Week 3 Quiz

Back to Week 3

10/10 points earned (100%)

Quiz passed!

1 / 1 points

1.

We modeled the gas mileage of 398 cars built in the 1970's and early 1980's using engine displacement (in cubic inches), year of manufacture in relation to 1970 (e.g. 4 means the car was built in 1974; 12 means built in 1982, etc.), and manufacturing site (domestic to the USA = 0; foreign to the USA = 1). The regression output is provided below. Note that domestic is the reference level for manufacturing site.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	26.86	0.87	30.75	0.0000
displacement	-0.04	0.00	-16.42	0.0000
year	0.72	0.06	12.48	0.0000
site:foreign	2.21	0.54	4.08	0.0001

Which of the following is the degrees of freedom associated with the p-value for site?

- 395

- 3
- 398
- 2
- 394 (selected)

Correct

This question refers to the following learning objective(s): Note that the p-values associated with each predictor are conditional on other variables being included in the model, so they can be used to assess if a given predictor is significant, given that all others are in the model. These p-values are calculated based on a t distribution with $n - k - 1$ degrees of freedom.

$$394 = n - k - 1 = 398 - 3 - 1$$

---

1 / 1 points

2.

We modeled the prices of 93 cars (in $1,000s) using its city MPG (miles per gallon) and its manufacturing site (foreign or domestic). The regression output is provided below. Note that domestic is the reference level for manufacturing site. Data are outdated so the prices may seem low.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	42.56	3.17	13.42	0.0000
city MPG	-1.14	0.14	-8.03	0.0000
site:foreign	5.26	1.59	3.30	0.0014

Which of the following is the correct predicted price (in $1,000s) **of a foreign car that gets 26 MPG**?

- 42.56−(1.14x26)
- 42.56-(1.14x26)+5.26

Correct

This question refers to the following learning objective(s): Define the multiple linear regression model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where there are k predictors (explanatory variables).

- 42.56+(1.14x26)+5.26
- (−1.14x26)+5.26

---

1 / 1 points

3.

We modeled the prices of 93 cars (in $1,000s) using its city MPG (miles per gallon) and its manufacturing site (foreign or domestic). The regression output is provided below. Note that domestic is the reference level for manufacturing site. Data are outdated so the prices may seem low.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	42.56	3.17	13.42	0.0000
city MPG	-1.14	0.14	-8.03	0.0000
site:foreign	5.26	1.59	3.30	0.0014

Which of the following is the **best** interpretation of the slope of manufacturing site?

- Manufacturing site can add up to $42,560 to the price of the car.

- All else held constant, the model predicts that domestic cars cost $5,260 more than foreign cars, on average.
- All else held constant, the model predicts that foreign made cars cost $5,260 more than domestic cars, on average.

Correct

This question refers to the following learning objective(s):

- Interpret the estimate for the intercept (b_0) as the expected value of y when all predictors are equal to 0, on average.

- Interpret the estimate for a slope (say b_1) as "All else held constant, for each unit increase in x_1, we would expect y to be higher/lower on average by b_1."

- As a car goes from being domestic to foreign its price increases by $5,260.
- All else held constant, the model predicts that foreign made cars cost $1,140 less than domestic cars, on average.

---

1 / 1 points

4.

We modeled the prices of 93 cars (in $1,000s) using its city MPG (miles per gallon) and its manufacturing site (foreign or domestic). The regression output is provided below. Note that domestic is the reference level for manufacturing site. Data are outdated so the prices may seem low.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	42.56	3.17	13.42	0.0000
city MPG	-1.14	0.14	-8.03	0.0000
site:foreign	5.26	1.59	3.30	0.0014

Which of the following is **false**?

- If we add another variable to the model, say highway MPG, the p-values associated with city MPG and manufacturing site may change.
- The 95% confidence interval for the slope of city MPG can be calculated as $-1.14 \pm (-8.03 * 0.14)$.

Correct

Use critical T score, not the calculated T score, for the confidence interval.

This question refers to the following learning objective(s):

-The significance of the model as a whole is assessed using an F-test.

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k$

HA: At least one $\beta_i \neq 0$.

- $df = n - k - 1$ degrees of freedom.

- Usually reported at the bottom of the regression output.

- Note that the p-values associated with each predictor are conditional on other variables being included in the model, so they can be used to assess if a given predictor is significant, given that all others are in the model.

- These p-values are calculated based on a t distribution with n − k − 1 degrees of freedom.

- The same degrees of freedom can be used to construct a confidence interval for the slope parameter of each predictor:

$$b_i \pm t^{\star}_{n-k-1} SE_{b_i}$$

- Manufacturing site is a significant predictor of car price, given information on the city MPG of the car.
- City MPG is a significant predictor of car price, given information on the manufacturing site of the car.

---

1 / 1 points

5.
R^2 will never decrease when a predictor is added to a linear model.

- True

Correct
This question refers to the following learning objective(s): Note that R^2 will increase with each explanatory variable added to the model, regardless of whether or not the added variable is a meaningful predictor of the response variable. Therefore we use adjusted R^2, which applies a penalty for the number of predictors included in the model, to better assess the strength of a multiple linear regression model:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

where n is the number of cases and k is the number of predictors.

- Note that R^2_{adj} will only increase if the added variable has a meaningful contribution to the amount of explained variability in y, i.e. if the gains from adding the variable exceeds the penalty.

False

1 / 1 points

6.

Which of the following is **false**?

R^2 is always greater than or equal to adjusted R^2.

In backwards model selection using p-value as the criterion, we start with the full model, and in the first step simultaneously drop all variables that are not significant.

Correct

We drop the variables one at a time.

This question refers to the following learning objective(s): The general idea behind backward-selection is to start with the full model and eliminate one variable at a time until the ideal model is reached.

- p-value method:

(i) Start with the full model.

(ii) Drop the variable with the highest p-value and refit the model.

(iii) Repeat until all remaining variables are significant.

- adjusted R^2 method:

(i) Start with the full model.

(ii) Refit all possible models omitting one variable at a time, and choose the model with the highest adjusted R^2.

(iii) Repeat until maximum possible adjusted R^2 is reached.

- In backwards model selection using p-value as the criterion, we start with the full model, and drop the variable with highest p-value, one at a time, until all the variables in the model are significant given the others.

- One of the consequences of collinearity in multiple regression is biased estimates on the slope coefficients.

---

1 / 1 points

7.

As part of the Second International Mathematics Study on 8th graders from randomly sam- pled classrooms in the US who completed mathematics achievement tests at the beginning and at the end of the academic year. Students also answered questioners regarding their attitudes toward mathematics. The linear model output predicts the gain score in this test (post-test - pretest score) using the following explanatory variables:

• pretest: score on the exam taken at the beginning of the semester

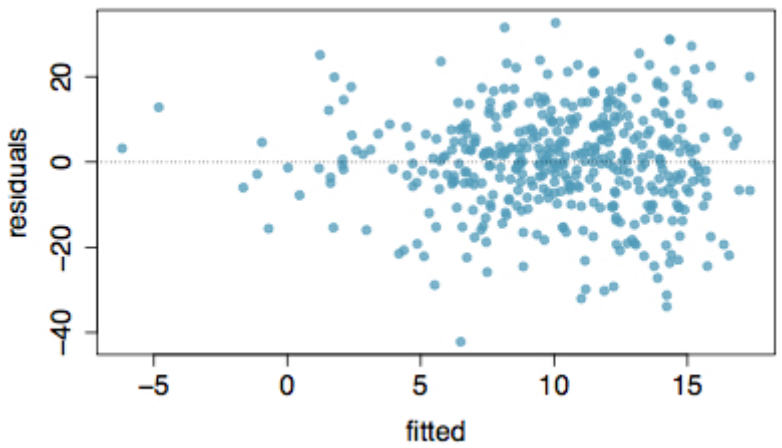
• gender: male or female

• more_ed: expected number of years for continued education (up to 2 years, 2 to 5 years, 5 to 6 years, 8 or more years)

• useful: Math is useful in everyday life (strongly disagree, disagree, undecided, agree, strongly agree)

• ethnic: ethnicity of student (African American, Anglo, Other)

The following is the residuals plot for this model. Which of the following conditions can this plot be used to check?



- Nearly normal residuals
- Non-collinear explanatory variables.

Constant variability of residuals

Correct

This question refers to the following learning objective(s): List the conditions for multiple linear regression as

(1) linear relationship between each (numerical) explanatory variable and the response - checked using scatterplots of y vs. each x, and residuals plots of residuals vs. each x

(2) nearly normal residuals with mean 0 - checked using a normal probability plot and histogram of residuals

(3) constant variability of residuals - checked using residuals plots of residuals vs. $\hat{y}$, and residuals vs. each x

(4) independence of residuals (and hence observations) - checked using a scatterplot of residuals vs. order of data collection (will reveal non-independence if data have time series structure)

Independent residuals

---

1 / 1 points

8.

Which of the following is the **best** definition of a parsimonious model?

The model with the least amount of collinearity between predictors.

The simplest model with the highest predictive power.

Correct

This question refers to the following learning objective(s): Note that we usually prefer simple (parsimonious) models over more complicated ones.

- The model with the most number of predictors.
- The model with the least number of predictors.

---

1 / 1 points

9.

A high correlation between two explanatory variables such that the two variables contribute redundant information to the model is known as

- multiple correlation
- homoscedasticity
- collinearity

Correct

This question refers to the following learning objective(s): Define collinearity as a high correlation between two independent variables such that the two variables con- tribute redundant information to the model – which is something we want to avoid in multiple linear regression.

- adjusted R^2
- heterogeneity
- multiple interaction

- homogeneity
- heteroscedasticity

---

1 / 1 points

10.
Suppose you have performed forward selection using adjusted R^2 as the criterion and have chosen a model with 6 predictors. Based on your studies of model selection, which of the following is most likely to be **true**?

- All 6 predictors will be significant in the model.
- The model you've arrived at is the most parsimonious model.
- Your final model has a higher adjusted R^2 than any of the smaller models you tried.

Correct
This question refers to the following learning objective(s): The general idea behind forward-selection is to start with only one variable and adding one variable at a time until the ideal model is reached.

- p-value method:

(i) Try all possible simple linear regression models predicting y using one explanatory variable at a time. Choose the model where the explanatory variable of choice has the lowest p-value.

(ii) Try all possible models adding one more explanatory variable at a time, and choose the model where the added explanatory variable has the lowest p-value.

(iii) Repeat until all added variables are significant.

- adjusted R^2 method:

(i) Try all possible simple linear regression models predicting y using one explanatory variable at a time. Choose the model with the highest adjusted R^2.

(ii) Try all possible models adding one more explanatory variable at a time, and choose the model with the highest adjusted R^2.

(iii) Repeat until maximum possible adjusted R^2 is reached.

- If any of your 6 predictors is not significant in the model, given the other predictors, then your slope coefficients will be biased.