MLB Payroll Analysis

Nicholas Ezyk

21 March 2015

```
#loading the payroll data from the Python document
payroll <- read.table("~/Documents/payroll.txt", header=TRUE, quote="\"")

summary(payroll)
```

```
##                 TeamName  PayrollMillions    X2014Wins
##  ArizonaDiamondbacks: 1   Min.   : 69.50   Min.   :64.0
##  AtlantaBraves      : 1   1st Qu.: 96.67   1st Qu.:73.0
##  BaltimoreOrioles   : 1   Median :116.35   Median :80.5
##  BostonRedSox       : 1   Mean   :123.74   Mean   :81.0
##  ChicagoCubs        : 1   3rd Qu.:143.00   3rd Qu.:88.0
##  ChicagoWhiteSox    : 1   Max.   :265.90   Max.   :98.0
##  (Other)            :24
```

```
bank <- payroll$PayrollMillions
wins <- payroll$X2014Wins

#displaying the mean and sd of payroll and wins (out of 162, of course)
mean(bank)
```

```
## [1] 123.74
```

```
sd(bank)
```

```
## [1] 44.0318
```

```
mean(wins)
```

```
## [1] 81
```

```
sd(wins)
```

```
## [1] 9.598851
```

```
#setting a linear regression
reg <- lm(wins ~ bank)
summary(reg)
```

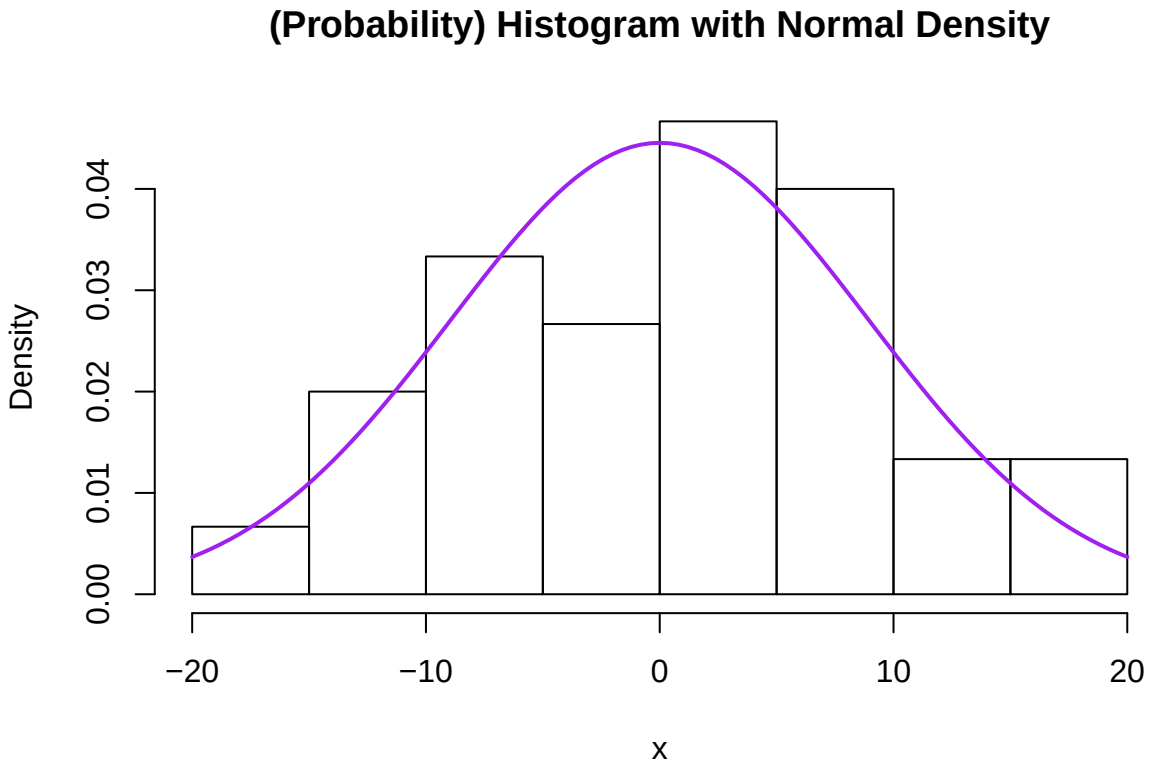
```
##
## Call:
## lm(formula = wins ~ bank)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -15.2053  -7.1873   0.1969   6.7572  15.2935
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.29010    5.03878  14.148 2.79e-14 ***
## bank         0.07847    0.03844   2.042   0.0507 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.114 on 28 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.09848
## F-statistic: 4.168 on 1 and 28 DF,  p-value: 0.05072
```

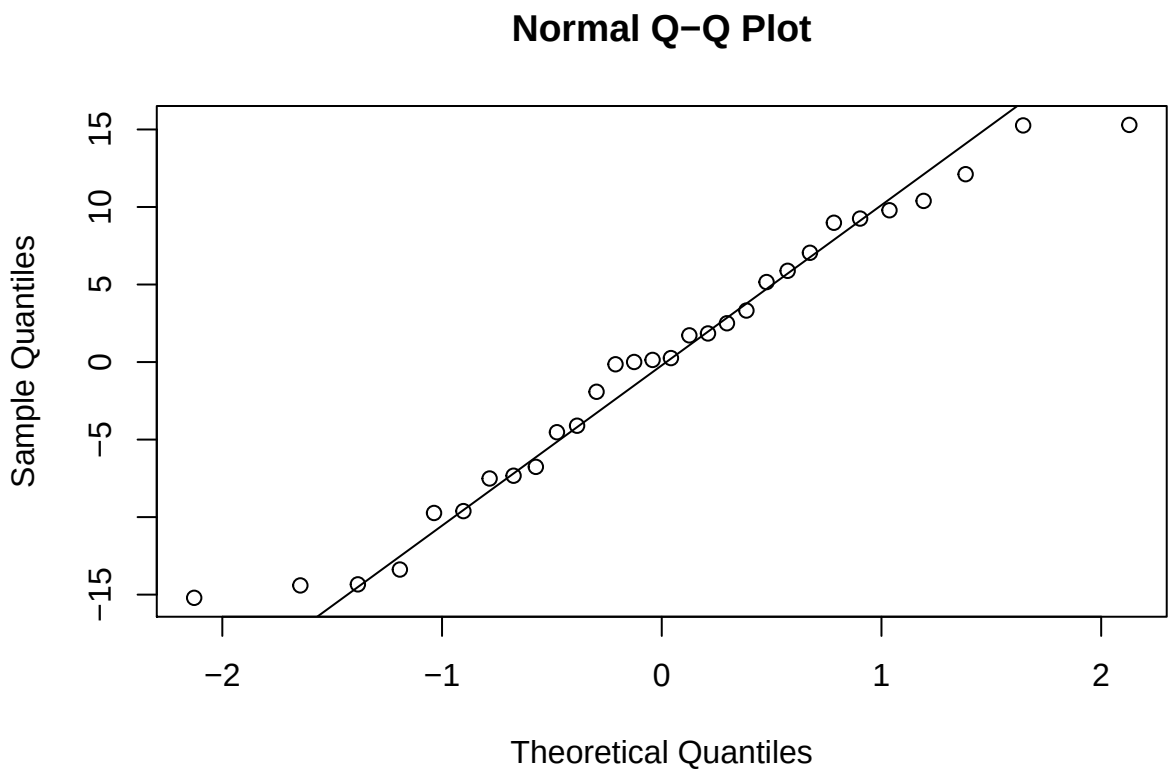
```r
#the regression is valid to significance < .10 (p-value .05072),
#but the R-squared is only .1296, a weak correlation

#a means of comparing the histogram to a normal distribution
histNorm <- function(x, densCol = "darkblue"){
  m <- mean(x)
  std <- sqrt(var(x))
  h <- max(hist(x,plot=FALSE)$density)
  d <- dnorm(x, mean=m, sd=std)
  maxY <- max(h,d)
  hist(x, prob=TRUE,
       xlab="x", ylim=c(0, maxY),
       main="(Probability) Histogram with Normal Density")
  curve(dnorm(x, mean=m, sd=std),
        col=densCol, lwd=2, add=TRUE)
}

#showing the histogram with normal distribution line
histNorm(reg$residuals, "purple")
```




```
#QQplots and Shapiro-Wilk test
qqnorm(reg$residuals)
qqline(reg$residuals)
```



```r
shapiro.test(reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.9637, p-value = 0.383
```

```r
#p-value is .383; this can be considered a normal distribution

plot(reg$fitted.values,reg$residuals)
abline(h = 0)
#variances are wide, but in a channel

install.packages("lmtest", repos="http://cran.rstudio.com/")
```

```
##
## The downloaded binary packages are in
##  /var/folders/jf/26_g5lyn0jz_5rktdj345bn00000gn/T//RtmpSMWR07/downloaded_packages
```

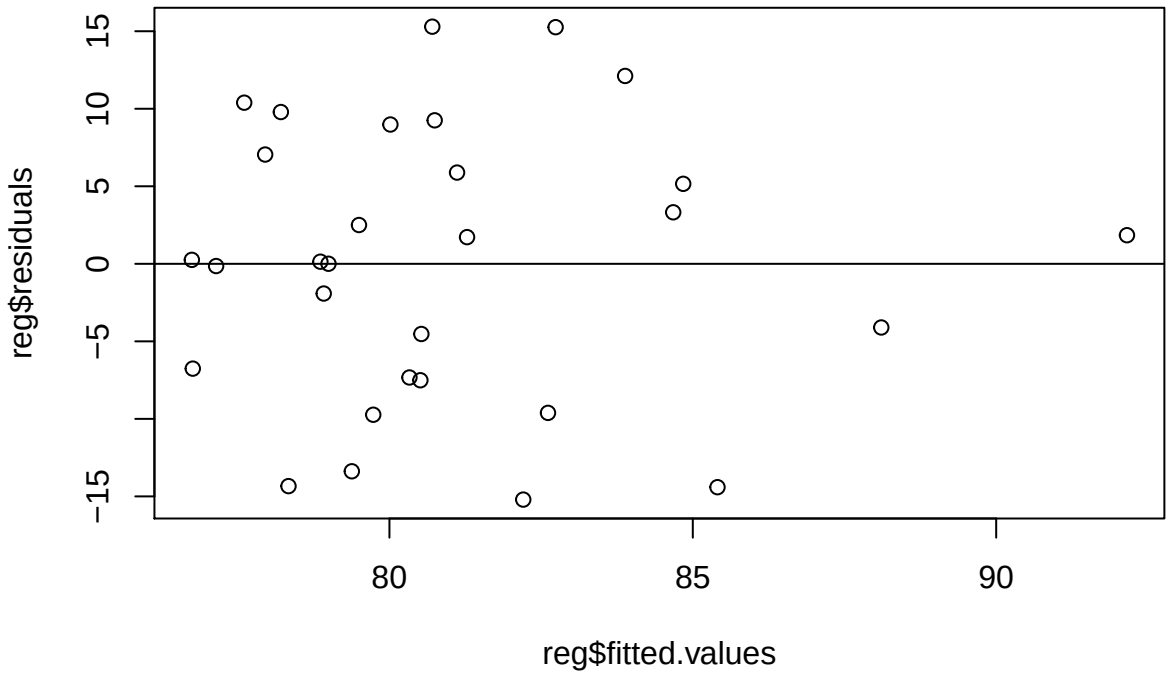
```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.1.2

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.1.2

##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```
bptest(reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 0.0362, df = 1, p-value = 0.849
```

```
#p-value of .849 give; we can assume variances are constant throughout the distribution

hats <- hatvalues(reg)

hatmu <- mean(hats)
hats[hats > 2 * hatmu]
```

```
##        14        19
## 0.3927712 0.1791952
```

```
#we get teams 14 and 19 with high leverage; the Dodgers and Yankees with their astronomical payrolls
```