Probability and Statistics

Will Pauls

February 19, 2019

## Machine Learning

### Hypothesis

: = 0 - Teams payroll rank has no linear relationship with their winning %  
: 0 - Teams payroll rank has a linear relationship with their winning %

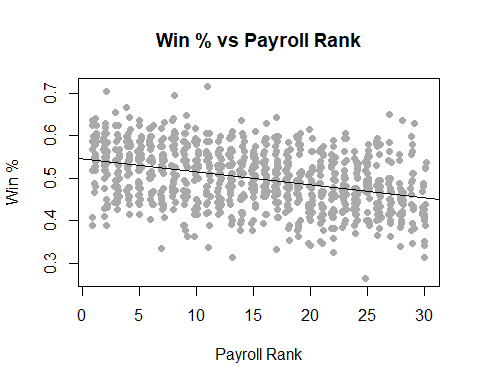
#### Linear regression summary with confidence interval

##   
## Call:  
## lm(formula = baseball$winpercent ~ baseball$payrank)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.204099 -0.046392 0.002225 0.044436 0.204081   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5455626 0.0043006 126.86 <2e-16 \*\*\*  
## baseball$payrank -0.0030585 0.0002518 -12.15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06378 on 916 degrees of freedom  
## Multiple R-squared: 0.1387, Adjusted R-squared: 0.1378   
## F-statistic: 147.6 on 1 and 916 DF, p-value: < 2.2e-16

## 2.5 % 97.5 %  
## (Intercept) 0.53712252 0.554002730  
## baseball$payrank -0.00355268 -0.002564411

The P-value is practically 0 so there is a signifcant result and we can reject the null hypothesis and conclude that there is a linear realtionship between a teams winning % and where they rank in Team Payroll.

plot(baseball$winpercent~jitter(baseball$payrank), main = "Win % vs Payroll Rank", xlab = "Payroll Rank", ylab = "Win %", pch = 16, col = "dark grey")  
abline(MLB)



### Testing Linear Regression Assumptions

#### Assessing Outliers

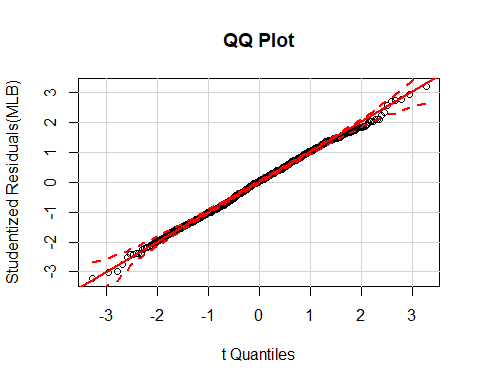
# Assessing Outliers  
  
outlierTest(MLB) # Bonferonni p-value for most extreme obs

##   
## No Studentized residuals with Bonferonni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferonni p  
## 306 -3.220594 0.0013244 NA

There are no outliers that would influence our model in a signifcant way.

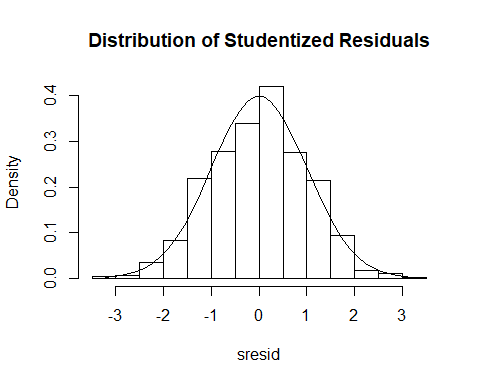
#### Normality of Residuals

# Normality of Residuals  
  
qqPlot(MLB, main="QQ Plot") #qq plot for studentized resid

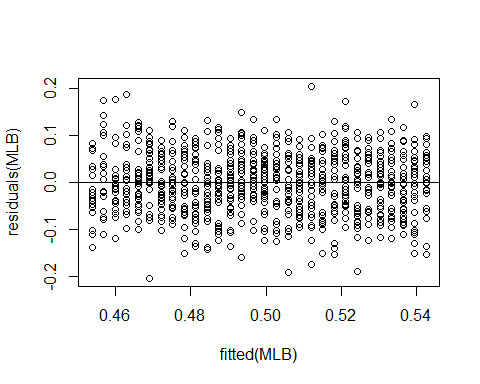


#### Distribution of studentized residuals

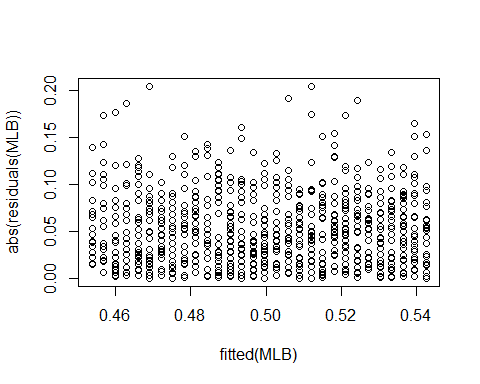
# distribution of studentized residuals  
  
sresid <- studres(MLB)   
hist(sresid, freq=FALSE,   
 main="Distribution of Studentized Residuals")  
xfit<-seq(min(sresid),max(sresid),length=40)   
yfit<-dnorm(xfit)   
lines(xfit, yfit)



# plot studentized residuals vs. fitted values   
plot(fitted(MLB), residuals(MLB))  
abline(h = 0)



plot(fitted(MLB), abs(residuals(MLB)))



summary(lm(abs(residuals(MLB))~fitted(MLB)))

##   
## Call:  
## lm(formula = abs(residuals(MLB)) ~ fitted(MLB))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.051472 -0.029895 -0.005111 0.022959 0.152936   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.046885 0.024313 1.928 0.0541 .  
## fitted(MLB) 0.009121 0.048562 0.188 0.8511   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.03763 on 916 degrees of freedom  
## Multiple R-squared: 3.851e-05, Adjusted R-squared: -0.001053   
## F-statistic: 0.03527 on 1 and 916 DF, p-value: 0.8511

After testing the residuals we can conclude that are data is an approximately normal distribution which is needed to fit are assumptions in our linear model.

#### Evaluation of homoscedasticity

# Evaluate homoscedasticity  
# non-constant error variance test  
ncvTest(MLB)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.03794686 Df = 1 p = 0.8455499

Since the P-value is .84 we can conclude that the standard deviations of the error terms are constant and do not depend on the x-value, which is an assumption needed for our linear model.

#### Global validation of linear model assumptions .

gvmodel <- gvlma(MLB)   
summary(gvmodel)

##   
## Call:  
## lm(formula = baseball$winpercent ~ baseball$payrank)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.204099 -0.046392 0.002225 0.044436 0.204081   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.5455626 0.0043006 126.86 <2e-16 \*\*\*  
## baseball$payrank -0.0030585 0.0002518 -12.15 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.06378 on 916 degrees of freedom  
## Multiple R-squared: 0.1387, Adjusted R-squared: 0.1378   
## F-statistic: 147.6 on 1 and 916 DF, p-value: < 2.2e-16  
##   
##   
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS  
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:  
## Level of Significance = 0.05   
##   
## Call:  
## gvlma(x = MLB)   
##   
## Value p-value Decision  
## Global Stat 2.9495 0.5663 Assumptions acceptable.  
## Skewness 0.6193 0.4313 Assumptions acceptable.  
## Kurtosis 1.2533 0.2629 Assumptions acceptable.  
## Link Function 0.1632 0.6862 Assumptions acceptable.  
## Heteroscedasticity 0.9137 0.3391 Assumptions acceptable.

All assumptions to our linear model are acceptable.

## Conclusion

The plot of team payroll rank and winning % showed a clear negative linear relationship. Meaning that as the Teams drop lower in the ranks of team payroll, their winning % is expected to also decrease. After performing all the diagnostic tests, the model passed each one confirming that it is a good model fit. The summary of the model showed a significant relationship between payroll rank and winning % at an = .05 level, therefore, we can reject the null hypothesis and conclude, that there is a relationship between where a team ranks in team payroll and their winning %. According to our model for every spot a team drops in team payroll rank, their winning % is expected to decrease by 0.3%, and over a 162 game season that would be alomost 15 less wins expected for the 30th ranked payroll vs the 1st ranked Payroll.