STA 207 HW-8 | Due Date: 12/6 by 09.00PM

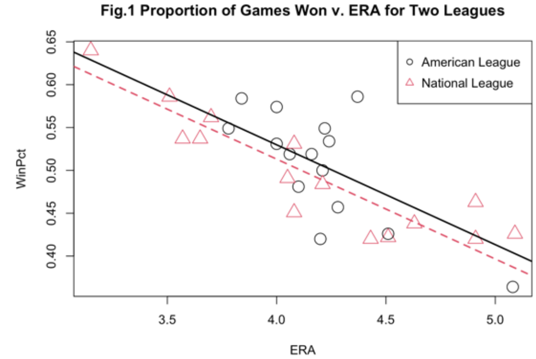
Derin Gezgin | Camel ID:00468038

Problem: 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 the presence of ERA. Show hypothesis test for the League predictor.**

Regress WinPct on ERA and League and report the fitted model.

Make a plot of ERA v. WinPct with separate lines per category.



$\widehat{Y} = \mathbf{0.996 - 0.117}\boldsymbol{x_1} - \mathbf{0.017}\boldsymbol{x_2}$ where:

Y→ Proportion of games won (WinPct)

x_1→ Earned run average (earned runs allowed per 9 innings)

x_2→ League: AL=American or NL=National

Interpret the regression coefficient for the League predictor

In the **R** analysis, the **American League** is the base category with an $\boldsymbol{x_2}$ **value of 0**.

In this case, the fitted model for the American League ($x_2 = 0$) would be;

$$\widehat{Y} = \mathbf{0.996 - 0.117}\boldsymbol{x_1}$$

0.996 is the estimated proportion of games won, for the American League when the earned run average is 0.

*On the other hand, the **National League** is the other category with an $\boldsymbol{x_2}$ **value of 1**.*

In this case, the fitted model for the National League ($x_2 = 1$) would be;

$$\widehat{Y} = \mathbf{0.979 - 0.117}\boldsymbol{x_1}$$

0.979 is the estimated proportion of games won, for the National League when the earned run average is 0.

As the intercept of the American League is higher, we can say that -in terms of proportion of games won- the American League performs better than the National League

General Interpretation of the Regression Coefficients

$\boldsymbol{\beta_1} = \mathbf{0.117}$→ Estimated change in proportion of games won for a unit increase in earned run average for either league.

$\boldsymbol{\beta_2} = \mathbf{0.017}$ → Estimated difference in proportion of games won for the American league, as compared to the national league for any earned run average.

$\boldsymbol{\beta_0} = \mathbf{0.996}$ → The estimated proportion of games won for the American League when the earned run average is 0.

$\boldsymbol{\beta_0 + \beta_2} = \mathbf{0.979}$ → The estimated proportion of games won for the National League when the earned run average is 0.

Is League a significant predictor of WinPct in the presence of ERA. Show hypothesis test for the League predictor.

Step 1: Formulate Null and Alternative Hypothesis

$\boldsymbol{H_0}$: $\boldsymbol{\beta_2 = 0}$ → League is not a significant predictor of proportion of games won (WinPct).

$\boldsymbol{H_A}$: $\boldsymbol{\beta_2 \neq 0}$ → League is a significant predictor of proportion of games won (WinPct).

Step 2: Set level of significance: 0.05

Step 3: Test statistic: -1.135

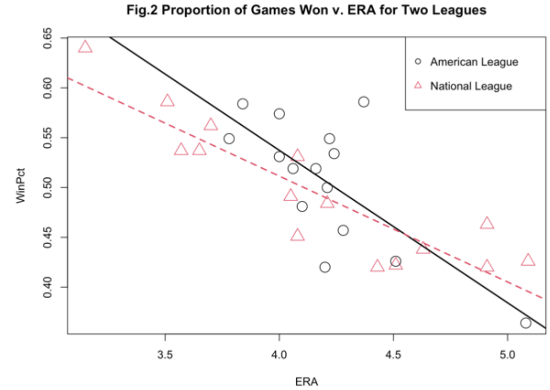
Step 4: P-Value: 0.266

Step 5: Conclusion:

P-Value is larger than our level of significance. This means that at 5% level of significance, we cannot reject the null hypothesis. In other words, *league* is not a significant predictor of proportion of games won (WinPct) in the presence of ERA.

b. [25 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.

Fig.2 Proportion of Games Won v. ERA for Two Leagues



Report the fitted model

$\hat{Y} = 1.149 - 0.153x_1 - 0.213x_2 + 0.047x_1x_2$

Y→ Proportion of games won (WinPct)

x_1→ Earned run average (earned runs allowed per 9 innings)

x_2→ League: AL=American or NL=National

x_1x_2 → Interaction term between ERA and League.

Fitted Model for the American League: x_2 value of 0

$$\boldsymbol{\hat{Y} = 1.149 - 0.153x_1}$$

Fitted Model for the National League: x_2 value of 1

$$\boldsymbol{\hat{Y} = 0.936 - 0.106x_1}$$

Interpretation of the Regression Coefficients

$\boldsymbol{\beta_1 = 0.153}$ → Estimated change in proportion of games won for a unit increase in earned run average for the American league.

$\boldsymbol{\beta_2 = 0.213}$ → Estimated difference in proportion of games won for the American league, as compared to the national league for a run average of 0.

$\boldsymbol{\beta_3 = 0.047}$→ Estimated difference of proportion of games won for an increase in earned run average, for the American league compared to the National league.

$\boldsymbol{\beta_1 + \beta_3 = 0.106}$ → Estimated change in proportion of games won for a unit increase in earned run average for the National League.

$\mathbf{\beta_0 = 1.149}$ → The estimated proportion of games won for the American League when the earned run average is 0.

$\mathbf{\beta_0 + \beta_2 = 0.936}$ → The estimated proportion of games won for the National League when the earned run average is 0.

Checking if the Interaction Term is Statistically Significant

Step 1: Formulate Null and Alternative Hypothesis

$\boldsymbol{H_0}$: $\mathbf{\beta_3 = 0}$ → Interaction Term is not a significant predictor.

$\boldsymbol{H_A}$: $\mathbf{\beta_3 \neq 0}$ → Interaction term is a significant predictor

Step 2: Set level of significance: 0.05

Step 3: Test statistic**:** 1.185

Step 4: P-Value: 0.247

Step 5: Conclusion:

P-Value is larger than our level of significance. This means that at 5% level of significance, we do not enough evidence to reject the null hypothesis and conclude that the interaction term between *League* and earned run average is not a significant predictor of proportion of games won (WinPct).

c. [25 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.

Report the Fitted Model

$$\hat{Y} = 0.464 - 0.073x_1 + 0.0006x_2 - 0.00004x_1x_2$$

$$\hat{Y} = 0.464 - (0.073 - 0.00004x_2)x_1 + 0.0006x_2$$

Y→ Proportion of games won

x_1→ Earned run average (earned runs allowed per 9 innings)

x_2→ Runs (Number of runs scored)

x_1x_2 → Interaction term between earned run average and runs scored.

Interpretation of the Coefficients

$\mathbf{\beta_0 = 0.464}$→Estimated proportion of games won when the earned run average and number of runs scored are both 0.

$\mathbf{\beta_1 = -0.073}$ →Estimated change in proportion of games won for a unit increase in earned run average when number of runs scored is 0.

$\mathbf{\beta_2 = 0.0006}$ → Estimated change in proportion of games won for a unit increase in number of runs scored when earned run average is 0.

$\mathbf{\beta_3 = -0.00004}$ → Estimate of the modification to the change in proportion of games won for a unit increase in earned run average in case of a certain number of runs scored.

Checking the statistical significance of the interaction term

Step 1: Formulate Null and Alternative Hypothesis

$\boldsymbol{H_0}$: $\boldsymbol{\beta_3 = 0}$ → Interaction Term is not a statistically significant predictor.

$\boldsymbol{H_A}$: $\boldsymbol{\beta_3 \neq 0}$ → Interaction term is a statistically significant predictor

Step 2: Set level of significance $(\boldsymbol{\alpha})$***:*** 0.05

Step 3: Test statistic**:** -0.198

Step 4: P-Value: 0.845

Step 5: Conclusion:

P-Value is larger than our level of significance. At 5% level of significance, we do not have enough evidence to reject the null hypothesis. In other words, this means that at 5% level of significance, the interaction term between ERA and Runs is not a significant predictor of WinPct.

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

Step 1: Formulate Null and Alternative Hypothesis

$\boldsymbol{H_0}$: $\boldsymbol{\beta_3 = 0}$ → Interaction Term is not a statistically significant predictor. Favoring the removed model

$\boldsymbol{H_A}$: $\boldsymbol{\beta_3 \neq 0}$ → Interaction term is a statistically significant predictor. Favoring the full model.

Step 2: Set level of significance $(\boldsymbol{\alpha})$***:*** 0.05

Step 3: Test statistic**:** 0.039

Step 4: P-Value: 0.8449

Step 5: Conclusion:

P-Value is larger than our level of significance. This means that at 5% level of significance, we do not have evidence to reject the null hypothesis. We can conclude the interaction term is not a statistically significant predictor and we would favor the removed model which does not include the interaction term.