HR Case: Written Solutions

Module 1:

1. What is the interpretation of the regression coefficient of satisfaction_level?
   Each regression tells you how much y increases when you increase x by one unit, holding other x variables fixed. For example, the coefficient on 'satisfaction_level' implies that the probability of leaving goes down by 64% when satisfaction increases from 0 to 100%.

2. What is the interpretation of the statistical significance of satisfaction_level?
   Statistical significance tells you how plausible it is that a coefficient is 0. Since satisfaction_level is highly significant, we can reject the notion that the coefficient is 0. On the other hand, the coefficient on salesIT is not significant, which means that a coefficient of 0 is plausible.

3. Do the variables that are not significant have an effect on 'left'?
   They likely do. Statistics will rarely estimate a coefficient of exactly 0. A lack of significance does not mean a variable has no effect.

4. Of the different job types (HR, IT), what are the top 3 most likely to leave?
   Looking at the coefficient estimates of the 'sales' variable, we can see that the highest values are hr, technical, and support. The next highest is accounting, which is the default value.

Module 2:

1. How do you model an interaction effect in R?
   You use the * symbol to interact two or more variables together.

2. What is the interpretation of the coefficients on the interaction effect?
   Interaction effects allow one variable to modify the relationship a different variable has on y. You could consider looking at the effect when one variable is 0, and seeing how the relationship changes when that variable increases. In this section, the first interaction allowed salary to change the relationship between satisfaction and left. When salary was high, the relationship was weak at -.237, but when salary was low the relationship was stronger at -.237 - .488 = -.725

Module 3:

1. I presented the best model using the BIC. Would you expect the best model according to the AIC to be larger or smaller?
   In general, the model selected by the AIC will be larger because the penalty term is larger

2. Can you interpret the resulting coefficients from this analysis as causal?
   No, because we have not considered omitted variable bias. This analysis even excluded variables were in the dataset to make the analysis more interpretable.

Module 4:

1. **What can we learn from the 7th interaction term? Keep in mind that \texttt{plotmo} deletes vowels to make the variable names more readable.**
   From the seventh interaction, those with no projects who have spent a long time at the company do not leave

Module 5:

1. Why do you think we are observing a consistent positive relationship between evaluations and leaving? What omitted variables might cause that?
   Potential omitted variables could include overall ability, ambition, and education. All three would be positively correlated with evaluations, and positively correlated with left, leading to a positive bias.
2. How do you decide what variables to include in a causal model? How does this differ from how you select variables in a predictive model?
   You select variables to be in a causal model based on their potential to cause bias. If a variable might be correlated with both X and Y, then it must be controlled for to get valid causal inferences. A predictive model will remove variables that reduce the quality of predictions.
3. This analysis did not control for the overall level of benefits that an employee received. In what way would you expect this omitted variable to bias the coefficient on salary? **Benefits are likely positively correlated with high salary, and negatively correlated with left, leading to a negative bias of the coefficient for high salary.**
4. Can the firm determine whether salary increases lead to an improvement in employee retainment from this dataset alone?
   No, because we lack data on overall ability, ambition, and education, all of which could cause a serious bias.