Linear Regression Part 2

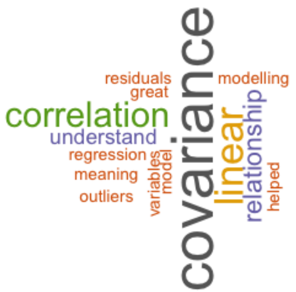
DATA 606 - Statistics & Probability for Data Analytics

Jason Bryer, Ph.D.

April 14, 2021

One Minute Paper Results

What was the most important thing you learned during this class?



What important question remains unanswered for you?

Homework Presentations

- 8.22 Tage Singh
- 8.25 Henry Owens

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('../course_data/NYSReportCard-Grade7Math.Rda')
names(reportCard)
```

```
##  [1] "BEDSCODE"      "School"        "NumTested2012" "Mean2012"      "Pass2012"      "Charter"       "GradeSubject
##  [8] "County"        "BOCES"         "NumTested2013" "Mean2013"      "Pass2013"
```

`reportCard` Data Frame

Show 3 entries Search:

BEDSCODE	School	NumTested2012	Mean2012	Pass2012	Charter	GradeSubject	County	BOCES	NumTested2013	Mean2013
010100010020	NORTH ALBANY ACADEMY	47	649	13	false	Grade 7 Math	Albany	BOCES ALBANY-SCHOH-SCHENECTADY-SARAT	45	268
010100010030	WILLIAM S HACKETT MIDDLE SCHOOL	212	652	30	false	Grade 7 Math	Albany	BOCES ALBANY-SCHOH-SCHENECTADY-SARAT	250	279
010100010045	STEPHEN AND HARRIET MYERS MIDDLE SCHOOL	262	670	50	false	Grade 7 Math	Albany	BOCES ALBANY-SCHOH-SCHENECTADY-SARAT	256	284

Showing 1 to 3 of 1,362 entries

Previous 1 2 3 4 5 … 454 Next

Descriptive Statistics

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

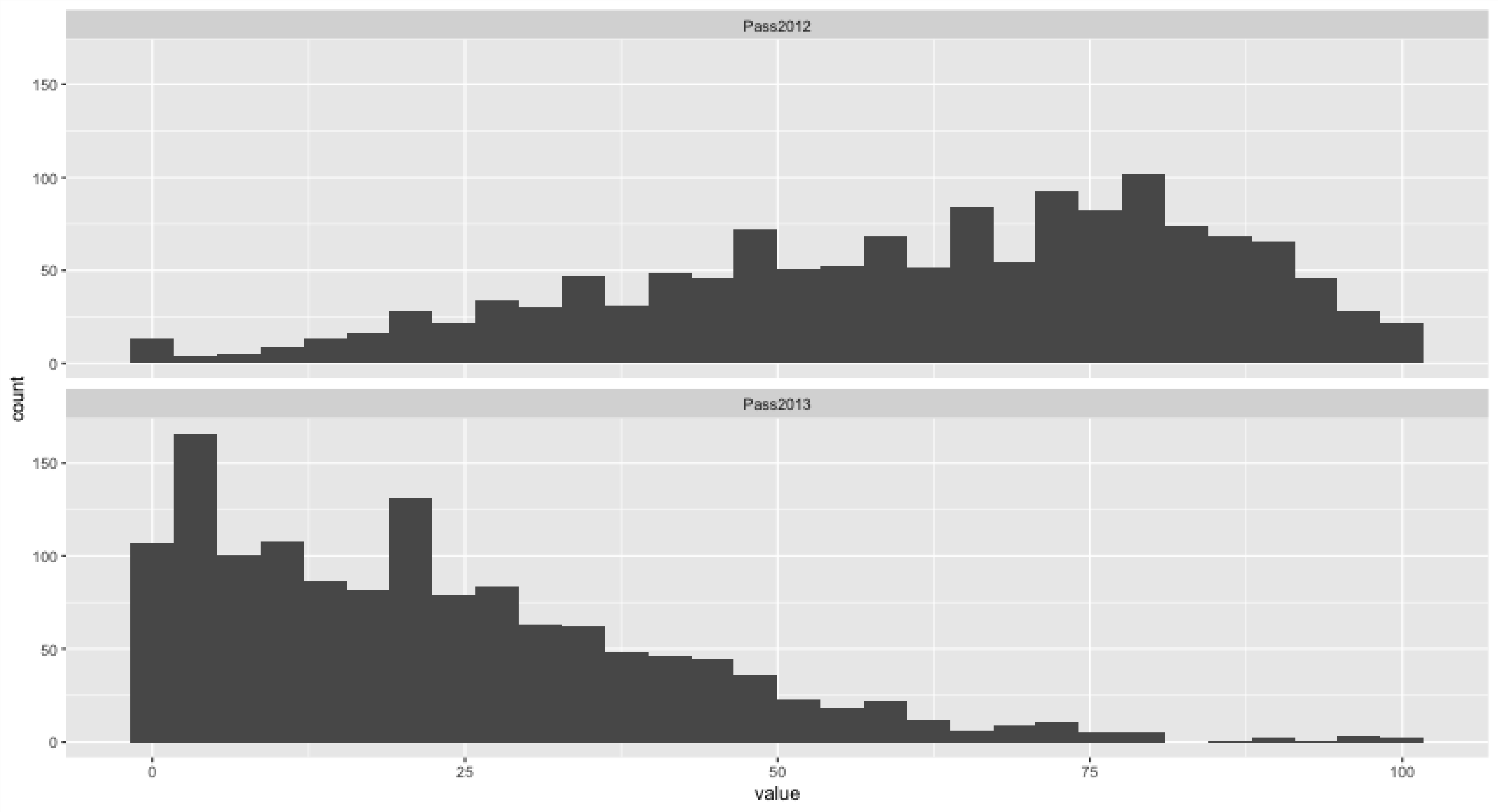
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.00   46.00   65.00   61.73   80.00  100.00
```

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

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

Histograms

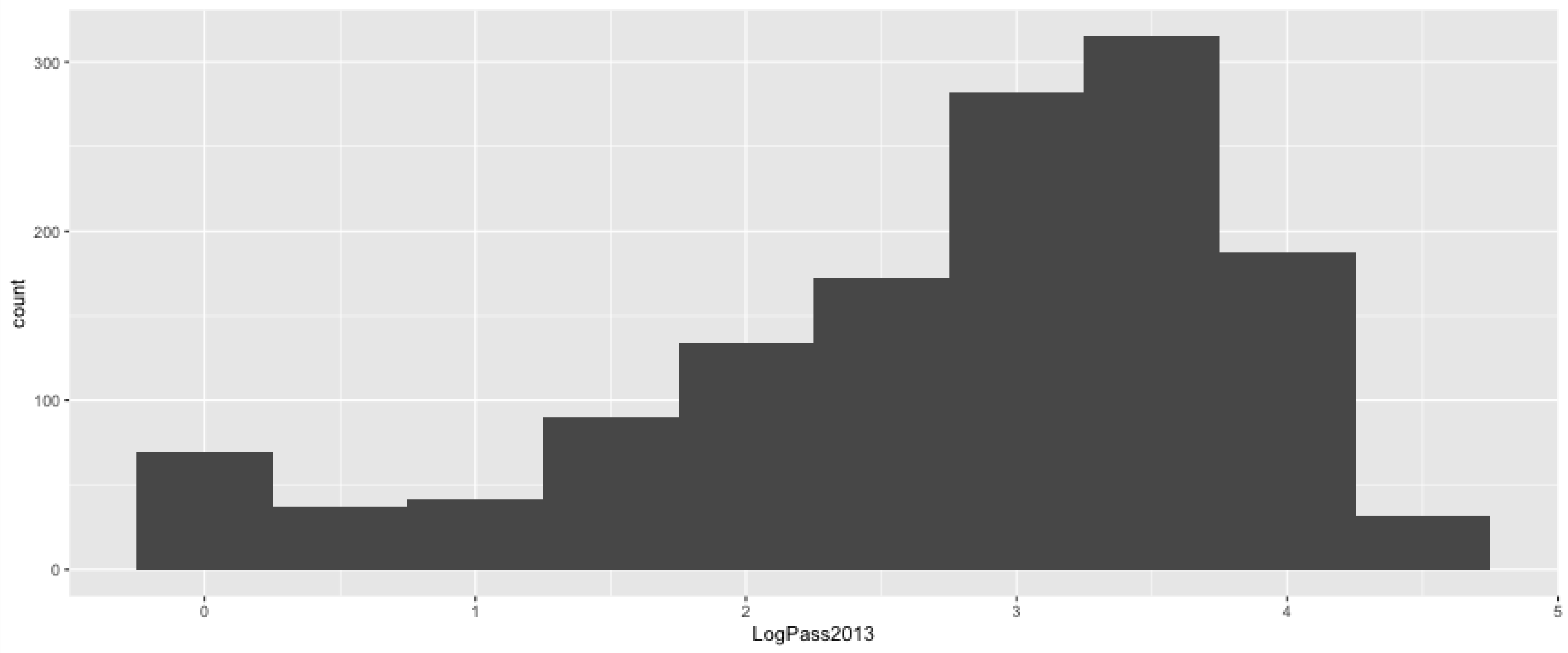
```r
melted <- melt(reportCard[,c('Pass2012', 'Pass2013')])
ggplot(melted, aes(x=value)) + geom_histogram() + facet_wrap(~ variable, ncol=1)
```



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)

```r
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%.

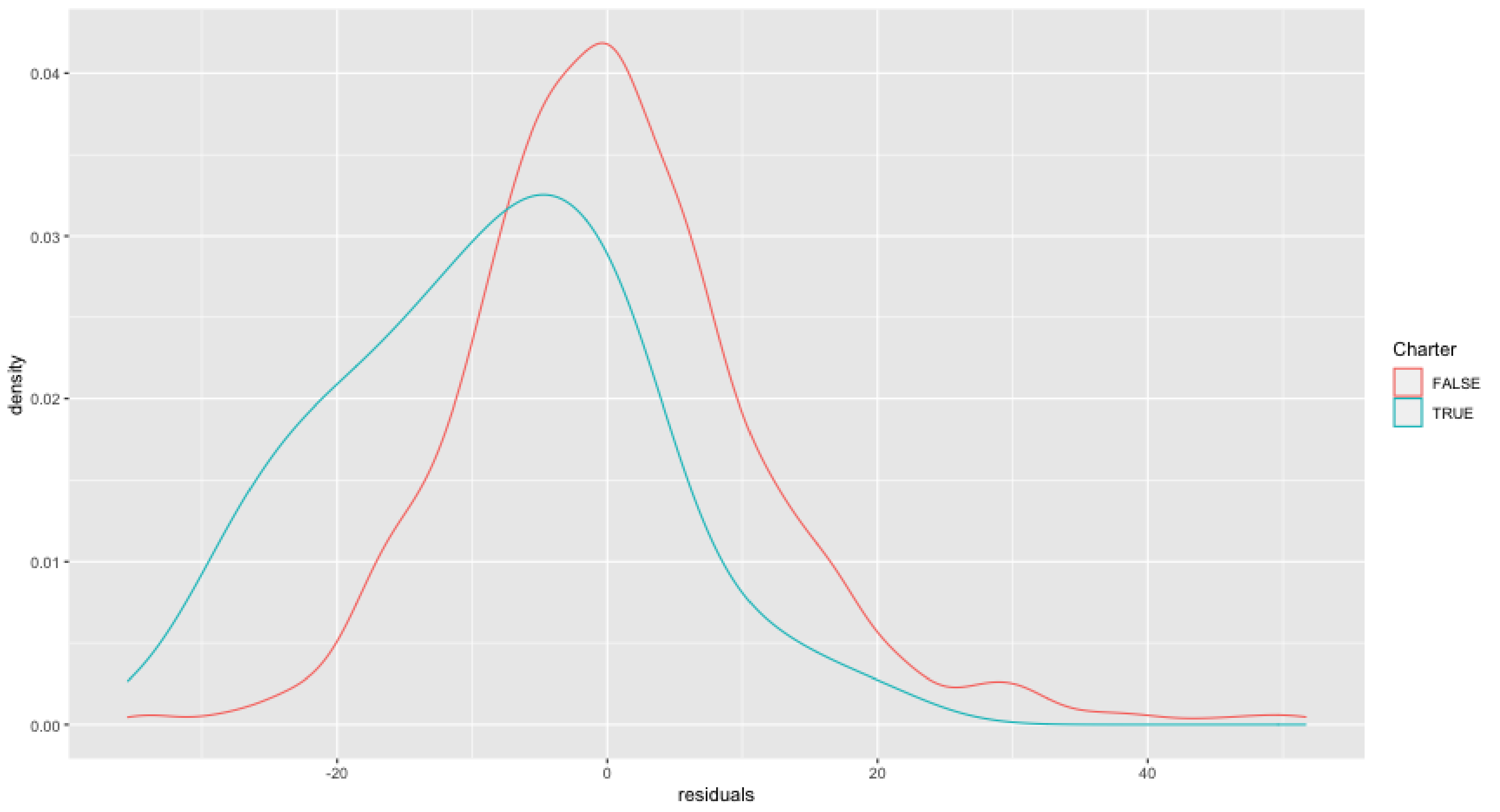
To answer the second question, whether the drops were different for public and charter schools, we'll look at the residuals.

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

DATA 606
Spring 2021

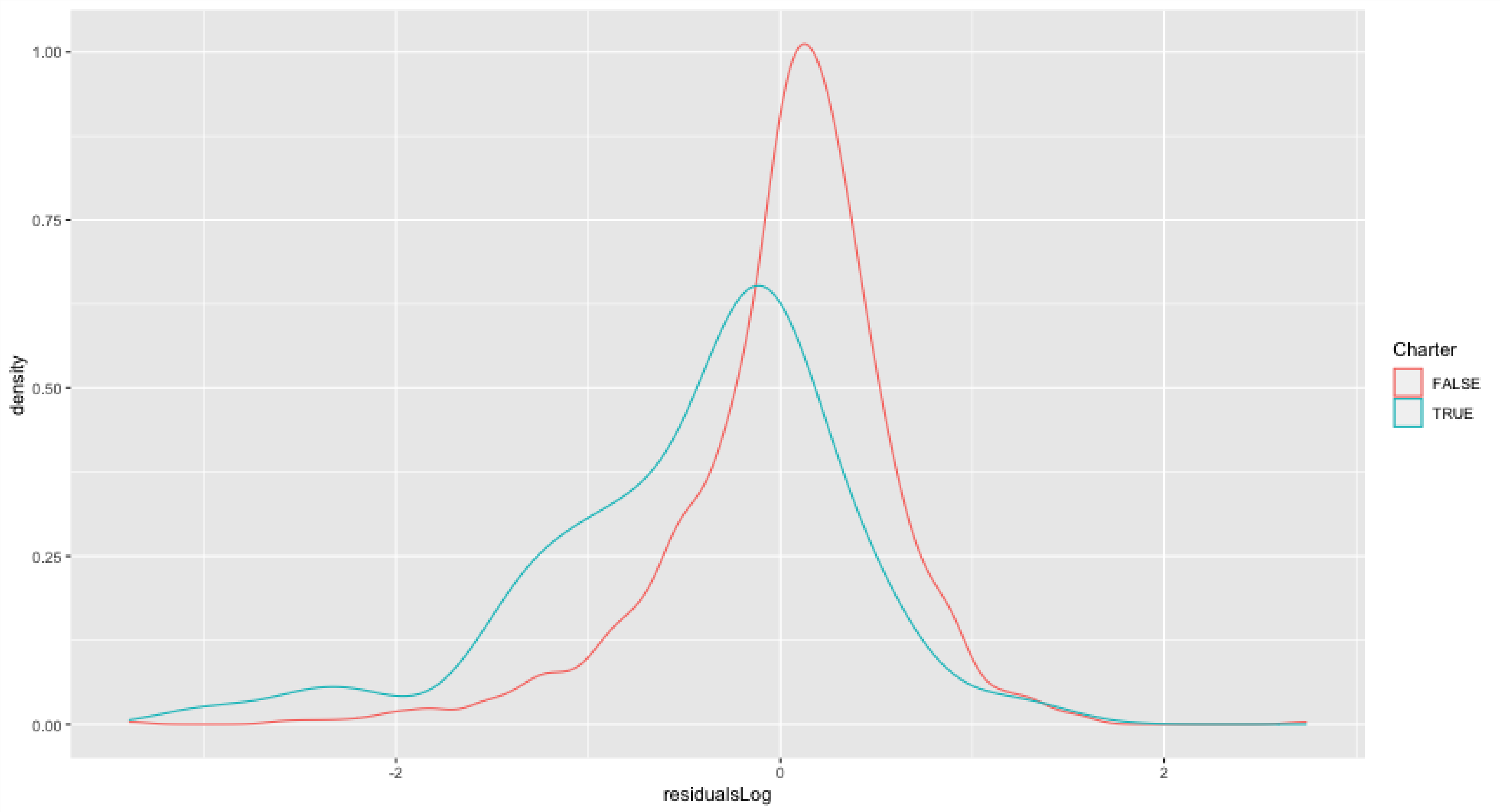
Distribution of Residuals

```
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
```

DATA 606 Spring 2021

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
```

Polynomial Models (e.g. Quadratic)

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

$$y = b_1 x^2 + b_2 x + b_0$$

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

```
## [1] 0.7065206
```

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

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

DATA 606
Spring 2021

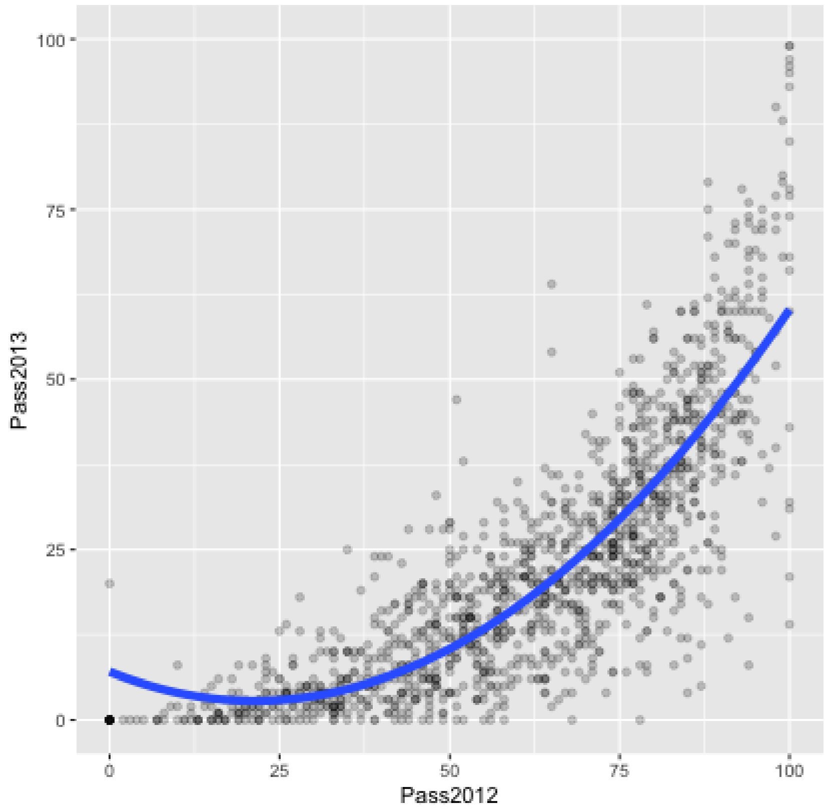
Quadratic Model

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

```
##
## Call:
## lm(formula = Pass2013 ~ I(Pass2012^2) + Pass2012, data = reportCard)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.258  -4.906  -0.507   5.430  43.509
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.0466153  1.4263773   4.940 8.77e-07 ***
## I(Pass2012^2)  0.0092937  0.0004659  19.946  < 2e-16 ***
## Pass2012      -0.3972481  0.0533631  -7.444 1.72e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 1359 degrees of freedom
## Multiple R-squared:  0.7065,    Adjusted R-squared:  0.7061
## F-statistic:  1636 on 2 and 1359 DF,  p-value: < 2.2e-16
```

Scatter Plot

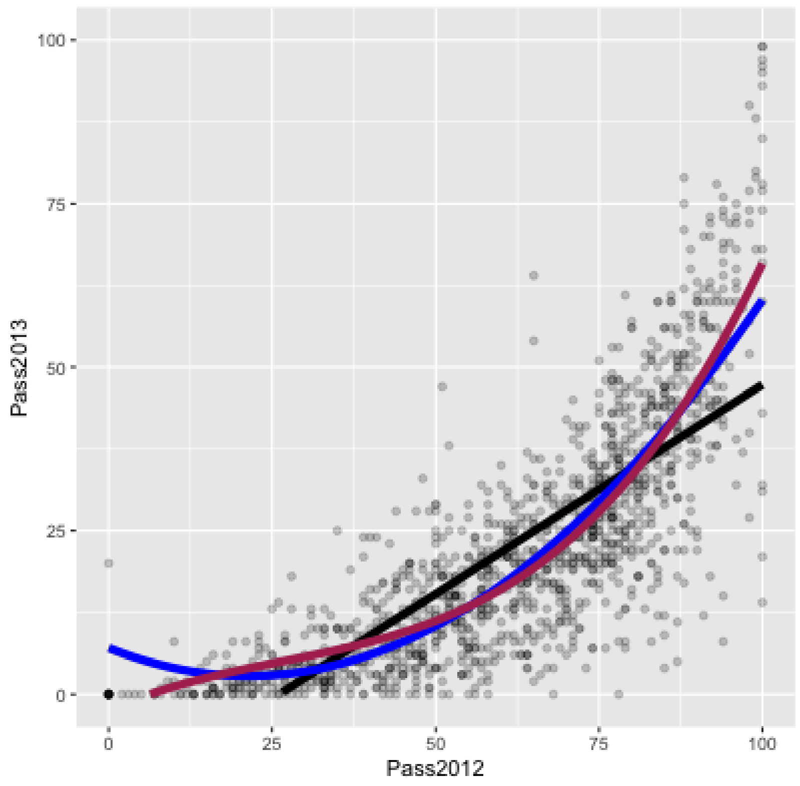
```
ggplot(reportCard, aes(x=Pass2012, y=Pass2013)) + geom_point(alpha=0.2) +
    geom_smooth(method='lm', formula=y ~ I(x^2) + x, size=2, se=FALSE) +
    coord_equal() + ylim(c(0,100)) + xlim(c(0,100))
```



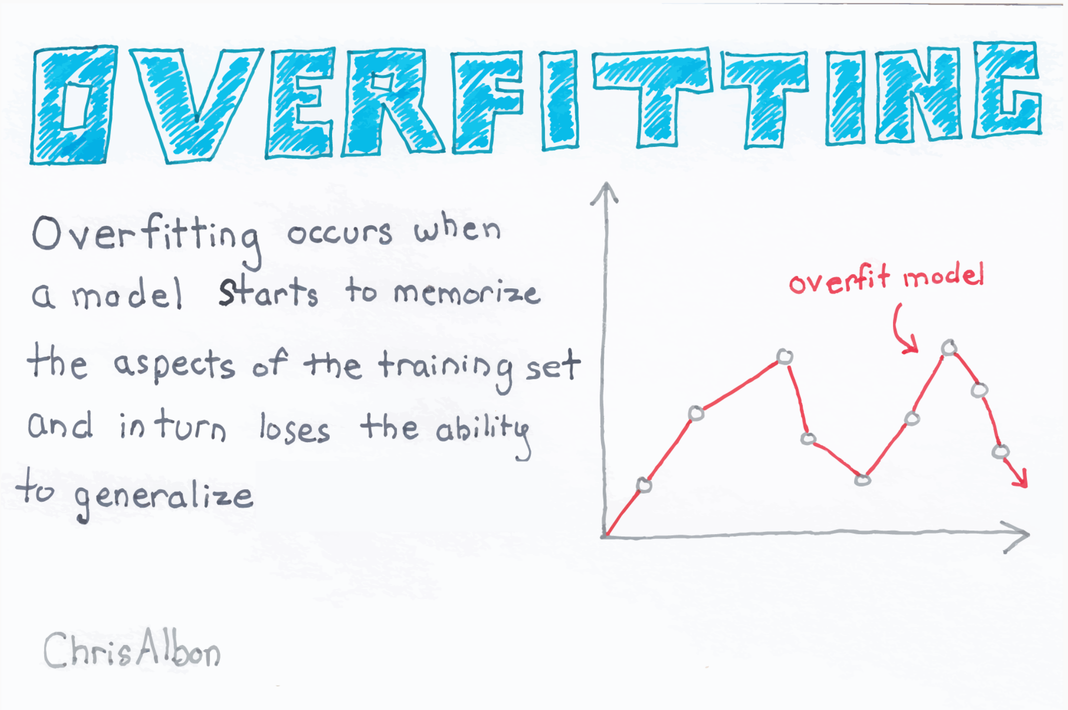
Let's go crazy, cubic!

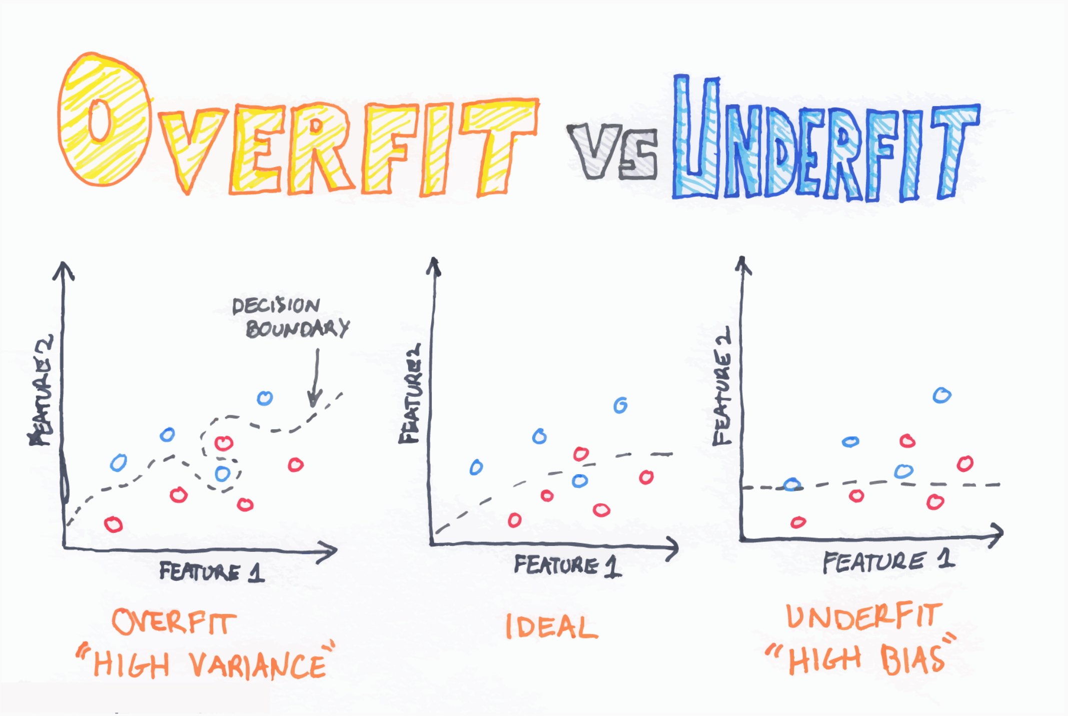
```
cube.out <- lm(Pass2013 ~ I(Pass2012^3) + I(Pass2012^2) + Pass2012, data=reportCard)
summary(cube.out)$r.squared
```

```
## [1] 0.7168206
```



Be careful of overfitting...

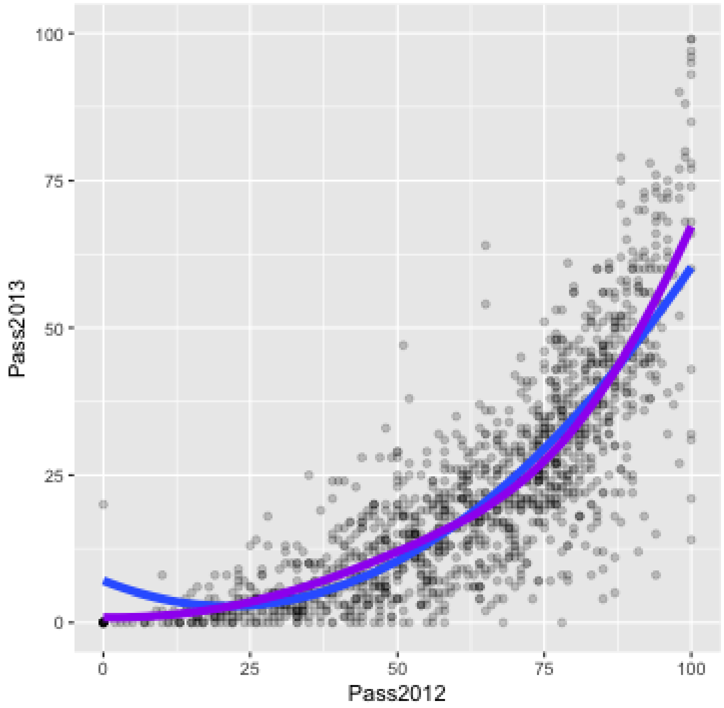




Source: Chris Albon @chrisalbon MachineLearningFlashCards.com

Loess Regression

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



See this site for more info: http://varianceexplained.org/files/loess.html

Shiny App

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

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

One Minute Paper

Complete the one minute paper:

https://forms.gle/gY9SeBCPggHEtZYw6

1. What was the most important thing you learned during this class?
2. What important question remains unanswered for you?