Stats Lab 9

Jack Wright

10/24/2020

Load Packages

```
library(tidyverse)
library(openintro)
library(stats)
```

1. What are the dimensions of the dataset?

```
dim<-dim(hfi)
cat("the dimensions of the hfi study are ",dim[1]," rows by ",dim[2]," columns.")
```

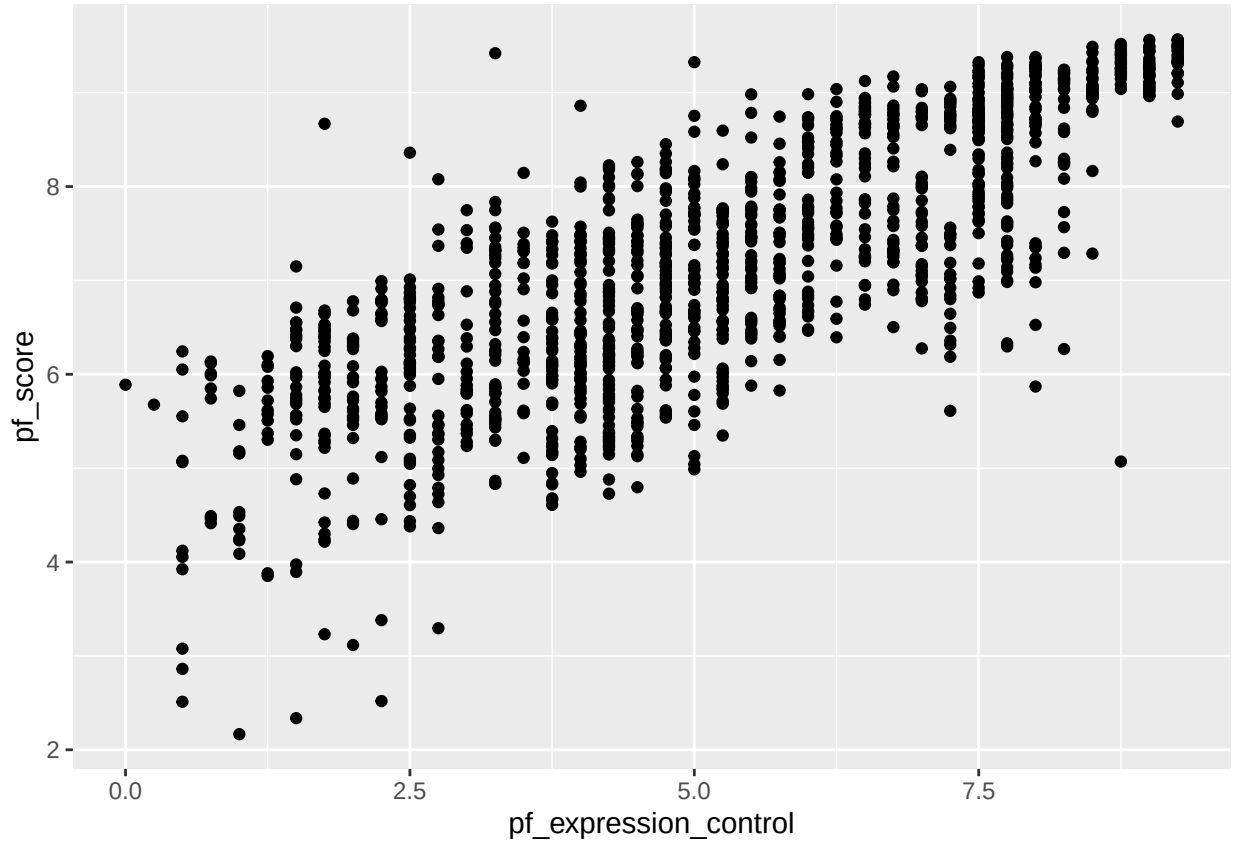
```
## the dimensions of the hfi study are  1458  rows by  123  columns.
```

1. What type of plot would you use to display the relationship between the personal freedom score, `pf_score`, and one of the other numerical variables? Plot this relationship using the variable `pf_expression_control` as the predictor. Does the relationship look linear? If you knew a country's `pf_expression_control`, or its score out of 10, with 0 being the most, of political pressures and controls on media content, would you be comfortable using a linear model to predict the personal freedom score?

I would use a scatter plot.

```
ggplot(data=hfi, aes(x=pf_expression_control, y=pf_score))+
  geom_point()
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



Yes it does look linear. I would be comfortable modeling with a linear regression.

```
hfi %>%
  summarise(cor(pf_expression_control, pf_score, use = "complete.obs"))
```

```
## # A tibble: 1 x 1
##   `cor(pf_expression_control, pf_score, use = "complete.obs")`
##                                                          <dbl>
## 1                                                        0.796
```

```
hfi<-hfi
```

Sum of squared residuals

1. Looking at your plot from the previous exercise, describe the relationship between these two variables. Make sure to discuss the form, direction, and strength of the relationship as well as any unusual observations.

The relationship looks to be positively correlated. The center looks to be around <5,6> (the means for each variable) and the data looks to be right skew

(removing na's from hifi)

```
hfi2<-hfi%>%
  filter_at(vars(pf_expression_control,pf_score),all_vars(!is.na(.)))
```

sum of squares

1: 1031 2: 1103 3: 958

1. Using `plot_ss`, choose a line that does a good job of minimizing the sum of squares. Run the function several times. What was the smallest sum of squares that you got? How does it compare to your neighbors?

The smallest sum of squares I got was 958. which compared favorably to the other tries I got, but was about 10% smaller on average.

The linear model

```
m1 <- lm(pf_score ~ pf_expression_control, data = hfi)
summary(m1)
```

```
##
## Call:
## lm(formula = pf_score ~ pf_expression_control, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8467 -0.5704  0.1452  0.6066  3.2060
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.61707    0.05745   80.36   <2e-16 ***
## pf_expression_control  0.49143    0.01006   48.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8318 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.6342, Adjusted R-squared:  0.634
## F-statistic:  2386 on 1 and 1376 DF,  p-value: < 2.2e-16
```

1. Fit a new model that uses `pf_expression_control` to predict `hf_score`, or the total human freedom score. Using the estimates from the R output, write the equation of the regression line. What does the slope tell us in the context of the relationship between human freedom and the amount of political pressure on media content?

```
m2<-lm(hf_score ~ pf_expression_control, data=hfi)
summary(m2)
```

```
##
## Call:
## lm(formula = hf_score ~ pf_expression_control, data = hfi)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6198 -0.4908  0.1031  0.4703  2.2933
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.153687   0.046070  111.87   <2e-16 ***
## pf_expression_control 0.349862   0.008067   43.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.667 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.5775, Adjusted R-squared:  0.5772
## F-statistic:  1881 on 1 and 1376 DF,  p-value: < 2.2e-16
```

The slope tells us that there is a positive correlation between political pressure on media content and the human freedom score.

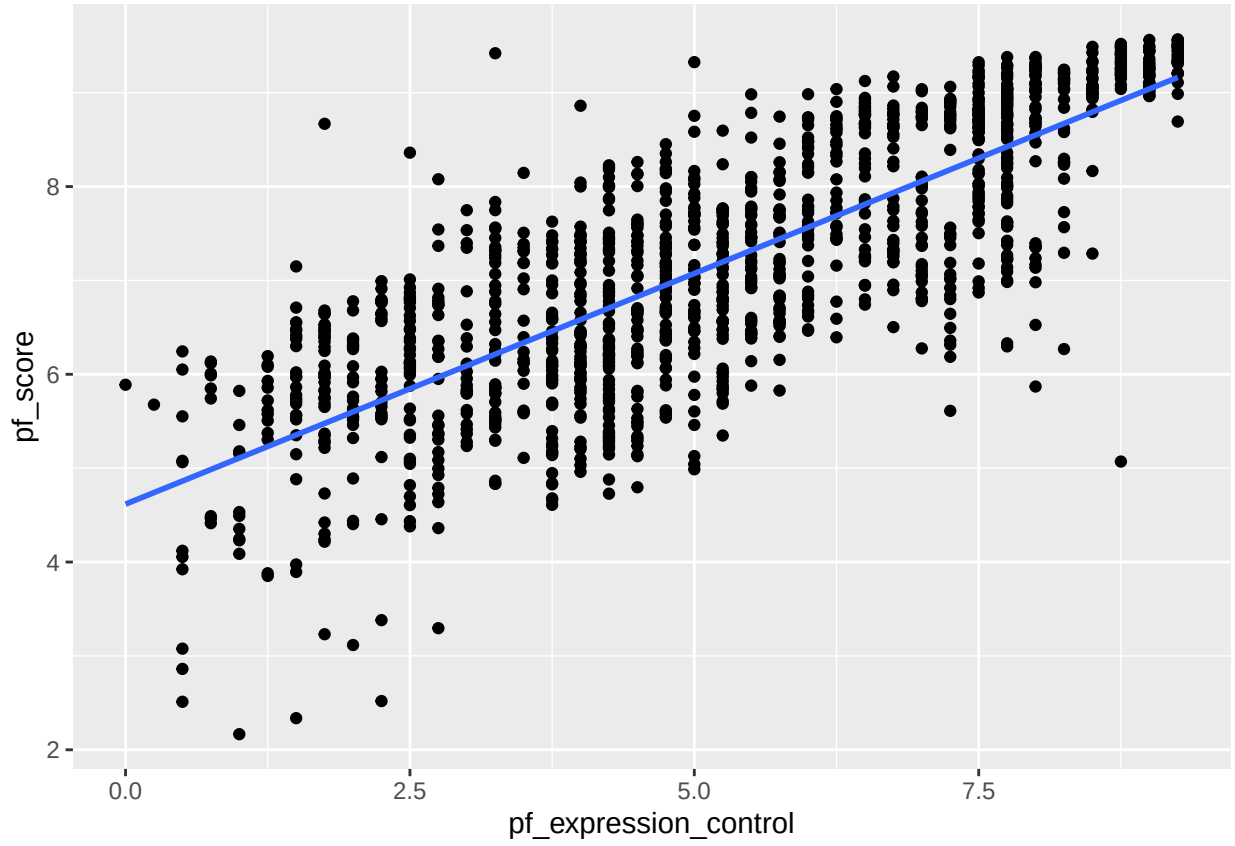
prediction and prediction errors

```
ggplot(data = hfi, aes(x = pf_expression_control, y = pf_score)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



1. If someone saw the least squares regression line and not the actual data, how would they predict a country's personal freedom school for one with a 6.7 rating for `pf_expression_control`? Is this an overestimate or an underestimate, and by how much? In other words, what is the residual for this prediction?

If they saw the regression line, they would say that the estimate of the pf_score would be 7.909.

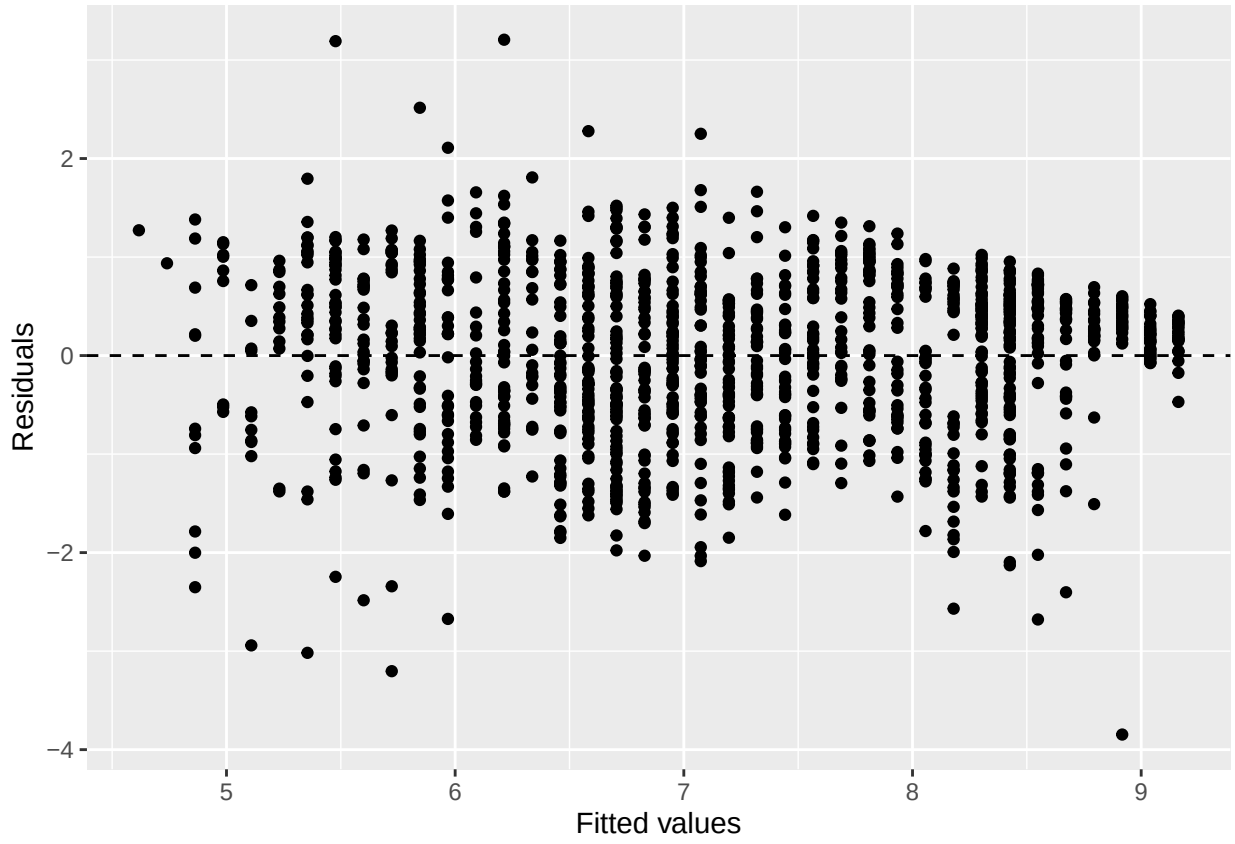
```
tmp<-hfi%>%
  filter(pf_expression_control>6.49 & pf_expression_control<6.76)%>%
  select(pf_score)

mean<-unname(unlist(summarise(tmp, mean(pf_score)) ))

residual<-7.909-mean

cat("the residual is ",residual)
```

```
## the residual is  -0.1883223
```

The difference between the result of the regression line and the mean of the data points between 6.5 and 7.5 is -.188.

Model diagnostics

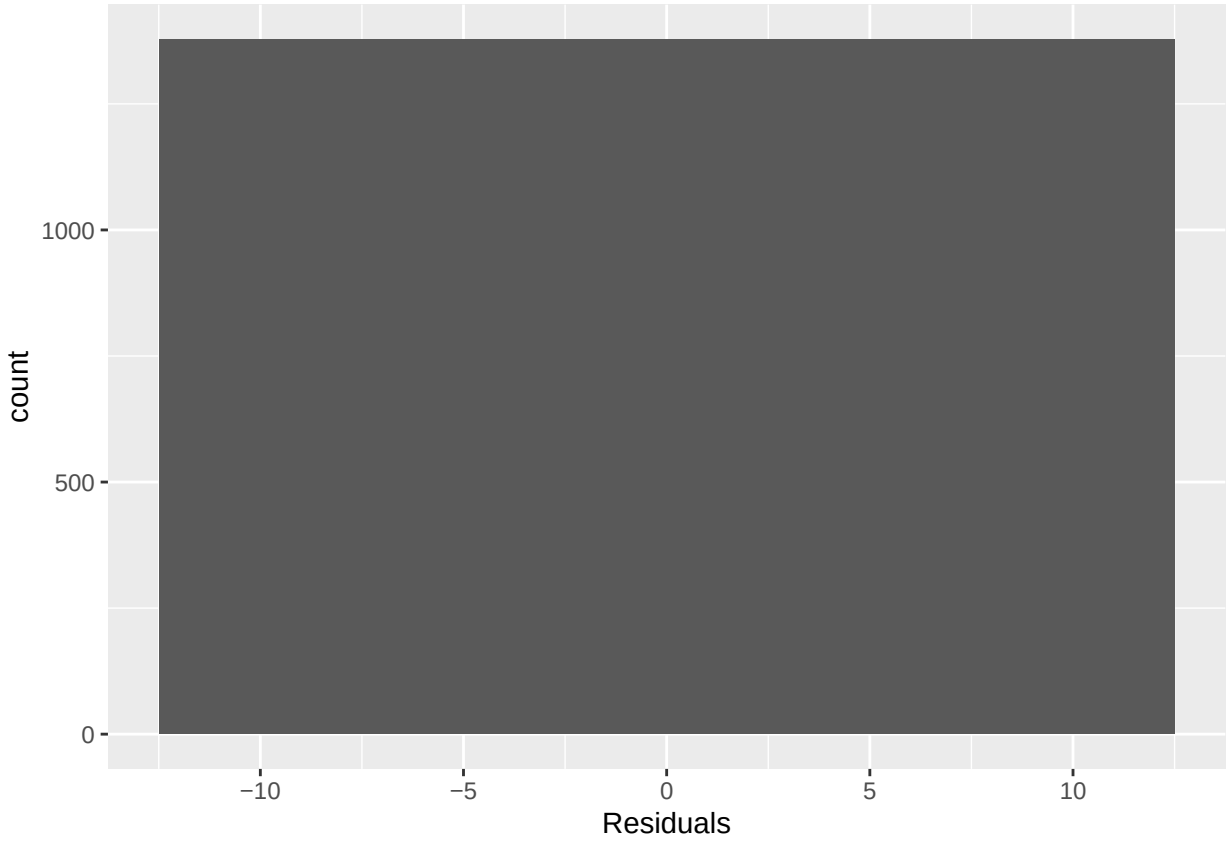
```
ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



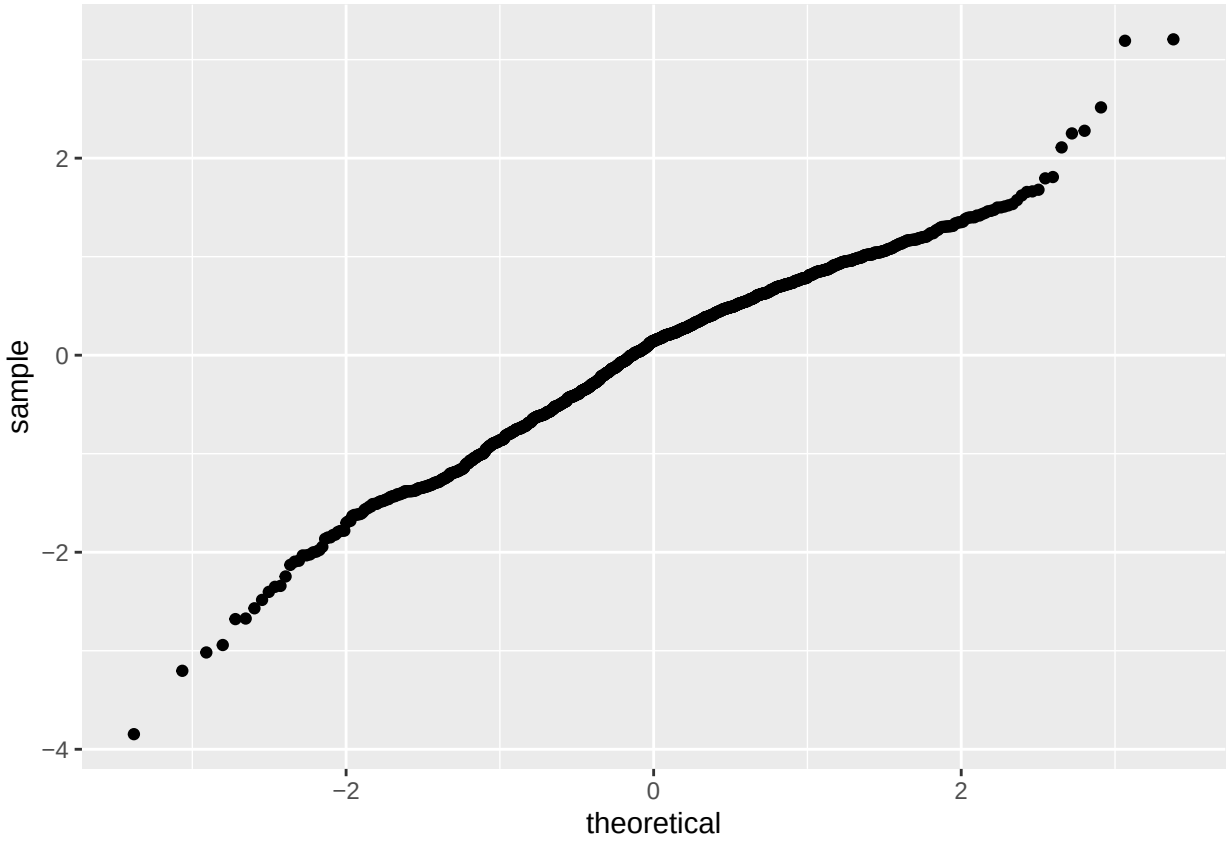
It does not look like there is a pattern in the residuals, which is good. If there were a pattern in the residuals, it would mean our model is "missing" some trend in the data. For example if the data was exponentail but we were fitting with a line we would see an increase in residuals as the exponent appreciated.

nearly normal residuals

```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = 25) +
  xlab("Residuals")
```




```
ggplot(data = m1, aes(sample = .resid)) +
  stat_qq()
```



1.

the data does seem to trend off of the qq line at the tails, which might represent some skew.

more practice

I want to look at the relationship between homicide and the pf_score.

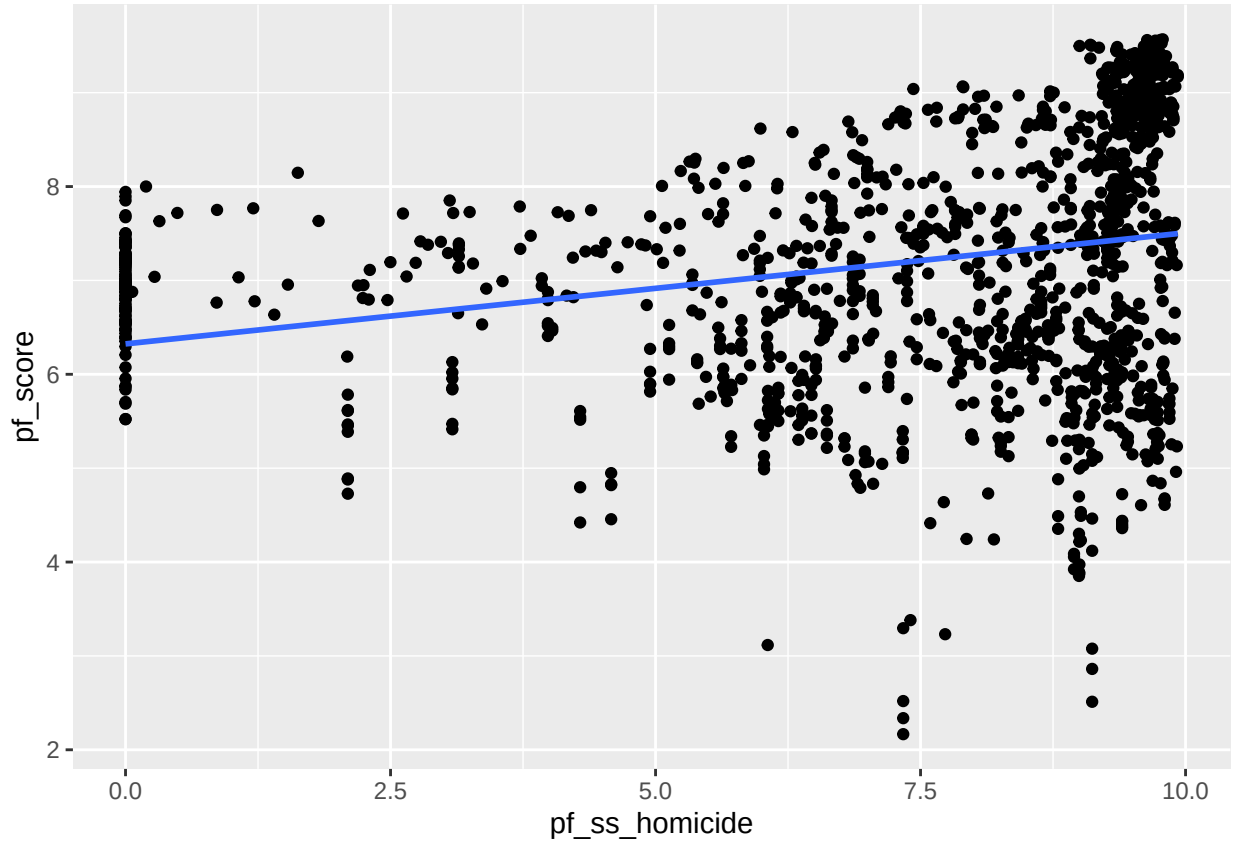
I believe there will be a strong negative corrolation between them.

```
ggplot(data = hfi, aes(x = pf_ss_homicide, y = pf_score)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 80 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 80 rows containing missing values (geom_point).
```



There does not appear to be a linear relationship. the fanning out of data as the homicide rate increases and the cluster of scores at zero keep this from having a linear relationship.

2.

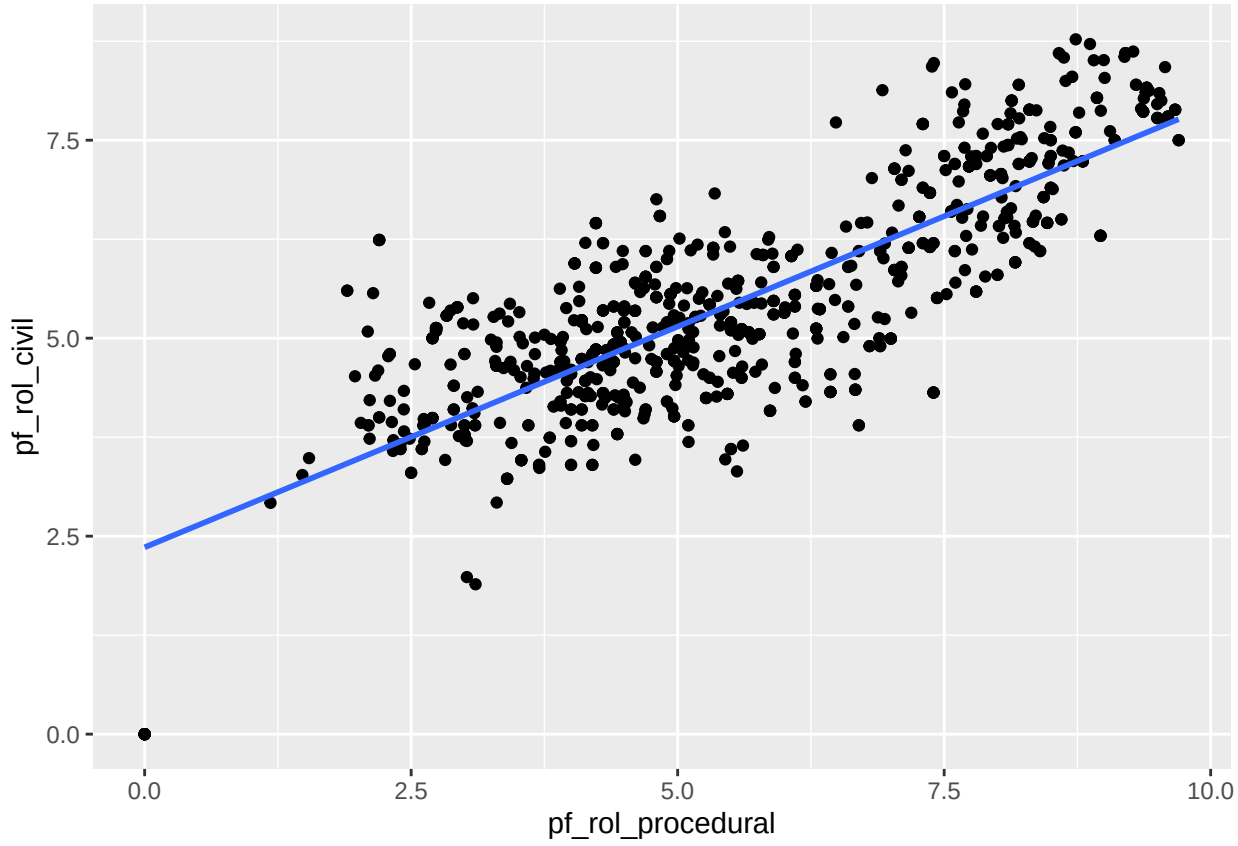
I want to look at the relationship between civil justice and procedural justice. I believe there will be a strong positive correlation

```
ggplot(data = hfi, aes(x = pf_rol_procedural, y = pf_rol_civil)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 578 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 578 rows containing missing values (geom_point).
```



It looks like the data is farily linear,although there is a high leverage point at <0,0>

```
m1<-lm(pf_rol_procedural ~ pf_rol_civil, data=hfi)
summary(m1)
```

```
##
## Call:
## lm(formula = pf_rol_procedural ~ pf_rol_civil, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2945 -0.7505  0.1179  0.8373  3.1838
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.8860     0.1624  -5.456 6.33e-08 ***
## pf_rol_civil   1.1828     0.0287  41.211  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.215 on 878 degrees of freedom
##   (578 observations deleted due to missingness)
## Multiple R-squared:  0.6592, Adjusted R-squared:  0.6588
## F-statistic:  1698 on 1 and 878 DF,  p-value: < 2.2e-16
```

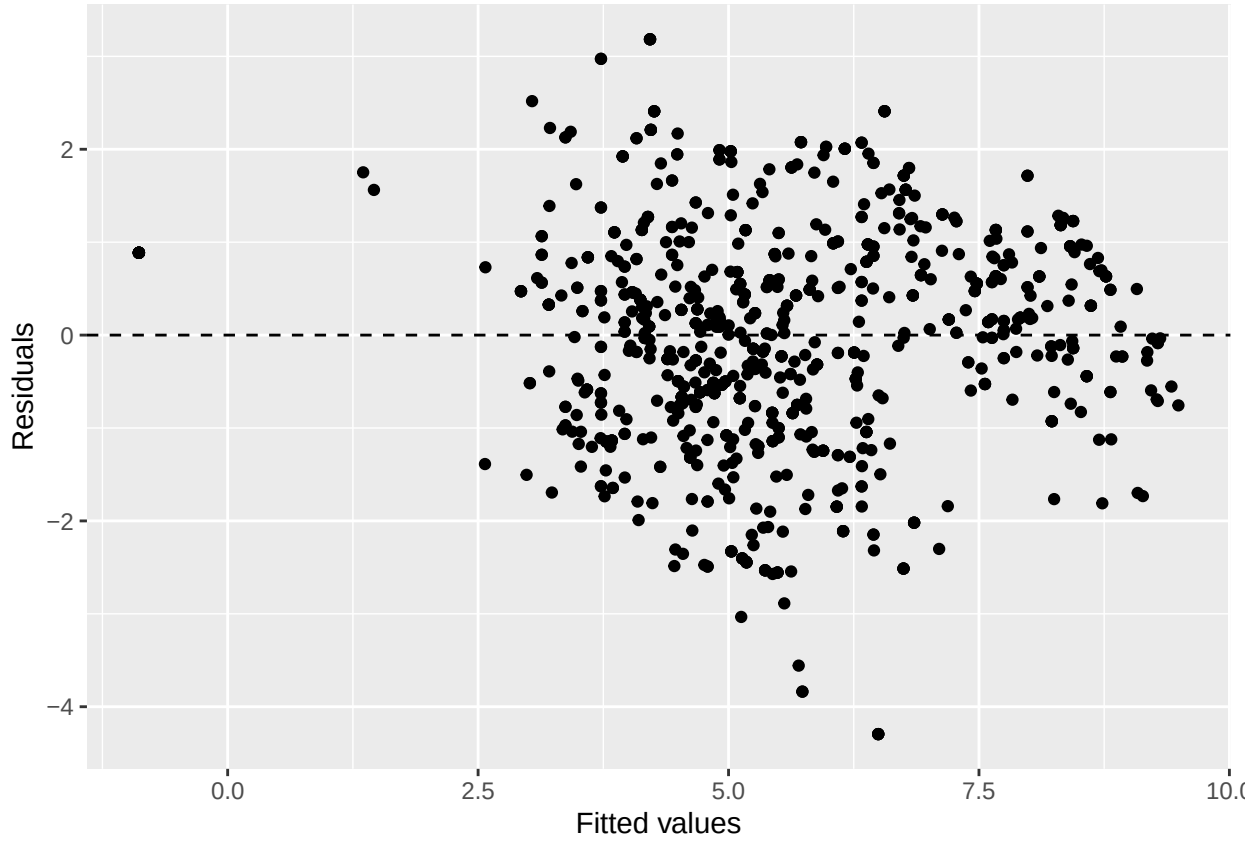
The R^2 for rule of law civil and procedural is .653, while the freedom score is .633. My independent variable

predicts my dependent one better, very slightly. Since the R^2 is higher, it is accounting for more of the variance than the example the assignment is based on.

3.

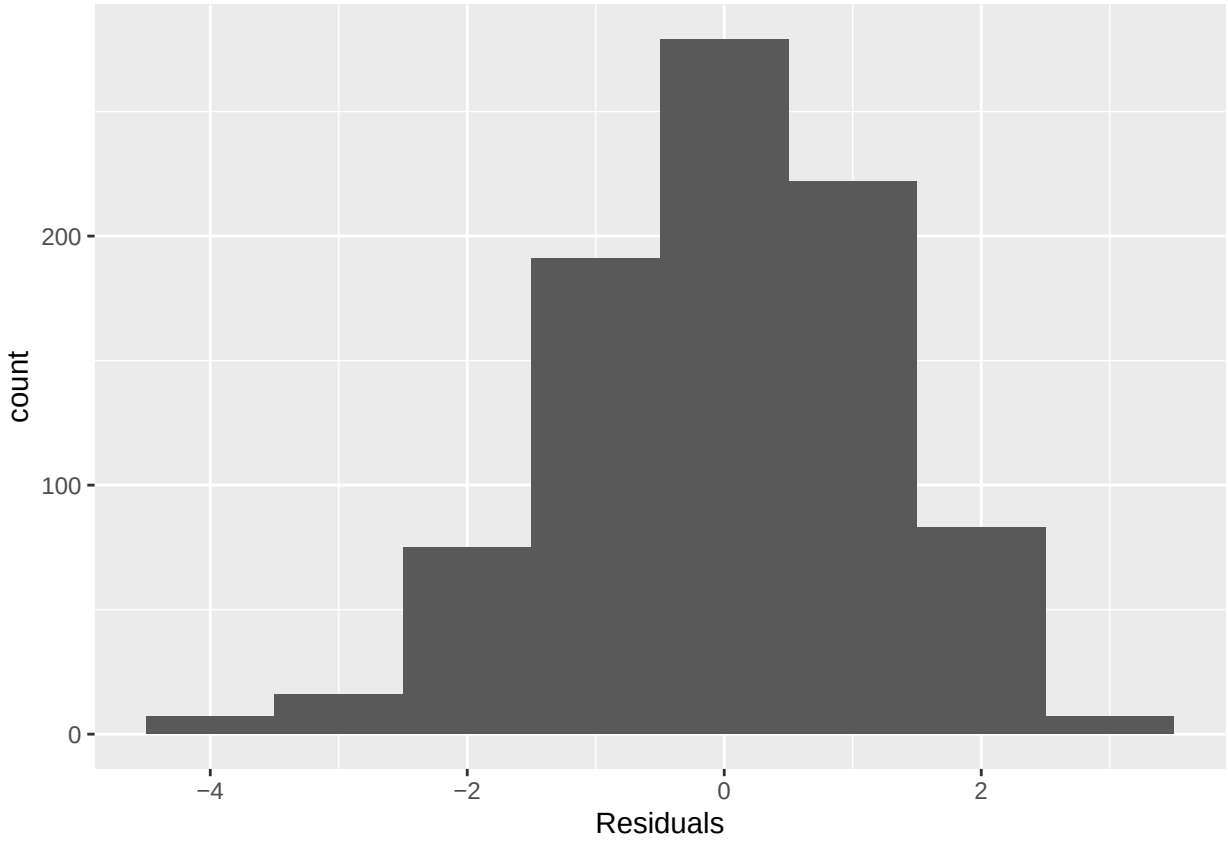
model diagnostics for rule of law civil vs procedural

```
ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



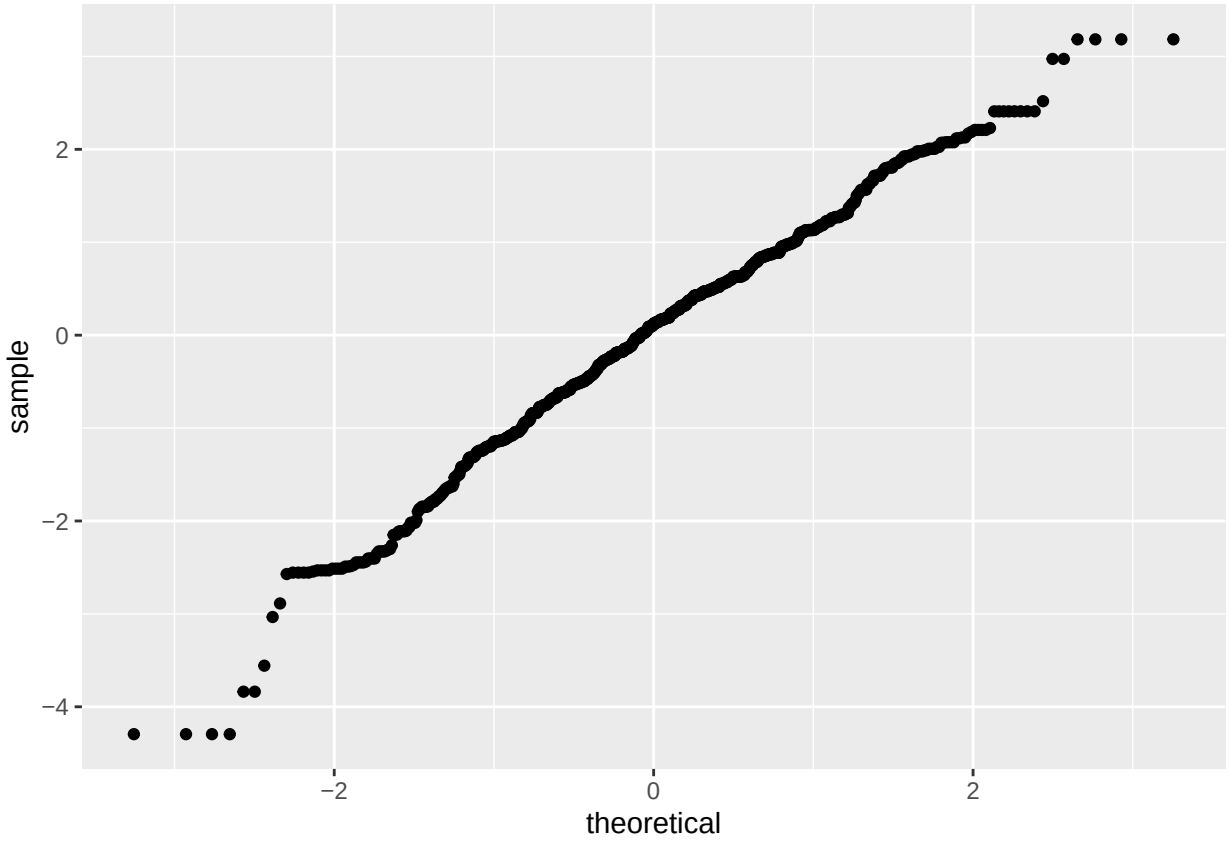
There seems to be no trend in the residuals which is good.

```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  xlab("Residuals")
```



the residuals are clustered around zero, without many outliers. which is good.

```
ggplot(data = m1, aes(sample = .resid)) +
  stat_qq()
```



The qq plot looks very linear, except for outliers at the tails, which is a result of the skew. but this data is normal.