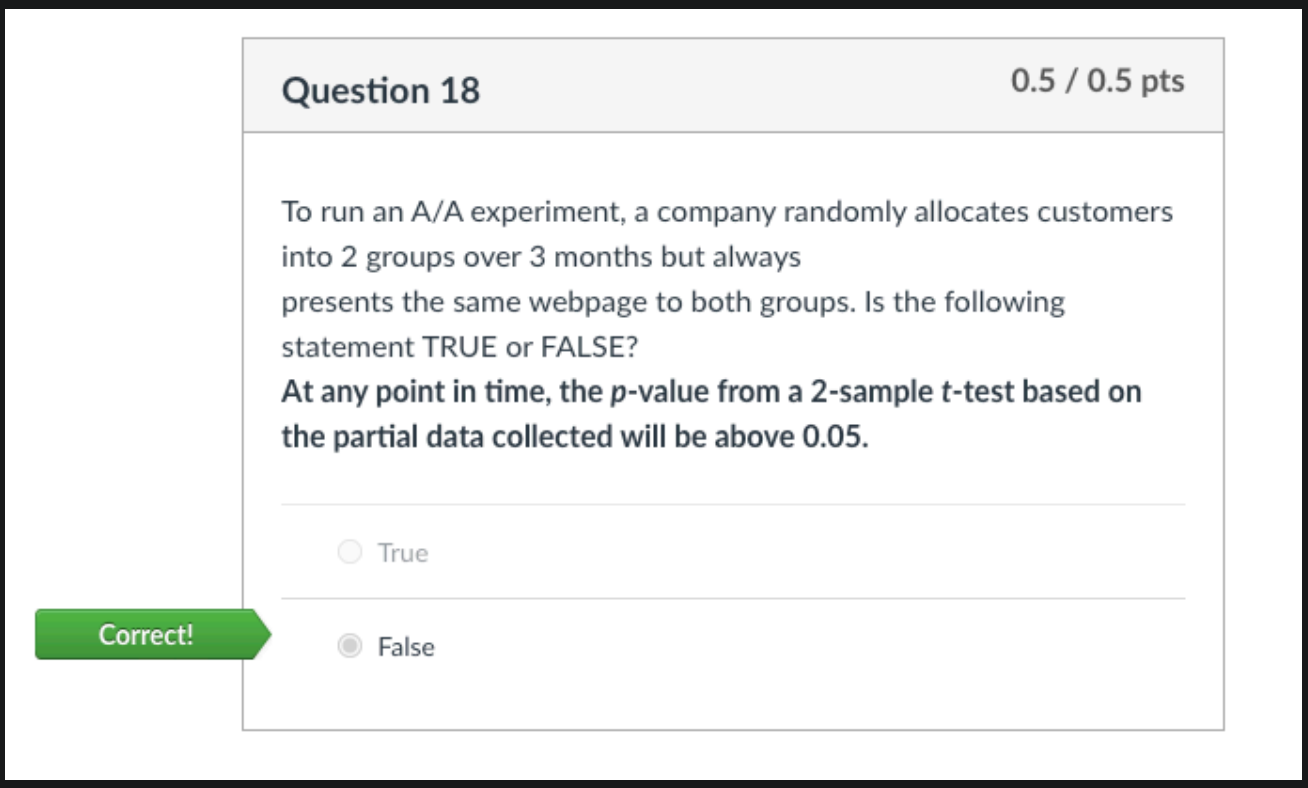
- reducing the significance level is the same making the p-value larger, both serve to reject less often

- question 17



- bit of a trivia question, but Pocock is not constant throughout

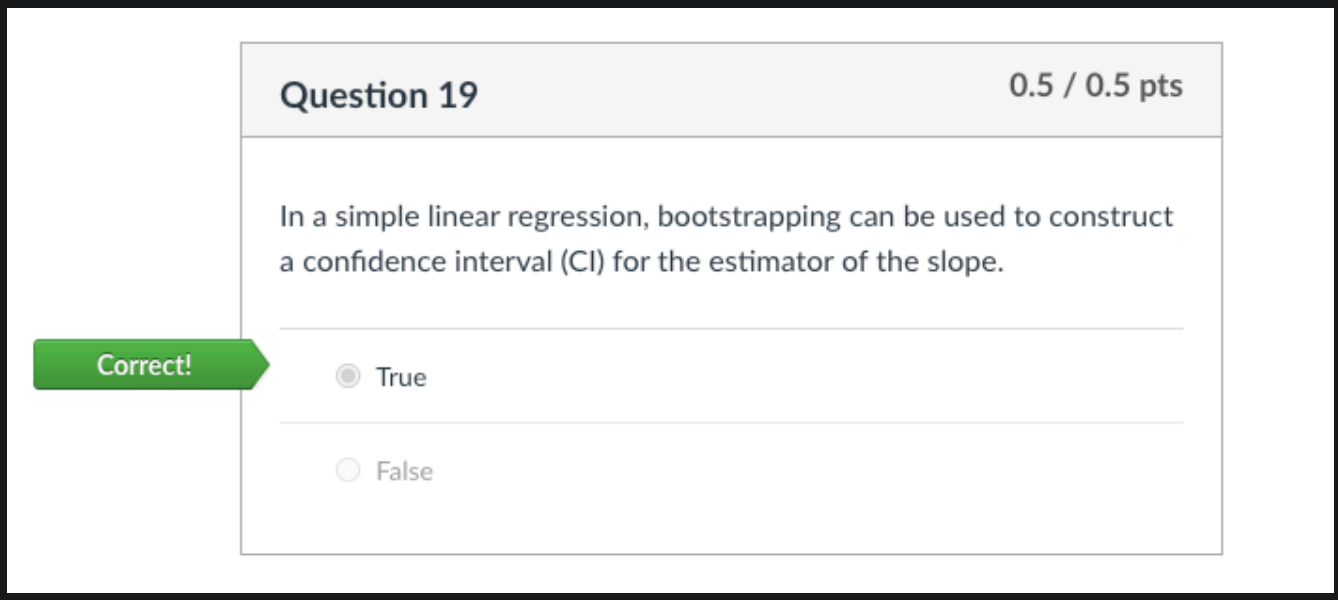
- question 18



- this is basically A/A testing
- at some point in time, there's always a possibility (possibly due to pure random chance) that our p-value might dip below 0.05, which is why we shouldn't just reject as soon as this happens

- question 19

.



- true, this is the alternate approach to `tidy()`