Moneyball

Julian Lucero

2024-07-02

```
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")
load("mlb11.RData")
```

```
### Best predicting variable for runs
Predicter <- lm(runs ~ new_onbase, data = mlb11)
summary(Predicter)
```

```
##
## Call:
## lm(formula = runs ~ new_onbase, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.270 -18.335   3.249  19.520  69.002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1118.4      144.5  -7.741 1.97e-08 ***
## new_onbase    5654.3      450.5  12.552 5.12e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.61 on 28 degrees of freedom
## Multiple R-squared:  0.8491, Adjusted R-squared:  0.8437
## F-statistic: 157.6 on 1 and 28 DF,  p-value: 5.116e-13
```

An R squared of ,8491 is observed. That is significant for a predictor.

```
Predicter <- lm(runs ~ new_slug, data = mlb11)
summary(Predicter)
```

```
##
## Call:
## lm(formula = runs ~ new_slug, data = mlb11)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.41 -18.66  -0.91  16.29  52.29
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -375.80      68.71   -5.47 7.70e-06 ***
## new_slug     2681.33     171.83   15.61 2.42e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.96 on 28 degrees of freedom
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8932
## F-statistic: 243.5 on 1 and 28 DF,  p-value: 2.42e-15
```

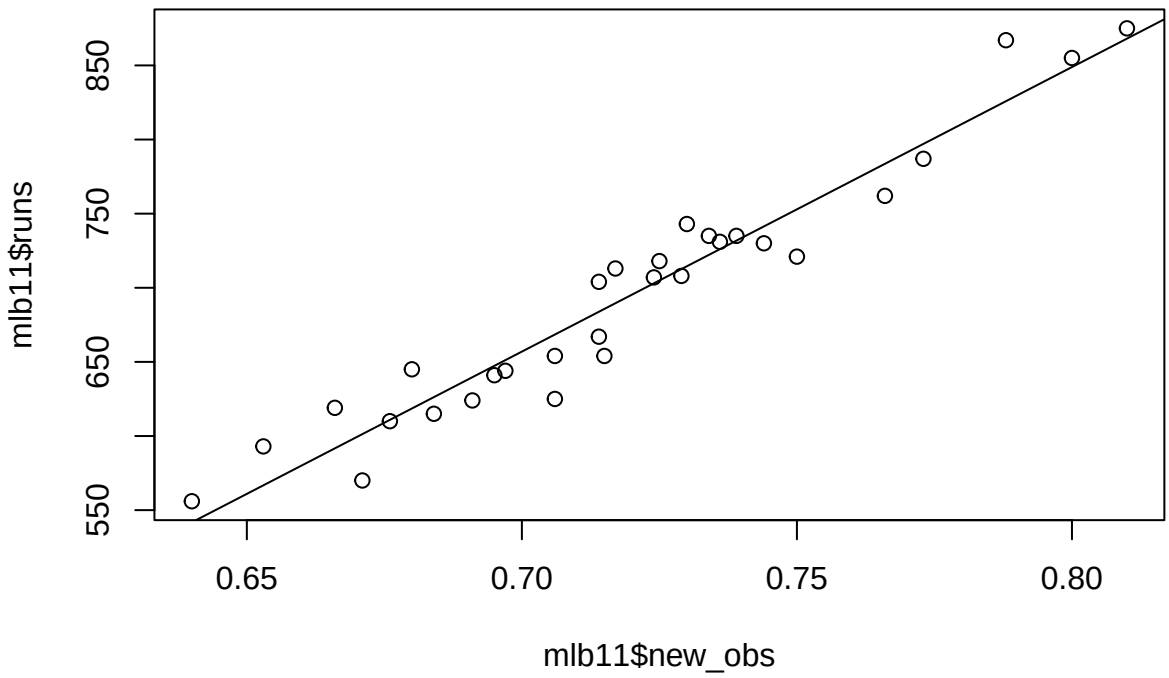
Slugging percentage is a greater fit to predict runs than on base percentage.

```
Predicter <- lm(runs ~ new_obs, data = mlb11)
summary(Predicter)
```

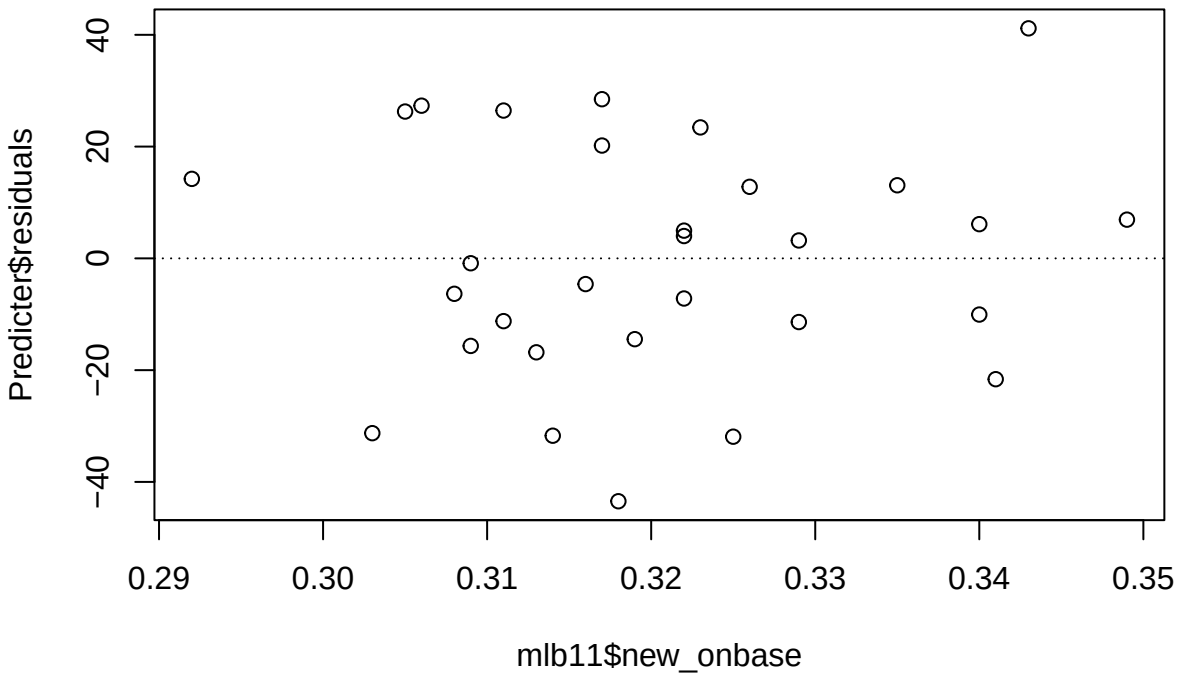
```
##
## Call:
## lm(formula = runs ~ new_obs, data = mlb11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.456 -13.690   1.165  13.935  41.156
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -686.61      68.93  -9.962 1.05e-10 ***
## new_obs      1919.36      95.70  20.057  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.41 on 28 degrees of freedom
## Multiple R-squared:  0.9349, Adjusted R-squared:  0.9326
## F-statistic: 402.3 on 1 and 28 DF,  p-value: < 2.2e-16
```

OBS is the best single predicter for runs. This is a combination of player's on base percentage and slugging percentage. This is adequate in generating runs and thus would increase the liklihood of a team's odds of winning.

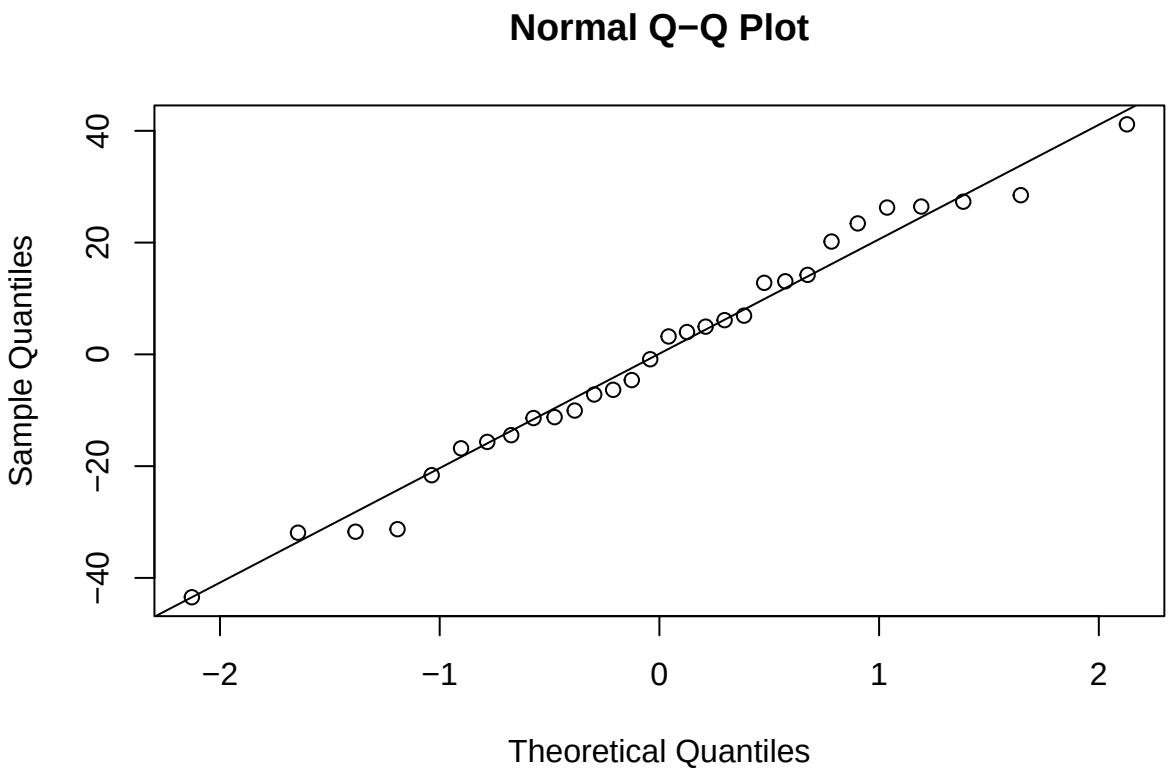
```
### Scatterplot
plot(mlb11$runs ~ mlb11$new_obs)
abline(Predicter)
```




```
### Residual Check
plot(Predicter$residuals ~ mlb11$new_onbase)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0
```




```
### Normal Probability
qqnorm(Predicter$residuals)
qqline(Predicter$residuals)
```




```
cor(mlb11$runs, mlb11$new_obs)
```

```
## [1] 0.9669163
```

Recommendations: Insert individuals within the starting lineup that have the highest OBS rather than just traditional batting average. This is heavily correlated with more runs per game (0.97). This can be very useful in the decision for designated hitters.