Linear Regression

Code

Linear Regression

'Basebal.csv' shows data from all 30 Major League Baseball teams. You will be examining the linear relationship between runs scored in a season and a number of other player statistics.

A data frame with 30 observations on the following 12 variables.

- team: Team name
- runs: Number of runs
- at_bats: Number of at bats
- hits: Number of hits
- homerun: Number of home runs
- bat_avg: Batting average
- strikeouts: Number of strikeouts
- stolen_bases: Number of stolen bases
- wins: Number of wins
- new_onbase: On base percentage, measure of how often a batter reaches base for any reason other than a fielding error, fielder's choice, dropped/uncaught third strike, fielder's obstruction, or catcher's interference
- new_slug: Slugging percentage, popular measure of the power of a hitter calculated as the total bases divided by at bats
- new_obs: On base plus slugging, calculated as the sum of these two variables

1. Plot a diagram to show the relationship between 'runs' and 'at_bats'. Consider 'at_bats' as the explanatory variable.
2. Can you quantify this relationship ?

Hide

```
cor(baseball$runs,baseball$at_bats)
```

3. A more efficient way to find the min Sum of Squares is to use the `lm` f unction in R to fit the linear model (a.k.a. regression line).

```
m1 <- lm ( runs ~ at_bats , data = baseball )
```

Here it can be read that we want to make a linear model of runs as a function of `at_bats` . The second argument specifies that R should look in the baseball data frame to find the runs and at_bats variables.

`summary(m1)` contains all of the information we need about the linear model that was just fit

What is the linear function that describes the relationship (Hint: look at the coefficients in the summary)

`r` r summary(m1)

```
```

```
Call:
lm(formula = runs ~ at_bats, data = baseball)

Residuals:
    Min      1Q  Median      3Q     Max
-125.58  -47.05  -16.59   54.40  176.87

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2789.2429   853.6957  -3.267 0.002871 **
at_bats         0.6305     0.1545   4.080 0.000339 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.47 on 28 degrees of freedom
Multiple R-squared:  0.3729,    Adjusted R-squared:  0.3505
F-statistic: 16.65 on 1 and 28 DF,  p-value: 0.0003388
```

So the formula is :

$ = 0.6305 * - 2789.24 $

4. Fit another model that uses `homeruns` to predict `runs`. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between success of a team and its home runs?

Create a scatter plot using the following commands

```
plot(baseball$ runs  ~  baseball$ at_bats )
abline(m1)
```

The abline can help you predict y at any value x. If the prediction is done for a value of x that is outside of the range of the dataset, what is the process is called ? What is the problem with it ?

5. If a team manager saw the least squares regression line and not the actual data, how many runs would he or she predict for a team with 5,579 `at_bats` ? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?
6. Choose another traditional variable from baseball daga that you think might be a good predictor of `runs`. Produce a scatterplot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?
7. How does this relationship compare to the relationship between runs and `at_bats` ? Use the R2 values from the two model summaries to compare. Does your variable seem to predict runs better than `at_bats` ? How can you tell?

*Adopted from openintro labs. http://openintro.org (http://openintro.org)