HW 2 solutions

Sports and statistics

For this lab, we will be using data provided in the `Lahman` package.

Question 1

Let's use the `Teams` data set (recall: to load this data set from the Lahman package, run the command `data(Teams)`). Using every season since 1970, fit a multiple regression model of runs as a function of singles, doubles, triples, home runs, and walks. (recall: you have to create the singles variable. See Lab 2.) Showing the code output is sufficient for this question.

```
Teams.1 <- filter(Teams, yearID >=1970)
Teams.1 <- mutate(Teams.1, X1B = H - X2B - X3B - HR)
fit1 <- lm(R ~ X1B + X2B + X3B + HR + BB, data = Teams.1 )
summary(fit1)
```

```
##
## Call:
## lm(formula = R ~ X1B + X2B + X3B + HR + BB, data = Teams.1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -89.982 -21.602  -2.002  19.046 100.389
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -173.06706   10.54241  -16.42   <2e-16 ***
## X1B            0.33256    0.01122   29.65   <2e-16 ***
## X2B            0.61465    0.02700   22.76   <2e-16 ***
## X3B            1.27017    0.09428   13.47   <2e-16 ***
## HR             1.35958    0.03001   45.30   <2e-16 ***
## BB             0.30295    0.01338   22.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.92 on 1228 degrees of freedom
## Multiple R-squared:  0.9046, Adjusted R-squared:  0.9042
## F-statistic:  2330 on 5 and 1228 DF,  p-value: < 2.2e-16
```

```
#Your code goes here
```

Question 2

Refer to the fit in question 1. Identify the y-intercept, as well as the slopes for singles, doubles, triples, home runs and walks (do not interpret).

```
fit1$coefficients
```

```
##  (Intercept)          X1B          X2B          X3B           HR
## -173.0670577    0.3325582    0.6146452    1.2701733    1.3595775
##           BB
##    0.3029453
```

The estimated y-intercept is -173.1, and the coefficients are 0.33, 0.61, 1.27, 1.36, and 0.3 for our coefficients.

Question 3

Refer to the fit in question 1. Interpret the slope coefficient estimate for triples.

Given the model with singles, doubles, HRs, and walks, each triple is worth an estimated 1.27 runs.

Question 4

Use the fit in question 1 to generate a set of predicted runs scored for each team in your data set.

```
Teams.1 <- mutate(Teams.1, predict.R = predict(fit1, Teams.1))
cor(predict.R ~ R, data = Teams.1)
```

```
## [1] 0.9511197
```

What is the correlation between your predicted runs and the number of actual runs? How does this compare to the correlation between created runs and actual runs that we found in Lab 2?

The correlation between runs and predicted runs is 0.95 - slightly higher than what we found with the other runs created formula in Lab 2 (which was 0.93)

Question 6

Using the output from Question 1, discuss the relative importance of each type of productive at bat (singles, doubles, triples, home runs, walks) with respect to run generation. Does anything surprise you?

No real surprises: doubles are worth roughly twice that of singles, with home runs worth nearly twice what doubles are. Interestingly, walks are valued nearly as much as singles.

Question 7

Pick another variable in the `Teams` data set, and add it to your regression model. Interpret it's slope. Also, does this new variable appear to be significantly associated with runs scored, given the other variables in the model?

Answers will vary.

Question 8

Bonus: Go to the lecture on Thursday night featuring Dan Turkenkopf, Director of Research for the Milwaukee Brewers.