DATA606 - Linear Regression Part 2

Jason Bryer, Ph.D.
November 14, 2018

Presentations

- Jeff Littlejohn (6.5) http://rpubs.com/jefflittlejohn/Data_606_Prob_6_5_Pres
- Asher Dvir-Djerassi (7.19)
- Adam Douglas (7.23) http://rpubs.com/lysanthus/DATA606HWPresentation
- Sergio Ortega-Cruz (7.41) http://rpubs.com/sortega78/439610

NYS Report Card

NYS publishes data for each school in the state. We will look at the grade 8 math scores for 2012 and 2013. 2013 was the first year the tests were aligned with the Common Core Standards. There was a lot of press about how the passing rates for most schools dropped. Two questions we wish to answer:

1. Did the passing rates drop in a predictable manner?
2. Were the drops different for charter and public schools?

```
load('../Data/NYSReportCard-Grade7Math.Rda')
head(reportCard, n=4)
```

```
##       BEDSCODE                                School NumTested2012 Mean2012 Pass2012 Charter
## 1 010100010020                  NORTH ALBANY ACADEMY            47      649       13   FALSE
## 2 010100010030         WILLIAM S HACKETT MIDDLE SCHOOL           212      652       30   FALSE
## 3 010100010045 STEPHEN AND HARRIET MYERS MIDDLE SCHOOL           262      670       50   FALSE
## 4 010100860867         KIPP TECH VALLEY CHARTER SCHOOL            61      684       85    TRUE
##   GradeSubject County                                BOCES NumTested2013 Mean2013 Pass2013
## 1 Grade 7 Math Albany BOCES ALBANY-SCHOH-SCHENECTADY-SARAT            45      268        0
## 2 Grade 7 Math Albany BOCES ALBANY-SCHOH-SCHENECTADY-SARAT           250      279        9
## 3 Grade 7 Math Albany BOCES ALBANY-SCHOH-SCHENECTADY-SARAT           256      284        8
## 4 Grade 7 Math Albany BOCES ALBANY-SCHOH-SCHENECTADY-SARAT            59      298        9
```

Descriptive Statistics

```
summary(reportCard$Pass2012)
```

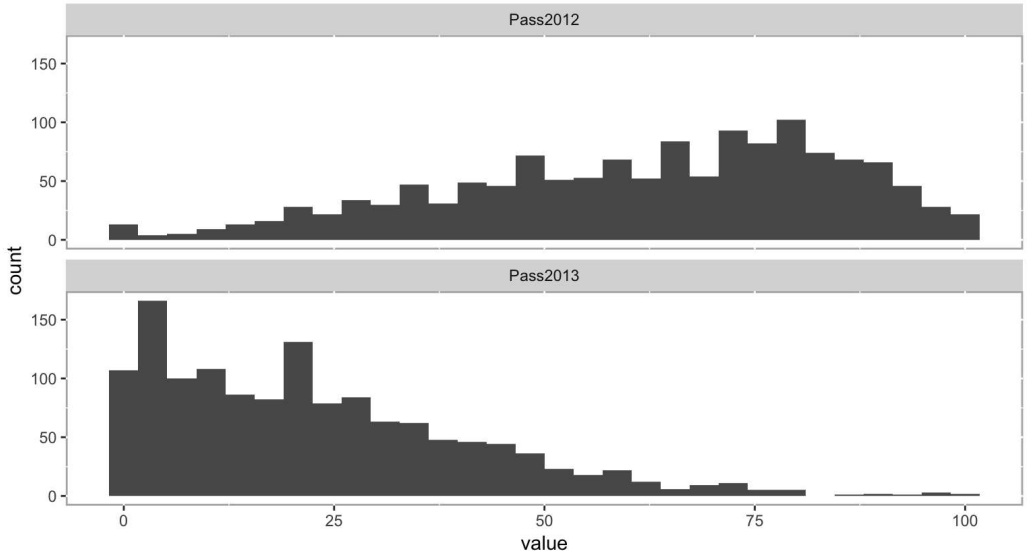
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   46.00   65.00   61.73   80.00  100.00
```

```
summary(reportCard$Pass2013)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   20.00   22.83   33.00   99.00
```

Histograms

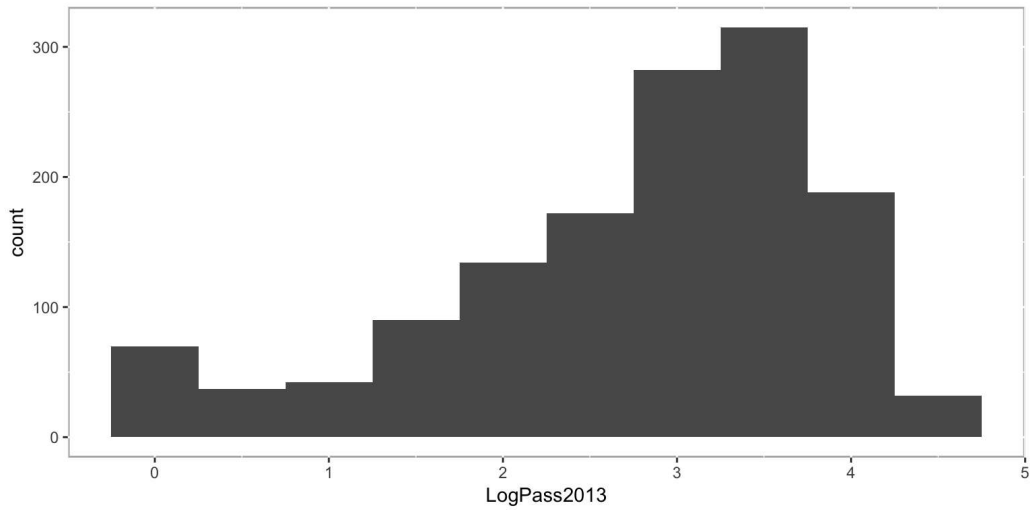
```
melted <- melt(reportCard[,c('Pass2012', 'Pass2013')])
ggplot(melted, aes(x=value)) + geom_histogram() + facet_wrap(~ variable, ncol=1)
```



Log Transformation

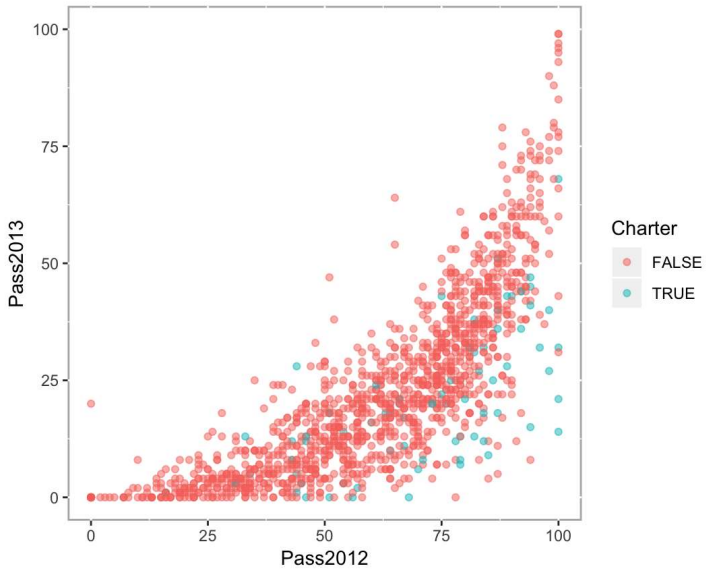
Since the distribution of the 2013 passing rates is skewed, we can log transfor that variable to get a more reasonably normal distribution.

```
reportCard$LogPass2013 <- log(reportCard$Pass2013 + 1)
ggplot(reportCard, aes(x=LogPass2013)) + geom_histogram(binwidth=0.5)
```



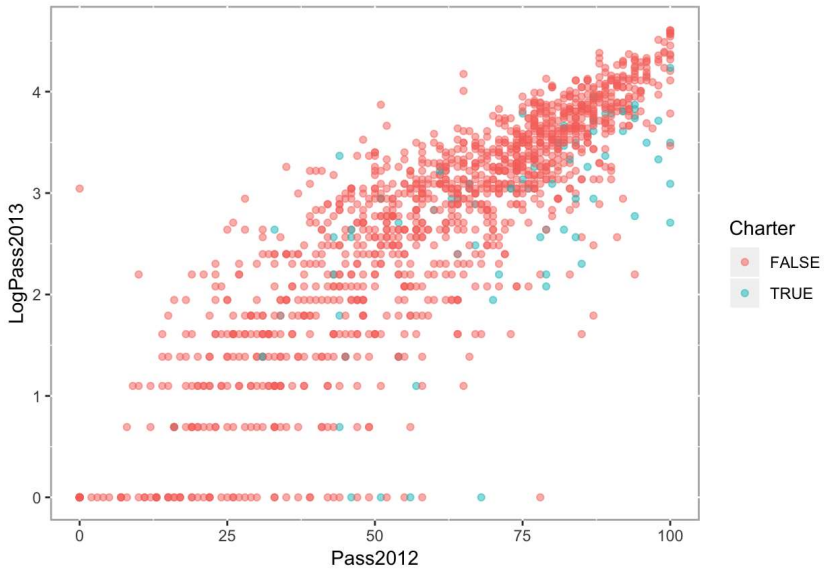
Scatter Plot

```
ggplot(reportCard, aes(x=Pass2012, y=Pass2013, color=Charter)) +
    geom_point(alpha=0.5) + coord_equal() + ylim(c(0,100)) + xlim(c(0,100))
```



Scatter Plot (log transform)

```
ggplot(reportCard, aes(x=Pass2012, y=LogPass2013, color=Charter)) +
    geom_point(alpha=0.5) + xlim(c(0,100)) + ylim(c(0, log(101)))
```



Correlation

```
cor.test(reportCard$Pass2012, reportCard$Pass2013)
```

```
##
##  Pearson's product-moment correlation
##
## data:  reportCard$Pass2012 and reportCard$Pass2013
## t = 47.166, df = 1360, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7667526 0.8071276
## sample estimates:
##       cor
## 0.7877848
```

Correlation (log transform)

```
cor.test(reportCard$Pass2012, reportCard$LogPass2013)
```

```
##
##  Pearson's product-moment correlation
##
## data:  reportCard$Pass2012 and reportCard$LogPass2013
## t = 56.499, df = 1360, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8207912 0.8525925
## sample estimates:
##       cor
## 0.8373991
```

Linear Regression

```
lm.out <- lm(Pass2013 ~ Pass2012, data=reportCard)
summary(lm.out)
```

```
##
## Call:
## lm(formula = Pass2013 ~ Pass2012, data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.484  -6.878  -0.478   5.965  51.675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -16.68965    0.89378  -18.67   <2e-16 ***
## Pass2012      0.64014    0.01357   47.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.49 on 1360 degrees of freedom
## Multiple R-squared:  0.6206, Adjusted R-squared:  0.6203
## F-statistic:  2225 on 1 and 1360 DF,  p-value: < 2.2e-16
```

Linear Regression (log transform)

```
lm.log.out <- lm(LogPass2013 ~ Pass2012, data=reportCard)
summary(lm.log.out)
```

```
##
## Call:
## lm(formula = LogPass2013 ~ Pass2012, data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3880 -0.2531  0.0776  0.3461  2.7368
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.307692   0.046030   6.685 3.37e-11 ***
## Pass2012    0.039491   0.000699  56.499  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5915 on 1360 degrees of freedom
## Multiple R-squared:  0.7012, Adjusted R-squared:  0.701
## F-statistic:  3192 on 1 and 1360 DF,  p-value: < 2.2e-16
```

Did the passing rates drop in a predictable manner?

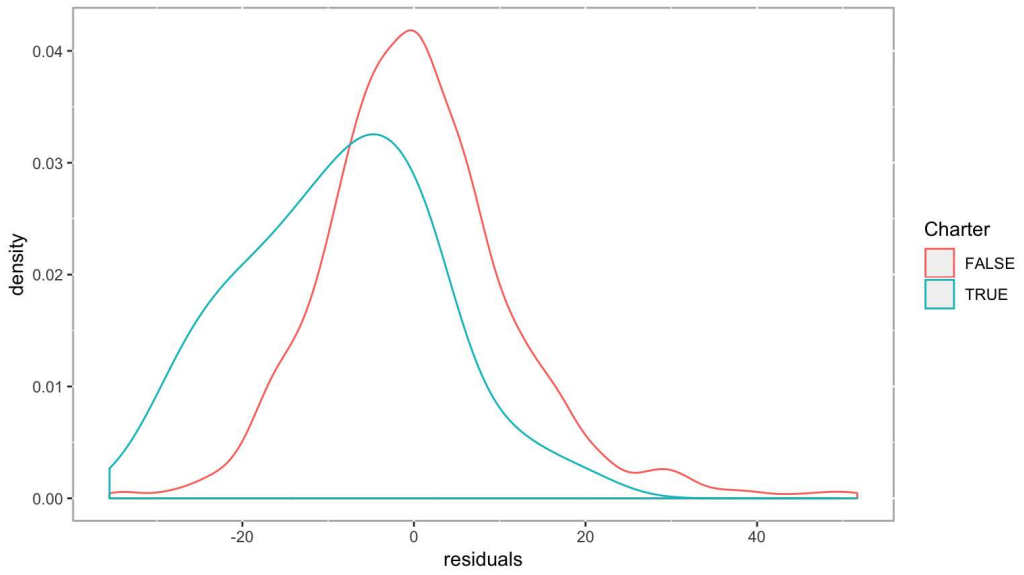
Yes! Whether we log tranform the data or not, the correlations are statistically significant with regression models with R^2 creater than 62%.

To answer the second question, whether the drops were different for public and charter schools, we'll look at the residuals.

```
reportCard$residuals <- resid(lm.out)
reportCard$residualsLog <- resid(lm.log.out)
```

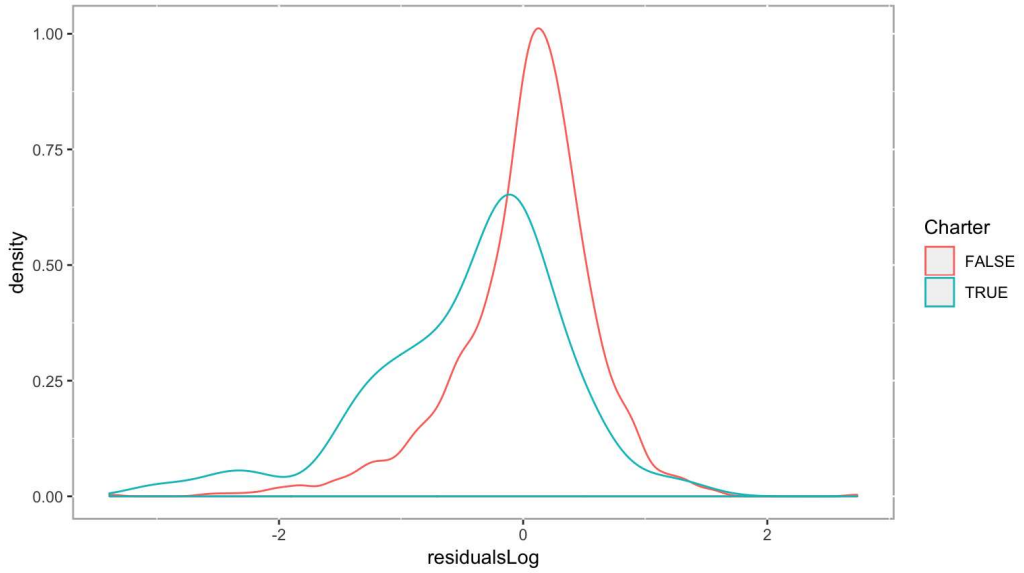
Distribution of Residuals

```
ggplot(reportCard, aes(x=residuals, color=Charter)) + geom_density()
```



Distribution of Residuals

```
ggplot(reportCard, aes(x=residualsLog, color=Charter)) + geom_density()
```



Null Hypothesis Testing

H_0: There is no difference in the residuals between charter and public schools.

H_A: There is a difference in the residuals between charter and public schools.

```
t.test(residuals ~ Charter, data=reportCard)
```

```
##
##  Welch Two Sample t-test
##
## data:  residuals by Charter
## t = 6.5751, df = 77.633, p-value = 5.091e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.411064 11.980002
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            0.479356           -8.716177
```

Null Hypothesis Testing (log transform)

```
t.test(residualsLog ~ Charter, data=reportCard)
```

```
##
##  Welch Two Sample t-test
##
## data:  residualsLog by Charter
## t = 4.7957, df = 74.136, p-value = 8.161e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2642811 0.6399761
## sample estimates:
## mean in group FALSE  mean in group TRUE
##          0.02356911         -0.42855946
```

Quadratic Models

It is possible to fit quatric models fairly easily in R, say of the following form:

$$y = b_1 x^2 + b_0$$

```
quad.out <- lm(Pass2013 ~ I(Pass2012^2), data=reportCard)
summary(quad.out)$r.squared
```

```
## [1] 0.6945532
```

```
summary(lm.out)$r.squared
```

```
## [1] 0.6206049
```

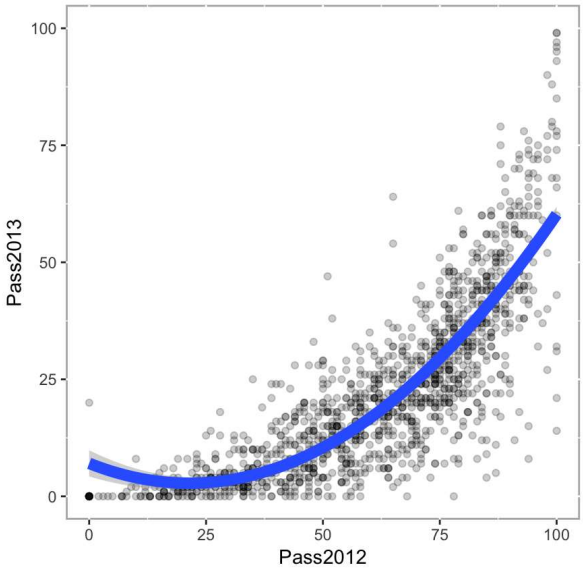
Quadratic Models

```
summary(quad.out)
```

```
##
## Call:
## lm(formula = Pass2013 ~ I(Pass2012^2), data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.315  -5.322   0.106   5.058  42.685
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.8155478  0.5391020  -5.223 2.04e-07 ***
## I(Pass2012^2)  0.0059130  0.0001063  55.610  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.31 on 1360 degrees of freedom
## Multiple R-squared:  0.6946, Adjusted R-squared:  0.6943
## F-statistic:  3092 on 1 and 1360 DF,  p-value: < 2.2e-16
```

Scatter Plot

```
ggplot(reportCard, aes(x=Pass2012, y=Pass2013)) +
    geom_point(alpha=0.2) + geom_smooth(method='lm', formula=y~poly(x,2,raw=TRUE), size=3) +
    coord_equal() + ylim(c(0,100)) + xlim(c(0,100))
```



Shiny App

```
shiny::runGitHub('NYSchools','jbryer',subdir='NYSReportCard')
```

See also the Github repository for more information: https://github.com/jbryer/NYSchools