In the heart of the roaring arenas and under the watchful eyes of thousands, the world of professional bull riding (PBR) unfolds, where grit, skill, and a dash of luck intertwine. In the pursuit of understanding how points are achieved by riders, we will dive into the data from the 2023 season of the Touring Pro Division.

1. Fit a model predicting riders Points, what can we notice about the output? Why is this happening? \*Hint: check dataset documentation\*

Rides + buckoff + Outs very collinear, showing evidence of multicollinearity, since two of those variables are used to determine the third, or vice versa. Therefore, when using all three of those in a model 1/3 of those variables will not be included in the final model output due to their strong correlation. *\*Check correlation matrix to help interpret this\**

2. Using the rider’s dataset, predict the points of a rider using avg. time for rider to fall off (`Avg Buckoff Time`), number of rides (Rides), and percent rides ridden (prop. Ridden) as predictors.

A. Assess model’s residual plots. What do we see?

Residuals do not seem to be linear about zero, the normal q-q plot seems to have points fraying off, along with there being some influential points on cook distance. Does not really pass LINE test too well.

B. Does model prove to be significance? Are there any insignificant predictors?

Yes, p-value is approximately zero with F = 573.3 and DF = 353.

Yes, when running Anova test for the predictors…. The `Avg Buckoff Time`

Has a p-value = 0.913615, which is quite insignificant

C. Now add an interaction between `Avg Buckoff Time` and Rides. Record R-code formula here.

lm(Points ~ `Avg Buckoff Time` + Rides + `Avg Buckoff Time`:Rides + prop.Ridden, data = riders)

D. Test whether having the interaction in the model is significant or not.

modRed <- lm(Points ~ `Avg Buckoff Time` + Rides + prop.Ridden, data = riders)

modFull <- lm(Points ~ `Avg Buckoff Time` + Rides + `Avg Buckoff Time`:Rides + prop.Ridden, data = riders)

anova(modRed, modFull, test = "Chisq")

Yes, does prove to be significant with P approximately equal to zero with F=352, df=1

3. Using the model from beginning of question 2….

A. What can we do to help improve the original model? Explain.

If we add a quadratic term for `Avg BuckOff time`, so have both original and quadratic. Also, square the proportion of rides ridden. Then we can see improvement in the models overall R-squared. Helps increased variability in model.

B. Record R-code here along with new R-squared.

lm(Points ~ poly(`Avg Buckoff Time`, 2) + Rides + I(prop.Ridden^2), data = riders)

C. Test which model, original or transformed, is more effective for the data?

modOG <- lm(Points ~ `Avg Buckoff Time` + Rides + prop.Ridden, data = riders)

modTrans <- lm(Points ~ poly(`Avg Buckoff Time`, 2) + Rides + I(prop.Ridden^2), data = riders)

anova(modOG, modTrans, test = "Chisq")

Transformed model is the more effective model for the data

( F =352, DF=1 and p-value = 0.0001273)