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? **1 / 1 point**

- log(*area*)
- log(*Lot.Area*)
- *Bedroom.AbvGr*
- *Overall.Qual*
- *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? **1 / 1 point**

- 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.
- 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.

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."

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? **1 / 1 point**

- OldTown
- StoneBr
- GrnHill
- IDOTRR

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.

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? **1 / 1 point**

- GrnHill
- BlueSte
- StoneBr
- MeadowV

Correct

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

Examine the residuals of a linear model.

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? **1 / 1 point**

- ○ 0
- ○ 46
- ◉ 64
- ○ 924

✓ **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.

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? **1 / 1 point**

- ○ 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.

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? **1 / 1 point**

- (●) 075: 2-1/2 story houses
- ( ) 060: 2 story, 1946 and Newer
- ( ) 120: 1 story planned unit development
- ( ) 090: Duplexes

✓ **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, $x^\star$, by plugging in $x^\star$ in the linear model

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 *lm.influence*. **1 / 1 point**

- ( ) 125
- (●) 268
- ( ) 640
- ( ) 832

✓ **Correct**

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

Identify outliers and high leverage points in a linear model.

9. Which of the following corresponds to a correct interpretation of the coefficient k of *Bedroom.AbvGr*, where $\log(price)$ is the dependent variable? **1 / 1 point**

- 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.
- 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.

✓ **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."

10. Which of the following sale condition categories shows significant differences from the normal selling condition? **1 / 1 point**

- Family
- Abnorm
- Partial
- Abnorm and Partial

✓ **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.
- List the conditions for multiple linear regression.

11. Subset ames_train to only include houses sold under normal sale conditions. What percent of the original observations remain? **1 / 1 point**

- 81.2%
- 83.4%
- 87.7%
- 91.8%

Correct

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

Use R commands to effectively manipulate data.

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 R^2). **1 / 1 point**

- True, restricting the model to only include observations with normal sale conditions increases the R^2 from 0.547 to 0.575.
- True, restricting the model to only include observations with normal sale conditions increases the 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 R^2 from 0.603 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 R^2 as the percentage of the variability in the response variable explained by the the explanatory variable.