Week 3 Quiz

Quiz, 10 questions

8/10 points (80%)

Congratulations! You passed!

Next Item

1 / 1 point

1.

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 Data are outdated so the prices may seem low.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 degrees of freedom associated with the p-value for city MPG?

- 90

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.

$$90 = n - k - 1 = 93 - 2 - 1$$

- 91
- 92
- 1

0 / 1 point

Week 3 Quiz

Quiz, 10 questions

8/10 points (80%)

2.

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 correct predicted gas mileage (in miles per gallon) for a **domestic car with an engine displacement of 350 cubic inches built in 1975**?

- $26.86 - 0.04 \times 350 + 0.72 \times 1975 + 2.21$

This should not be selected

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).

Also note: The year should be measured as years since 1970, and domestic is the reference level of site.

- $-0.04 \times 350 + 0.72 \times 1975$
- $-0.04 \times 350 + 0.72 \times 5 + 2.21$
- $26.86 - 0.04 \times 350 + 0.72 \times 5$

✖ 0 / 1 point

3.

Week 3 Quiz

Quiz, 10 questions

8/10 points (80%)

The data in this question come from the Second International Mathematics Study on 8th graders from randomly sampled classrooms in the US who completed mathematics achievement tests at the beginning and at the end of the academic year. Students also answered questions regarding their attitudes toward mathematics. The linear model output below is for predicting the gain score in this test (posttest - 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)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.9529	6.8446	-0.43	0.666
pretest	-0.1629	0.0372	-4.38	1.5e-05
gender:male	0.5586	1.1910	0.47	0.639
more_ed:2 to 5 years	0.0731	6.2351	0.01	0.991
more_ed:5 to 6 years	5.8558	6.1974	0.94	0.345
more_ed:8 or more years	6.6138	6.2726	1.05	0.292
useful:disagree	7.2809	4.3065	1.69	0.092
useful:undecided	7.7716	3.8461	2.02	0.044
useful:agree	8.5578	3.6693	2.33	0.020
useful:strongly agree	9.2262	3.7946	2.43	0.015
ethnic:Anglo	6.4974	2.1779	2.98	0.003
ethnic:Other	5.3995	2.8049	1.92	0.055

What does the **intercept** in this model represent?

- An African American male student who scored 0 on the pretest, expects to continue their education for up to 2 years, who strongly disagrees with the statement on usefulness of math.

This should not be selected

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."

- Any student who scored 0 on the pretest.

A student who scored 0 on the pretest and did not answer the other questions on expected years of education, usefulness of math, and ethnicity.

Week 3 Quiz

Quiz, 10 questions

8/10 points (80%)

An African American female student who scored 0 on the pretest, expects to continue their education for up to 2 years, who strongly disagrees with the statement on usefulness of math.

---

1 / 1 point

4.

A random sample of 200 women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, were tested for diabetes according to World Health Organi- zation criteria. The model below is used for predicting their plasma glucose concentration based on their diastolic blood pressure (bp, in mmHg), age (age, in years), and whether or not they are diabetic (type, Yes and No). Which of the following is **false**?

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	76.00	12.24	6.21	0.00
bp	0.35	0.18	1.94	0.05
age	0.43	0.20	2.12	0.04
type:Yes	26.57	4.37	6.08	0.00

Residual standard error: 27 on 196 degrees of freedom
Multiple R-squared: 0.28, Adjusted R-squared: 0.27
F-statistic: 25 on 3 and 196 DF, p-value: 1e-13

The model as a whole is significant, even though one of the variables (blood pressure) may not be.

The predicted difference in blood glucose levels of two 25 year old females who don't have diabetes one of whom has a blood pressure of 70mmHg and the other 75 mmHG is $0.35 * 5 = 1.75$.

The model predicts that women without diabetes have blood glucose levels that are on average 26.57 higher than those who have diabetes, given that they are similar in terms of their blood pressure and age.

Correct

False, "no" is the reference level, therefore those with diabetes have higher blood glucose levels by 26.57.

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

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

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

Week 3 Quiz

Quiz, 10 questions

8/10 points (80%)

H_A: 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:

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

The model explains 28% of variability in blood glucose levels of these women.

---

1 / 1 point

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

Week 3 Quiz 1 point

Quiz, 10 questions

8/10 points (80%)

6.

Consider the following output from a multiple linear regression model with 10 predictors.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	2.58	0.44	5.48	0.0000
to_multiple	0.24	0.06	0.29	0.0058
winner	-0.50	0.20	-6.80	0.0359
format	-0.15	0.91	-1.44	0.0032
re_subj	-3.75	0.50	-3.59	0.0000
exclaim_subj	5.87	0.13	10.54	0.5503
cc	-0.67	0.17	-0.10	0.0185
attach	-0.78	0.23	-5.36	0.0000
dollar	3.69	0.32	4.65	0.7154
inherit	-1.32	0.71	-0.75	0.0044
password	0.30	0.89	2.94	0.6342

If you were doing backwards selection on this model using p-value as the criterion, which of the following would be an acceptable next step?

- Remove the variables “password” and “dollar” because their high p-values indicate collinearity with other variables.
- Remove the variable “dollar” because it has the highest p-value.

Correct

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.

Remove any **one** of the variables with a high p-value which as long as removing the variable does not cause the adjusted R^2 to decrease in the re-fitted model.

Remove **one** of the variables " re subj" or " attach" because they both have the lowest p-values.

---

1 / 1 point

7.

Which of the following is **false** about conditions for multiple linear regression?

Explanatory variables should have strong relationships with each other.

Correct

This would result in collinearity, which is something we want to avoid in multiple linear regression.

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)

Residuals should be normally distributed around 0.

Residuals should have constant variance.

Explanatory variables should have linear relationship with the response variable.

---

1 / 1 point

8.

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

The model with the least number of predictors.

The model with the least amount of collinearity between predictors.

Week 3 Quiz

Quiz, 10 questions

8/10 points (80%)

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.

---

1 / 1 point

9.
A high correlation between two explanatory variables such that the two variables contribute redundant information to the model is known as

heteroscedasticity

homogeneity

multiple interaction

heterogeneity

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

multiple correlation

homoscedasticity

---

1 / 1 point

10.

Week 3 Quiz

Quiz, 10 questions

A model selection method where we start with an empty model and add variables one at a time until no other important variables are found is called

8/10 points (80%)

- [x] forward selection

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.

- [ ] bootstrapping
- [ ] multiple testing
- [ ] backwards elimination
- [ ] ANOVA