

Multiple Regression Model

Zachary Palmore

4/12/2021

Prompt

Using R, build a multiple regression model for data that interests you. Include in this model at least one quadratic term, one dichotomous term, and one dichotomous vs. quantitative interaction term. Interpret all coefficients. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Data

Before I begin, it might be useful to define the kind of variables required. This way we can evaluate whether the data fits the definitions as I understand them and attempt to conduct an analysis based on the same understanding of the requirements. Those terms given in the prompt are defined as follows:

- Quadratic term - a variable that appears in the form ax^2
- Dichotomous term - relating to a variable that contains only two parts
- Dichotomous vs. quantitative interaction term - an interaction between a variable that splits into two parts and one that is discrete or continuous but numeric by nature

```
# Packages
library(tidyverse)
library(ggpubr)
library(kableExtra)
library(reshape2)
library(corrplot)
library(ggstatsplot)
library(MASS)
library(bestNormalize)
library(Metrics)
library(VGAM)
theme_set(theme_minimal())

tdata <- read.csv(
  "https://raw.githubusercontent.com/palmorezm/msds/main/621/HW1/moneyball-training-data.csv")
# Basic statistics from the set
total.obs <- count(tdata)
avg.wins <- mean(tdata$TARGET_WINS)
max.wins <- max(tdata$TARGET_WINS)
sd.wins <- sd(tdata$TARGET_WINS)
colnames <- colnames(tdata)
missing.values <- (sum(is.na(tdata)))
```

Transforming some value into a quadratic form.

```
head(tdata, 3)
```

```
##      INDEX TARGET_WINS TEAM_BATTING_H TEAM_BATTING_2B TEAM_BATTING_3B
## 1         1          39          1445           194           39
## 2         2          70          1339           219           22
## 3         3          86          1377           232           35
##      TEAM_BATTING_HR TEAM_BATTING_BB TEAM_BATTING_SO TEAM_BASERUN_SB
## 1              13          143           842             NA
## 2             190          685          1075             37
## 3             137          602           917             46
##      TEAM_BASERUN_CS TEAM_BATTING_HBP TEAM_PITCHING_H TEAM_PITCHING_HR
## 1              NA             NA          9364             84
## 2              28             NA          1347            191
## 3              27             NA          1377            137
##      TEAM_PITCHING_BB TEAM_PITCHING_SO TEAM_FIELDING_E TEAM_FIELDING_DP
## 1              927          5456          1011             NA
## 2              689          1082           193            155
## 3              602           917           175            153
```

Analysis

One interest of mine is baseball. Thankfully, there is plenty of data on this topic which is why I have decided to use this sports data for analysis. Our end goal is to predict.