STA 207 HW-8 solution

Problem-1 [55 points]: For the MLBStandings2016 data, do the following:

a. [35 points] Regress WinPct on ERA and League and report the fitted model. Interpret the regression coefficient for the League predictor. Make a plot of ERA vs WinPct with separate lines for the two leagues. Is League a significant predictor of WinPct in

First thing is we check the class of League variable in R. It is a factor in R (meaning categorical predictor) and the levels are AL for American League and NL for the National League.

Fitted Model [7.5’]:

$\hat{Y}$=0.996 – 0.116 $X_1 - 0.017\ X_2$,

where

Y: proportion of games won,

$X_1$is the ERA

X_2 is a dummy variable with a value 0 for reference league (American) and 1 for National League.

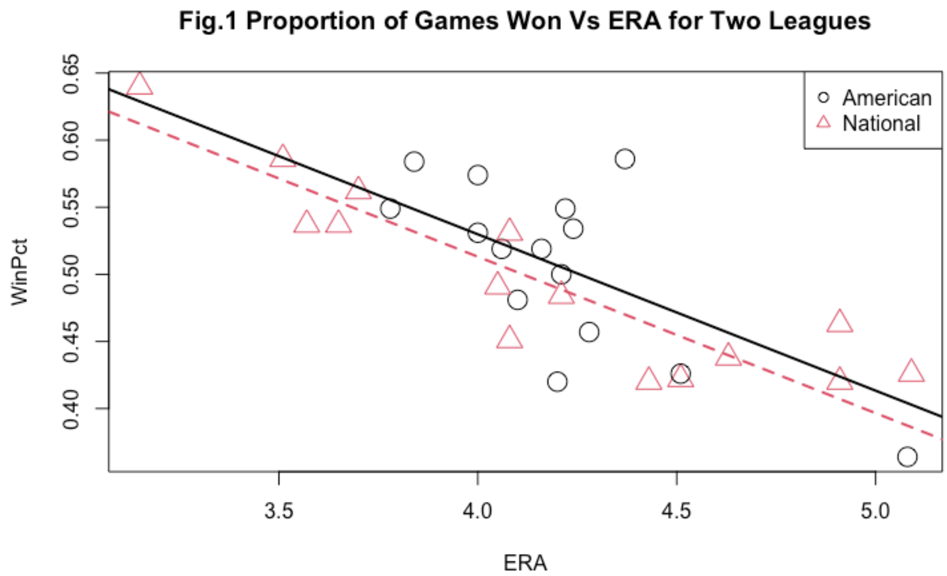
Interpretations [10]:

For a unit increase in ERA with fixed league the expected decrease in proportion of games won is 0.116. The slope is same for either league.

When ERA is 0, the expected proportion of games won in the American League (reference) is 0.996.

When ERA is 0, the expected proportion of games won in the National League (reference) is 0.979.

The plot in Fig.1 below shows the two separate regression lines for the two leagues. **[7.5’]**



Hypothesis Test for League: [10']

Step1→ $H_0: \beta_2 = 0, H_1: \beta_2 \neq 0$

Step2→ $\alpha = 0.05$

Step3→ test statistic=t=-1.35

Step4→ P-value=0.266

Step5→ P-value > α so we cannot reject the null hypothesis.

Conclusion: At 5% significance level, League is not a statistically significant predictor for predicting proportion of games won when ERA is accounted for.

b. [30 points] For the model regressing WinPct on the predictors ERA and League include the interaction term. Report the fitted model and interpret the coefficients. Make a plot for the two leagues with different intercepts and slopes. Check if the interaction term is statistically significant.

Fitted Model [5']:

$\hat{Y}$=1.149 – 0.153 $X_1 - 0.213\ X_2 + 0.047 X_1 X_2$,

Where Y: proportion of games won,

X_1 is the ERA

X_2 is a dummy variable with a value 0 for reference league (American) and 1 for National League.

$X_1 X_2$ is the interaction between ERA and League variable

Interpretations:[10']

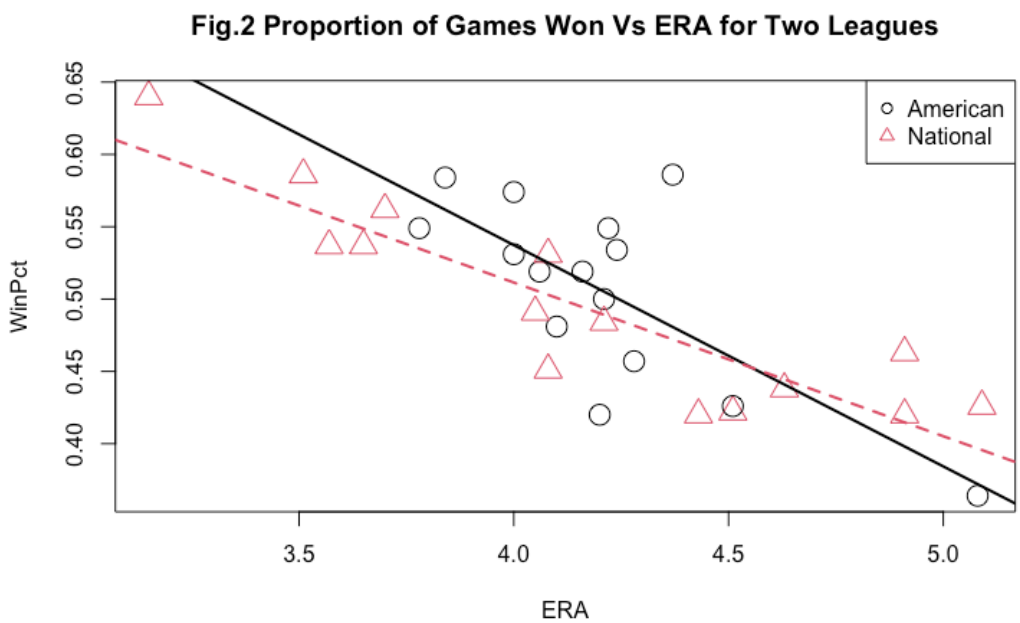
When ERA is 0, the expected proportion of games won in the American League (reference) is 1.149.

When ERA is 0, the expected proportion of games won in the National League (reference) is 0.936.

For a unit increase in ERA with American (reference) League the expected decrease in proportion of games won is 0.153.

For a unit increase in ERA with National League the expected decrease in proportion of games won is 0.106.

Figure 2 shows the two regression lines with different intercepts and slopes for the two leagues representing relation between ERA and Proportion of games won **[10].**



Hypothesis Test for interaction term [5']:

Step1→ $H_0: \beta_3 = 0, H_1: \beta_3 \neq 0$

Step2→ $\alpha = 0.05$

Step3→ test statistic=t=1.185

Step4→ P-value=0.247

Step5→ P-value > α so we cannot reject the null hypothesis.

Conclusion: At a 5% significance level, interaction between ERA and League is not statistically significant as a predictor for the proportion of games won.

c. [20 points] Regress WinPct on the predictors ERA and Runs including their interaction term. Report the fitted model and interpret the coefficients. Check if the interaction term is statistically significant.

Fitted Model [5']:

$\hat{Y}$=0.464 – 0.073 $X_1 + 0.0007\ X_2 - 0.00004 X_1 X_2$,

Where Y: proportion of games won,

X_1 is the ERA

X_2 is Runs

$X_1 X_2$ is the interaction between ERA and Runs

Interpretations [10']:

- $\hat{\beta}_0 = 0.464$ is the estimated average proportion of games won by a team with 0 ERA and 0 Runs.
- $\hat{\beta}_1 = -0.073$ is the estimated decrease in proportion of games won for an increase of 1 unit in ERA, for a team with 0 Runs.
- $\hat{\beta}_2 = +0.0007$ is the estimated change in proportion of games won for an increase in 1 unit in Runs, for a team with 0 ERA
- $\hat{\beta}_3 = -0.00004$ is an estimate of the modification to the change in proportion of games won for an increase in ERA, for a team of a certain Runs (or vice versa).

Hypothesis Test for interaction term [5']:

Step1→ $H_0: \beta_3 = 0, H_1: \beta_3 \neq 0$

Step2→ $\alpha = 0.05$

Step3→ test statistic=t=-0.198

Step4→ P-value=0.845

Step5→ P-value > α so we cannot reject the null hypothesis.

Conclusion: At a 5% significance level, the interaction term between ERA and Runs is not statistically significant as a predictor for proportion of games won.

d. [15 points] Perform ANOVA with FM as model with ERA, Runs, and their interaction term and RM as the ERA and Runs model.

ANOVA

Step1→ H_0: $\beta_3 = 0$, meaning that the reduced model (RM) without the interaction is enough.

H_1: $\beta_3 \neq 0$, meaning that we need the full model (FM) with interaction term.

Step2→ $\alpha = 0.05$

Step3→ test statistic=F= 0.0391

Step4→ P-value=0.845

Step5→ P-value > α so we cannot reject the null hypothesis.

Conclusion: Reduced model is enough and full model is not required. In other words, the interaction term between ERA and Runs is not statistically significant when ERA and Runs are in the model to predict the proportion of games won.