Data 606 - Lab 8 - Introduction to linear regression

Avery Davidowitz

2022-11-12

Import Libraries

```
library(tidyverse)
library(openintro)
data('hfi', package='openintro')
```

Exercise 1

What are the dimensions of the dataset?

```
glimpse(hfi)
```

```
## Rows: 1,458
## Columns: 123
## $ year                                <dbl> 2016, 2016, 2016, 2016, 2016, 2016,~
## $ ISO_code                            <chr> "ALB", "DZA", "AGO", "ARG", "ARM", ~
## $ countries                           <chr> "Albania", "Algeria", "Angola", "Ar~
## $ region                              <chr> "Eastern Europe", "Middle East & No~
## $ pf_rol_procedural                   <dbl> 6.661503, NA, NA, 7.098483, NA, 8.4~
## $ pf_rol_civil                        <dbl> 4.547244, NA, NA, 5.791960, NA, 7.5~
## $ pf_rol_criminal                     <dbl> 4.666508, NA, NA, 4.343930, NA, 7.3~
## $ pf_rol                              <dbl> 5.291752, 3.819566, 3.451814, 5.744~
## $ pf_ss_homicide                      <dbl> 8.920429, 9.456254, 8.060260, 7.622~
## $ pf_ss_disappearances_disap          <dbl> 10, 10, 5, 10, 10, 10, 10, 10, 10, ~
## $ pf_ss_disappearances_violent        <dbl> 10.000000, 9.294030, 10.000000, 10.~
## $ pf_ss_disappearances_organized      <dbl> 10.0, 5.0, 7.5, 7.5, 7.5, 10.0, 10.~
## $ pf_ss_disappearances_fatalities     <dbl> 10.000000, 9.926119, 10.000000, 10.~
## $ pf_ss_disappearances_injuries       <dbl> 10.000000, 9.990149, 10.000000, 9.9~
## $ pf_ss_disappearances                <dbl> 10.000000, 8.842060, 8.500000, 9.49~
## $ pf_ss_women_fgm                     <dbl> 10.0, 10.0, 10.0, 10.0, 10.0, 10.0,~
## $ pf_ss_women_missing                 <dbl> 7.5, 7.5, 10.0, 10.0, 5.0, 10.0, 10~
## $ pf_ss_women_inheritance_widows      <dbl> 5, 0, 5, 10, 10, 10, 10, 5, NA, 0, ~
## $ pf_ss_women_inheritance_daughters   <dbl> 5, 0, 5, 10, 10, 10, 10, 10, NA, 0,~
## $ pf_ss_women_inheritance             <dbl> 5.0, 0.0, 5.0, 10.0, 10.0, 10.0, 10~
## $ pf_ss_women                         <dbl> 7.500000, 5.833333, 8.333333, 10.00~
## $ pf_ss                               <dbl> 8.806810, 8.043882, 8.297865, 9.040~
## $ pf_movement_domestic                <dbl> 5, 5, 0, 10, 5, 10, 10, 5, 10, 10, ~
## $ pf_movement_foreign                 <dbl> 10, 5, 5, 10, 5, 10, 10, 5, 10, 5, ~
## $ pf_movement_women                   <dbl> 5, 5, 10, 10, 10, 10, 10, 5, NA, 5,~
```

```
## $ pf_movement                          <dbl> 6.666667, 5.000000, 5.000000, 10.00~
## $ pf_religion_estop_establish          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_religion_estop_operate            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_religion_estop                    <dbl> 10.0, 5.0, 10.0, 7.5, 5.0, 10.0, 10~
## $ pf_religion_harassment               <dbl> 9.566667, 6.873333, 8.904444, 9.037~
## $ pf_religion_restrictions             <dbl> 8.011111, 2.961111, 7.455556, 6.850~
## $ pf_religion                          <dbl> 9.192593, 4.944815, 8.786667, 7.795~
## $ pf_association_association           <dbl> 10.0, 5.0, 2.5, 7.5, 7.5, 10.0, 10.~
## $ pf_association_assembly              <dbl> 10.0, 5.0, 2.5, 10.0, 7.5, 10.0, 10~
## $ pf_association_political_establish   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_association_political_operate     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_association_political             <dbl> 10.0, 5.0, 2.5, 5.0, 5.0, 10.0, 10.~
## $ pf_association_prof_establish        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_association_prof_operate          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_association_prof                  <dbl> 10.0, 5.0, 5.0, 7.5, 5.0, 10.0, 10.~
## $ pf_association_sport_establish       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_association_sport_operate         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ pf_association_sport                 <dbl> 10.0, 5.0, 7.5, 7.5, 7.5, 10.0, 10.~
## $ pf_association                       <dbl> 10.0, 5.0, 4.0, 7.5, 6.5, 10.0, 10.~
## $ pf_expression_killed                 <dbl> 10.000000, 10.000000, 10.000000, 10~
## $ pf_expression_jailed                 <dbl> 10.000000, 10.000000, 10.000000, 10~
## $ pf_expression_influence              <dbl> 5.0000000, 2.6666667, 2.6666667, 5.~
## $ pf_expression_control                <dbl> 5.25, 4.00, 2.50, 5.50, 4.25, 7.75,~
## $ pf_expression_cable                  <dbl> 10.0, 10.0, 7.5, 10.0, 7.5, 10.0, 1~
## $ pf_expression_newspapers             <dbl> 10.0, 7.5, 5.0, 10.0, 7.5, 10.0, 10~
## $ pf_expression_internet               <dbl> 10.0, 7.5, 7.5, 10.0, 7.5, 10.0, 10~
## $ pf_expression                        <dbl> 8.607143, 7.380952, 6.452381, 8.738~
## $ pf_identity_legal                    <dbl> 0, NA, 10, 10, 7, 7, 10, 0, NA, NA,~
## $ pf_identity_parental_marriage        <dbl> 10, 0, 10, 10, 10, 10, 10, 10, 10, ~
## $ pf_identity_parental_divorce         <dbl> 10, 5, 10, 10, 10, 10, 10, 10, 10, ~
## $ pf_identity_parental                 <dbl> 10.0, 2.5, 10.0, 10.0, 10.0, 10.0, ~
## $ pf_identity_sex_male                 <dbl> 10, 0, 0, 10, 10, 10, 10, 10, 10, 1~
## $ pf_identity_sex_female               <dbl> 10, 0, 0, 10, 10, 10, 10, 10, 10, 1~
## $ pf_identity_sex                      <dbl> 10, 0, 0, 10, 10, 10, 10, 10, 10, 1~
## $ pf_identity_divorce                  <dbl> 5, 0, 10, 10, 5, 10, 10, 5, NA, 0, ~
## $ pf_identity                          <dbl> 6.2500000, 0.8333333, 7.5000000, 10~
## $ pf_score                             <dbl> 7.596281, 5.281772, 6.111324, 8.099~
## $ pf_rank                              <dbl> 57, 147, 117, 42, 84, 11, 8, 131, 6~
## $ ef_government_consumption            <dbl> 8.232353, 2.150000, 7.600000, 5.335~
## $ ef_government_transfers              <dbl> 7.509902, 7.817129, 8.886739, 6.048~
## $ ef_government_enterprises            <dbl> 8, 0, 0, 6, 8, 10, 10, 0, 7, 10, 7,~
## $ ef_government_tax_income             <dbl> 9, 7, 10, 7, 5, 5, 4, 9, 10, 10, 8,~
## $ ef_government_tax_payroll            <dbl> 7, 2, 9, 1, 5, 5, 3, 4, 10, 10, 8, ~
## $ ef_government_tax                    <dbl> 8.0, 4.5, 9.5, 4.0, 5.0, 5.0, 3.5, ~
## $ ef_government                        <dbl> 7.935564, 3.616782, 6.496685, 5.346~
## $ ef_legal_judicial                    <dbl> 2.6682218, 4.1867042, 1.8431292, 3.~
## $ ef_legal_courts                      <dbl> 3.145462, 4.327113, 1.974566, 2.930~
## $ ef_legal_protection                  <dbl> 4.512228, 4.689952, 2.512364, 4.255~
## $ ef_legal_military                    <dbl> 8.333333, 4.166667, 3.333333, 7.500~
## $ ef_legal_integrity                   <dbl> 4.166667, 5.000000, 4.166667, 3.333~
## $ ef_legal_enforcement                 <dbl> 4.3874441, 4.5075380, 2.3022004, 3.~
## $ ef_legal_restrictions                <dbl> 6.485287, 6.626692, 5.455882, 6.857~
## $ ef_legal_police                      <dbl> 6.933500, 6.136845, 3.016104, 3.385~
## $ ef_legal_crime                       <dbl> 6.215401, 6.737383, 4.291197, 4.133~
```

```
## $ ef_legal_gender                    <dbl> 0.9487179, 0.8205128, 0.8461538, 0.~
## $ ef_legal                           <dbl> 5.071814, 4.690743, 2.963635, 3.904~
## $ ef_money_growth                    <dbl> 8.986454, 6.955962, 9.385679, 5.233~
## $ ef_money_sd                        <dbl> 9.484575, 8.339152, 4.986742, 5.224~
## $ ef_money_inflation                 <dbl> 9.743600, 8.720460, 3.054000, 2.000~
## $ ef_money_currency                  <dbl> 10, 5, 5, 10, 10, 10, 10, 5, 0, 10,~
## $ ef_money                           <dbl> 9.553657, 7.253894, 5.606605, 5.614~
## $ ef_trade_tariffs_revenue           <dbl> 9.626667, 8.480000, 8.993333, 6.060~
## $ ef_trade_tariffs_mean              <dbl> 9.24, 6.22, 7.72, 7.26, 8.76, 9.50,~
## $ ef_trade_tariffs_sd                <dbl> 8.0240, 5.9176, 4.2544, 5.9448, 8.0~
## $ ef_trade_tariffs                   <dbl> 8.963556, 6.872533, 6.989244, 6.421~
## $ ef_trade_regulatory_nontariff      <dbl> 5.574481, 4.962589, 3.132738, 4.466~
## $ ef_trade_regulatory_compliance     <dbl> 9.4053278, 0.0000000, 0.9171598, 5.~
## $ ef_trade_regulatory                <dbl> 7.489905, 2.481294, 2.024949, 4.811~
## $ ef_trade_black                     <dbl> 10.00000, 5.56391, 10.00000, 0.0000~
## $ ef_trade_movement_foreign          <dbl> 6.306106, 3.664829, 2.946919, 5.358~
## $ ef_trade_movement_capital          <dbl> 4.6153846, 0.0000000, 3.0769231, 0.~
## $ ef_trade_movement_visit            <dbl> 8.2969231, 1.1062564, 0.1106256, 7.~
## $ ef_trade_movement                  <dbl> 6.406138, 1.590362, 2.044823, 4.697~
## $ ef_trade                           <dbl> 8.214900, 4.127025, 5.264754, 3.982~
## $ ef_regulation_credit_ownership     <dbl> 5, 0, 8, 5, 10, 10, 8, 5, 10, 10, 5~
## $ ef_regulation_credit_private       <dbl> 7.295687, 5.301526, 9.194715, 4.259~
## $ ef_regulation_credit_interest      <dbl> 9, 10, 4, 7, 10, 10, 10, 9, 10, 10,~
## $ ef_regulation_credit               <dbl> 7.098562, 5.100509, 7.064905, 5.419~
## $ ef_regulation_labor_minwage        <dbl> 5.566667, 5.566667, 8.900000, 2.766~
## $ ef_regulation_labor_firing         <dbl> 5.396399, 3.896912, 2.656198, 2.191~
## $ ef_regulation_labor_bargain        <dbl> 6.234861, 5.958321, 5.172987, 3.432~
## $ ef_regulation_labor_hours          <dbl> 8, 6, 4, 10, 10, 10, 6, 6, 8, 8, 10~
## $ ef_regulation_labor_dismissal      <dbl> 6.299741, 7.755176, 6.632764, 2.517~
## $ ef_regulation_labor_conscription   <dbl> 10, 1, 0, 10, 0, 10, 3, 1, 10, 10, ~
## $ ef_regulation_labor                <dbl> 6.916278, 5.029513, 4.560325, 5.151~
## $ ef_regulation_business_adm         <dbl> 6.072172, 3.722341, 2.758428, 2.404~
## $ ef_regulation_business_bureaucracy <dbl> 6.000000, 1.777778, 1.333333, 6.666~
## $ ef_regulation_business_start       <dbl> 9.713864, 9.243070, 8.664627, 9.122~
## $ ef_regulation_business_bribes      <dbl> 4.050196, 3.765515, 1.945540, 3.260~
## $ ef_regulation_business_licensing   <dbl> 7.324582, 8.523503, 8.096776, 5.253~
## $ ef_regulation_business_compliance  <dbl> 7.074366, 7.029528, 6.782923, 6.508~
## $ ef_regulation_business             <dbl> 6.705863, 5.676956, 4.930271, 5.535~
## $ ef_regulation                      <dbl> 6.906901, 5.268992, 5.518500, 5.369~
## $ ef_score                           <dbl> 7.54, 4.99, 5.17, 4.84, 7.57, 7.98,~
## $ ef_rank                            <dbl> 34, 159, 155, 160, 29, 10, 27, 106,~
## $ hf_score                           <dbl> 7.568140, 5.135886, 5.640662, 6.469~
## $ hf_rank                            <dbl> 48, 155, 142, 107, 57, 4, 16, 130, ~
## $ hf_quartile                        <dbl> 2, 4, 4, 3, 2, 1, 1, 4, 2, 2, 4, 2,~
```

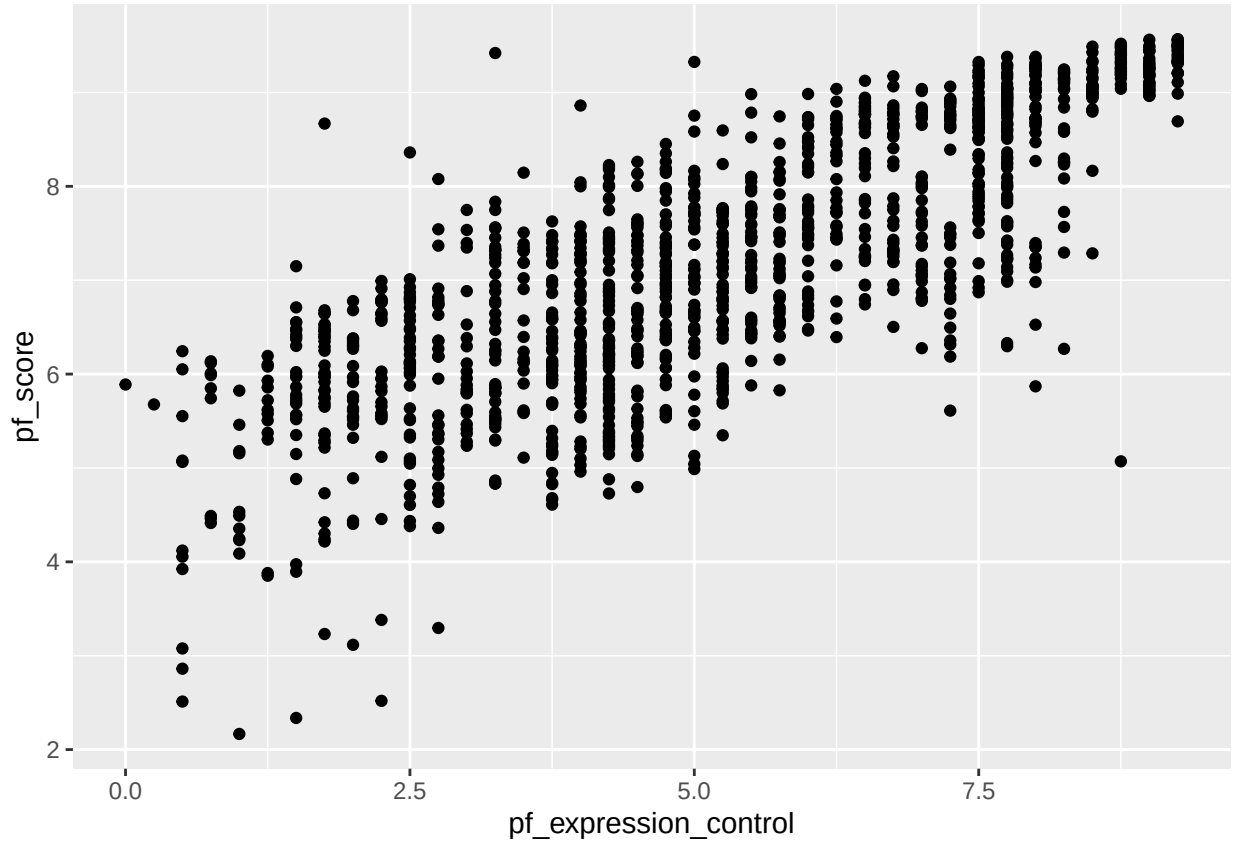
The data set is Rows: 1,458 by Columns: 123

Exercise 2

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?

```
ggplot2::ggplot(hfi, aes(x=pf_expression_control,y=pf_score)) + geom_point()
```



You would use a scatter plot to see a linear relationship between variables. There does appear to be a linear relationship in the data. I would be comfortable using a linear model to predict pf scores if other conditions such as normal residual requirements are also met. This is also supported by the relatively high correlation.

```
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
```

Exercise 3

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.

Given the previous plot you can see a relatively strong positive correlation between the variables. There is a slightly wider spread in the points closer to 0 which may indicate that a log transform would tighten the linearity (this effect is pretty weak).

Exercise 4

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?

I run the functions and get: Error in oSide[oSide < LLim | oSide > RLim] <- c(x + r)[oSide < LLim | : NAs are not allowed in subscripted assignments DATA606::plot_ss(x = pf_expression_control, y = pf_score, data = hfi, showSquares = TRUE) Error in DATA606::plot_ss(x = pf_expression_control, y = pf_score, data = hfi, : unused argument (data = hfi)

Exercise 5

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
```

hf_score_hat = 5.153687 + .349862 * pf_expression_control The slope tells us the line will have a slight positive rise. Meaning for every 1 units of pf_expression_control we expect hf_score to go up by about 1/3.

Exercise 6

If someone saw the least squares regression line and not the actual data, how would they predict a country's personal freedom school[sic] 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?

```
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
```

The prediction for 6.7 would lie on the regression line and produce a y value of:

```
4.61707 + .49143 * 6.7
```

```
## [1] 7.909651
```

Knowing the data, we could compute the error or residual for the point for a pf_expression_control with value 6.7. However, there are no such values in our dataset.

```
dplyr::filter(hfi,pf_expression_control==6.7)
```

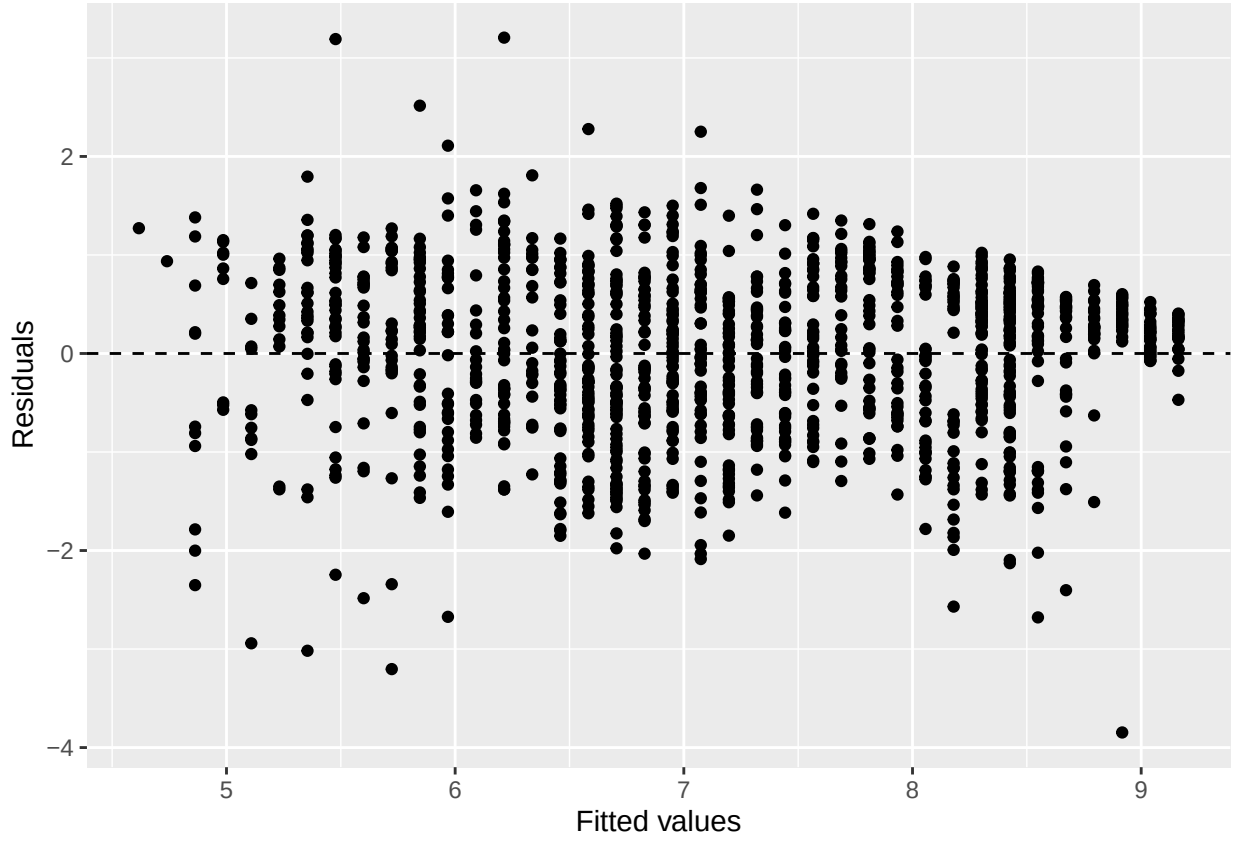
```
## # A tibble: 0 x 123
## # ... with 123 variables: year <dbl>, ISO_code <chr>, countries <chr>,
## #   region <chr>, pf_rol_procedural <dbl>, pf_rol_civil <dbl>,
## #   pf_rol_criminal <dbl>, pf_rol <dbl>, pf_ss_homicide <dbl>,
## #   pf_ss_disappearances_disap <dbl>, pf_ss_disappearances_violent <dbl>,
## #   pf_ss_disappearances_organized <dbl>,
## #   pf_ss_disappearances_fatalities <dbl>, pf_ss_disappearances_injuries <dbl>,
## #   pf_ss_disappearances <dbl>, pf_ss_women_fgm <dbl>, ...
```

The expected error associated with a prediction would be the Residual standard error: 0.8318 for each point predicted with the model. A reasonable assumption is that the actual value of a pf_score for a pf_expression_control==6.7 would be 7.909651 +/- 0.8318

Exercise 7

Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between the two variables?

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

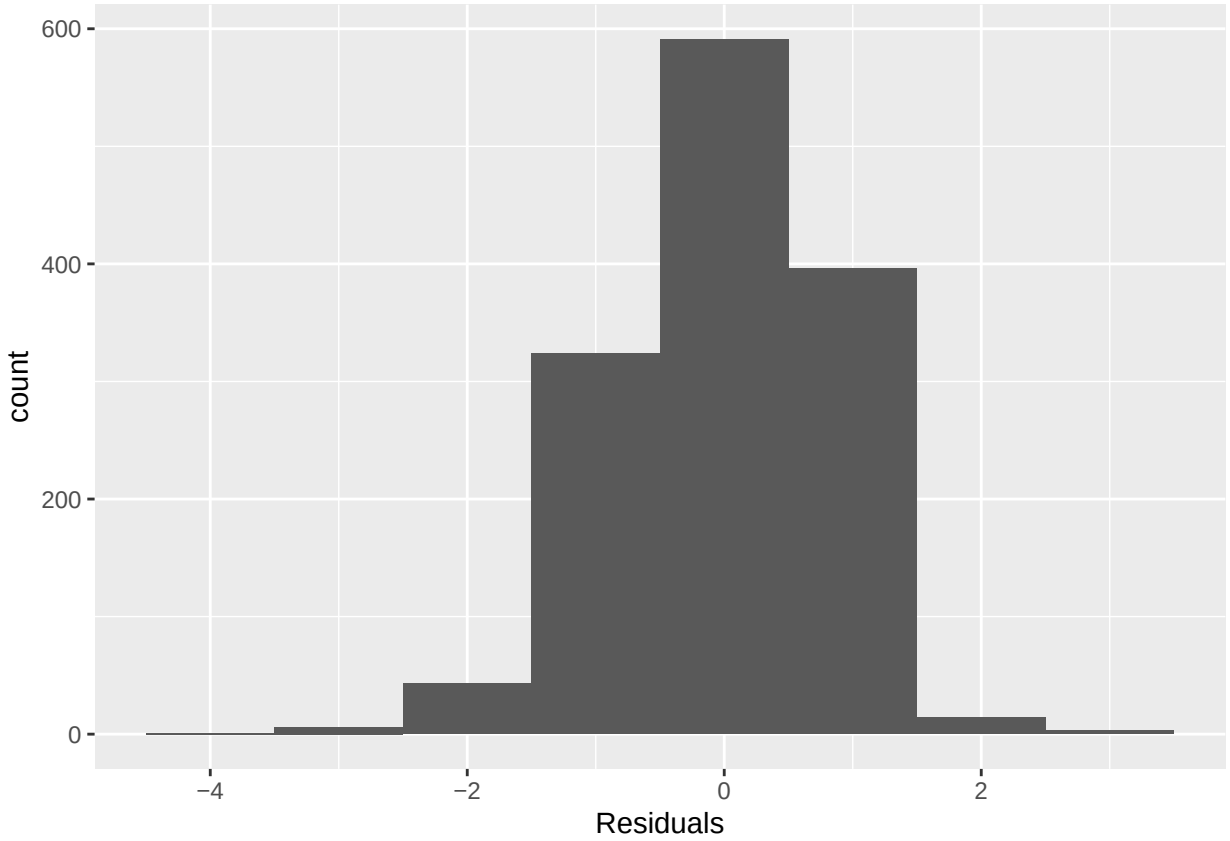


The residuals seem pretty evenly distributed both about 0 and along the range of fitted values which indicates linearity and constant variability requirements for linear regression models are met.

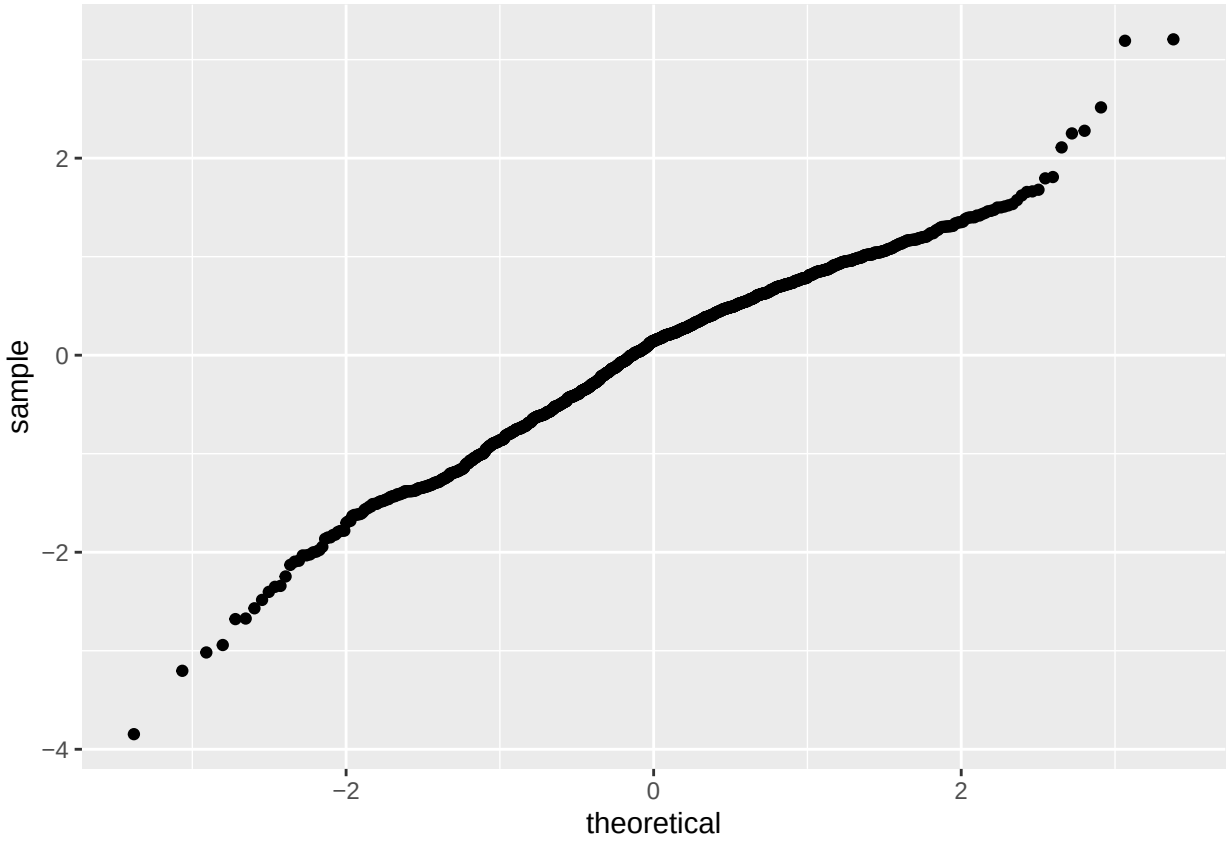
Exercise 8

Based on the histogram and the normal probability plot, does the nearly normal residuals condition appear to be met?

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




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



The QQ plot indicates that the distribution of the residuals is nearly normal. The histogram also indicates a nearly normal distribution with no extreme outliers.

Exercise 9

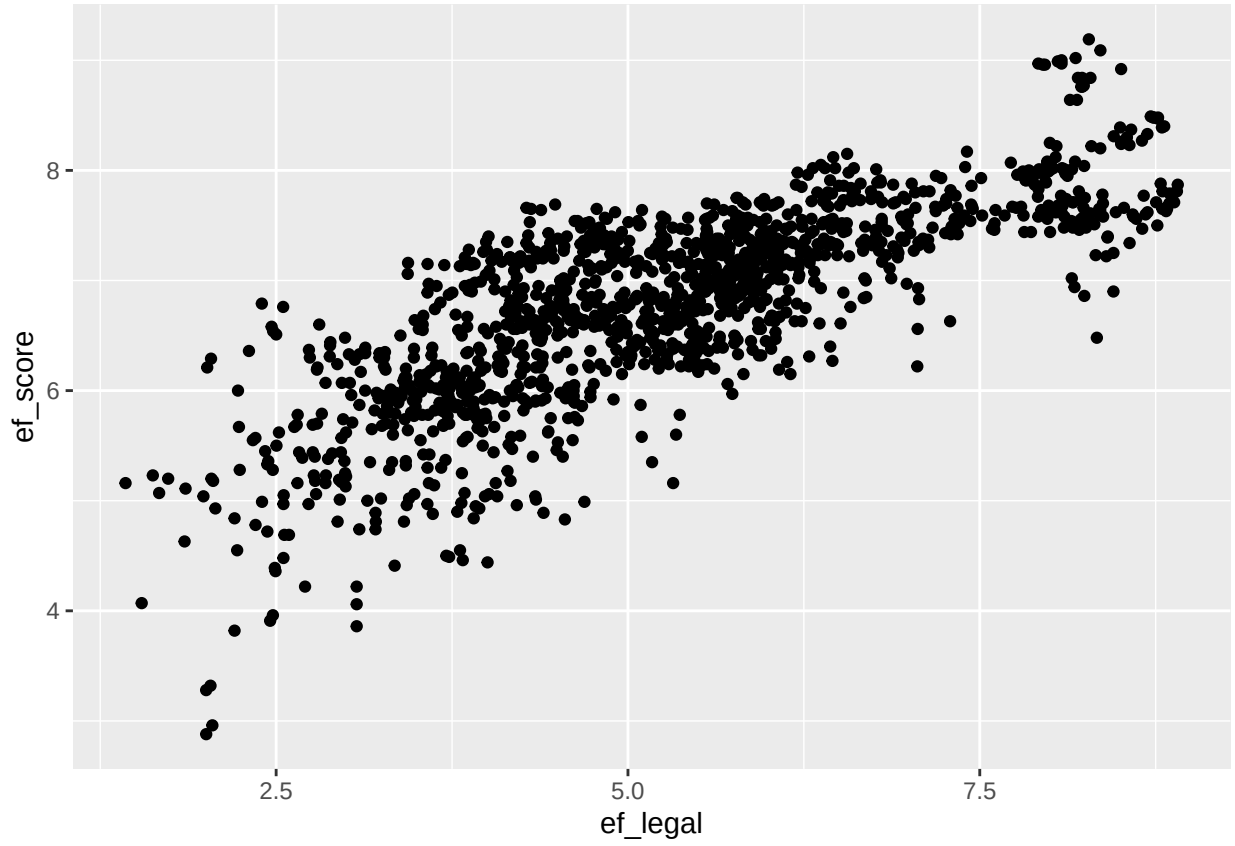
Based on the residuals vs. fitted plot, does the constant variability condition appear to be met?

Yes. Based on the plot the constant variability condition appears to be met.

More Practice

1.Choose another freedom variable and a variable you think would strongly correlate with it.. Produce a scatterplot of the two variables and fit a linear model. At a glance, does there seem to be a linear relationship?

```
ggplot2::ggplot(hfi, aes(x=ef_legal,y=ef_score)) + geom_point()
```




```
hfi %>%
  summarise(cor(ef_legal, ef_score, use = "complete.obs"))
```

```
## # A tibble: 1 x 1
##   `cor(ef_legal, ef_score, use = "complete.obs")`
##                                             <dbl>
## 1                                           0.781
```

There does appear to be a strong linear relationship between the variables ef_legal and overall ef_score. The correlation also supports the linearity.

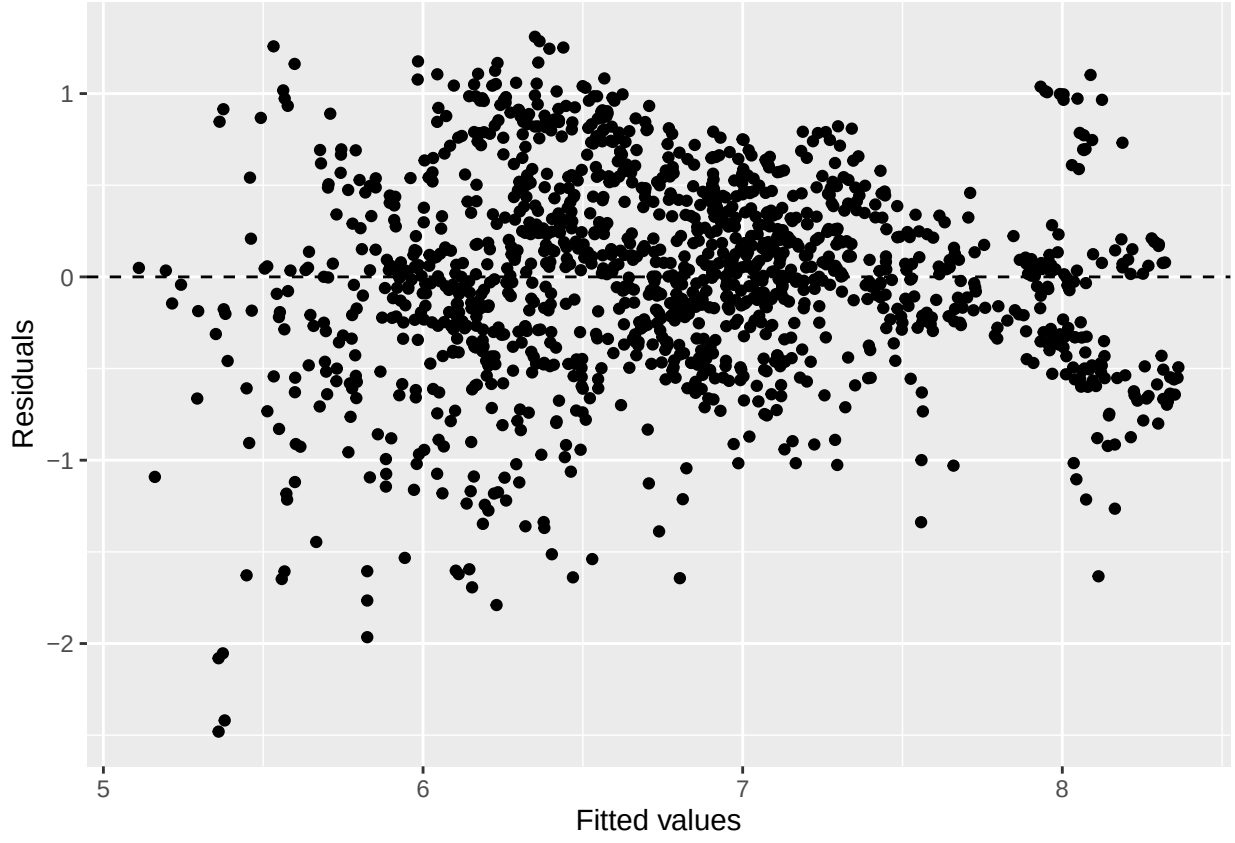
2.How does this relationship compare to the relationship between pf_expression_control and pf_score? Use the R2 values from the two model summaries to compare. Does your independent variable seem to predict your dependent one better? Why or why not?

```
m3 <- lm(ef_score ~ ef_legal, data = hfi)
summary(m3)
```

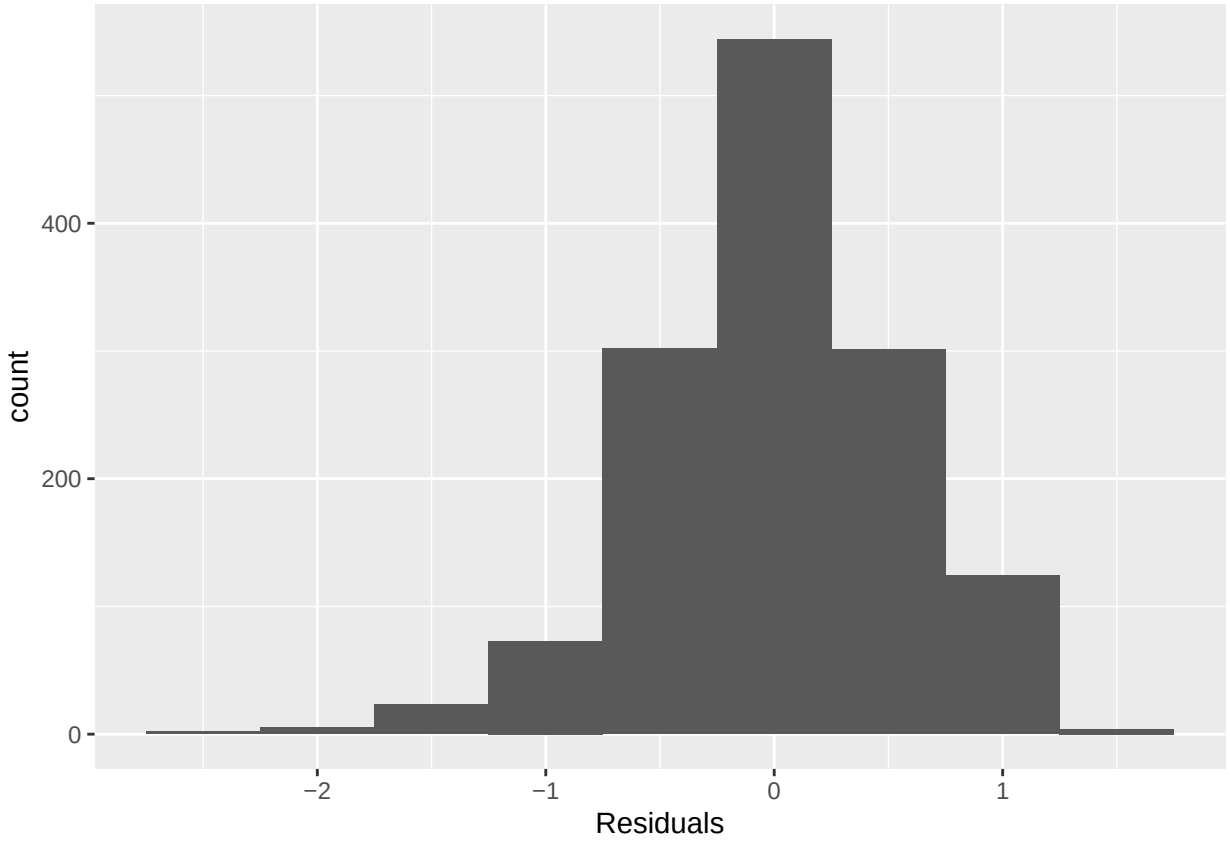
```
##
## Call:
## lm(formula = ef_score ~ ef_legal, data = hfi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48042 -0.33010  0.02116  0.35665  1.30988
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.489294   0.051687   86.86   <2e-16 ***
## ef_legal    0.434927   0.009376   46.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.552 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:   0.61,  Adjusted R-squared:  0.6097
## F-statistic:  2152 on 1 and 1376 DF,  p-value: < 2.2e-16
```

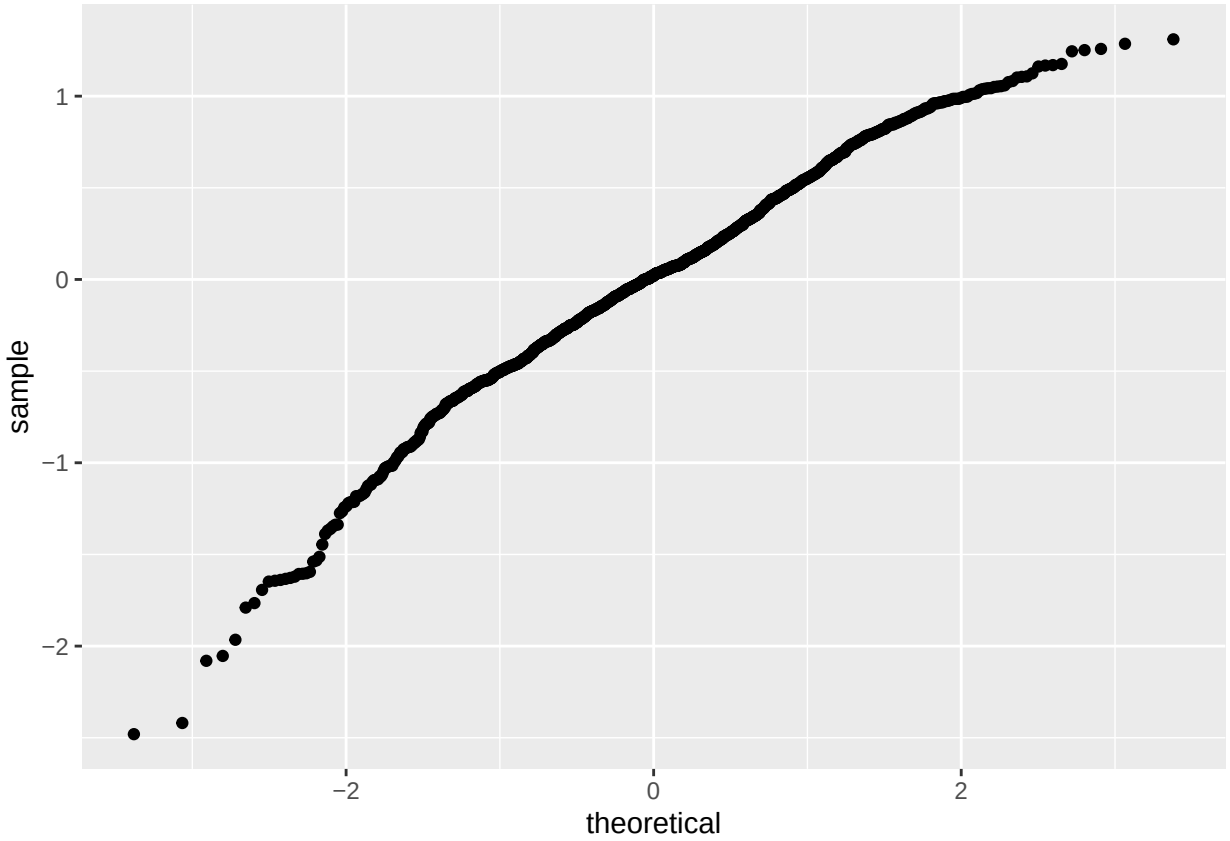
```
ggplot(data = m3, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = m3, aes(x = .resid)) +
  geom_histogram(binwidth = .5) +
  xlab("Residuals")
```




```
ggplot(data = m3, aes(sample = .resid)) +
  stat_qq()
```



Both of these models perform similarly with R squared values of .634 for m1 compared to .61 for this model.

3.What's one freedom relationship you were most surprised about and why? Display the model diagnostics for the regression model analyzing this relationship.

```
m4 <- lm(pf_score ~ ef_score, data = hfi)
summary(m4)
```

```
##
## Call:
## lm(formula = pf_score ~ ef_score, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6373 -0.7334  0.1090  0.8353  2.9840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.51812    0.22223   2.331   0.0199 *
## ef_score     0.98490    0.03248  30.326   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.065 on 1376 degrees of freedom
##   (80 observations deleted due to missingness)
## Multiple R-squared:  0.4006, Adjusted R-squared:  0.4002
## F-statistic: 919.7 on 1 and 1376 DF,  p-value: < 2.2e-16
```

I'm surprised that the relationship between overall economic freedom and personal freedom is not more highly associated.