Lab 8 Pricilla Nakyazze Introduction to linear regression

2025-03-24

```
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
install.packages('openintro')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

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

glimpse(hfi)

names(hfi)

Exercise 1

What are the dimensions of the dataset?

```
dim(hfi)
```

```
## [1] 1458  123
```

There 1458 observations/rows and 123 dimensions/columns

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?

I would be comfortable using a linear model to describe the relationship between the independent and dependent variable and assuming a linear relationship between pf_score and personal freedom score. With a best fit plane that minimises the difference between the two or three variables

#3 variables

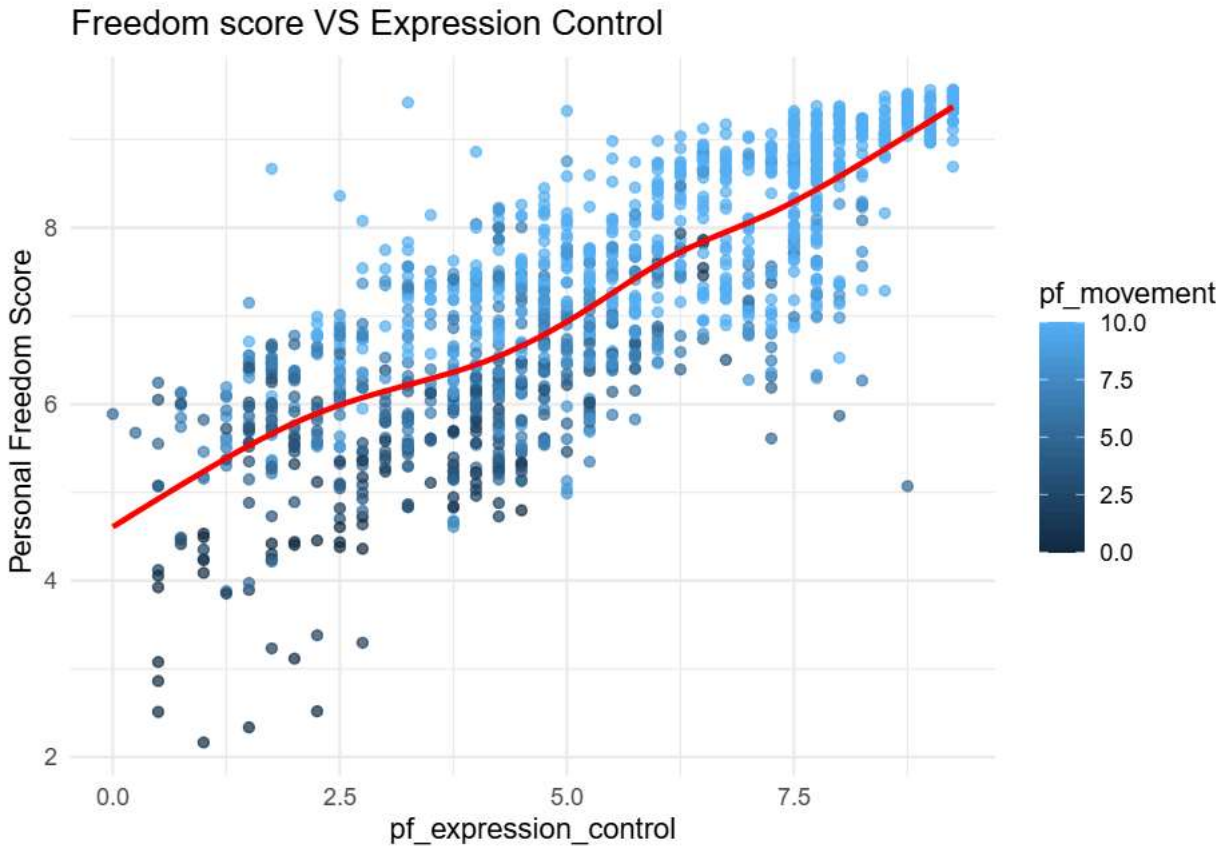
```
ggplot(data = hfi, aes(x= pf_expression_control, y = pf_score)) +
  geom_point(aes(color = pf_movement), alpha = .7)+
  geom_smooth(methond = 'lm',
              se = FALSE,
              color = "red")+
  labs(title = "Freedom score VS Expression Control",
  X = "Expression control score (0-10)",
  y  = " Personal Freedom Score") +
theme_minimal()
```

```
## Warning in geom_smooth(methond = "lm", se = FALSE, color = "red"): Ignoring
## unknown parameters: `methond`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



#2 variables

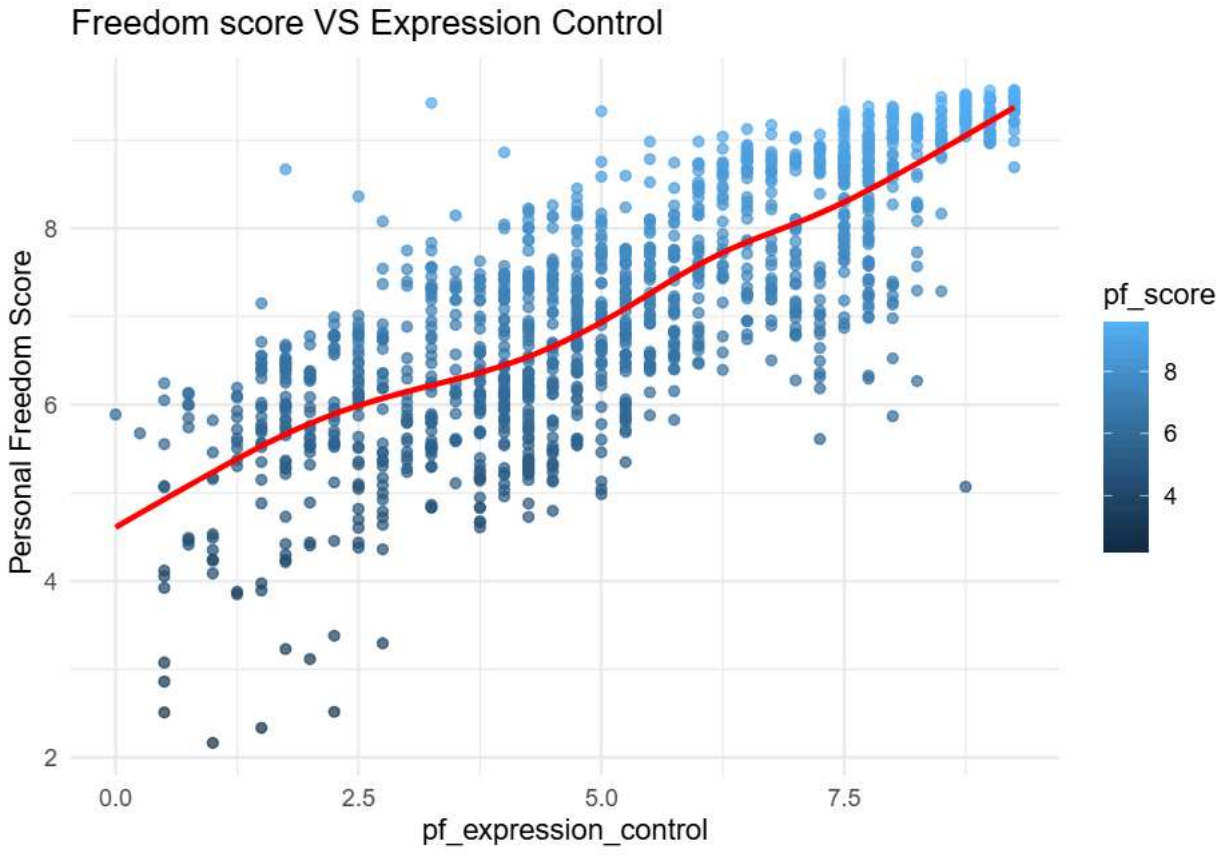
```
ggplot(data = hfi, aes(x= pf_expression_control, y = pf_score)) +
  geom_point(aes(color = pf_score), alpha = .7)+
  geom_smooth(methond = 'lm',
              se = FALSE,
              color = "red")+
  labs(title = "Freedom score VS Expression Control",
  X = "Expression control score (0-10)",
  y  = " Personal Freedom Score") +
theme_minimal()
```

```
## Warning in geom_smooth(methond = "lm", se = FALSE, color = "red"): Ignoring
## unknown parameters: `methond`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
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.

The relationship and correlation between the two variables is positive. As one increase the other increases as well because of the increasing slope.There some outliers on the bottom left of the plot.

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?

```
hfi2 <- hfi %>%
  select(pf_expression_control,pf_score)%>%
  na.omit()
ggplot(data = hfi, aes(x = pf_expression_control, y = pf_score, showSquares = TRUE, showSquares = TRUE)
geom_point(aes(color = pf_score), alpha = .7)+
  geom_smooth(methond = 'lm',
              se = FALSE,
              color = "orange")+
  labs(title = "Freedom score VS Expression Control",
  X = "Expression control score (0-10)",
  y  = " Personal Fredom Score") +
theme_minimal()
```

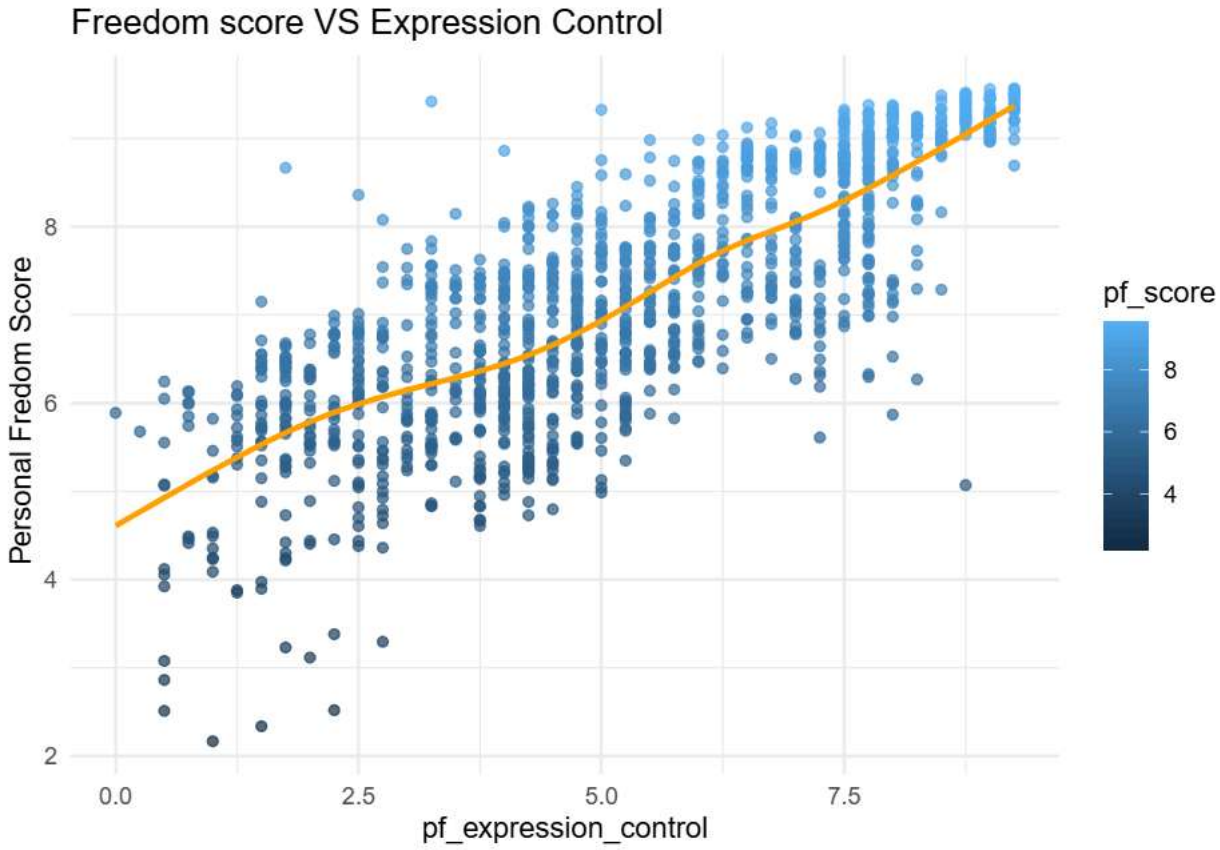
```
## Warning: Duplicated aesthetics after name standardisation: showSquares
```

```
## Warning in geom_smooth(methond = "lm", se = FALSE, color = "orange"): Ignoring
## unknown parameters: `methond`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 80 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



call

lm(formula = y-x, data = pts)

coefficients:

(intercept) x

4.6171 0.4914

sum of Squares: 952.153

The smallest sum of squares is 952.153

```
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
```

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?

y = 5.153687 + .349862 * pf_expression_control

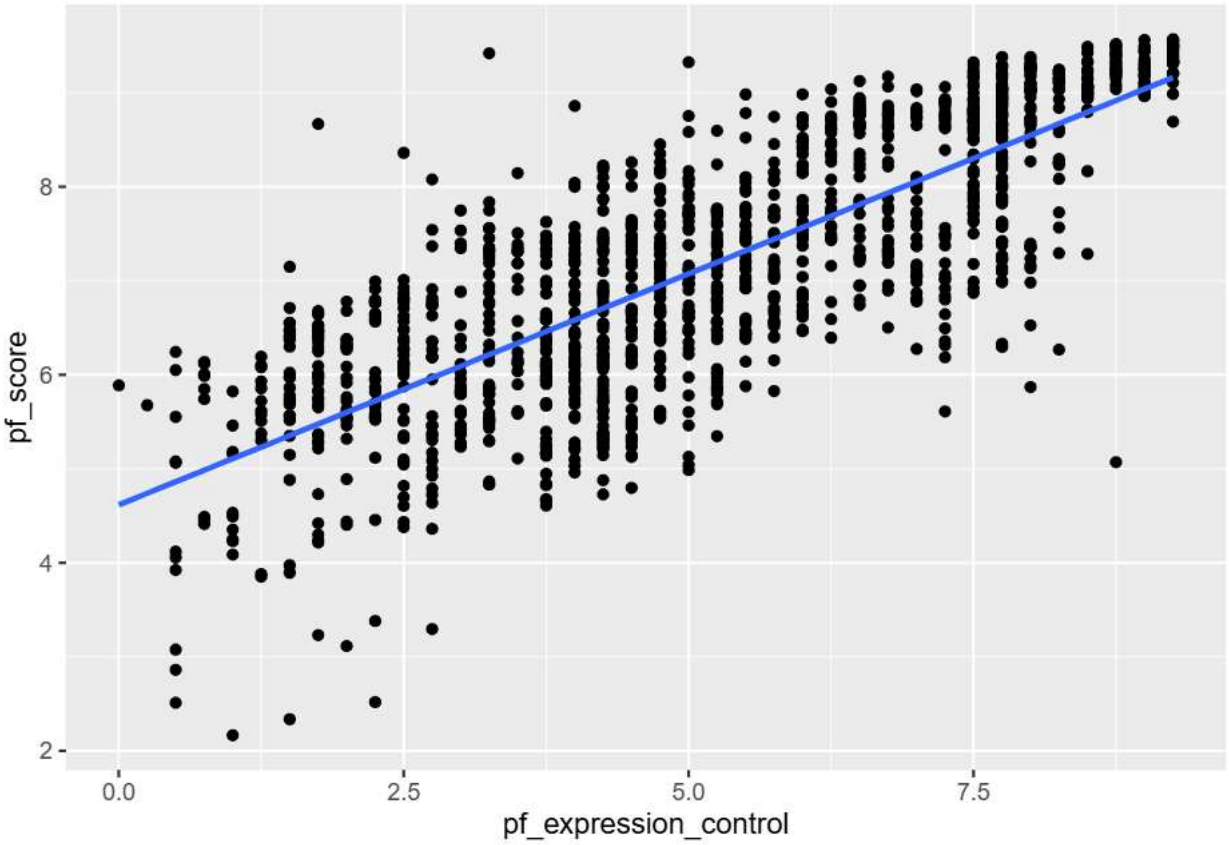
The positive correlation slope implies human freedom declines as political pressure on media content increases.

```
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 outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 80 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



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 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?

To calculate the residual prediction. We find the difference between the actual value and predicted value. (y-y) for any given point.

predicted value

```
Pv <- 4.61707 + 0.4913 * 6.7
```

find any actual value of pf_score with pf_expression control that is = 6.7

```
hfi3 <- hfi%>%
  filter(pf_expression_control >= 6.7)%>%
select(pf_expression_control,pf_score,countries)
hfi3
```

```
## # A tibble: 437 x 3
##    pf_expression_control pf_score countries
##                    <dbl>    <dbl> <chr>
##  1                  7.75     9.18 Australia
##  2                  8        9.25 Austria
##  3                  7.25     7.45 Bahamas
##  4                  7.5      7.71 Barbados
##  5                  9.25     8.99 Belgium
##  6                  6.75     7.43 Belize
##  7                  7.25     7.50 Benin
##  8                  7        7.46 Burkina Faso
##  9                  8.25     9.15 Canada
## 10                  7.75     7.99 Cape Verde
## # i 427 more rows
```

find any actual value of pf_score with pf_expression control that is = 6.7

```
Actual <- hfi %>%
   filter(pf_expression_control == 6.7) %>%
  select(pf_expression_control, pf_score)
Actual
```

```
## # A tibble: 0 x 2
## # i 2 variables: pf_expression_control <dbl>, pf_score <dbl>
```

Prediction residual for Belize

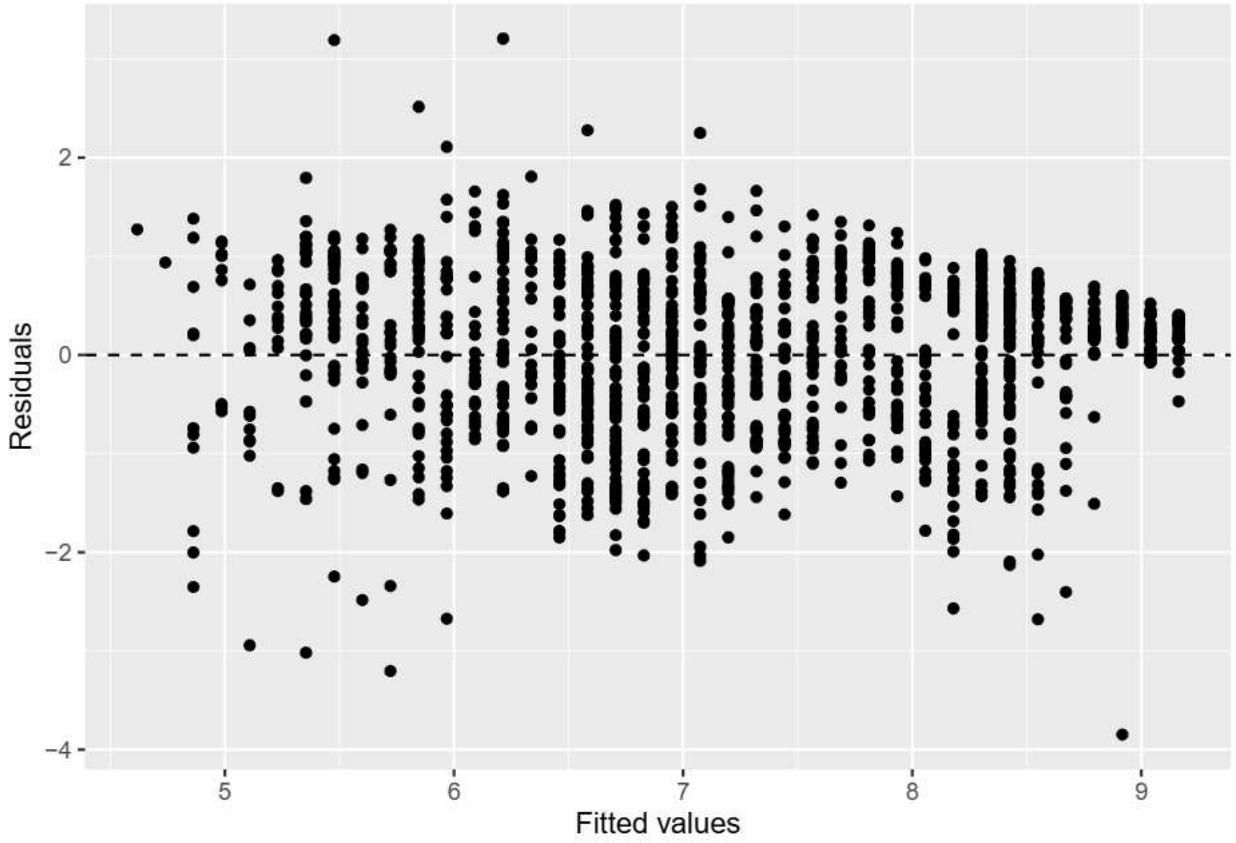
```
Prediction_ex <- 7.430864 - Pv
Prediction_ex
```

```
## [1] -0.477916
```

In this case the residual would be slightly overestimated.

```
ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
```

```
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```

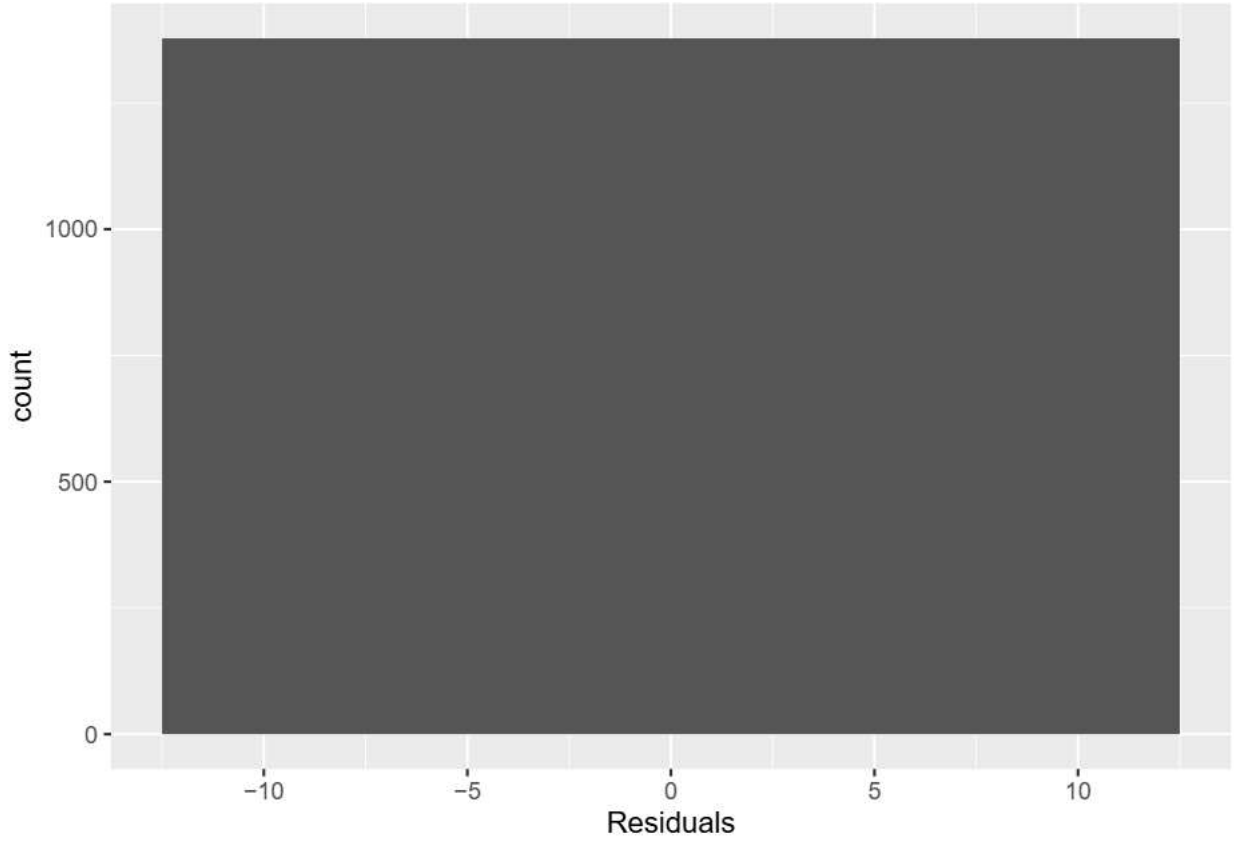


Exercise 7

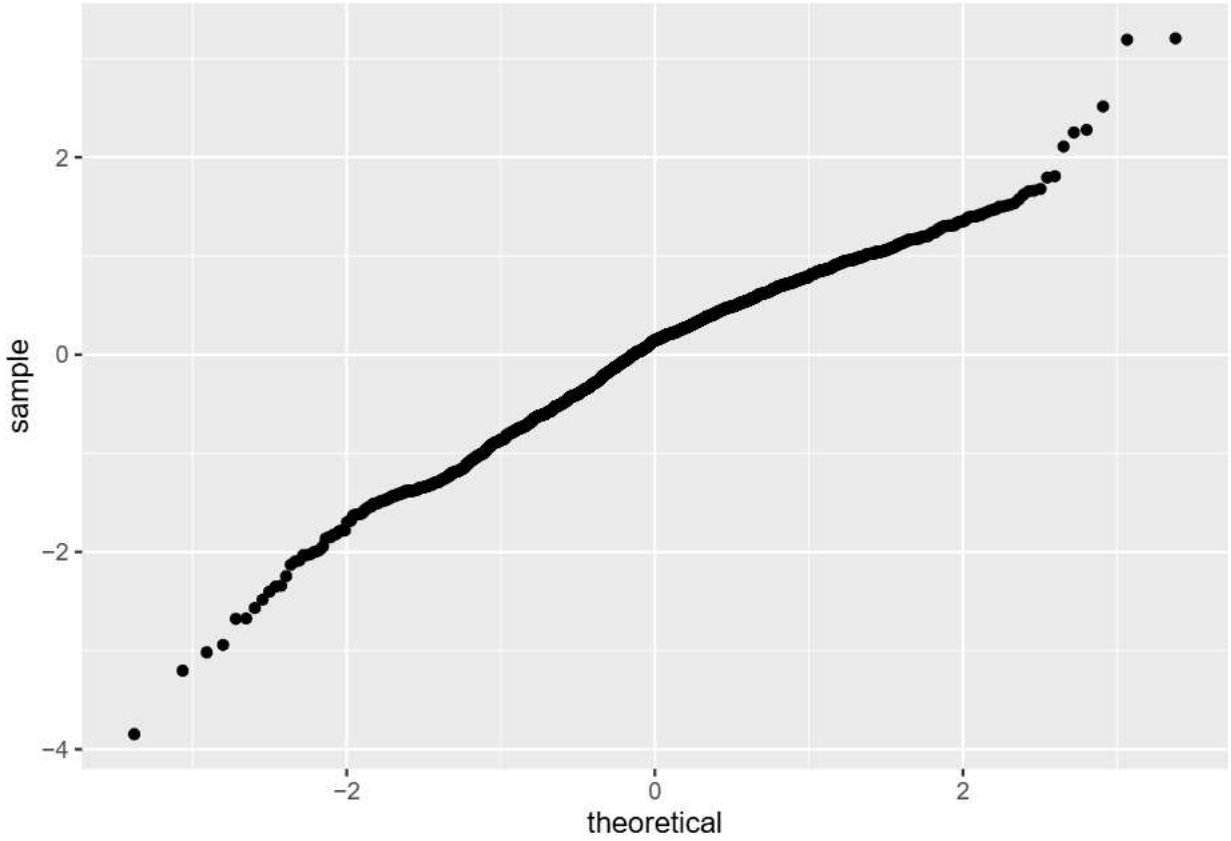
Is there any apparent pattern in the residuals plot? What does this indicate about the linearity of the relationship between the two variables?

The horinzontal line signifies that the independent variable x has no predictive power for the dependent variable y. Meaning the slope of regression line is 0. So a change in x does not lead to a change in y.

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




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



Exercise 8

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

Yes. The residuals are considered normally distributed because the difference between the sample values and theoretical values are normally distributed.

Exercise 9

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

Yes constant variability is met because the scatterring residuals are only around 0 and at the top with no trend in the residual plot.

More Practice

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?

The plot below shows the relationship between pf_ss_women_inheritance_widows and pf_score. It appears the correlation type is non-linear.

```
hfi2 <- hfi %>%
  select(pf_ss_women_inheritance_widows,pf_score)%>%
  na.omit()
ggplot(data = hfi, aes(x = pf_ss_women_inheritance_widows, y = pf_score, showSquares = TRUE, showSquare
geom_point(aes(color = pf_score), alpha = .7)+
  geom_smooth(methond = 'lm',
              se = FALSE,
              color = "orange")+
  labs(title = "women_inheritance_widows VS Expression Control",
```

```
  X = "women inheritance widows (0-10)",
  y  = " Personal Fredom Score") +
theme_minimal()
```

```
## Warning: Duplicated aesthetics after name standardisation: showSquares
```

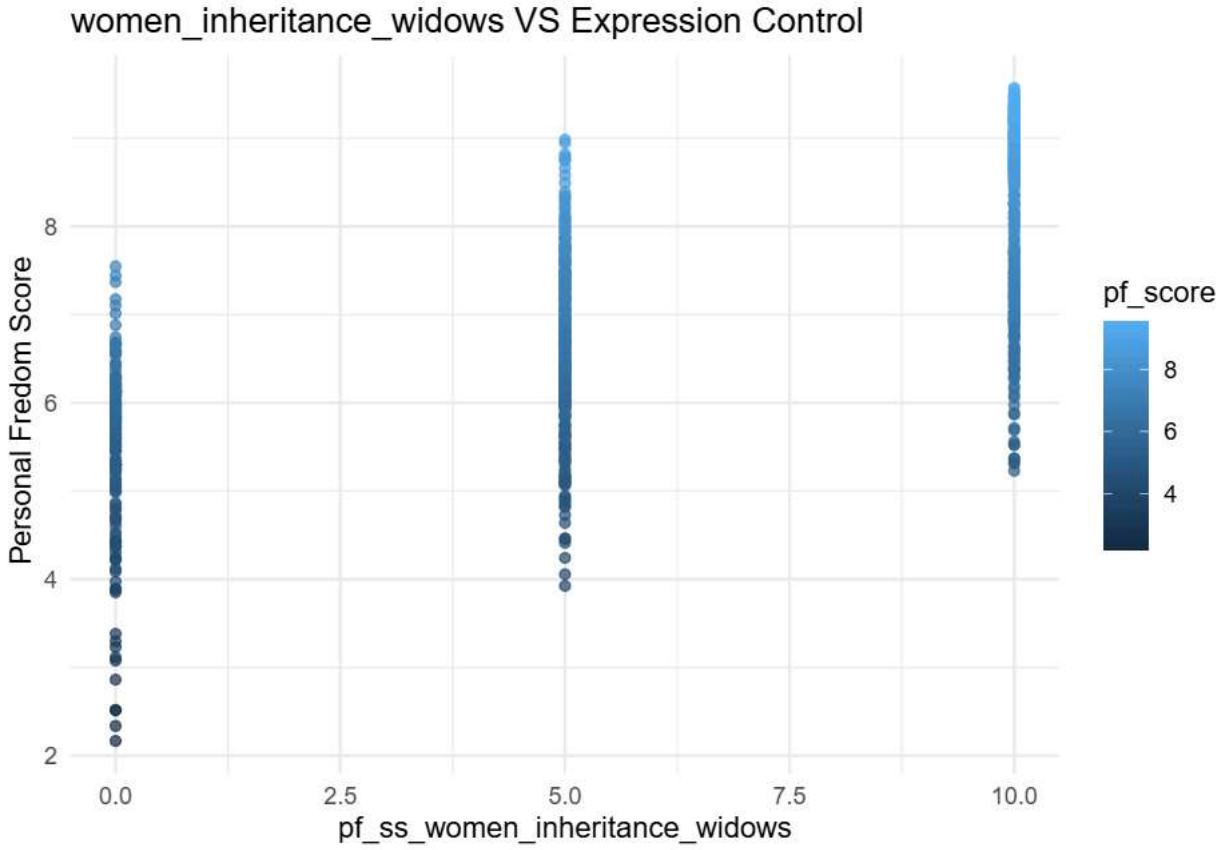
```
## Warning in geom_smooth(methond = "lm", se = FALSE, color = "orange"): Ignoring
## unknown parameters: `methond`
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 541 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Failed to fit group 1.
## Caused by error in `smooth.construct.cr.smooth.spec()`:
## ! x has insufficient unique values to support 10 knots: reduce k.
```

```
## Warning: Removed 541 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
hfi %>%
  summarise(cor(pf_ss_women_inheritance_widows, pf_score, use ='complete.obs'))
```

```
## # A tibble: 1 x 1
##   `cor(pf_ss_women_inheritance_widows, pf_score, use = "complete.obs")`
##                                                                   <dbl>
## 1                                                                 0.715
```

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?

The R2 values from the relationship between pf_expression_control and pf_score is 0.6342. And the R2 values from the relationship between pf_ss_women_inheritance_widows and pf_score is 0.5105. No my independent variable does not predict the dependent because 51% of the variance is explained by Women movement. Indicating a weak relationship.

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

```
##
## 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
```

```
summary(lm(pf_score ~ pf_ss_women_inheritance_widows, data = hfi))
```

```
##
## Call:
## lm(formula = pf_score ~ pf_ss_women_inheritance_widows, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2211 -0.6915  0.1729  0.7712  2.2242
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5.387661   0.064718   83.25   <2e-16 ***
## pf_ss_women_inheritance_widows 0.274355   0.008881   30.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9986 on 915 degrees of freedom
##   (541 observations deleted due to missingness)
## Multiple R-squared:  0.5105, Adjusted R-squared:   0.51
## F-statistic: 954.4 on 1 and 915 DF,  p-value: < 2.2e-16
```

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

With divorce, women movement increases, the r^2 value is 0.3531.The relationship between Divorce and women movement is weaker than the relationship between pf_expression_control and pf_score that has a

r^2 value of 0.6342.

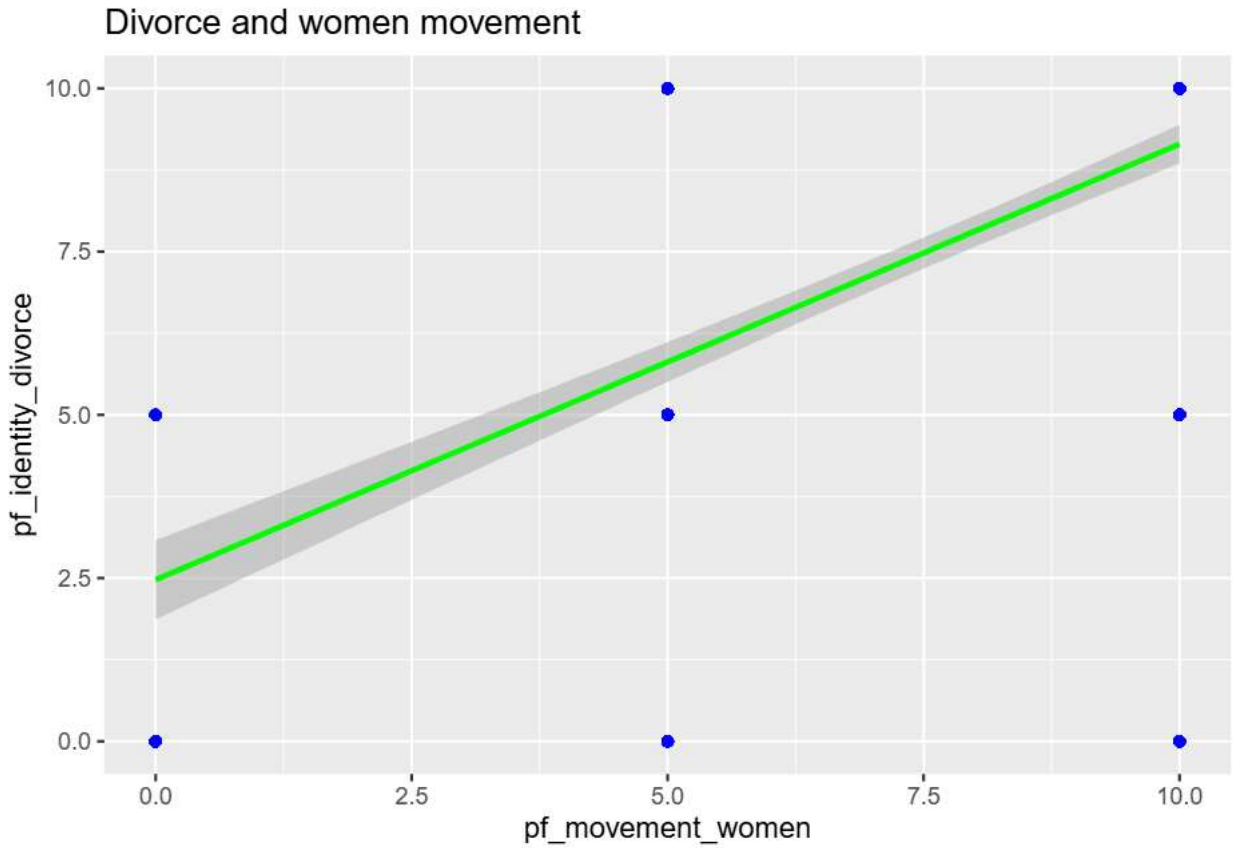
```
summary(lm(pf_identity_divorce ~ pf_movement_women, data = hfi))
```

```
##
## Call:
## lm(formula = pf_identity_divorce ~ pf_movement_women, data = hfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1466 -0.8106  0.8534  0.8534  4.1894
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.4747     0.3087   8.016 5.97e-15 ***
## pf_movement_women   0.6672     0.0374  17.838  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.934 on 583 degrees of freedom
##   (873 observations deleted due to missingness)
## Multiple R-squared:  0.3531, Adjusted R-squared:  0.352
## F-statistic: 318.2 on 1 and 583 DF,  p-value: < 2.2e-16
```

```
hfi %>% ggplot(aes(x = pf_movement_women, y = pf_identity_divorce)) +
  geom_point(col="blue", na.rm = TRUE)+
  geom_smooth(method = 'lm', Se = TRUE, col="green", na.rm = TRUE) +
  labs(title = 'Divorce and women movement')
```

```
## Warning in geom_smooth(method = "lm", Se = TRUE, col = "green", na.rm = TRUE):
## Ignoring unknown parameters: `Se`
```

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


Divorce and women movement
pf_identity_divorce
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
pf_movement_women