431 Class 10

Thomas E. Love

2018-09-27

Today's Agenda

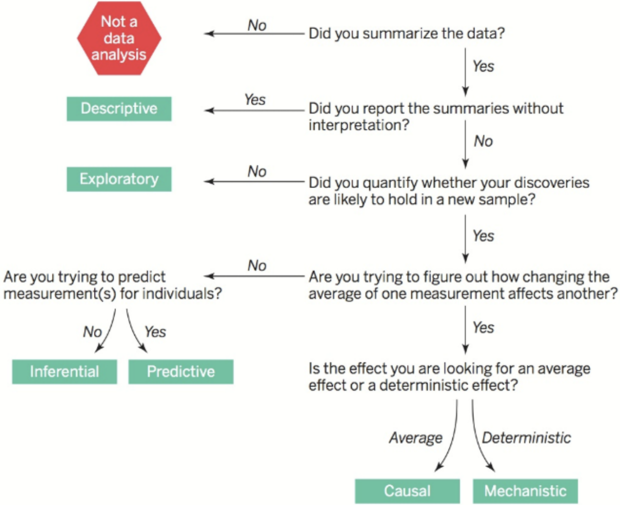
1. Elements of Data Analytic Style: Chapters 1-4 and 12
2. Association, Correlation, Linear Models (Notes: Ch 11)
   - A study of von Hippel-Lindau disease
   - Associations, Correlation and Scatterplots
   - Fitting a Linear Model
3. Getting started on Project Study 1 (Class Survey)

The 15 Questions Starting Project Study 1

1. Were you born in the United States?
2. Is English the language you speak better than any other?
3. Do you identify as female?
4. Do you wear prescription glasses or contact lenses?
5. Before taking 431, had you ever used R before?
6. Are you currently married or in a stable domestic relationship?
7. Have you smoked 100 cigarettes or more in your entire life?
8. In what year were you born?
9. How would you rate your current health overall (Excellent, Very Good, Good, Fair, Poor)
10. For how long, in months, have you lived in Northeast Ohio?
11. What is your height in inches?
12. What is your weight in pounds?
13. What is your pulse rate, in beats per minute?
14. Last week, on how many days did you exercise? (0 - 7)
15. Last night, how many hours of sleep did you get?

Jeff Leek: Chapters 1-4 and 12

- Chapter 1: Introduction
- Chapter 2: The Data Analytic Question (See next slide)
- Chapter 3: Tidying the Data
- Chapter 4: Checking the Data
- Chapter 12: Reproducibility


Not a data analysis
No
Did you summarize the data?
Yes
Descriptive
Yes
Did you report the summaries without interpretation?
No
Exploratory
No
Did you quantify whether your discoveries are likely to hold in a new sample?
Yes
Are you trying to predict measurement(s) for individuals?
No
Are you trying to figure out how changing the average of one measurement affects another?
No
Yes
Yes
Inferential
Predictive
Is the effect you are looking for an average effect or a deterministic effect?
Average
Deterministic
Causal
Mechanistic

Source: Leek JT Peng RD *Science* "What is the question?" 2015-03-20, linked at http://bit.ly/leek-peng-whatisthequestion

Studying the Association of Quantities

R setup for Today

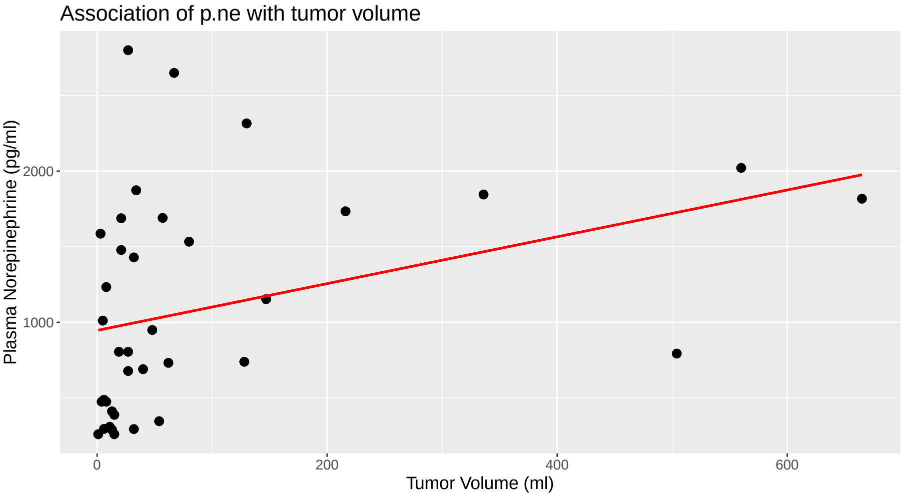
```
library(tidyverse)

VHL <- read.csv("vonHippel-Lindau.csv") %>% tbl_df

VHL
```

```
# A tibble: 37 x 4
      id disease  p.ne tumorvol
   <int>   <int> <int>    <int>
 1   101       0   289       13
 2   102       1   294       32
 3   103       0  2799       27
 4   104       0  2649       67
 5   105       0   346       54
 6   106       0  1690       57
 7   107       0   805       19
 8   108       1  1153      147
 9   109       0   678       27
```

Scatterplot with Linear Fit



The Linear Model

```
model1 <- lm(p.ne ~ tumorvol, data = VHL)

broom::tidy(model1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic      p.value
  <chr>          <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)   946.      130.        7.25 0.0000000181
2 tumorvol        1.55      0.708     2.19 0.0356
```

Correlation Coefficients

Two key types of correlation coefficient to describe an association between quantities.

- The one most often used is called the *Pearson* correlation coefficient, symbolized r or sometimes rho (ρ).
- Another is the Spearman rank correlation coefficient, also symbolized by ρ, or sometimes ρ_s.

```
cor(VHL$p.ne, VHL$tumorvol)
```

```
[1] 0.3465646
```

```
cor(VHL$p.ne, VHL$tumorvol, method = "spearman")
```

```
[1] 0.5414319
```

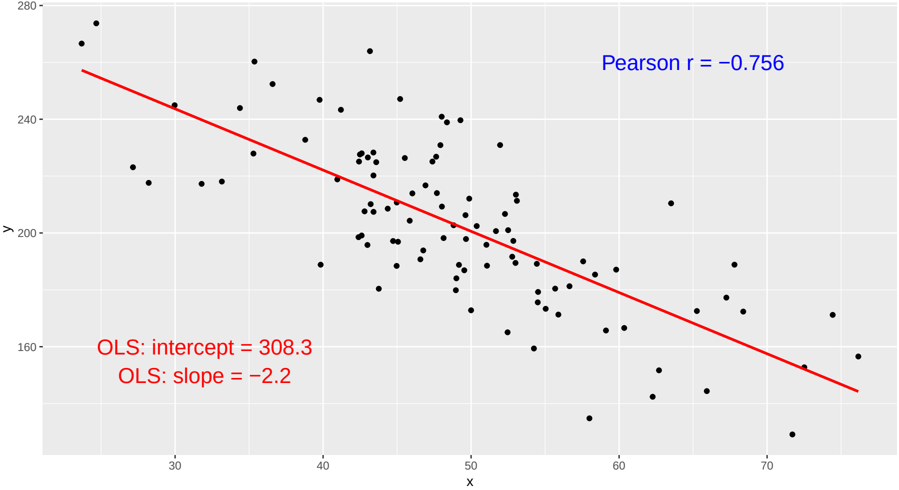
Meaning of Pearson Correlation

The Pearson correlation coefficient assesses how well the relationship between X and Y can be described using a linear function.

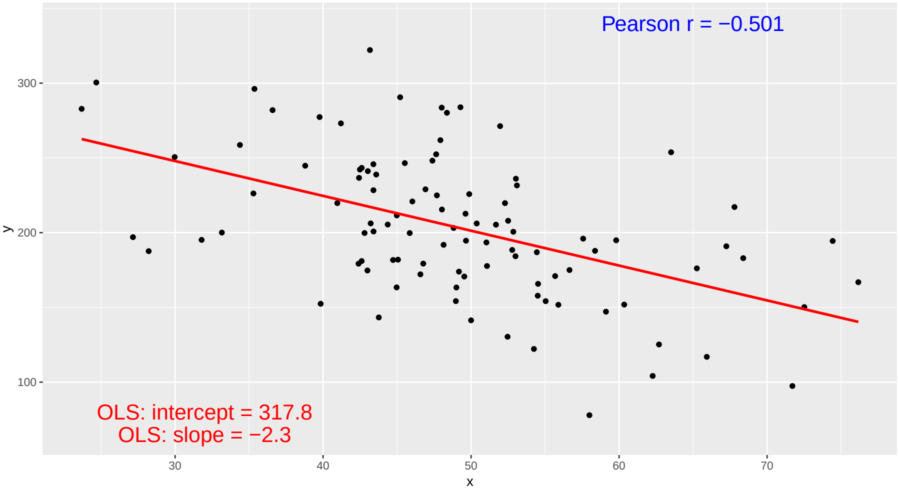
- The Pearson correlation is dimension-free.
- It falls between -1 and $+1$, with the extremes corresponding to situations where all the points in a scatterplot fall exactly on a straight line with negative and positive slopes, respectively.
- A Pearson correlation of zero corresponds to the situation where there is no linear association.
- Unlike the estimated slope in a regression line, the sample correlation coefficient is symmetric in x and y, so it does not depend on labeling one of them (y) the response variable, and one of them (x) the predictor.

$$r_{XY} = \frac{1}{n-1} \Sigma_{i=1}^{n} (\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

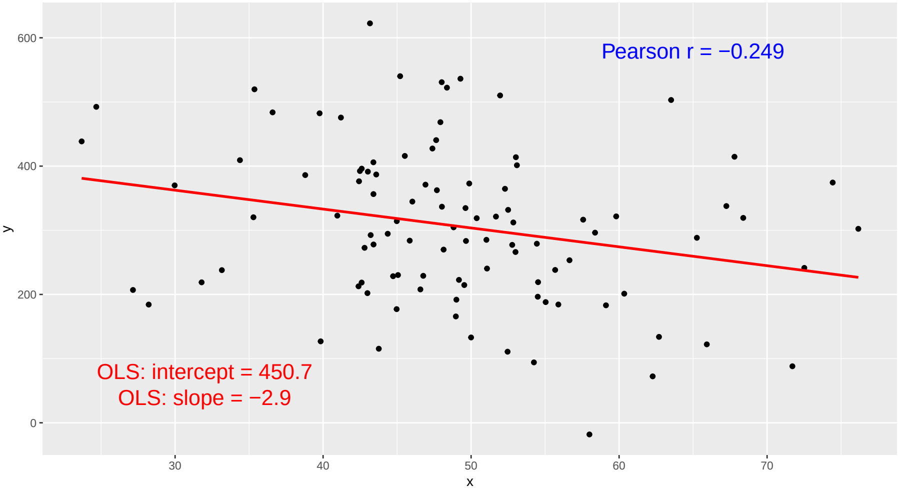
Simulated Example 1



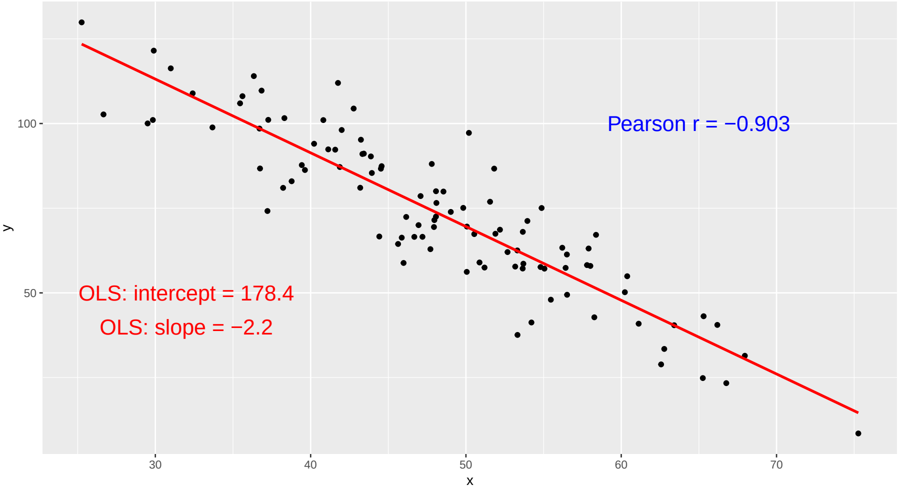
Simulated Example 2



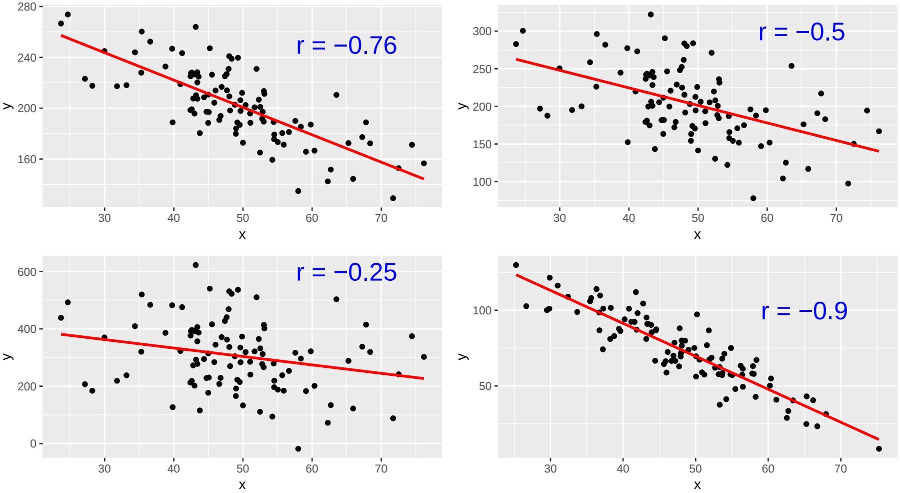
Simulated Example 3



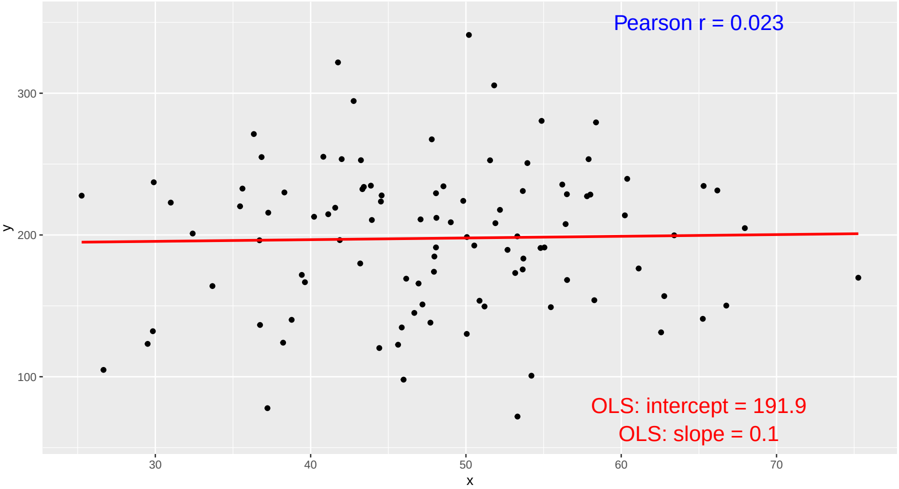
Simulated Example 4



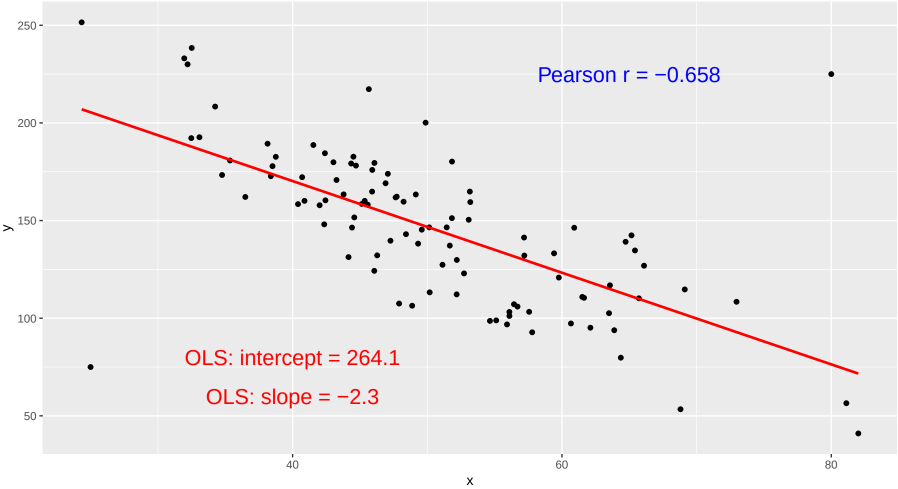
Calibrate Yourself on Correlation Coefficients



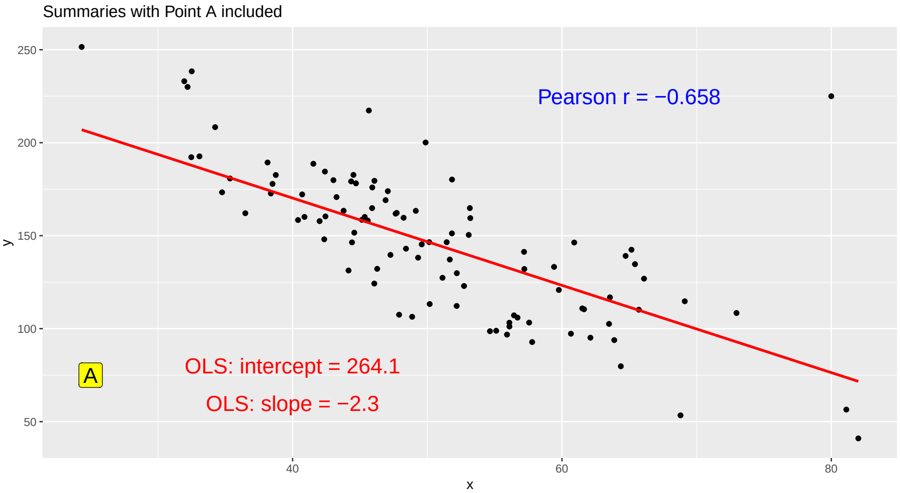
Simulated Example 5



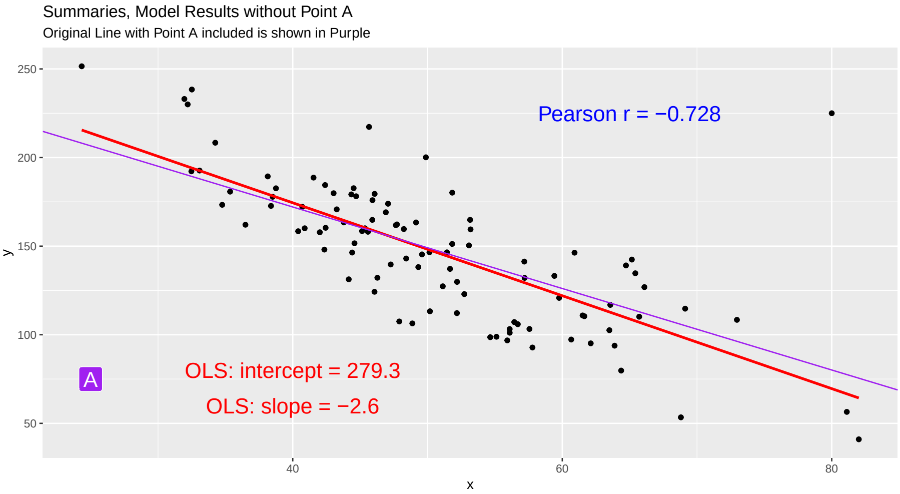
Simulated Example 6



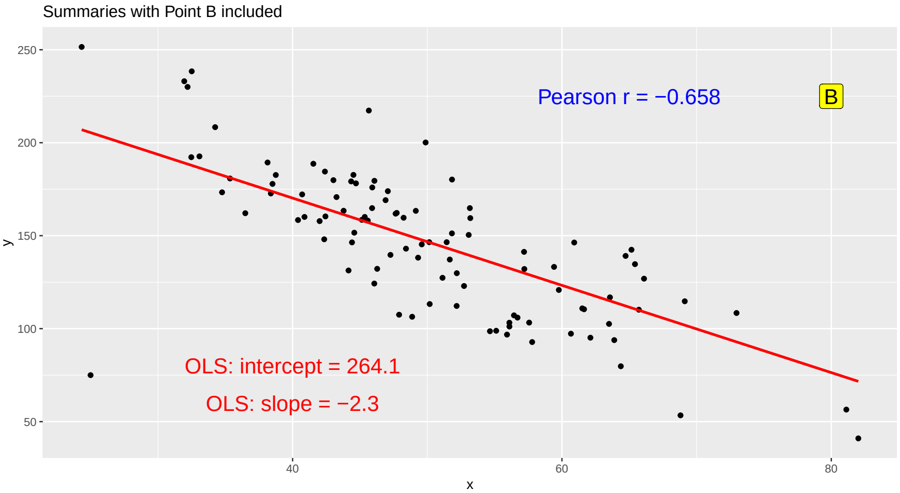
Example 6: What would happen if we omit Point A?



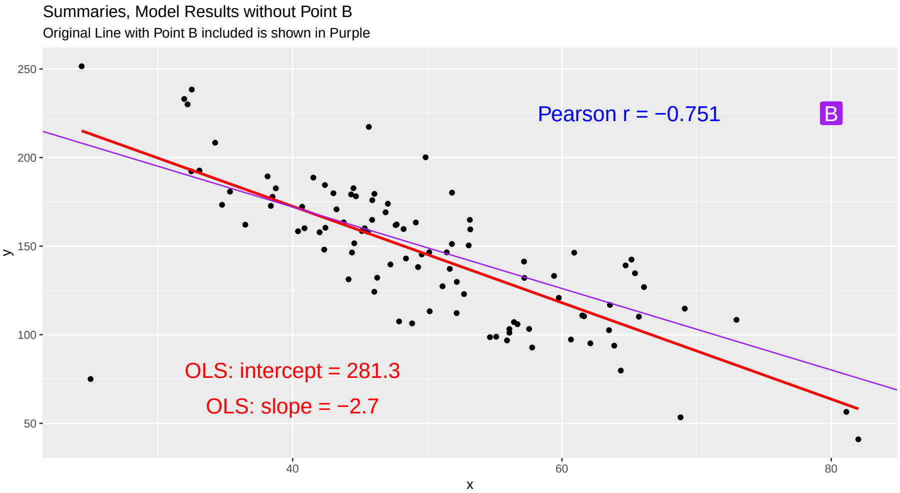
Example 6: Result if we omit Point A



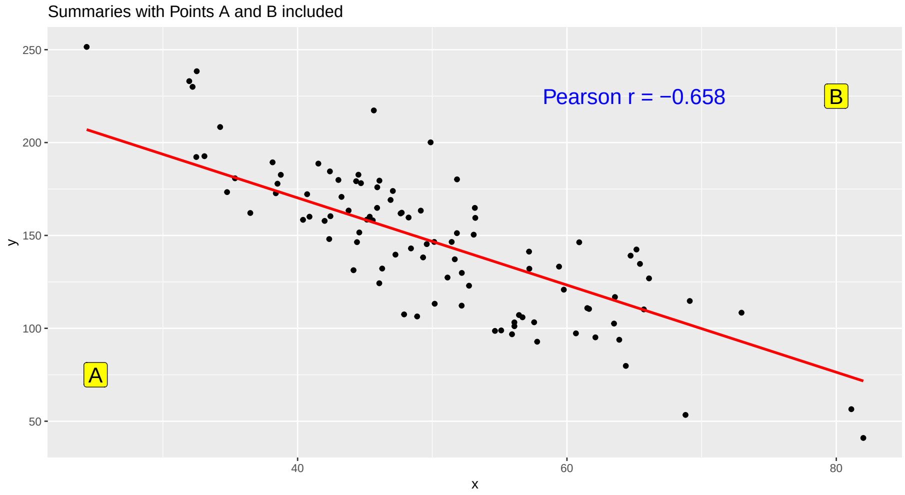
Example 6: What would happen if we omit Point B?



Example 6: Result if we omit Point B



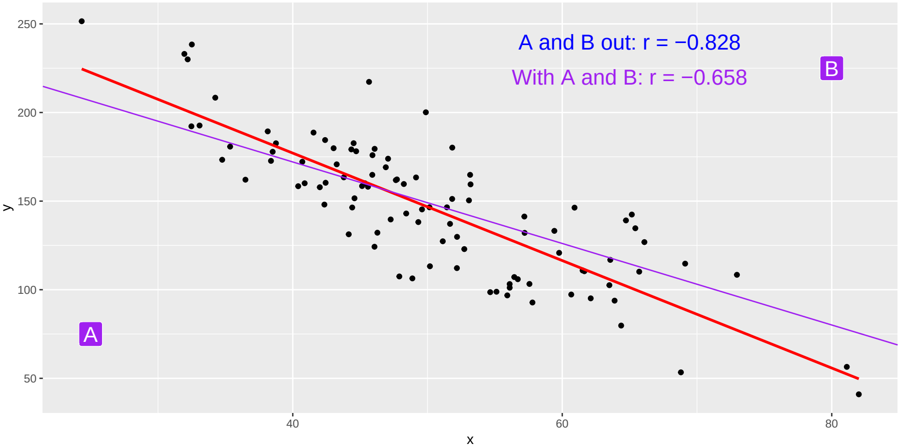
Example 6: What if we omit Point A AND Point B?



Example 6: Result if we omit Points A and B

Summaries, Model Results without A or B

Original Line with Points A and B included is shown in Purple

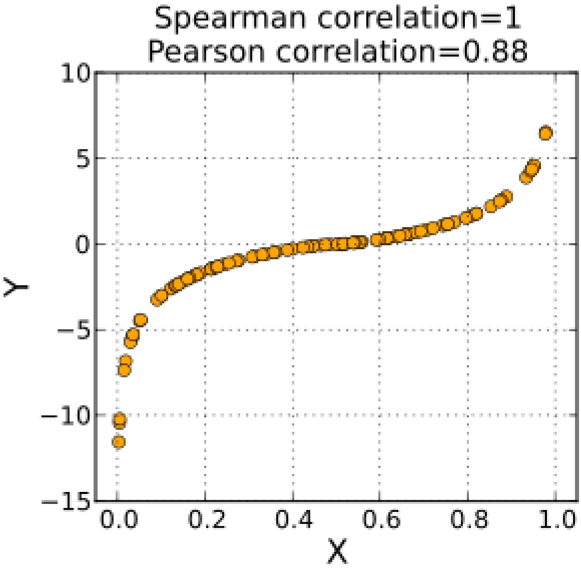


The Spearman Rank Correlation

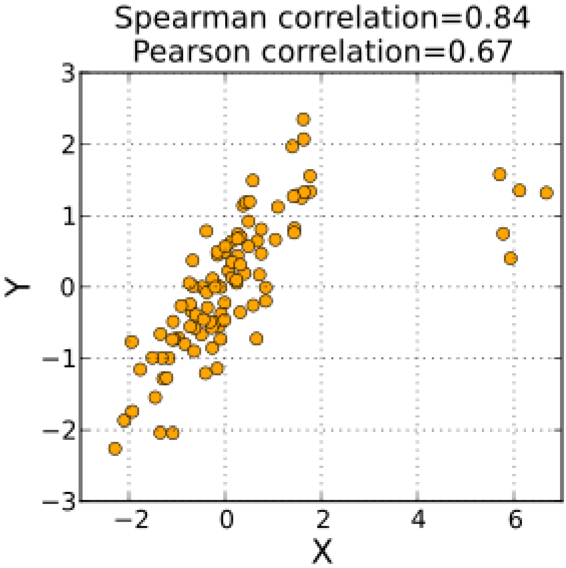
The Spearman rank correlation coefficient assesses how well the association between X and Y can be described using a **monotone function** even if that relationship is not linear.

- A monotone function preserves order - that is, Y must either be strictly increasing as X increases, or strictly decreasing as X increases.
- A Spearman correlation of 1.0 indicates simply that as X increases, Y always increases.
- Like the Pearson correlation, the Spearman correlation is dimension-free, and falls between -1 and +1.
- A positive Spearman correlation corresponds to an increasing (but not necessarily linear) association between X and Y, while a negative Spearman correlation corresponds to a decreasing (but again not necessarily linear) association.

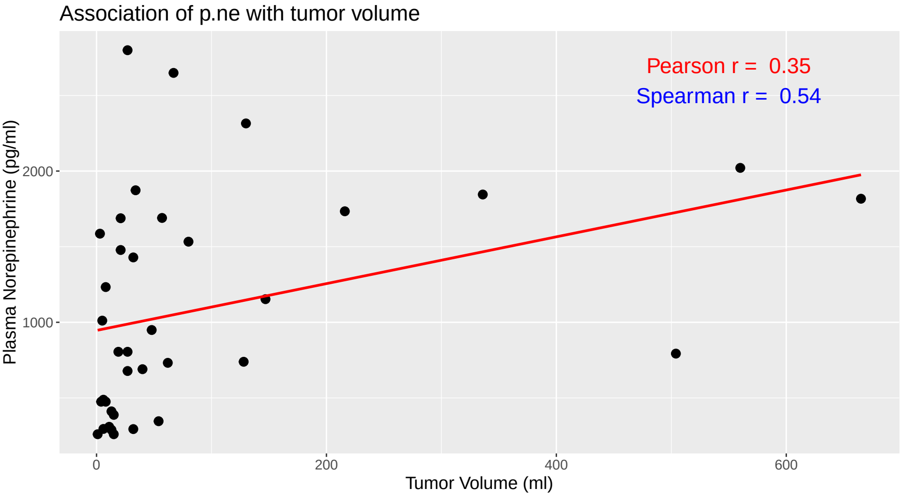
Monotone Association (Source: Wikipedia)



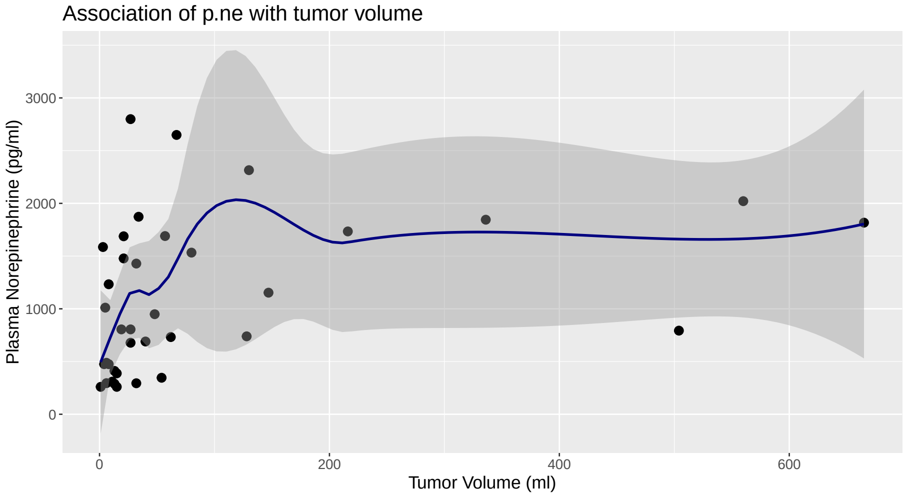
Spearman correlation reacts less to outliers



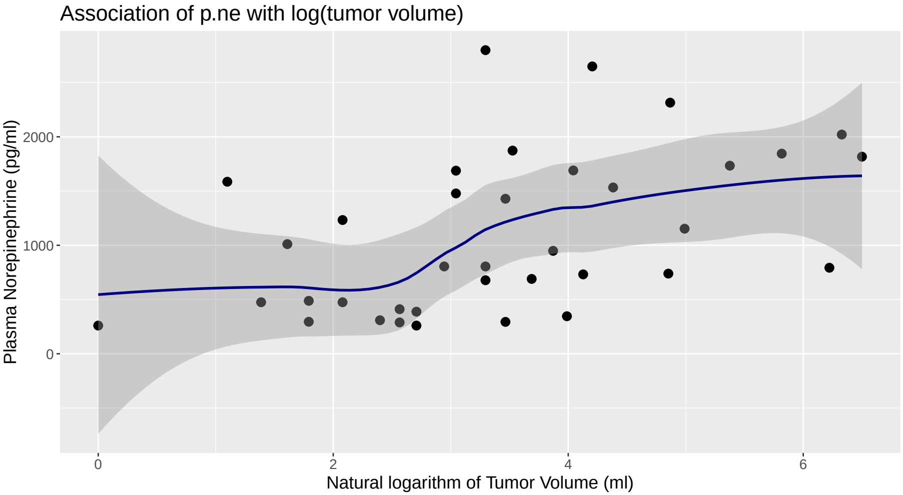
Our Key Scatterplot again



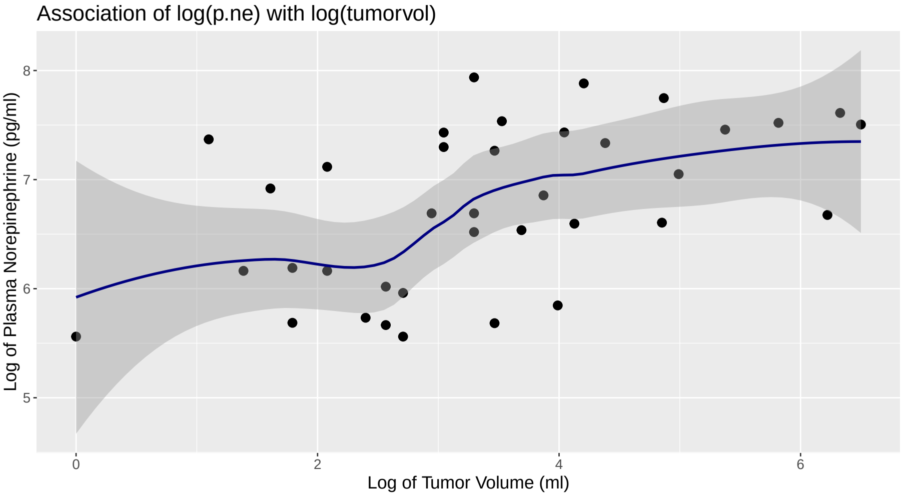
Smoothing using loess, instead



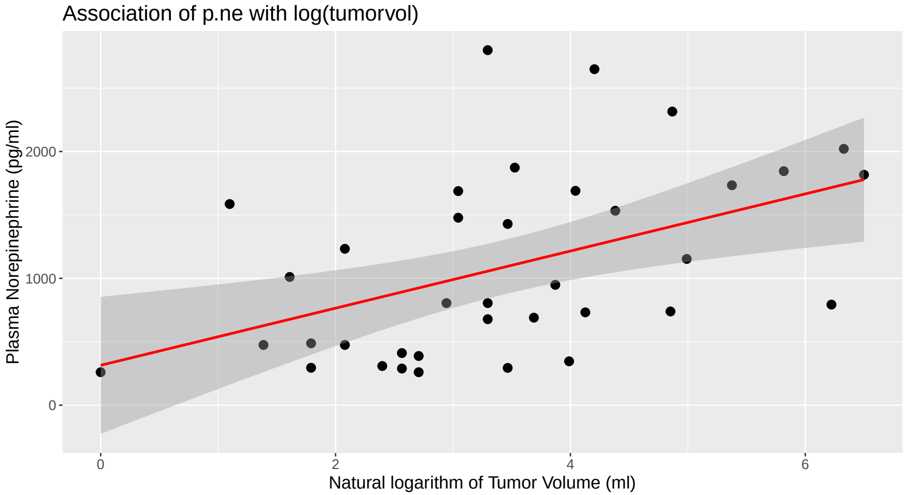
Using the Log transform to spread out the Volumes



Does a Log-Log model seem like a good choice?



Linear Model for p.ne using log(tumor volume)



Creating a Factor to represent disease diagnosis

We want to add a new variable, specifically a factor, called `diagnosis`, which will take the values `von H-L` or `neoplasia`.

- Recall `disease` is a numeric 1/0 variable (0 = von H-L, 1 = neoplasia)
- Use `fct_recode` from the `forcats` package...

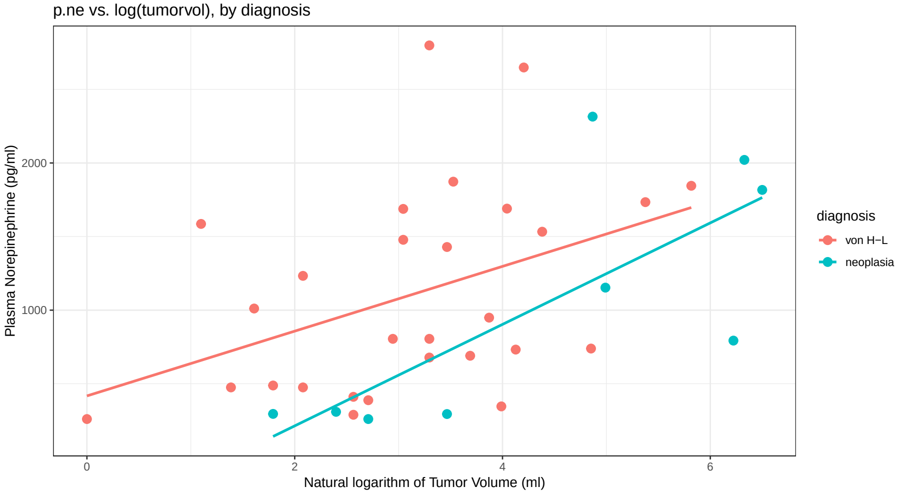
```r
VHL <- VHL %>%
  mutate(diagnosis = fct_recode(factor(disease),
                                "neoplasia" = "1",
                                "von H-L" = "0")
  )
```

Now, what does VHL look like?

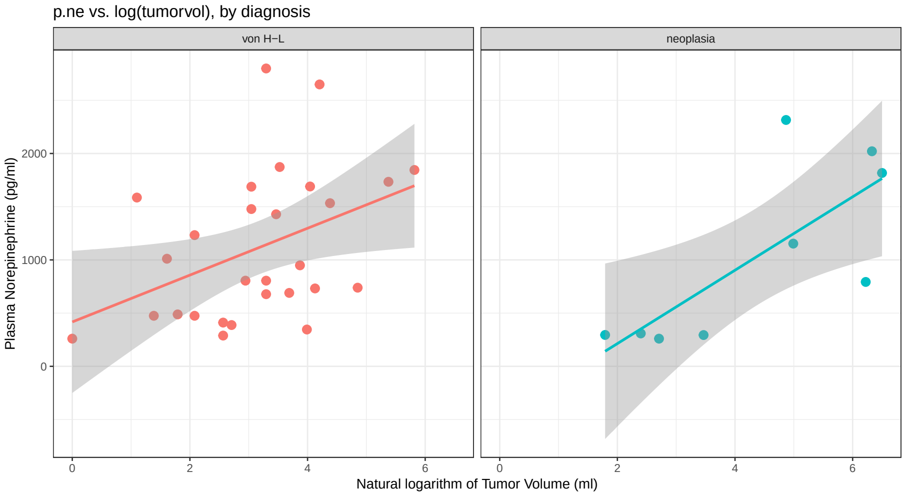
```
VHL
```

```
# A tibble: 37 x 5
      id disease  p.ne tumorvol diagnosis
   <int>   <int> <int>    <int> <fct>
 1   101       0   289       13 von H-L
 2   102       1   294       32 neoplasia
 3   103       0  2799       27 von H-L
 4   104       0  2649       67 von H-L
 5   105       0   346       54 von H-L
 6   106       0  1690       57 von H-L
 7   107       0   805       19 von H-L
 8   108       1  1153      147 neoplasia
 9   109       0   678       27 von H-L
10   110       1  1817      665 neoplasia
# ... with 27 more rows
```

Compare the patients by diagnosis



Facetted Scatterplots by diagnosis



Model accounting for different slopes and intercepts

```
model2 <- lm(p.ne ~ log(tumorvol) * diagnosis, data = VHL)
model2
```

```

Call:
lm(formula = p.ne ~ log(tumorvol) * diagnosis, data = VHL)

Coefficients:
                     (Intercept)
                           417.2
                   log(tumorvol)
                           220.0
              diagnosisneoplasia
                          -893.3
log(tumorvol):diagnosisneoplasia
                           124.8
```

Model 2 results

`p.ne` = 417 + 220 log(`tumorvol`) - 893 (`diagnosis = neoplasia`) + 125 (`diagnosis = neoplasia`)*log(`tumorvol`)

where the indicator variable (`diagnosis = neoplasia`) = 1 for neoplasia subjects, and 0 for other subjects...

- Model for `p.ne` in von H-L patients:
    - 417 + 220 log(`tumorvol`)
- Model for `p.ne` in neoplasia patients:
    - (417 - 893) + (220 + 125) log(`tumorvol`)
    - -476 + 345 log(`tumorvol`)

Model 2 Predictions

What is the predicted `p.ne` for a single new subject with `tumorvol` = 200 ml (so log(tumorvol) = 5.3) in each diagnosis category?

```
predict(model2, newdata = data_frame(tumorvol = 200,
        diagnosis = "neoplasia"), interval = "prediction")
```

```
       fit       lwr     upr
1 1350.896 -28.0571 2729.85
```

```
predict(model2, newdata = data_frame(tumorvol = 200,
        diagnosis = "von H-L"), interval = "prediction")
```

```
       fit      lwr      upr
1 1583.079 208.6489 2957.509
```

Tidying the Model 2 coefficients, with broom

```r
broom::tidy(model2)
```

```
# A tibble: 4 x 5
  term            estimate std.error statistic p.value
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)         417.      318.      1.31   0.199
2 log(tumorvol)       220.       93.6     2.35   0.0248
3 diagnosisneopl~    -893.      659.     -1.36   0.184
4 log(tumorvol):~     125.      155.      0.807  0.425
```

Model 2, summarized at a glance, with broom

```
broom::glance(model2)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df
*     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>
1     0.290         0.226  634.      4.50 0.00937     4
# ... with 5 more variables: logLik <dbl>, AIC <dbl>,
#   BIC <dbl>, deviance <dbl>, df.residual <int>
```

Compare this to model 1…

```
broom::glance(model1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic p.value    df
*     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>
1     0.120        0.0950  685.      4.78  0.0356     2
# ... with 5 more variables: logLik <dbl>, AIC <dbl>,
#   BIC <dbl>, deviance <dbl>, df.residual <int>
```

Group Discussion

1. As a group, specify a research question using only the set of questions I have already decided to include in the data set (list on next slide, and in section 2.2.2 of the Project Instructions)
2. As a group, brainstorm three additional questions you would like to include in the survey. Be sure one of them produces a quantitative result and at least one produces a categorical result. For the categorical question(s), be sure to specify each possible category into which a response could fall.
3. Now, specify a new research question which can be addressed using at least two of your three new questions specified in task 2.

Form to present your discussion is at http://bit.ly/431-2018-brainstorm-10

The 15 Questions We'll Start With...

1. Were you born in the United States?
2. Is English the language you speak better than any other?
3. Do you identify as female?
4. Do you wear prescription glasses or contact lenses?
5. Before taking 431, had you ever used R before?
6. Are you currently married or in a stable domestic relationship?
7. Have you smoked 100 cigarettes or more in your entire life?
8. In what year were you born?
9. How would you rate your current health overall (Excellent, Very Good, Good, Fair, Poor)
10. For how long, in months, have you lived in Northeast Ohio?
11. What is your height in inches?
12. What is your weight in pounds?
13. What is your pulse rate, in beats per minute?
14. Last week, on how many days did you exercise? (0 - 7)
15. Last night, how many hours of sleep did you get?