Statistics with R Specialization
Course 3: Linear Regression and Modeling

Linear Regression

LO 1. Define the explanatory variable as the independent variable (predictor), and the response variable as the dependent variable (predicted).

LO 2. Plot the explanatory variable (x) on the x-axis and the response variable (y) on the y-axis, and fit a linear regression model

$$y = \beta_0 + \beta_1 x,$$

where β_0 is the intercept, and β_1 is the slope.

- Note that the point estimates (estimated from observed data) for β_0 and β_1 are b_0 and b_1, respectively.

LO 3. When describing the association between two numerical variables, evaluate

- direction: positive ($x \uparrow, y \uparrow$), negative ($x \downarrow, y \uparrow$)
- form: linear or not
- strength: determined by the scatter around the underlying relationship

LO 4. Define correlation as the *linear* association between two numerical variables.

- Note that a relationship that is nonlinear is simply called an association.

LO 5. Note that correlation coefficient (R, also called Pearson's R) has the following properties:

- the magnitude (absolute value) of the correlation coefficient measures the strength of the linear association between two numerical variables
- the sign of the correlation coefficient indicates the direction of association
- the correlation coefficient is always between -1 and 1, -1 indicating perfect negative linear association, +1 indicating perfect positive linear association, and 0 indicating no linear relationship
- the correlation coefficient is unitless
- since the correlation coefficient is unitless, it is not affected by changes in the center or scale of either variable (such as unit conversions)
- the correlation of X with Y is the same as of Y with X
- the correlation coefficient is sensitive to outliers

LO 6. Recall that correlation does not imply causation.

LO 7. Define residual (e) as the difference between the observed (y) and predicted ($\hat{y}$) values of the response variable.

$$e_i = y_i - \hat{y}_i$$

LO 8. Define the least squares line as the line that minimizes the sum of the squared residuals, and list conditions necessary for fitting such line:

1. linearity
2. nearly normal residuals
3. constant variability

LO 9. Define an indicator variable as a binary explanatory variable (with two levels).

LO 10. Calculate the estimate for the slope (b_1) as

$$b_1 = R\frac{s_y}{s_x},$$

where R is the correlation coefficient, s_y is the standard deviation of the response variable, and s_x is the standard deviation of the explanatory variable.

LO 11. Interpret the slope as

- when x is numerical: "For each unit increase in x, we would expect y to be lower/higher on average by $|b_1|$ units"
- when x is categorical: "The value of the response variable is predicted to be $|b_1|$ units higher/lower between the baseline level and the other level of the explanatory variable."
- Note that whether the response variable increases or decreases is determined by the sign of b_1.

LO 12. Note that the least squares line always passes through the average of the response and explanatory variables ($\bar{x}, \bar{y}$).

LO 13. Use the above property to calculate the estimate for the intercept (b_0) as

$$b_0 = \bar{y} - b_1 \bar{x},$$

where b_1 is the slope, $\bar{y}$ is the average of the response variable, and $\bar{x}$ is the average of explanatory variable.

LO 14. Interpret the intercept as

- "When $x = 0$, we would expect y to equal, on average, b_0." when x is numerical.
- "The expected average value of the response variable for the reference level of the explanatory variable is b_0." when x is categorical.

More about Linear Regression

LO 1. Define a leverage point as a point that lies away from the center of the data in the horizontal direction.

LO 2. Define an influential point as a point that influences (changes) the slope of the regression line.

- This is usually a leverage point that is away from the trajectory of the rest of the data.

LO 3. Do not remove outliers from an analysis without good reason.

LO 4. Be cautious about using a categorical explanatory variable when one of the levels has very few observations, as these may act as influential points.

LO 5. Determine whether an explanatory variable is a significant predictor for the response variable using the t-test and the associated p-value in the regression output.

LO 6. Set the null hypothesis testing for the significance of the predictor as $H_0 : \beta_1 = 0$, and recognize that the standard software output yields the p-value for the two-sided alternative hypothesis.

- Note that $\beta_1 = 0$ means the regression line is horizontal, hence suggesting that there is no relationship between the explanatory and response variables.

LO 7. Calculate the T score for the hypothesis test as

$$T_{df} = \frac{b_1 - \text{null value}}{SE_{b_1}}$$

with $df = n - 2$.

- Note that the T score has $n - 2$ degrees of freedom since we lose one degree of freedom for each parameter we estimate, and in this case we estimate the intercept and the slope.

LO 8. Note that a hypothesis test for the intercept is often irrelevant since it's usually out of the range of the data, and hence it is usually an extrapolation.

LO 9. Calculate a confidence interval for the slope as

$$b_1 \pm t^\star_{df} SE_{b_1},$$

where $df = n - 2$ and $t^\star_{df}$ is the critical score associated with the given confidence level at the desired degrees of freedom.

- Note that the standard error of the slope estimate SE_{b_1} can be found on the regression output.

Regression with multiple predictos

LO 1. 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).

LO 2. Interpret the estimate for the intercept (b_0) as the expected value of y when all predictors are equal to 0, on average.

LO 3. 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."

LO 4. Define collinearity as a high correlation between two independent variables such that the two variables contribute redundant information to the model -- which is something we want to avoid in multiple linear regression.

LO 5. 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.

LO 6. Define model selection as identifying the best model for predicting a given response variable.

LO 7. Note that we usually prefer simpler (parsimonious) models over more complicated ones.

LO 8. Define the full model as the model with all explanatory variables included as predictors.

Inference for multiple regression and model selection

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.