431 Class 08

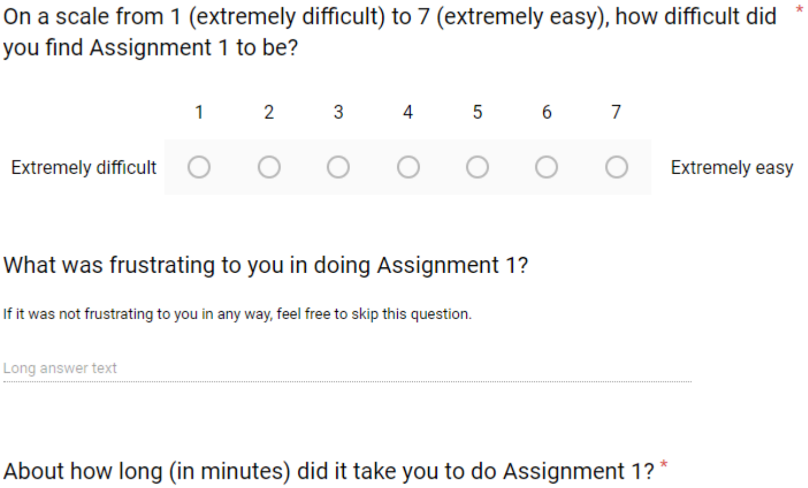
Thomas E. Love

2017-09-21

Today's Agenda

1. Comments on the Google Form & Assignment 1
2. More on Transformations (Notes: Ch 9, 11)
3. Summaries within subgroups (Notes: Ch 10)
4. Associations, Using Linear Models (Notes: Ch 11)
    - A study of von Hippel-Lindau disease
    - Associations, Correlation and Scatterplots
    - Fitting a Linear Model

Comments on The Google Form (re: Assignment 1)



On a scale from 1 (extremely difficult) to 7 (extremely easy), how difficult did you find Assignment 1 to be? *

	1	2	3	4	5	6	7	
Extremely difficult	○	○	○	○	○	○	○	Extremely easy

What was frustrating to you in doing Assignment 1?

If it was not frustrating to you in any way, feel free to skip this question.

Long answer text

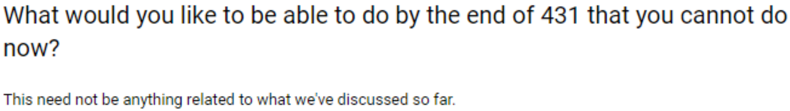
About how long (in minutes) did it take you to do Assignment 1? *

Is there a spell-check in R Studio?

Sure. Just hit F7, or the abc key with a check mark.

Comments on The Google Form (final item)

What did you folks want to be able to predict?



What would you like to be able to do by the end of 431 that you cannot do now?

This need not be anything related to what we've discussed so far.

R Setup

```r
library(viridis); library(gridExtra); library(ggridges)
library(knitr); library(pander)
library(tidyverse)

source("Love-boost.R")

nyfs1 <- read.csv("nyfs1.csv") %>% tbl_df
names(nyfs1)
```

```
[1] "subject.id"       "sex"
[3] "age.exam"         "bmi"
[5] "bmi.cat"          "waist.circ"
[7] "triceps.skinfold"
```

Why Transform?

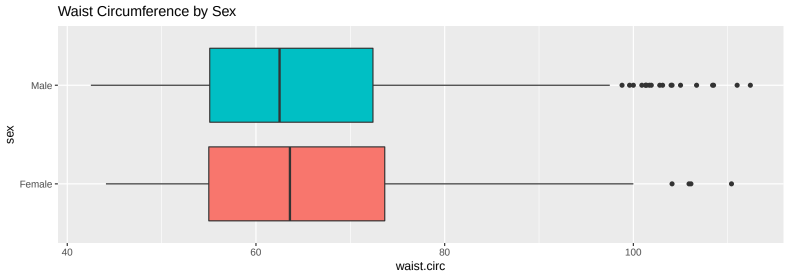
When we have unimodal but skewed data, we will often **transform** the data using a log, inverse, square root, square, etc. in order to obtain a new distribution which is closer to the Normal.

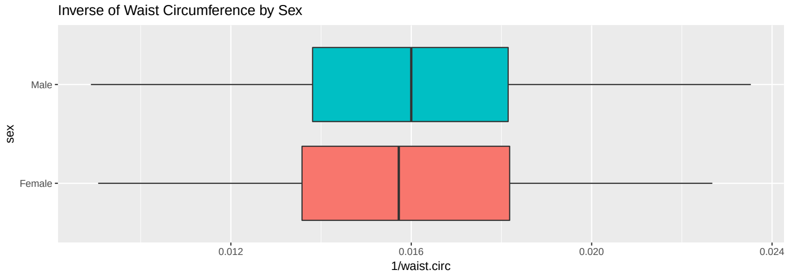
1. Sometimes we do this to facilitate comparisons.
    - Example: t-test to compare mean waist circumference among male children to female children
    - t-test requires that the distribution of the outcome in each sex be approximately Normal

Does a Transformation Help with Comparison?

```
p1 <- ggplot(nyfs1, aes(x = sex, y = waist.circ,
                  fill = sex)) +
  geom_boxplot() +
  coord_flip() +
  guides(fill = FALSE) +
  labs(title = "Waist Circumference by Sex")

p2 <- ggplot(nyfs1, aes(x = sex, y = 1/waist.circ,
                  fill = sex)) +
  geom_boxplot() +
  coord_flip() +
  guides(fill = FALSE) +
  labs(title = "Inverse of Waist Circumference by Sex")

gridExtra::grid.arrange(p1, p2)
```

Boxplots of Waist Circumference by Sex


Waist Circumference by Sex


Inverse of Waist Circumference by Sex

Why Transform?

When we have unimodal but skewed data, we will often **transform** the data using a log, inverse, square root, square, etc. in order to obtain a new distribution which is closer to the Normal.

2. Sometimes we do this to facilitate model-building.
    - What is the association of waist circumference with triceps skinfold?
    - Transformations that "normalize" the distributions of skewed variables also can "linearize" an apparent association.

A Quick Interlude: Ohio's population, 1850-1970

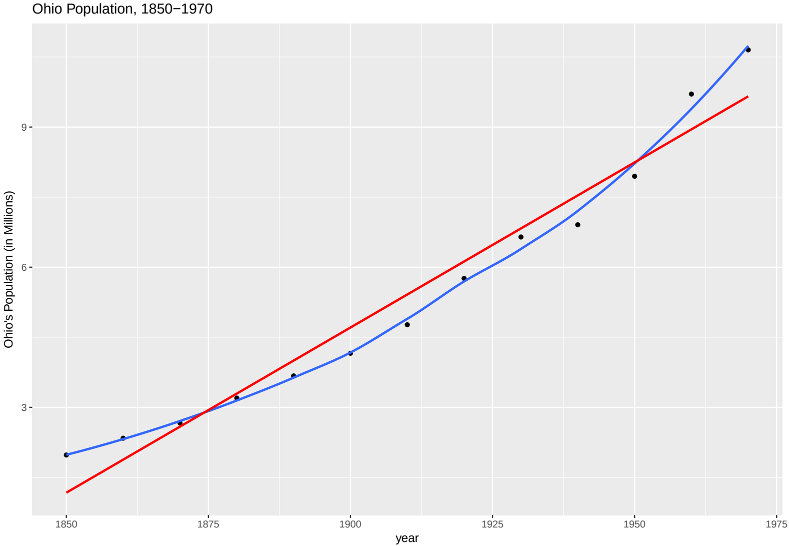
Source: http://population.us/oh/

```r
ohio1 <- data_frame(
  year = seq(from = 1850, to = 1970, by = 10),
  pop_oh = c(1980329, 2339511, 2665260,
             3198062, 3672329, 4157545,
             4767121, 5759394, 6646697,
             6907612, 7946627, 9706397, 10652017))
```

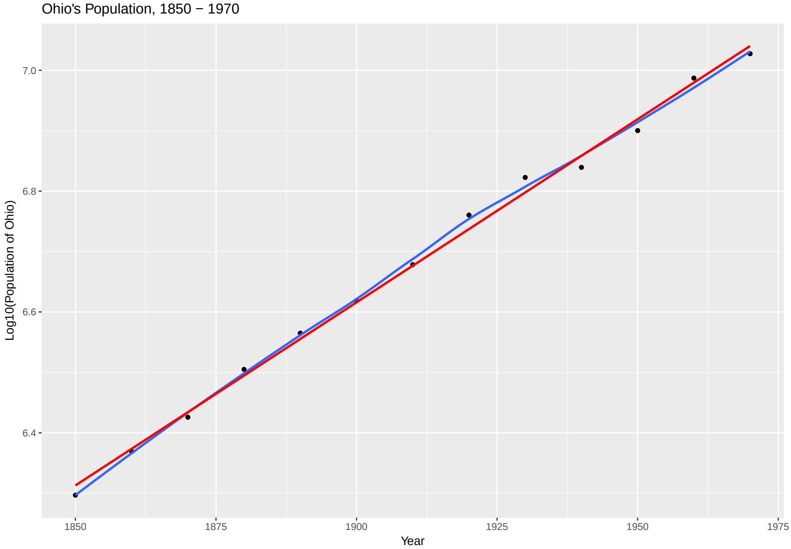
Was Ohio's population growth linear? (1850-1970)

```r
ggplot(ohio1, aes(x = year, y = pop_oh/1000000)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  labs(title = "Ohio Population, 1850-1970",
       y = "Ohio's Population (in Millions)")
```

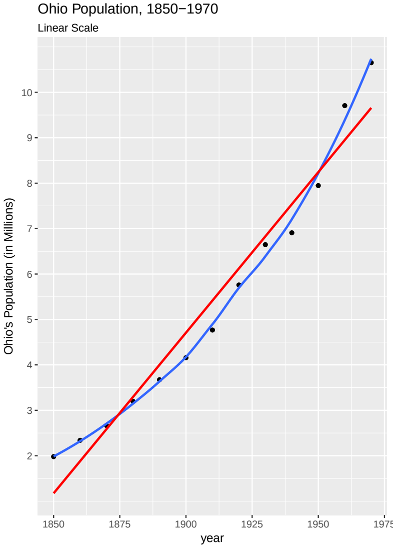
Was Ohio's population growth linear? (1850-1970)

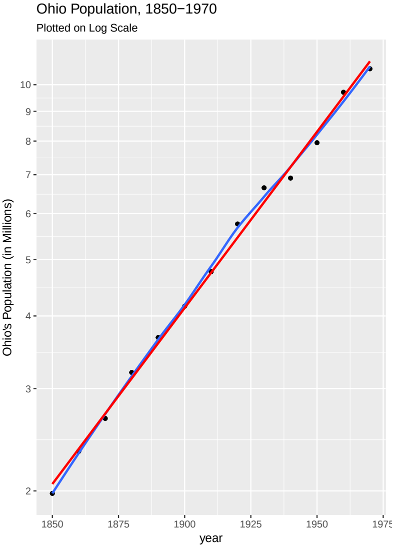


Was Ohio's log(population) linear in time?



Comparing the Linear to the Log Scale



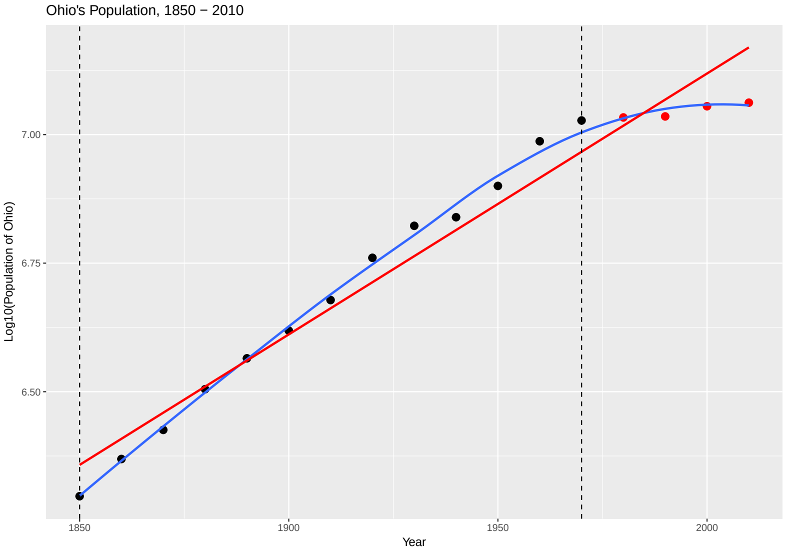


Ohio's population grew at a rate of 2% per decade

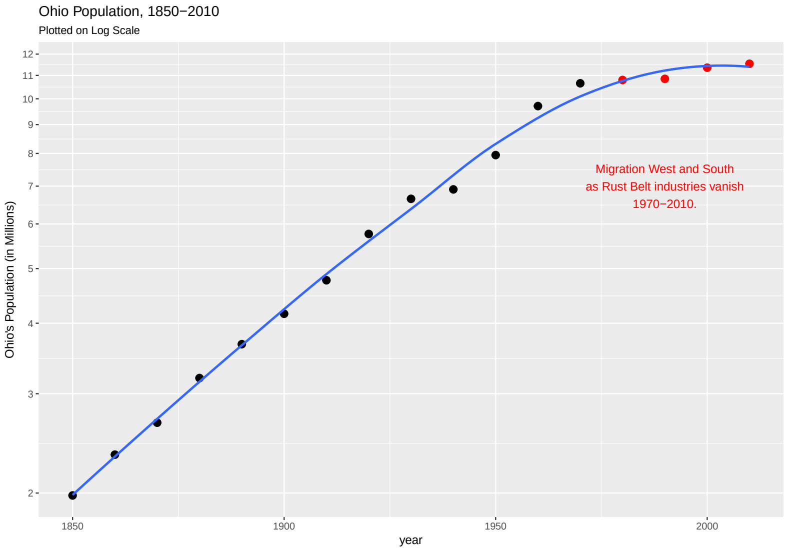
- If population grows **exponentially** over time, then log(population) will be **linear** in time.

What happened starting in 1970 in Ohio?

What about 1980-2010?



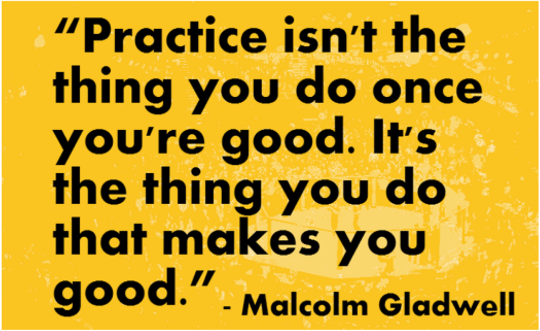
1850-2010 Ohio Population on Logarithmic Scale



Code for Previous Slide

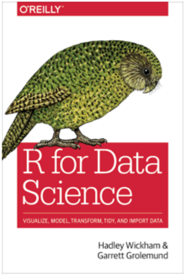
```r
ggplot(ohio2, aes(x = year, y = pop_oh/1000000)) +
  geom_point(col = ohio2$post, size = 3) +
  scale_y_log10(breaks = 2:12) +
  geom_smooth(method = "loess", se = FALSE) +
  annotate("text", x = 1990, y = 7, col = "red",
           label = "Migration West and South\nas Rust Belt inc
  labs(title = "Ohio Population, 1850-2010",
       subtitle = "Plotted on Log Scale",
       y = "Ohio's Population (in Millions)")
```

How do we learn how to build plots like that?



"Practice isn't the thing you do once you're good. It's the thing you do that makes you good." - Malcolm Gladwell

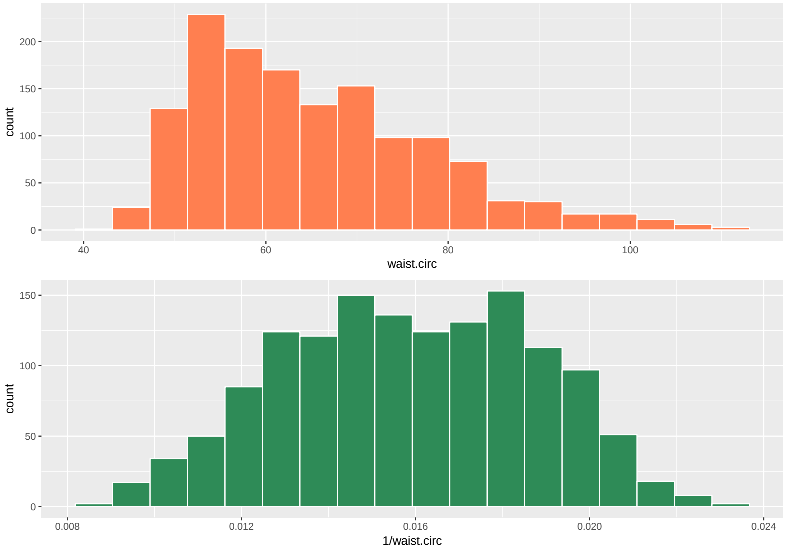
And there's more than a little of this, too. . .



Transforming the Waist Circumference Data

```r
p1 <- ggplot(nyfs1,
             aes(x = waist.circ)) +
  geom_histogram(bins = 18,
                 fill = "coral", col = "white")

p2 <- ggplot(nyfs1,
             aes(x = 1/waist.circ)) +
  geom_histogram(bins = 18,
                 fill = "seagreen", col = "white")

gridExtra::grid.arrange(p1, p2)
```

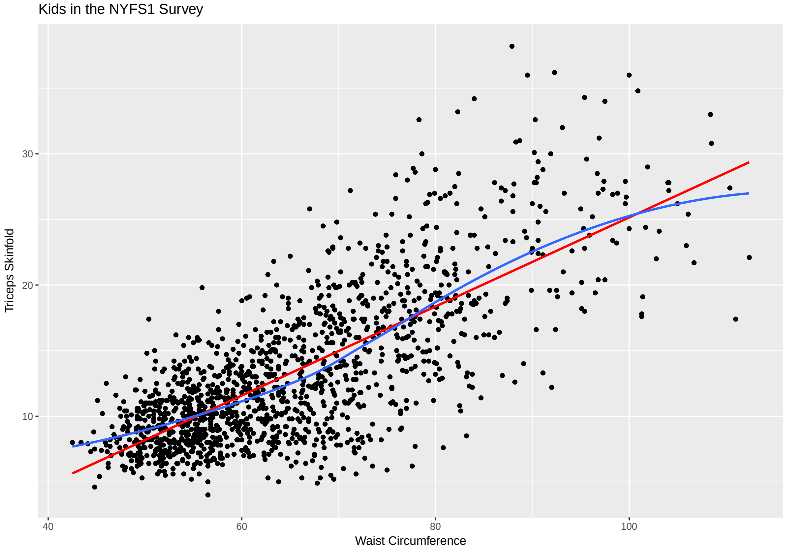
The Resulting Plot Array



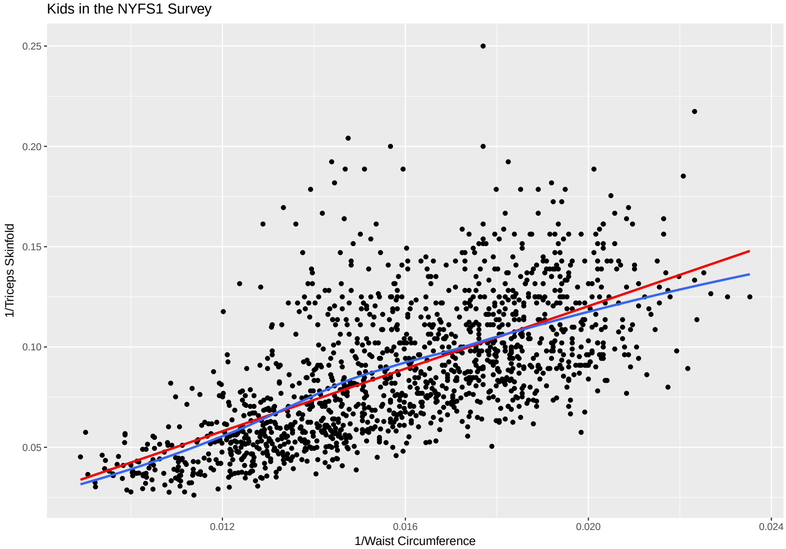
Is Waist Circumference related to Triceps Skinfold?

```
ggplot(nyfs1, aes(x = waist.circ, y = triceps.skinfold)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "red") +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Waist Circumference",
       y = "Triceps Skinfold",
       title = "Kids in the NYFS1 Survey")
```

Is Waist Circumference related to Triceps Skinfold?



After Inverse Transformations (no real help here)



BMI categories

```
nyfs1 %>%
  count(bmi.cat)
```

```
# A tibble: 4 x 2
          bmi.cat     n
           <fctr> <int>
1   1 Underweight    42
2 2 Normal weight   926
3    3 Overweight   237
4         4 Obese   211
```

Use group_by and summarize together

```
nyfs1 %>%
  group_by(bmi.cat) %>%
  summarize(mean = round(mean(waist.circ),1),
            median = median(waist.circ),
            sd = round(sd(waist.circ),1),
            skew1 = round(skew1(waist.circ),2))
```

```
# A tibble: 4 x 5
          bmi.cat  mean median    sd skew1
           <fctr> <dbl>  <dbl> <dbl> <dbl>
1   1 Underweight  54.9   53.9   7.6  0.14
2 2 Normal weight  61.0   59.2   9.1  0.19
3    3 Overweight  71.1   72.0  11.8 -0.08
4         4 Obese  79.9   79.9  15.0  0.00
```

Using knitr::kable to present the table

```
nyfs1 %>%
  group_by(bmi.cat) %>%
  summarize(mean = round(mean(waist.circ),1),
            median = median(waist.circ),
            sd = round(sd(waist.circ),1),
            skew1 = round(skew1(waist.circ),2)) %>%
  knitr::kable()
```

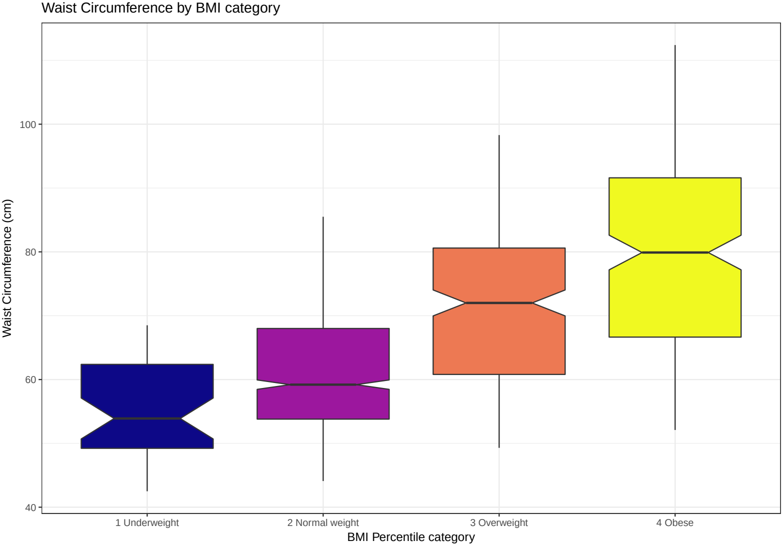
bmi.cat	mean	median	sd	skew1
1 Underweight	54.9	53.9	7.6	0.14
2 Normal weight	61.0	59.2	9.1	0.19
3 Overweight	71.1	72.0	11.8	-0.08
4 Obese	79.9	79.9	15.0	0.00

Using pander::pander to present the table

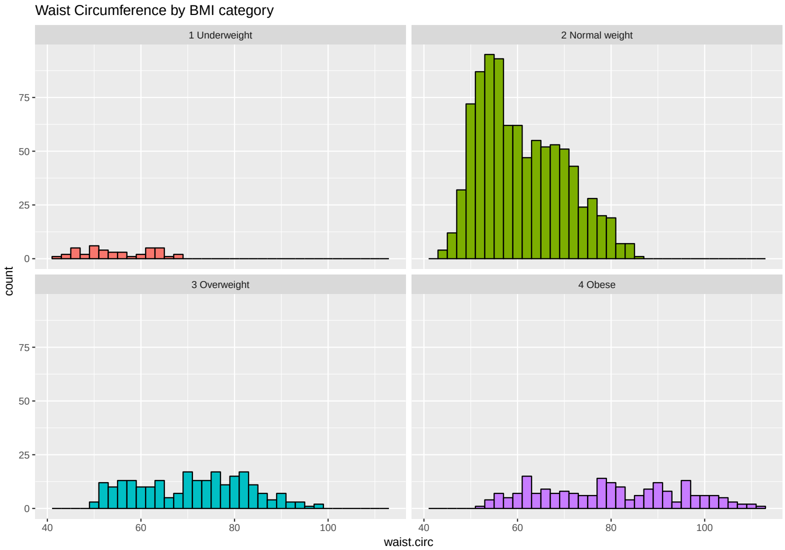
```
nyfs1 %>%
  group_by(bmi.cat) %>%
  summarize(mean = round(mean(waist.circ),1),
            median = median(waist.circ),
            sd = round(sd(waist.circ),1),
            skew1 = round(skew1(waist.circ),2)) %>%
  pander::pander()
```

bmi.cat	mean	median	sd	skew1
1 Underweight	54.9	53.9	7.6	0.14
2 Normal weight	61	59.2	9.1	0.19
3 Overweight	71.1	72	11.8	-0.08
4 Obese	79.9	79.9	15	0

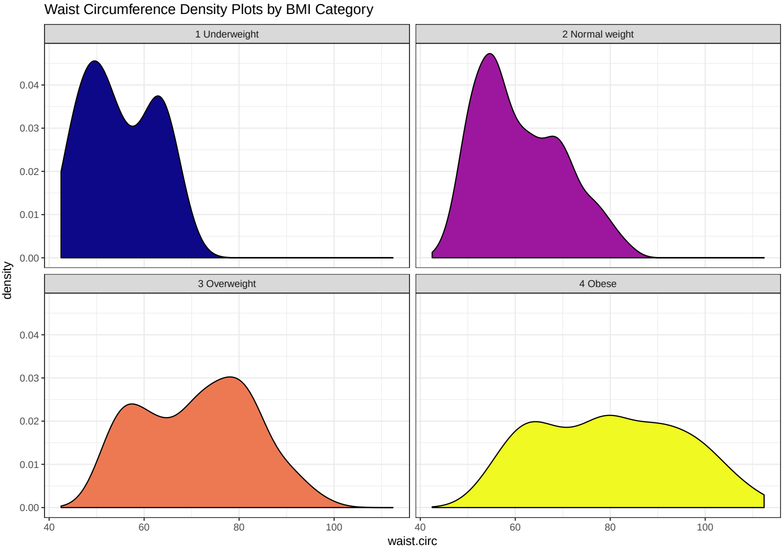
Comparison Boxplots with Notches



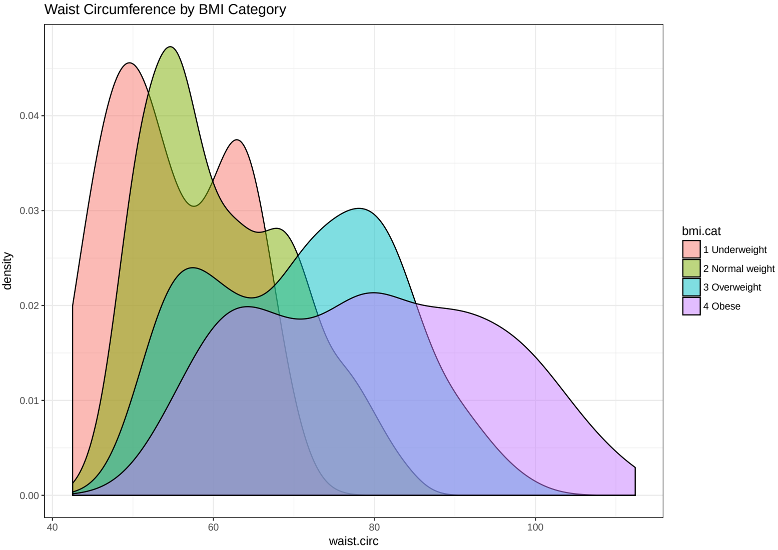
Comparing Distributions with Faceted Histograms



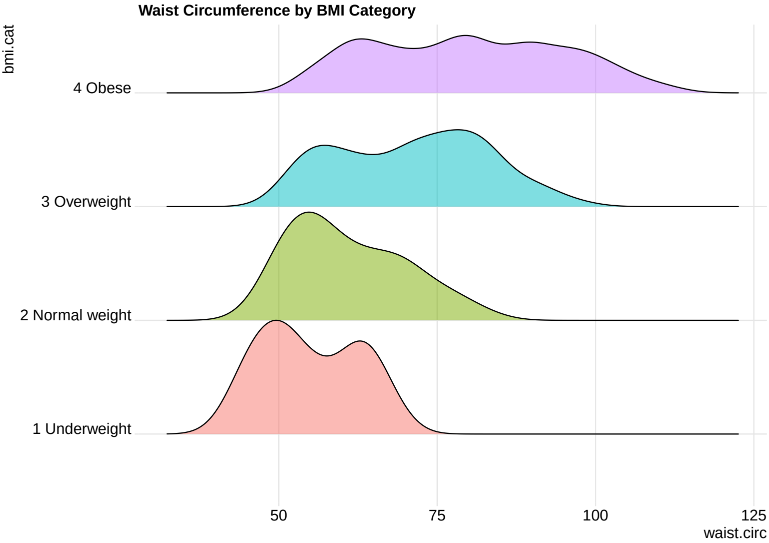
Density Plots, Faceted

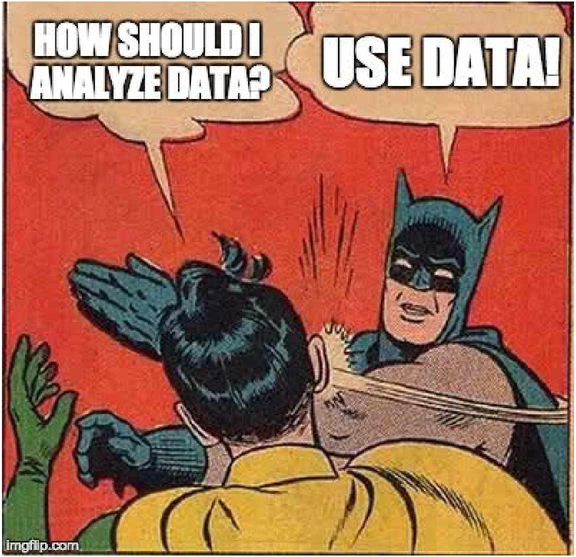


Density Plots, Overlapping



Ridgeline Plots (formerly Joy Plots)




HOW SHOULD I ANALYZE DATA?
USE DATA!
imgflip.com

A study of von Hippel-Lindau disease

Eisenhofer et al.[1] (1999) investigated the use of plasma normetanephrine and metanephrine for detecting pheochromocytoma in 35 subjects. 9 of the patients were diagnosed with multiple endocrine neoplasia type 2 and the rest with von Hippel-Lindau disease.

Our first goal is to understand the association between plasma norepinephrine and tumor volume across all of the subjects.

Then, we'll be interested in addressing the impact of diagnosis on this association.

I've stored the data in the `vonHippel-Lindau.csv` file on the data site.

[1] Reference: Eisenhofer GJ et al. (1999) "Plasma normetanephrine and metanephrine for detecting pheochromocytoma in von Hippel-Lindau disease and multiple endocrine neoplasia type 2" NEJM 340(24): 1872-9. My Source: http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/index.html

Looking over the data

```
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
10   110       1  1817      665
# ... with 27 more rows
```

Basic Numerical Summaries

```
    disease            p.ne         tumorvol     
 Min.   :0.0000   Min.   : 260   Min.   :  1.00  
 1st Qu.:0.0000   1st Qu.: 475   1st Qu.: 13.00  
 Median :0.0000   Median : 805   Median : 27.00  
 Mean   :0.2432   Mean   :1090   Mean   : 93.03  
 3rd Qu.:0.0000   3rd Qu.:1688   3rd Qu.: 67.00  
 Max.   :1.0000   Max.   :2799   Max.   :665.00  
```

Codebook

- `disease` = 1 for patients with multiple endocrine neoplasia type 2
- `disease` = 0 for patients with von Hippel-Lindau disease
- `p.ne` = plasma norepinephrine (pg/ml)
- `tumorvol` = tumor volumne (ml)

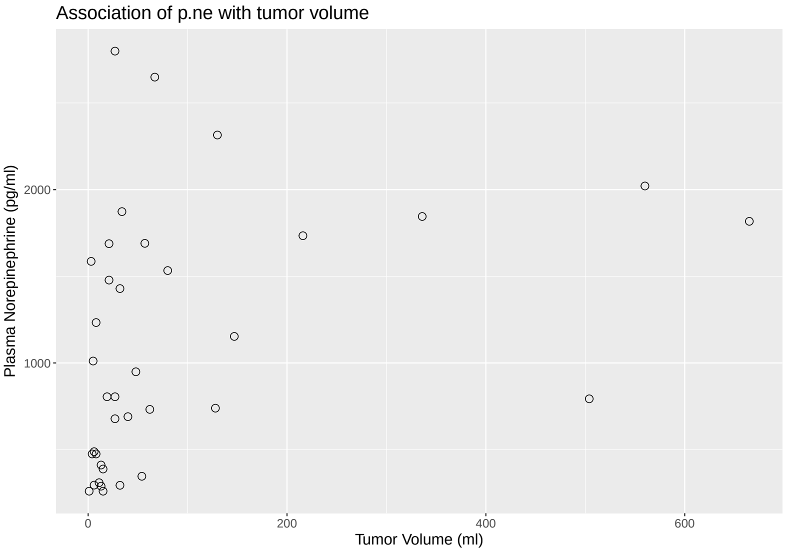
Creating a Factor to represent disease information

Label the disease data (0s and 1s) appropriately, in a new *factor*

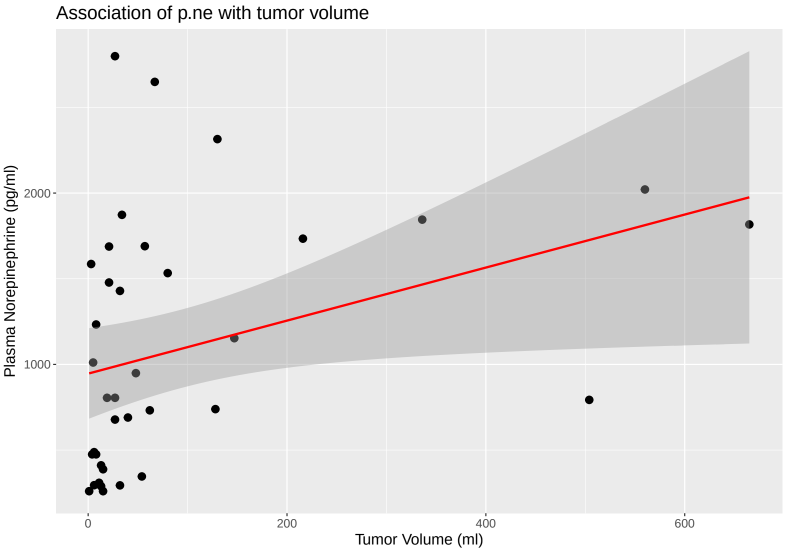
```
VHL$diagnosis <- factor(VHL$disease,
                        labels = c("von H-L", "neoplasia"))
table(VHL$diagnosis, VHL$disease)
```

```
            0  1
  von H-L   28  0
  neoplasia  0  9
```

Plotting an Association across all 37 subjects



Adding a Linear Model



The Linear Model

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

```

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

Coefficients:
(Intercept)     tumorvol
    946.185        1.547
```

- What does this model predict? Using what predictor?

The Linear Model

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

```

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

Coefficients:
(Intercept)     tumorvol
    946.185        1.547
```

- What does this model predict? Using what predictor?
- Predict `p.ne` for a subject with tumor volume = 100 ml.

Using the model to make predictions…

1. A 95% **prediction interval** for a single subject with volume 100 ml.

```
predict(model1, newdata = data_frame(tumorvol = 100),
        interval = "prediction", level = 0.95)
```

```
       fit       lwr      upr
1 1100.925 -308.7478 2510.598
```

- Can we make a prediction for all subjects in the population with a particular tumor volume, not just an individual?

Using the model to make predictions. . .

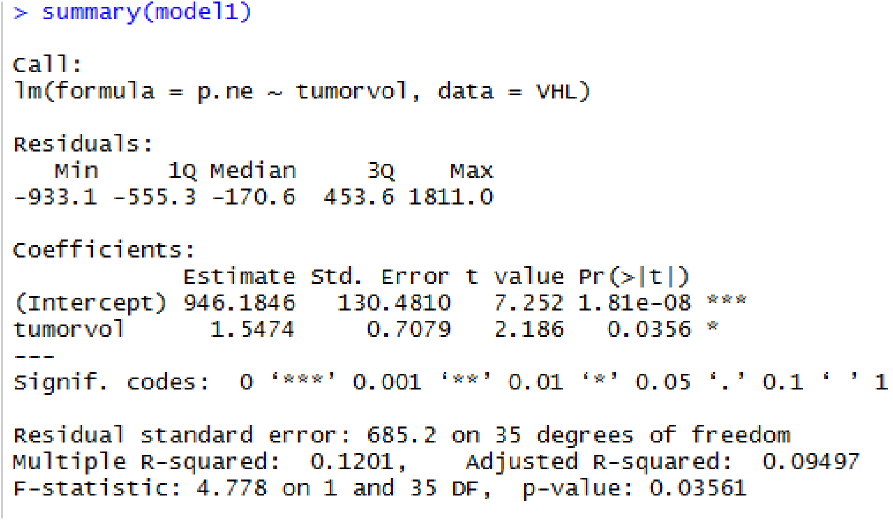
2. 95% **confidence interval** for the average of the population of all subjects with volume 100 ml, as well as for the population of subjects with volume 50 ml.

```
newvals <- data_frame(tumorvol = c(100, 50))

predict(model1, newdata = newvals,
        interval = "confidence", level = 0.95)
```

```
       fit      lwr      upr
1 1100.925 872.0323 1329.818
2 1023.555 786.6676 1260.442
```

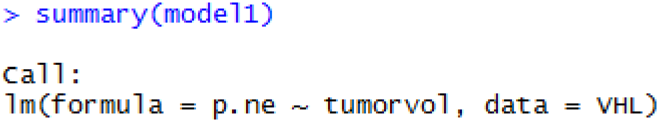
Summary of our Linear Model

```
Call:
lm(formula = p.ne ~ tumorvol, data = VHL)

Residuals:
   Min     1Q Median     3Q    Max 
-933.1 -555.3 -170.6  453.6 1811.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 946.1846   130.4810   7.252 1.81e-08 ***
tumorvol      1.5474     0.7079   2.186   0.0356 *  
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 685.2 on 35 degrees of freedom
```



```
> summary(model1)

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

Residuals:
   Min     1Q Median     3Q    Max 
-933.1 -555.3 -170.6  453.6 1811.0 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 946.1846   130.4810   7.252 1.81e-08 ***
tumorvol      1.5474     0.7079   2.186   0.0356 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 685.2 on 35 degrees of freedom
Multiple R-squared:  0.1201,	Adjusted R-squared:  0.09497 
F-statistic: 4.778 on 1 and 35 DF,  p-value: 0.03561
```

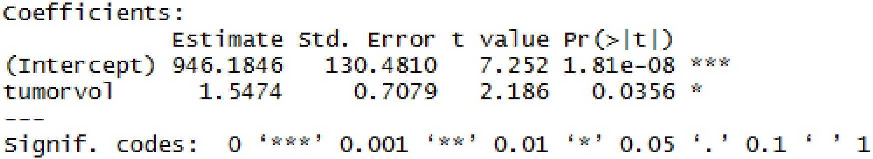
Key Elements of the Summary



```
> summary(model1)

Call:
lm(formula = p.ne ~ tumorvol, data = VHL)
```

- The outcome variable in this model is `p.ne`, and the predictor variable is `tumorvol`.
- The straight line model for these data fitted by least squares is p.ne $=$ 946 $+$ 1.55 `tumorvol`.

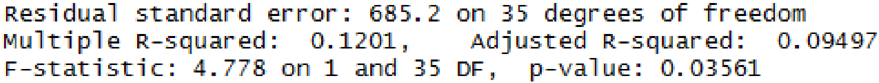
Key Elements of the Summary



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 946.1846   130.4810   7.252 1.81e-08 ***
tumorvol      1.5474     0.7079   2.186   0.0356 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- The slope of `tumorvol` is positive, which indicates that as `tumorvol` increases, we expect that `p.ne` will also increase.
- Specifically, we expect that for every additional ml of `tumorvol`, the `p.ne` is increased by 1.55 pg/ml.

Key Elements of the Summary



```
Residual standard error: 685.2 on 35 degrees of freedom
Multiple R-squared:  0.1201,    Adjusted R-squared:  0.09497
F-statistic: 4.778 on 1 and 35 DF,  p-value: 0.03561
```

- The multiple R-squared (squared correlation coefficient) is 0.12, which implies that 12% of the variation in `p.ne` is explained using this linear model with `tumorvol`.
- It also implies that the Pearson correlation between `p.ne` and `tumorvol` is the square root of 0.12, or 0.347.

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

```
[1] 0.3465646
```

Correlation Coefficients

Two key types of correlation coefficient to describe the association.

- 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.

```
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.

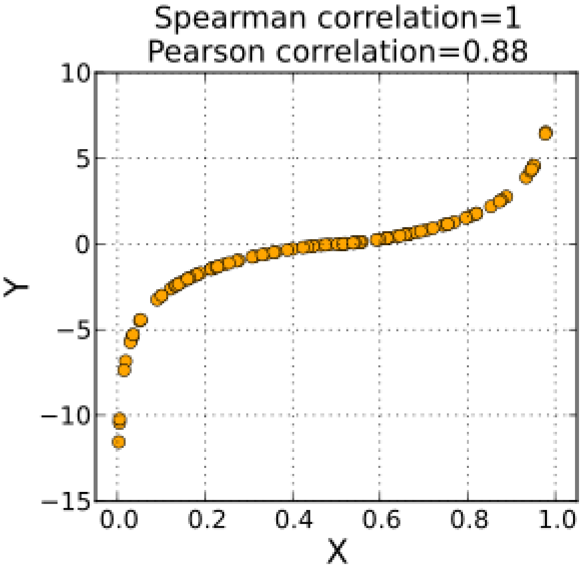
$$r_{XY} = \frac{1}{n-1}\Sigma_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y})$$

The Spearman Rank Correlation

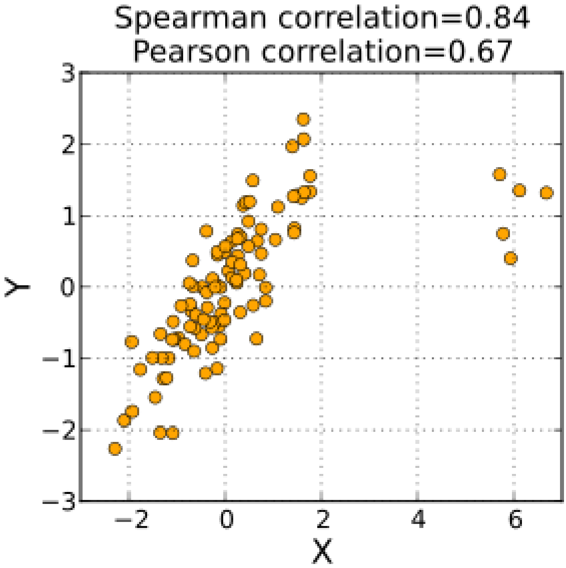
The Spearman rank correlation coefficient assesses how well the association between X and Y can be described using a **monotone function** even if that relationship is not linear.

- 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.

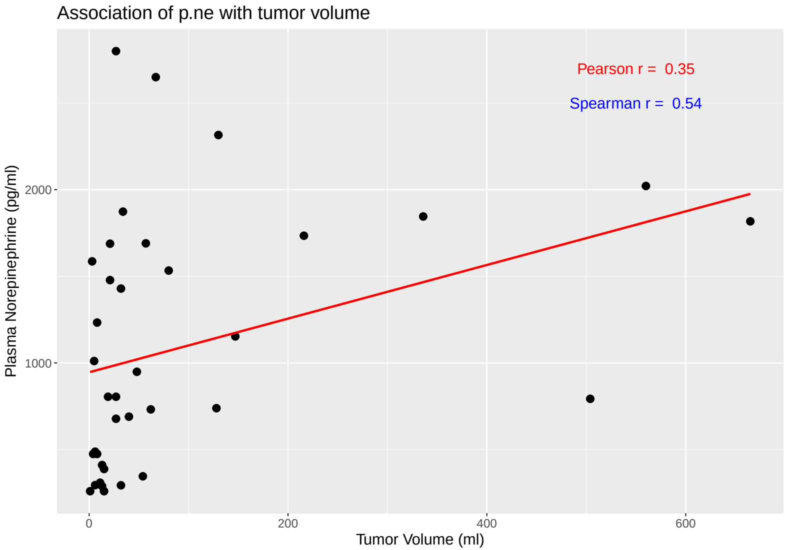
Monotone Association (Source: Wikipedia)



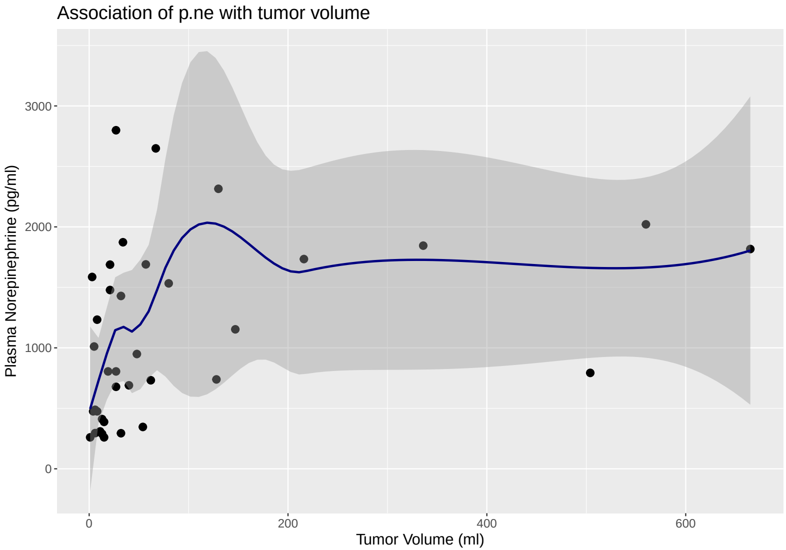
Spearman correlation reacts less to outliers



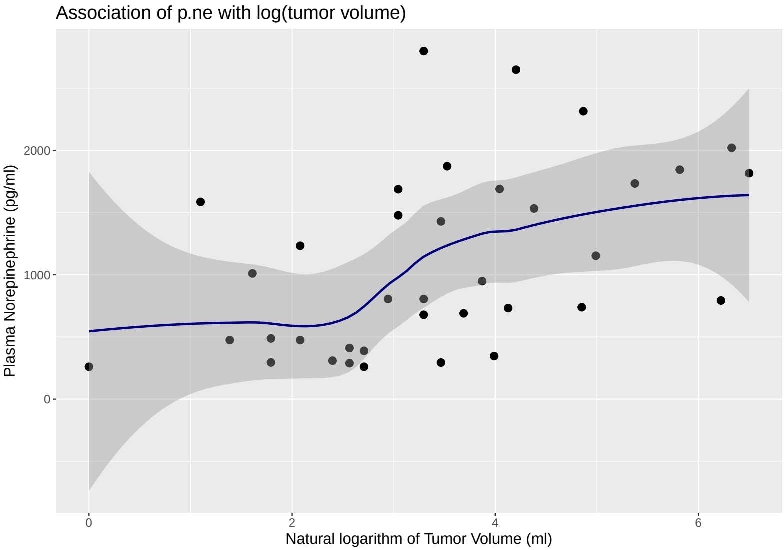
Our Key Scatterplot again



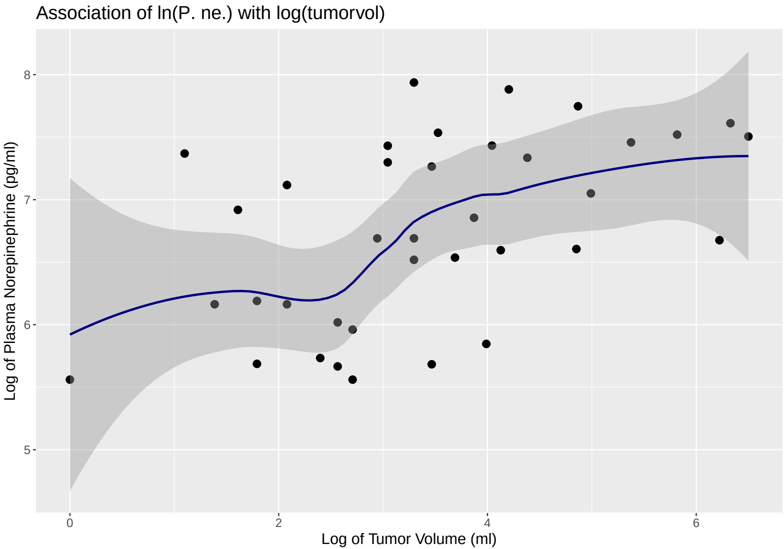
Smoothing using loess, instead



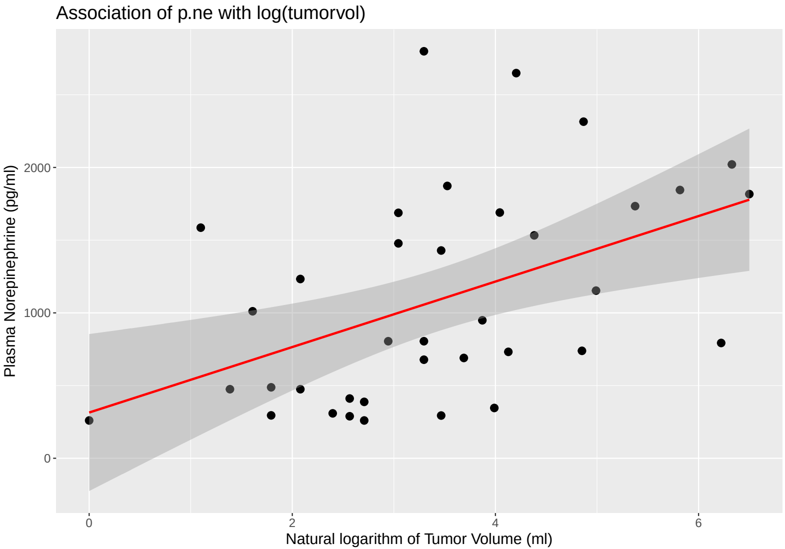
Using the Log transform to spread out the Volumes



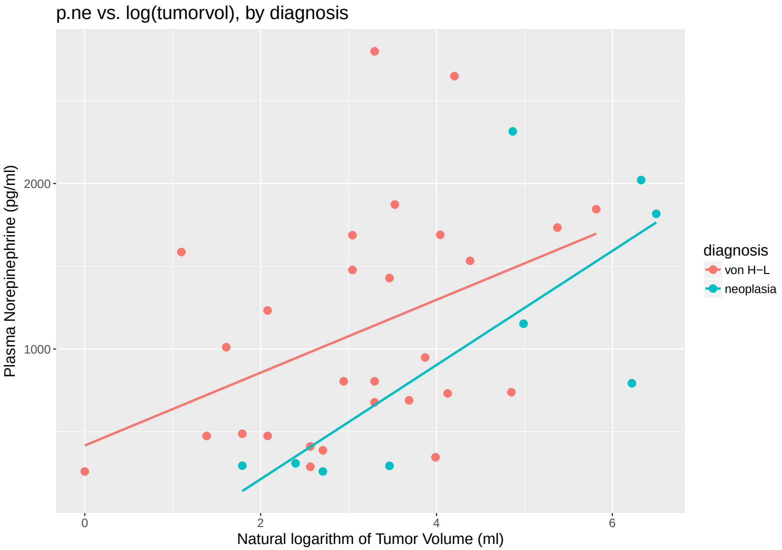
Does a Log-Log model seem like a good choice?



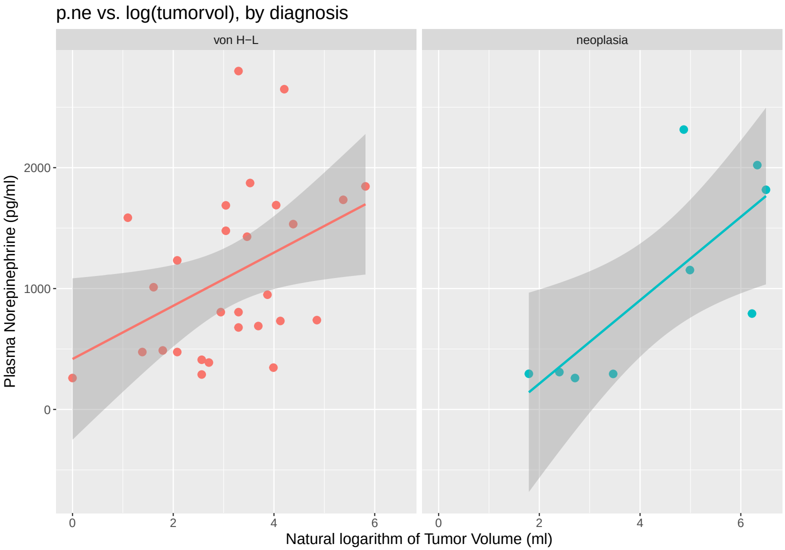
Linear Model for p.ne using log(tumor volume)



Compare the patients by diagnosis



Facetted Scatterplots by diagnosis



Model accounting for different slopes and intercepts

```r
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` = 55 ml (so log(tumorvol) = 4.01) in each diagnosis category?

```
predict(model2, newdata = data_frame(tumorvol = 55,
        diagnosis = "neoplasia"), interval = "prediction")
```

```
       fit       lwr      upr
1 905.7322 -456.1596 2267.624
```

```
predict(model2, newdata = data_frame(tumorvol = 55,
        diagnosis = "von H-L"), interval = "prediction")
```

```
       fit       lwr      upr
1 1299.003 -23.21001 2621.215
```

Don't forget to get Assignment 2 in!

Canvas, by noon Friday.