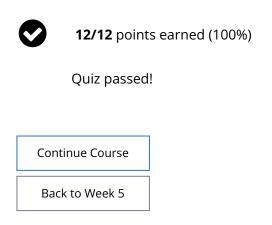
# Model Selection and Diagnostics Quiz





1/1 points

1.

Suppose you are regressing log(*price*) on log(*area*), log(*Lot.Area*), *Bedroom.AbvGr*, *Overall.Qual*, and *Land.Slope*. Which of the following variables are included with stepwise variable selection using AIC but not BIC?

0	log( <i>area</i> )
0	log( <i>Lot.Area</i> )
0	Bedroom.AbvGr
0	Overall.Qual
0	Land.Slope

# Correct

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

Use principled statistical methods to select a single parsimonious model.

2.

When regressing  $\log(price)$  on Bedroom.AbvGr, the coefficient for Bedroom.AbvGr is strongly positive. However, once  $\log(area)$  is added to the model, the coefficient for Bedroom.AbvGr becomes strongly negative. Which of the following best explains this phenomenon?

- The original model was misspecified, biasing our coefficient estimate for Bedroom.AbvGr
- Bedrooms take up proportionally less space in larger houses, which increases property valuation.
- Larger houses on average have more bedrooms and sell for higher prices. However, holding constant the size of a house, the number of bedrooms decreases property valuation.

# Correct

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

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  $b_1$ ."

Since the number of bedrooms is a statistically insignificant predictor of housing price, it is unsurprising that the coefficient changes depending on which variables are included.



1/1 points

3.

Run a simple linear model for log(price), with log(area) as the independent variable. Which of the following neighborhoods has the highest average residuals?

O OldTown

StoneBr

GrnHill

Correct

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

•	Identify the assumptions of linear regression and assess when a model
	may need to be improved.

• Examine the residuals of a linear model.





1/1 points

4.

We are interested in determining how well the model fits the data for each neighborhood. The model from Question 3 does the worst at predicting prices in which of the following neighborhoods?



GrnHill

#### Correct

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

Examine the residuals of a linear model.

BlueSte

StoneBr

MeadowV



1/1 points

5.

Suppose you want to model log(price) using only the variables in the dataset that pertain to quality: *Overall.Qual, Basement.Qual,* and *Garage.Qual.* How many observations must be discarded in order to estimate this model?

O 0

**)** 46

#### Correct

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

Identify the assumptions of linear regression and assess when a model may need to be improved.



924



1/1 points

6.

*NA* values for *Basement.Qual* and *Garage.Qual* correspond to houses that do not have a basement or a garage respectively. Which of the following is the best way to deal with these *NA* values when fitting the linear model with these variables?

- Drop all observations with *NA* values for *Basement.Qual* or *Garage.Qual* since the model cannot be estimated otherwise.
- Recode all *NA* values as the category TA since we must assume these basements or garages are typical in the absence of all other information.
- Recode all *NA* values as a separate category, since houses without basements or garages are fundamentally different than houses with both basements and garages.

#### Correct

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

Check the assumptions of a linear model.



1/1 points

7.

Run a simple linear model with log(price) regressed on *Overall.Cond* and *Overall.Qual*. Which of the following subclasses of dwellings (*MS.SubClass*) has the highest median predicted prices?

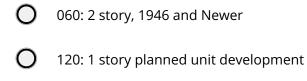


075: 2-1/2 story houses

#### Correct

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

Predict the value of the response variable for a given value of the explanatory variable,  $x^*$ , by plugging in  $x^*$  in the linear model





1/1 points

090: Duplexes

8.

Using the model from Question 7, which observation has the highest leverage or potential influence on the regression model? Hint: use hatvalues, hat or *Im.influence*.



### Correct

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

Identify outliers and high leverage points in a linear model.

O	640
0	832

9. Which of the following corresponds to a correct interpretation of the coefficient kof Bedroom.AbvGr, where log(price) is the dependent variable? Holding constant all other variables in the dataset, on average, an additional bedroom will increase housing price by k percent. Holding constant all other variables in the model, on average, an additional bedroom will increase housing price by k percent. Correct This question refers to the following learning objective(s): 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  $b_1$ ." Holding constant all other variables in the dataset, on average, an additional bedroom will increase housing price by k dollars. Holding constant all other variables in the model, on average, an additional bedroom will increase housing price by k dollars. 1/1 points 10. Which of the following sale condition categories shows significant differences from the normal selling condition? Family Abnorm **Partial** 

Correct

Abnorm and Partial

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

- Be cautious about using a categorical explanatory variable when one of the levels has very few observations, as these may act as influential points.
- List the conditions for multiple linear regression.



1/1 points

# 11.

Subset ames\_train to only include houses sold under normal sale conditions. What percent of the original observations remain?

0

81.2%



83.4%



#### Correct

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

Use R commands to effectively manipulate data.

 $\bigcirc$ 

87.7%



91.8%



1/1 points

# 12.

Now re-run the simple model from question 3 on the subsetted data. True or False: Modeling only the normal sales results in a better model fit than modeling all sales (in terms of  $\mathbb{R}^2$ ).



True, restricting the model to only include observations with normal sale conditions increases the  $\mathbb{R}^2$  from 0.547 to 0.575.



#### Correct

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

- Be cautious about using a categorical explanatory variable when one of the levels has very few observations, as these may act as influential points.
- Define  $\mathbb{R}^2$  as the percentage of the variability in the response variable explained by the the explanatory variable.

$\bigcirc$	True, restricting the model to only include observations with normal sale
	conditions increases the $\mathbb{R}^2$ from 0.575 to 0.603.

- False, restricting the model to only include observations with normal sale conditions decreases the  $R^2$  from 0.575 to 0.547.
- False, restricting the model to only include observations with normal sale conditions decreases the  $\mathbb{R}^2$  from 0.603 to 0.575.

