LO 1. The significance of the model as a whole is assessed using an F-test.

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
- 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.

LO 2. 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.

- $H_0 : \beta_1 = 0$, given all other variables are included in the model
- $H_A : \beta_1 \neq 0$, given all other variables are included 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:

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

LO 3. Stepwise model selection (backward or forward) can be done based on p-values (drop variables that are not significant) or based on adjusted R^2 (choose the model with higher adjusted R^2).

LO 4. 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:

1. Start with the full model.
2. Drop the variable with the highest p-value and refit the model.
3. Repeat until all remaining variables are significant.

- adjusted R^2 method:

1. Start with the full model.

2. Refit all possible models omitting one variable at a time, and choose the model with the highest adjusted R2.
3. Repeat until maximum possible adjusted R^2 is reached.

LO 5. 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:

(1) 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.

(2) 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.

(3) Repeat until all added variables are significant.

- adjusted R^2 method:

1. 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.
2. Try all possible models adding one more explanatory variable at a time, and choose the model with the highest adjusted R^2.
3. Repeat until maximum possible adjusted R^2 is reached.

LO 6. Adjusted R^2 method is more computationally intensive, but it is more reliable, since it doesn't depend on an arbitrary significance level.

LO 7. 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)

LO 8. Note that no model is perfect, but even imperfect models can be useful.

Mark as completed