Test A Solutions - Short-Answer

Solutions for Output A (NBA data)

11. Use the regression output from `lm1` to conduct a test of $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Based on the p-value from the previous question, make a conclusion in context.

Since the p-value is approximately 0, we have strong evidence of a linear relationship between minutes per game (Var6) and total points (Var1).

12. Using `lm1`, a 90% prediction interval is given in the output. Interpret this interval in context.

We are 90% confident that a particular player with 34 minutes per game will have total points between 800.2 and 2009.5. (Or: For all player with 34 minutes per game, 90% of them will score between 800.2 and 2009.5 points.)

14. We wish to predict `Var1` using `Var6`. Previous analysis indicates that $Var6^2$ might be useful in the model, so we fit this model, called `lm2`. Based on this output, was including $Var6^2$ a good idea? $Var6^2$ is significant, and leads to a 4% increase in adjusted R^2. So inclusion is reasonable.

15. We wish to predict *Var1* using *Var6*, *Var4*, and *Var3*; this is called `lm3`. Interpret the coefficient on *Var3* in context.

Holding minutes per game (Var6) and rebounds (Var4) constant, a one-game increase in games started (Var3) is associated with an increase in points (Var1) of 2.674.

30. Based only on the output from lm3, would you remove any of the variables from the model? Briefly justify your answer.

Probably not, since they are all at least moderately significant. However, some might argue tht Var3 is not worth it.

31. Using the output below model `lm3`, report and interpret a confidence interval for the coefficient on *Var4*.

Holding minutes per game (Var6) and games started (Var3) constant, we are 95% confident that one additional rebound (Var4) will increase a player's total points (Var1) by 0.457 to 0.950.

33. Given the other variables in model `lm3`, is *Var4* a useful predictor of *Var1*? Justify your answer.

With Var6 and Var3 in the model, the p-value on Var4 is 6.52*10^-8 So we have very strong evidence that rebounds is a statistically significant predictor of total points.

35. Are you concerned about multicollinearity in the `lm3` model? Justify your answer.

No, because all of the VIF values for `lm3` are small (below 5).

37. Consider model `lm4`. Interpret the value 71.7264 (it's one of the coefficient estimates) in context.

Players under 30 have an intercept that is 71.7 units higher That is, holding all other variables constant, we expect players under 30 to score 71.7 more points than players over 30.

39. Of the four models given in the output, which would you choose as "best", and why? (If you feel that an alternative model would be better, feel free to suggest it, but you should also choose one of the existing ones as your current "favorite" and justify that choice.)

The strongest argument is forr `lm3`. It has a much larger adjusted R^2 than `lm1` or `lm2`, and `lm4` is not an improvement over `lm3`. From the plots we have, the conditions are met. You may wish to delete Var3 from `lm3` and try a model with just Var6 and Var4.

40. Based on the information provided about this data set, are you comfortable with the two conditions for inference that can't be assessed by residual plots? Discuss both conditions and why you think they are or are not met.

The two conditions are independence and randomness/representativeness.

Random/representative: This is data for the "best" 193 NBA players in the 2018-19 season. So, these players were definitely not randomly chosen and are not representative of all players. But is this data representative of the "best" players in other years? Was this year randomly chosen? Are the stats from this season representative of past/future years? Without more information or knowledge of recent NBA history, we can't answer this question.

Independence: Are these outcomes independent of each other? Players on the same team are definitely not: minutes per game, starts, and rebounds are definitely dependent on the other members of my team. It's hard to say how much this dependence will effect stats over an entire season.

Solutions for Output B (NFL data)

11. Use the regression output from `lm1` to conduct a test of $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Based on the p-value from the previous question, make a conclusion in context.

Since the p-value is very small, we have strong evidence of a linear relationship between net points (Var6) and Win Percentage (Var1).

12. Using `lm1`, a 90% prediction interval is given in the output. Interpret this interval in context.

We are 90% confident that a particular team with net points = 4.0 will have a win percentage bewteen 33.6% and 68.1%. (Or: For all teams with net points=4.0, 90% of them will have a win percentage bewteen 33.6% and 68.1%.)

15. We wish to predict *Var1* using *Var6*, *Var4*, and *Var3*; this is called `lm3`. Interpret the coefficient on *Var3* in context.

Holding net points (Var6) and YardsFor (Var4) constant, a one yard increase in yards scored against the team (Var3) is associated with an increase in win proportion (Var1) of 0.000071 (0.0071%).

30. Based only on the output from lm3, would you remove any of the variables from the model? Briefly justify your answer.

You should remove *Var3* because it is not statistically significant at any reasonable level.

31. Using the output below model `lm3`, report and interpret a confidence interval for the coefficient on *Var4*.

Holding net points (Var6) and YardsAgainst (Var3) constant, we are 95% confident that a one-yard increase in yards scored by the team (Var4) will decrease a team's win proportion (Var1) between 0.00021 (0.021%) and 0.0000088 (0.00088%).

33. Given the other variables in model `lm3`, is *Var4* a useful predictor of *Var1*? Justify your answer.

With Var6 and Var3 in the model, the p-value on Var4 is 0.0342. So we have some evidence that YardsFor is a statistically significant predictor of win percentage, but it's not terribly strong. (I accepted "Yes" or "Somewhat" as an answer to 32.)

35. Are you concerned about multicollinearity in the `lm3` model? Justify your answer.

No, because all of the VIF values for `lm3` are small (below 5).

37. Consider model `lm4`. Interpret the value -9.065*10ˆ-3 (it's one of the coefficient estimates) in context.

NFC teams have an intercept that is 0.0091-units lower. That is, holding all other variables constant, we expect NFC teams to have a win percentage that is 0.91% lower than NFC teams.

39. Of the four models given in the output, which would you choose as "best", and why? (If you feel that an alternative model would be better, feel free to suggest it, but you should also choose one of the existing ones as your current "favorite" and justify that choice.)

You could make a strong argument for either `lm1` or `lm3`. `lm3` has a larger adjusted Rˆ2 than `lm1` by 3%. You may (or may not) feel that the additional variables are worth the additional complexity. The conditions are comparable for both models. A reasonable next step would be to delete Var3 from `lm3` and try a model with just Var6 and Var4.

40. Based on the information provided about this data set, are you comfortable with the two conditions for inference that can't be assessed by residual plots? Discuss both conditions and why you think they are or are not met.

The two conditions are independence and randomness/representativeness.

Random/representative: This is data for ALL 16 NFL teams from 2016. The question is, was this year randomly chosen? Are the stats from this season representative of past/future years? Without more information or knowledge of recent NFL history, we can't answer this question.

Independence: The outcomes are most certainly NOT independent of each other. This is because teams play each other, and if Team A wins, Team B must (by definition) lose. This is how sports work. So Team A's WinPct will increase and Team B's WinPct must decrease as a direct result. Overall for the whole NFL, the mean win percentage *must* be 50%. Teams that play each other a lot will be "less independent" than teams who don't play each other very much.