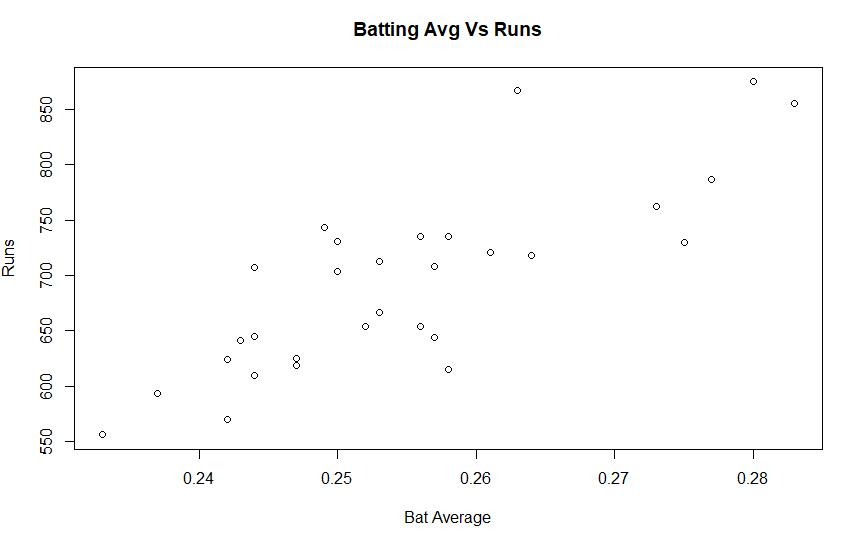
1. Choose another traditional variable from mlb11 that you think might be a good predictor of runs. Produce a scatterplot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?

Let select bat\_avg to check

|  |
| --- |
| load("C:/Users/thai/Documents/workspace-new/hu/502/linear-regression/mlb11.csv")  plot(mlb11$bat\_avg,mlb11$runs,xlab="Bat Average",ylab="Runs",main="Batting Avg Vs Runs") |



It seems linear relationship between Runs and Bat\_avg

|  |
| --- |
| > nBatAvg=lm(runs~bat\_avg,data=mlb11)  > summary(nBatAvg)  Call:  lm(formula = runs ~ bat\_avg, data = mlb11)  Residuals:  Min 1Q Median 3Q Max  -94.676 -26.303 -5.496 28.482 131.113  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -642.8 183.1 -3.511 0.00153 \*\*  bat\_avg 5242.2 717.3 7.308 5.88e-08 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 49.23 on 28 degrees of freedom  Multiple R-squared: 0.6561, Adjusted R-squared: 0.6438  F-statistic: 53.41 on 1 and 28 DF, p-value: 5.877e-08 |

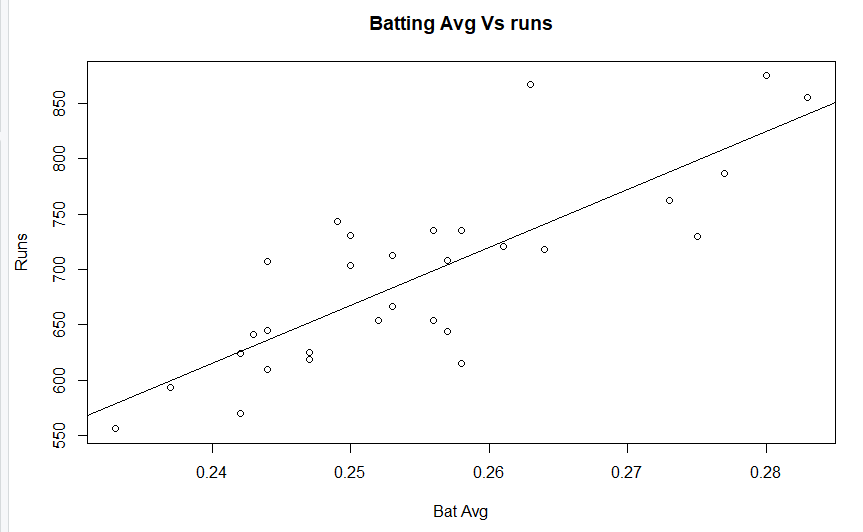
1. How does this relationship compare to the relationship between runs and at\_bats? Use the R22 values from the two model summaries to compare. Does your variable seem to predict runs better than at\_bats? How can you tell?

|  |
| --- |
| > nAtBats=lm(runs~at\_bats,data=mlb11)  > summary(nAtBats)  Call:  lm(formula = runs ~ at\_bats, data = mlb11)  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 |

|  |
| --- |
| Because R2 of relationship between runs and at\_bats is 0.3729, R2 of relationship between runs and bat\_avg is 0.6561. So bat\_avg is predicting better. |

1. Now that you can summarize the linear relationship between two variables, investigate the relationships between runs and each of the other five traditional variables. Which variable best predicts runs? Support your conclusion using the graphical and numerical methods we’ve discussed (for the sake of conciseness, only include output for the best variable, not all five).

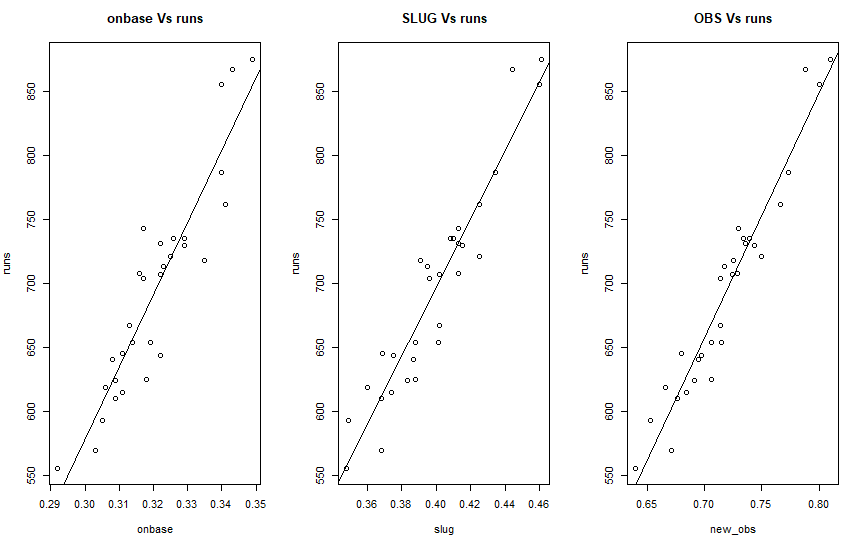
|  |
| --- |
| model=lm(runs~bat\_avg,data=mlb11) plot(mlb11$bat\_avg,mlb11$runs,xlab="Bat Avg",ylab="Runs",main="Batting Avg Vs runs")  abline(model) |



|  |
| --- |
| > summary(model)  Call:  lm(formula = runs ~ bat\_avg, data = mlb11)  Residuals:  Min 1Q Median 3Q Max  -94.676 -26.303 -5.496 28.482 131.113  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -642.8 183.1 -3.511 0.00153 \*\*  bat\_avg 5242.2 717.3 7.308 5.88e-08 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 49.23 on 28 degrees of freedom  Multiple R-squared: 0.6561, Adjusted R-squared: 0.6438  F-statistic: 53.41 on 1 and 28 DF, p-value: 5.877e-08  > sum(model$residuals^2)  [1] 67849.52 |

1. Now examine the three newer variables. These are the statistics used by the author of Moneyball to predict a teams success. In general, are they more or less effective at predicting runs that the old variables? Explain using appropriate graphical and numerical evidence. Of all ten variables we’ve analyzed, which seems to be the best predictor of runs? Using the limited (or not so limited) information you know about these baseball statistics, does your result make sense?

|  |
| --- |
| n1=lm(runs~new\_onbase,data=mlb11) n2=lm(runs~new\_slug,data=mlb11) n3=lm(runs~new\_obs,data=mlb11) par(mfrow=c(1,3)) plot(mlb11$new\_onbase,mlb11$runs,xlab="onbase",ylab="runs",main="onbase Vs runs")  abline(n1) plot(mlb11$new\_slug,mlb11$runs,xlab="slug",ylab="runs",main="SLUG Vs runs")  abline(n2) plot(mlb11$new\_obs,mlb11$runs,xlab="new\_obs",ylab="runs",main="OBS Vs runs") abline(n3) |



|  |
| --- |
| > summary(n1)  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 |

|  |
| --- |
| > summary(n2)  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 |

|  |
| --- |
| > summary(n3)  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 |

|  |
| --- |
| > sum(n1$residuals^2)  [1] 29768.7  > sum(n2$residuals^2)  [1] 20345.54  > sum(n3$residuals^2)  [1] 12837.65 |

We can see the new variable is giving better prediction than old one new\_obs with R2 0.9349, the best predictor.

1. Check the model diagnostics for the regression model with the variable you decided was the best predictor for runs.

new\_obs has highest R2 0.9349

|  |
| --- |
| > summary(n3)  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  > sum(n3$residuals^2)  [1] 12837.65 |