431 Class 10

github.com/THOMASELOVE/2019-431

2019-09-26

Today's Agenda (Notes, Chapters 11-13)

1. Measuring Association with Correlations
   - Pearson and Spearman approaches
   - Thinking about the impact of transformations
2. Adding a categorical predictor (factor) to a model
   - Using `fct_recode` from `forcats` (tidyverse)
   - Interpreting an indicator variable regression

Today's Packages and Loading the VHL Data

```
library(magrittr); library(janitor); library(patchwork)
library(broom); library(tidyverse)
```

```
VHL <- read_csv("vonHippel-Lindau.csv")
```

VHL Variables

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

Model 1

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

tidy(model1, conf.int = TRUE, conf.level = 0.9) %>%
  select(term, estimate, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

term	estimate	conf.low	conf.high
(Intercept)	946.18	725.73	1166.64
tumorvol	1.55	0.35	2.74

```
glance(model1) %>% select(r.squared, sigma) %>%
  knitr::kable(digits = 2)
```

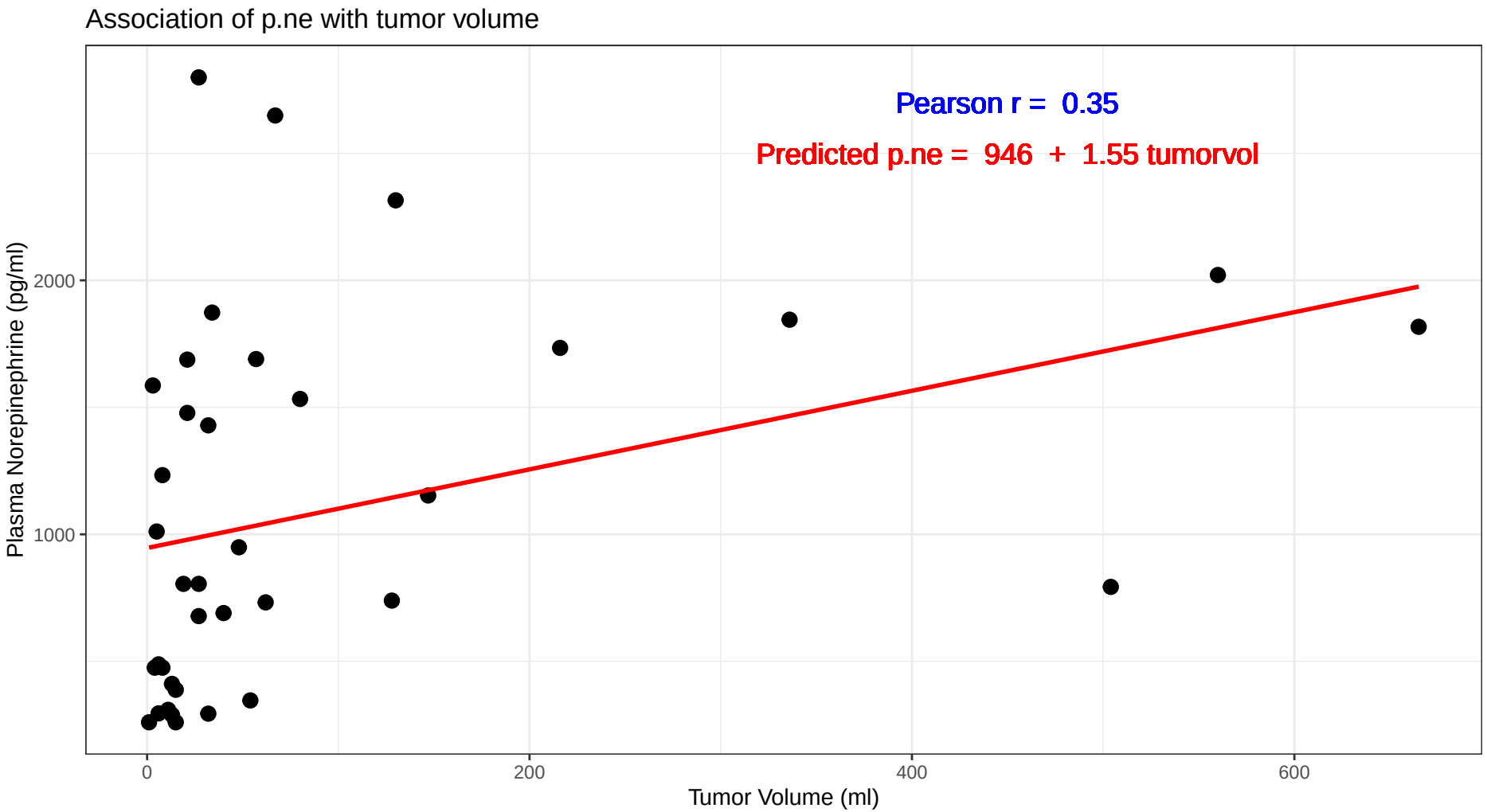
r.squared	sigma
0.12	685.17

Residuals from `model1`

```
model1_aug <- augment(model1)

head(model1_aug,3)
```

```
# A tibble: 3 x 9
   p.ne tumorvol .fitted .se.fit .resid   .hat .sigma
  <dbl>    <dbl>   <dbl>   <dbl>  <dbl>  <dbl>  <dbl>
1   289       13    966.    126.  -677. 0.0339   685.
2   294       32    996.    121.  -702. 0.0310   684.
3  2799       27    988.    122.  1811. 0.0317   619.
# ... with 2 more variables: .cooksd <dbl>,
#   .std.resid <dbl>
```

Predicting p.ne using tumorvol

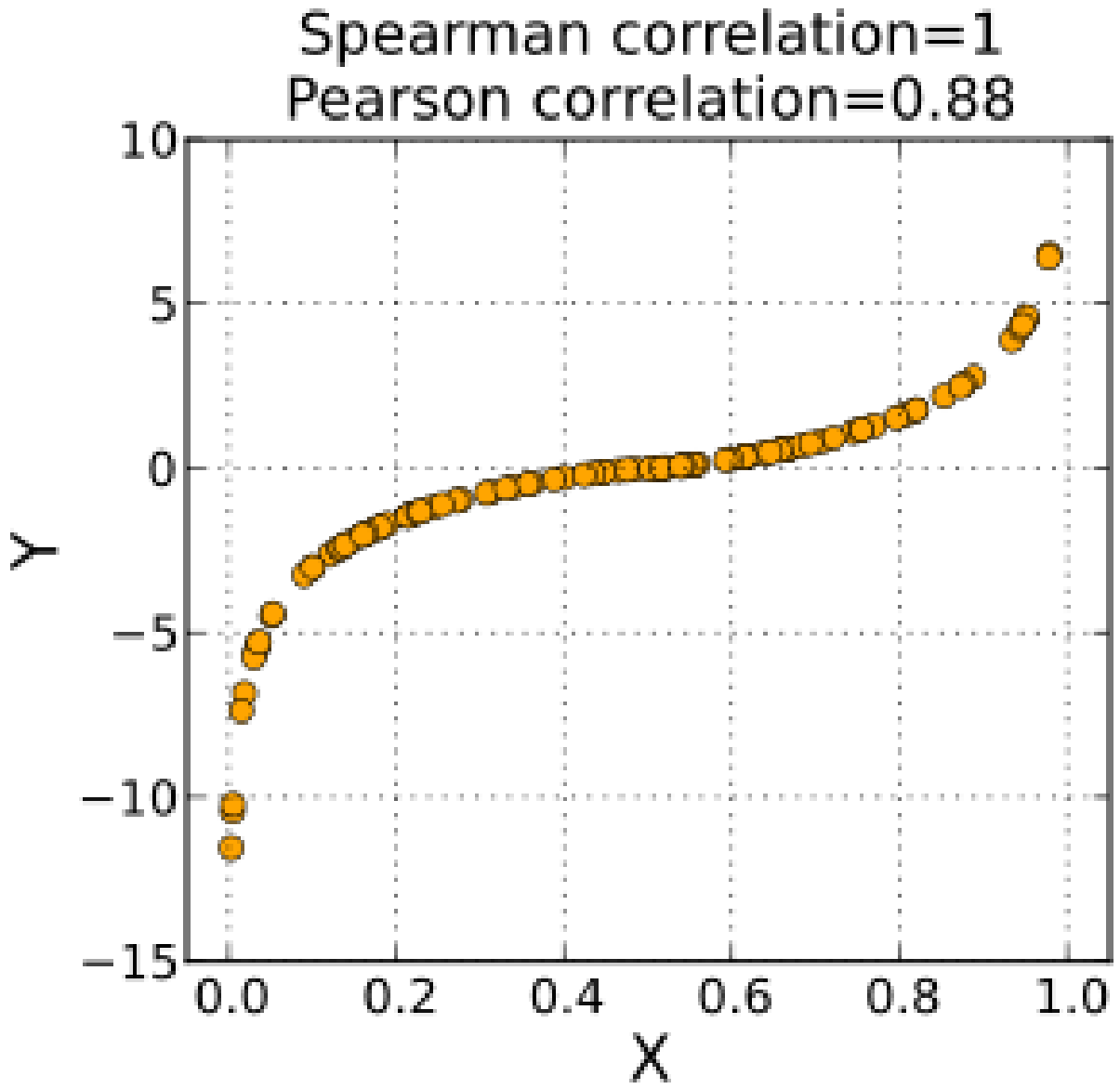


The Spearman Rank Correlation

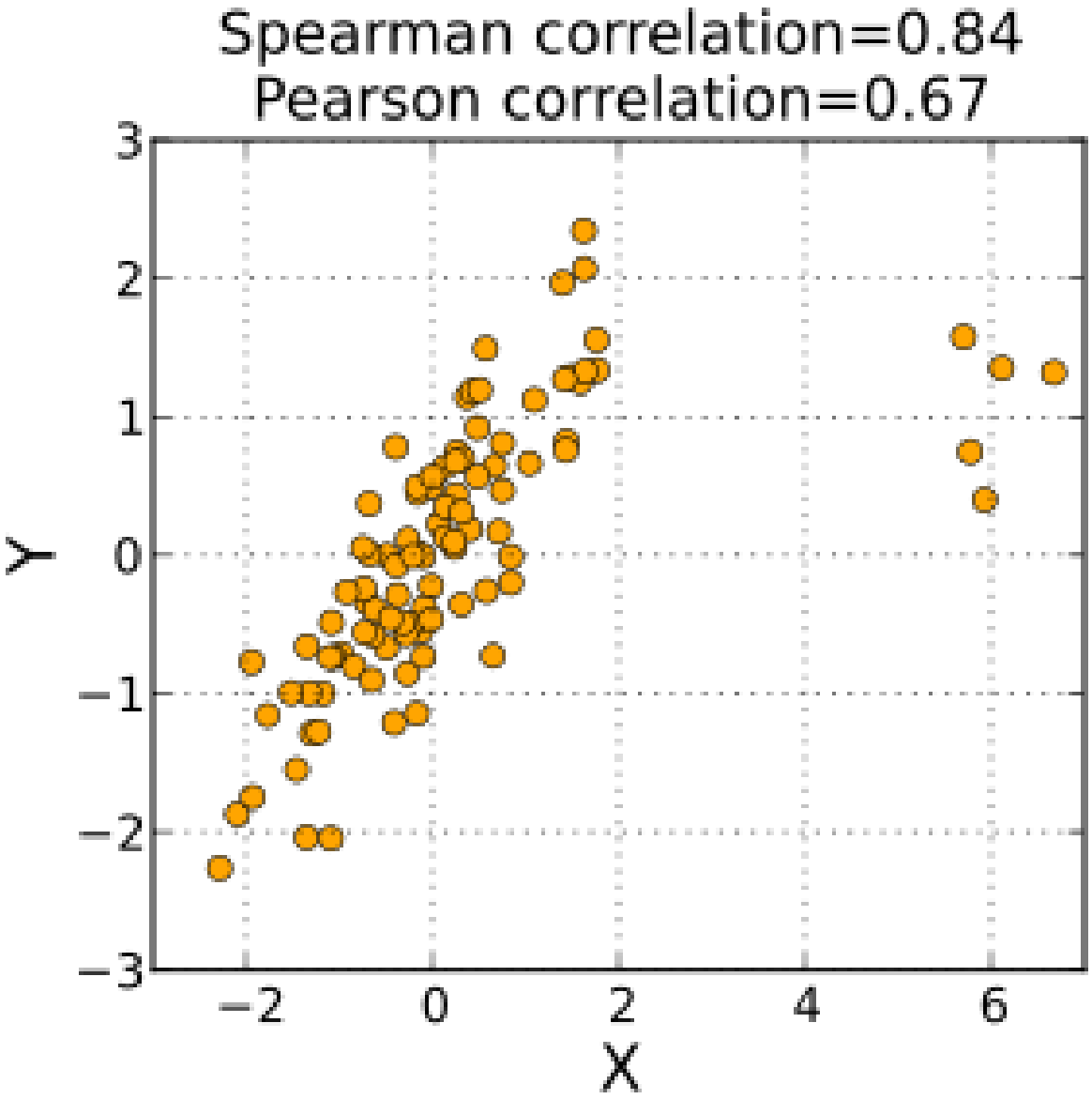
The Spearman rank correlation coefficient assesses how well the association between X and Y can be described using a **monotone function** even if that relationship is not linear.

- 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.

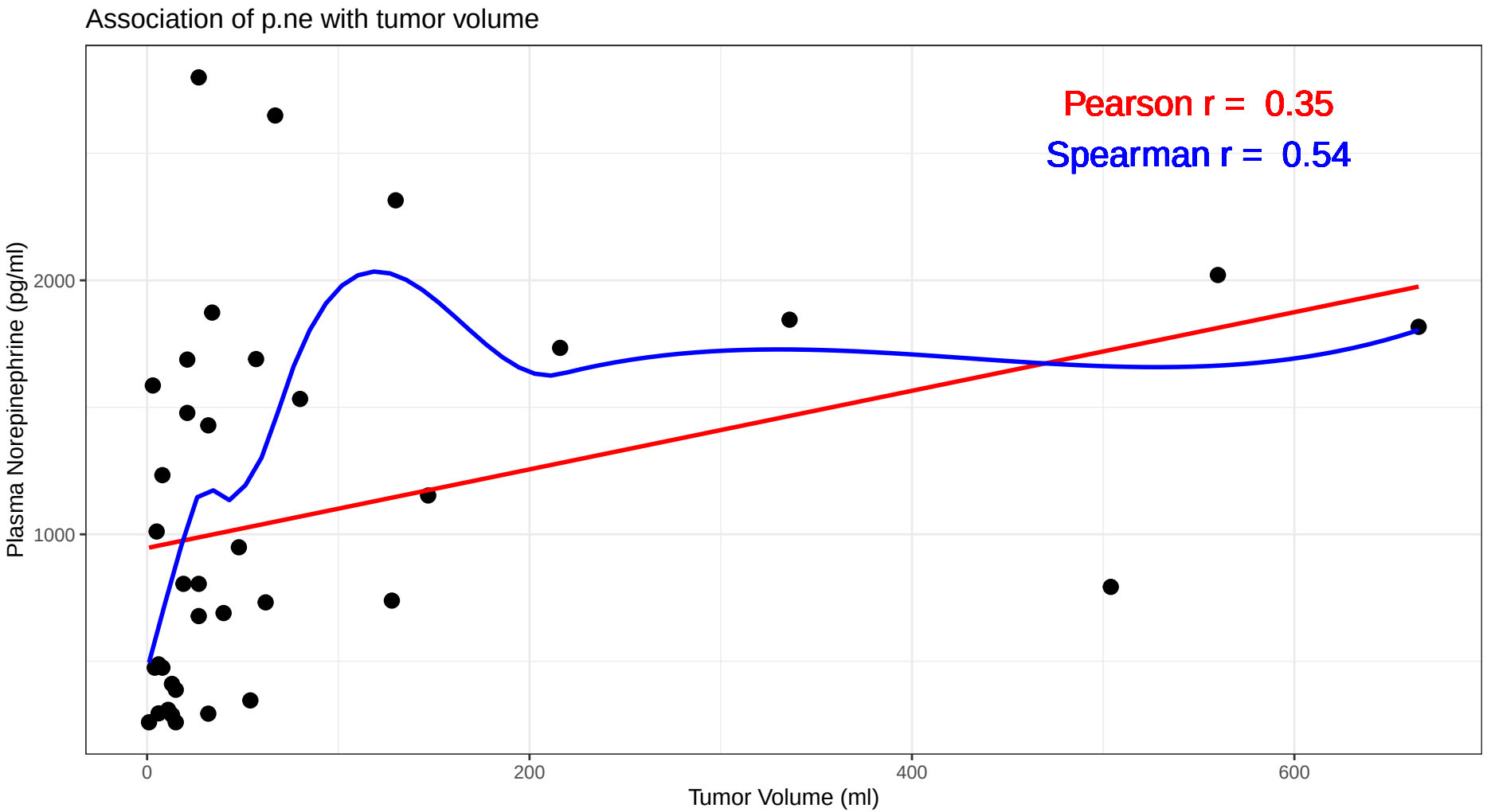
Monotone Association (Source: Wikipedia)



Spearman correlation reacts less to outliers

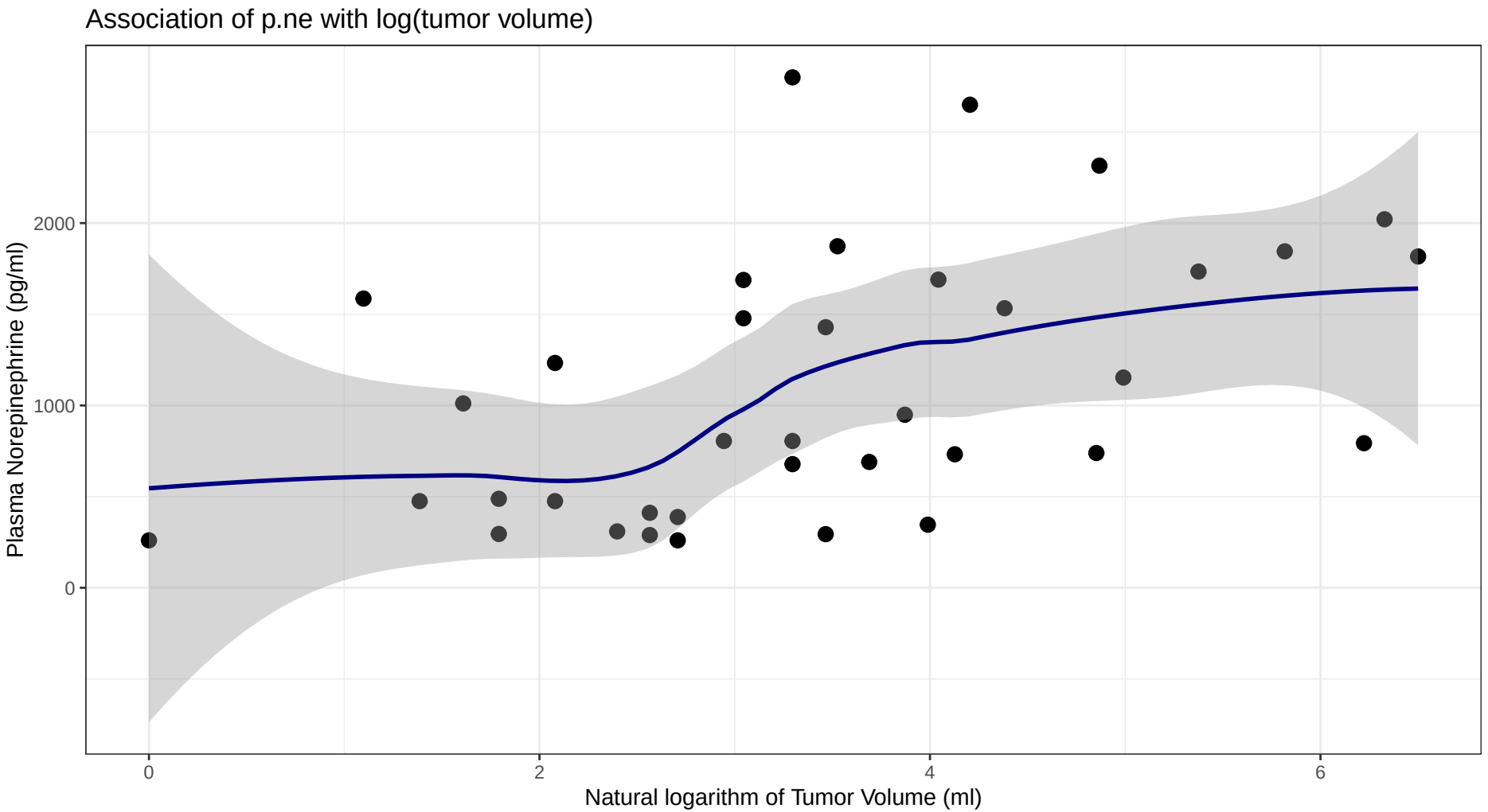


Our Key Scatterplot again

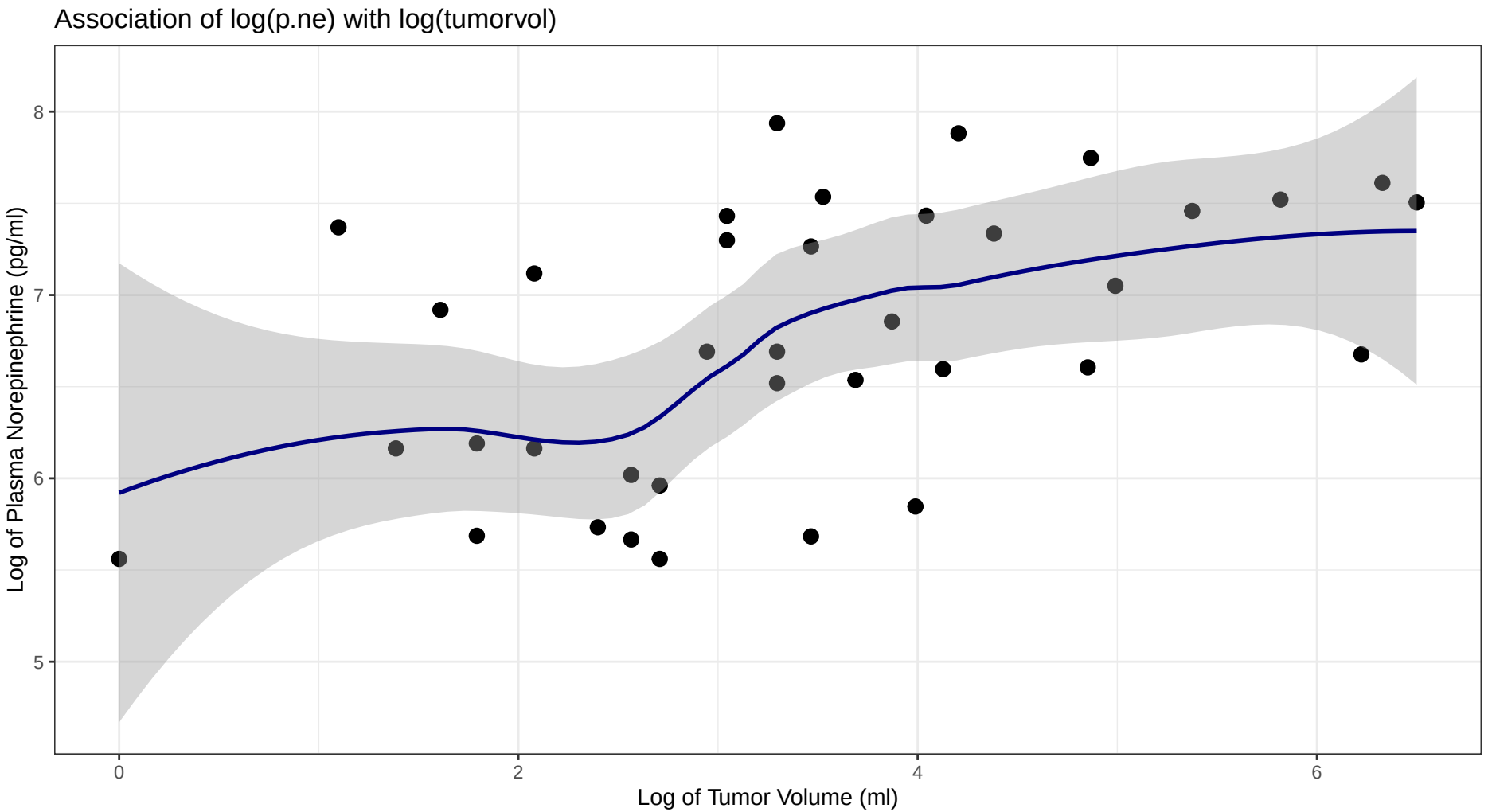


Can we transform X or Y to get to something more linear?

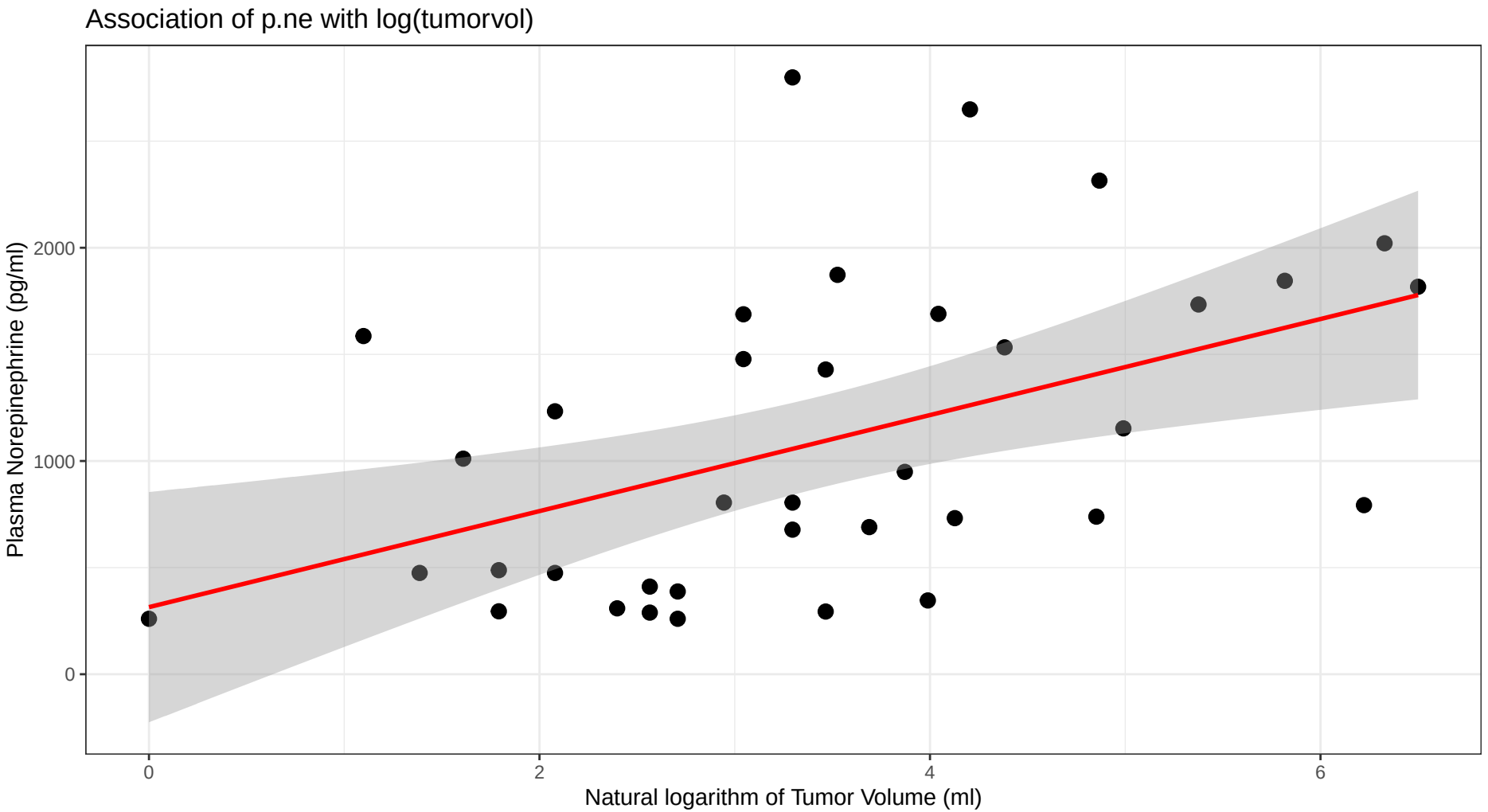
Using the `log` transform to spread out the Volumes



Does a `log-log` model seem like a good choice?



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



Fitting the m1log model (p.ne using log(tumorvol))

```
m1log <- lm(p.ne ~ log(tumorvol), data = VHL)

tidy(m1log, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, std.error, conf.low, conf.high) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	conf.low	conf.high
(Intercept)	314.6	265.95	-134.74	763.93
log(tumorvol)	225.2	70.85	105.49	344.92

Glancing at the model fit

```
m1log <- lm(p.ne ~ log(tumorvol), data = VHL)

glance(m1log) %>%
  select(r.squared, adj.r.squared, sigma) %>%
  knitr::kable(digits = 3)
```

r.squared	adj.r.squared	sigma
0.224	0.202	643.454

Summarizing the model's fit

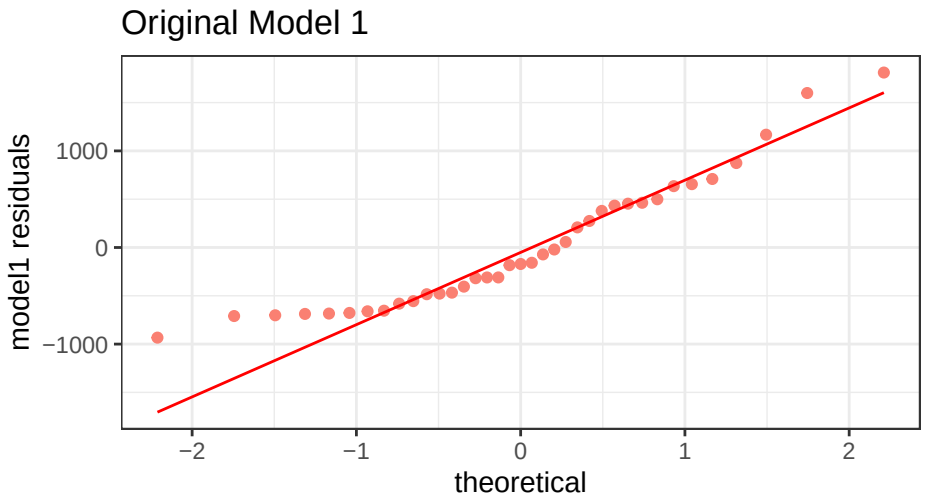
```
> summary(m1log)

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

Residuals:
   Min     1Q Median     3Q    Max
-922.9 -481.2 -172.7  333.9 1742.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    314.60     265.95   1.183  0.24481
log(tumorvol)  225.20      70.85   3.178  0.00309 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 643.5 on 35 degrees of freedom
Multiple R-squared:  0.224,     Adjusted R-squared:  0.2018
F-statistic:  10.1 on 1 and 35 DF,  p-value: 0.003092
```

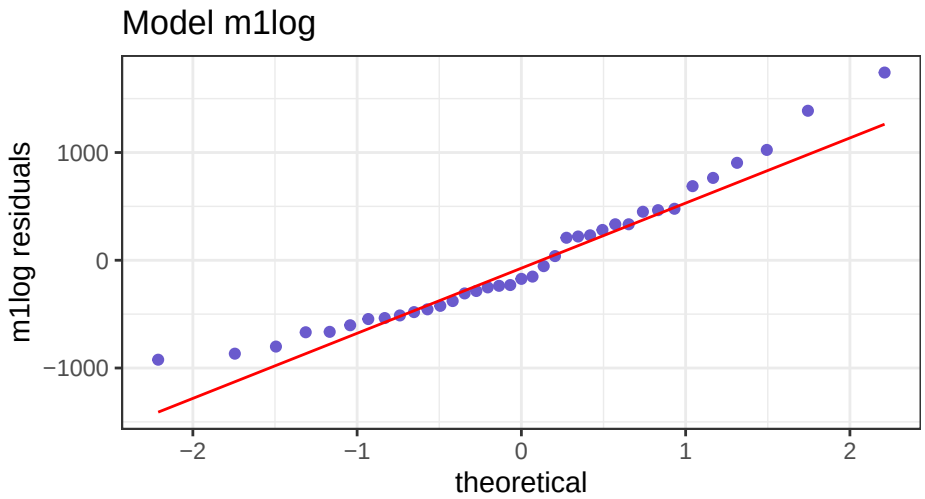
Residuals from m1log

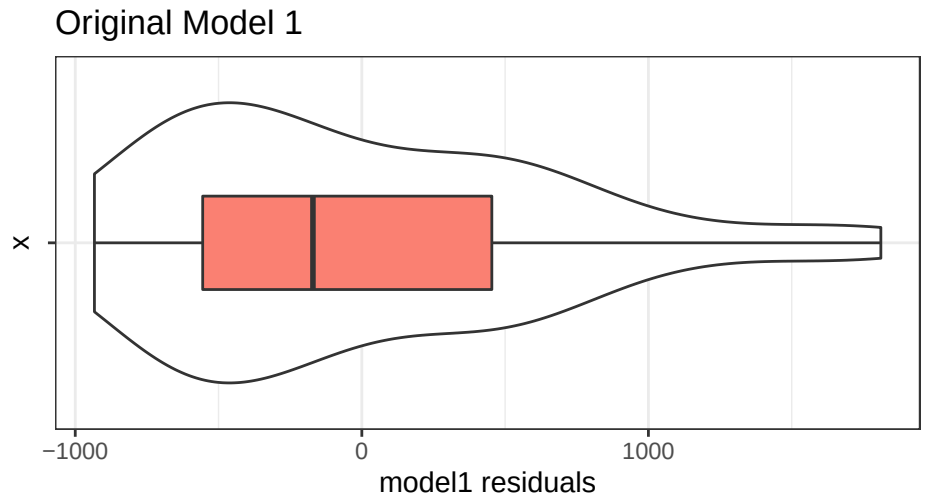
```
m1log_aug <- augment(m1log)

head(m1log_aug,3)
```

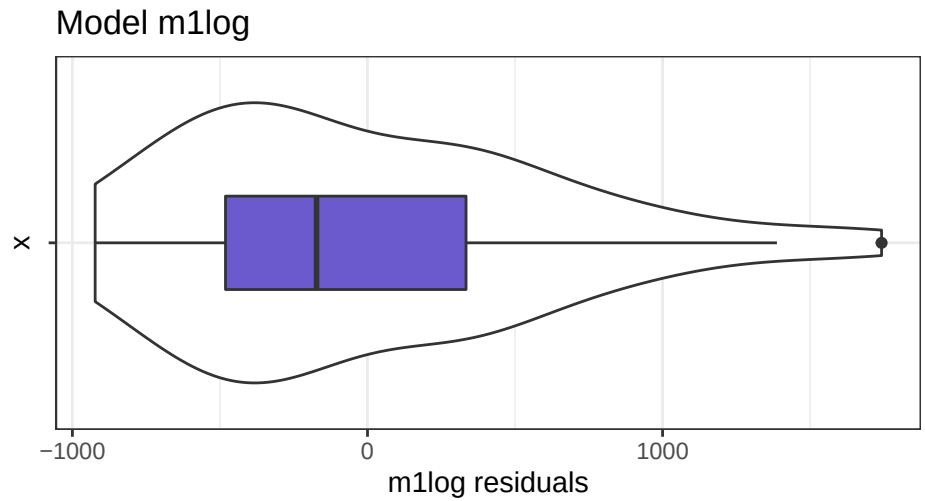
```
# A tibble: 3 x 9
   p.ne log.tumorvol. .fitted .se.fit .resid   .hat
  <dbl>         <dbl>   <dbl>   <dbl>  <dbl>  <dbl>
1   289          2.56    892.    123.  -603. 0.0364
2   294          3.47   1095.    106.  -801. 0.0270
3  2799          3.30   1057.    106.  1742. 0.0273
# ... with 3 more variables: .sigma <dbl>,
#   .cooksd <dbl>, .std.resid <dbl>
```

`m1log` residuals: Normally distributed?

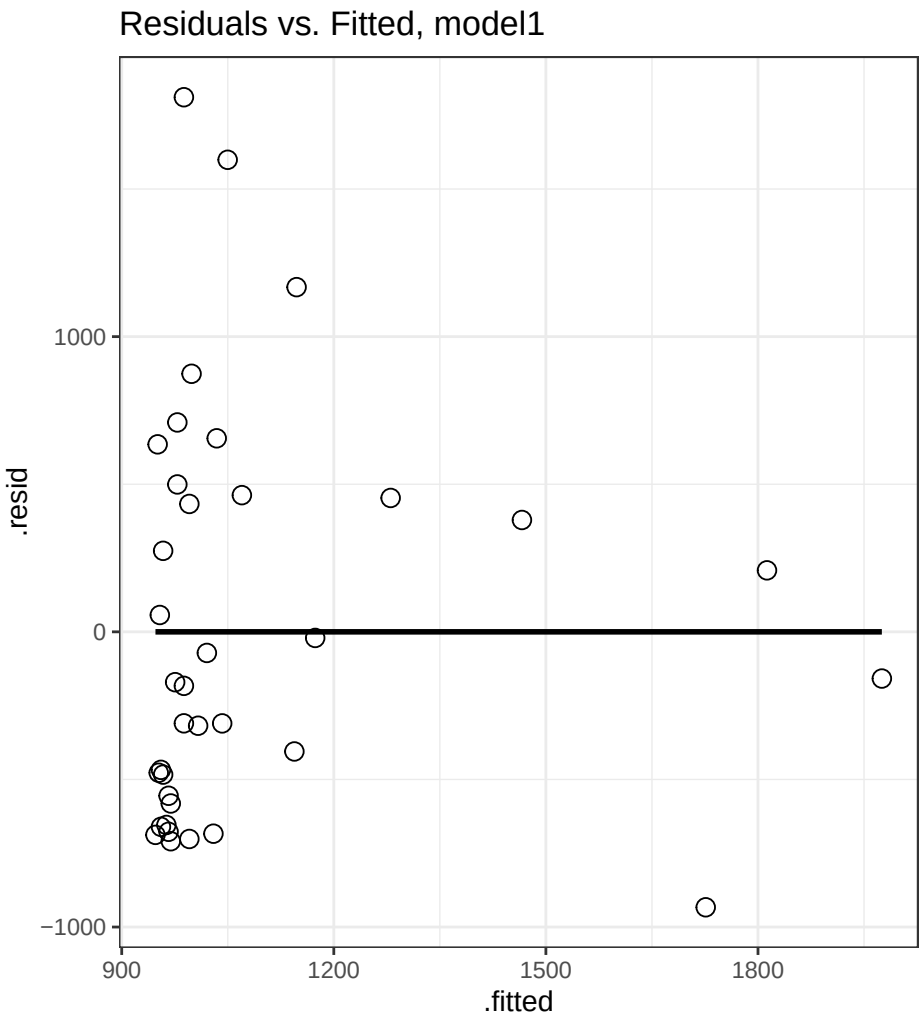


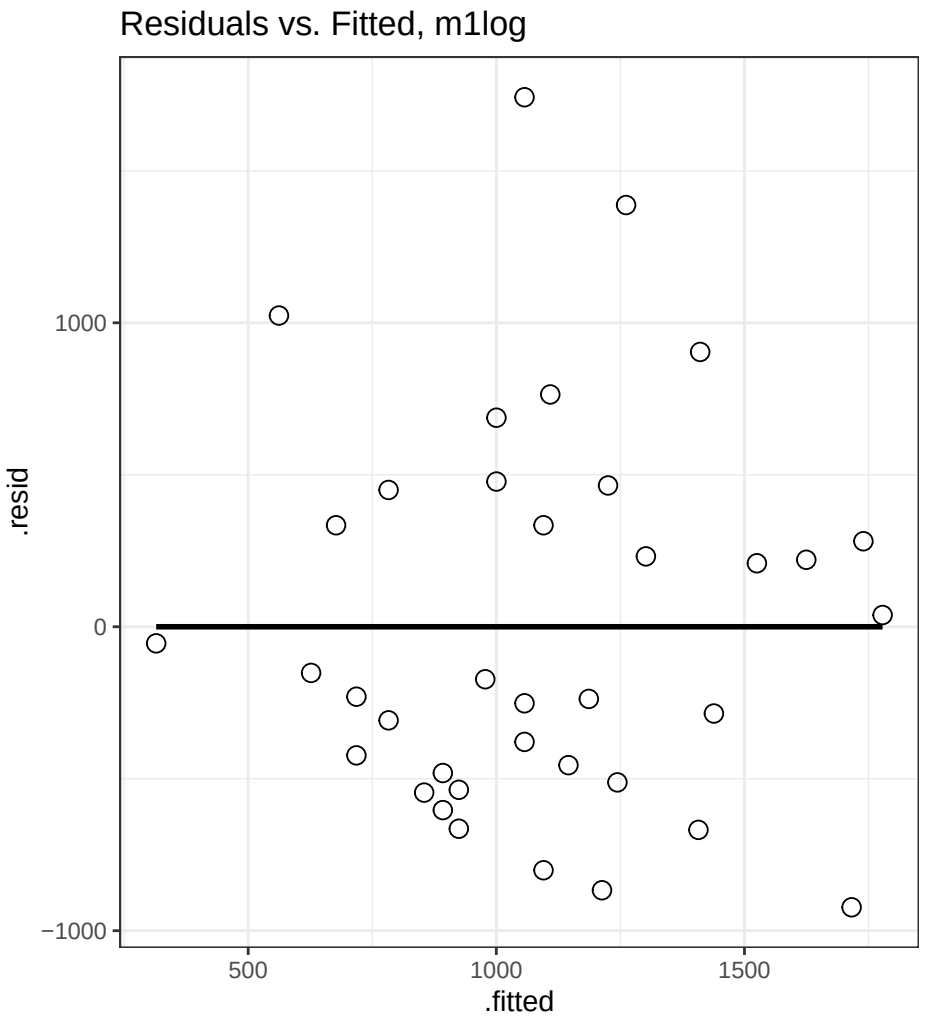






Residuals vs. Fitted plots (model1 and m1log)





Adding diagnosis to our model

Creating a Factor to represent disease category

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. . .

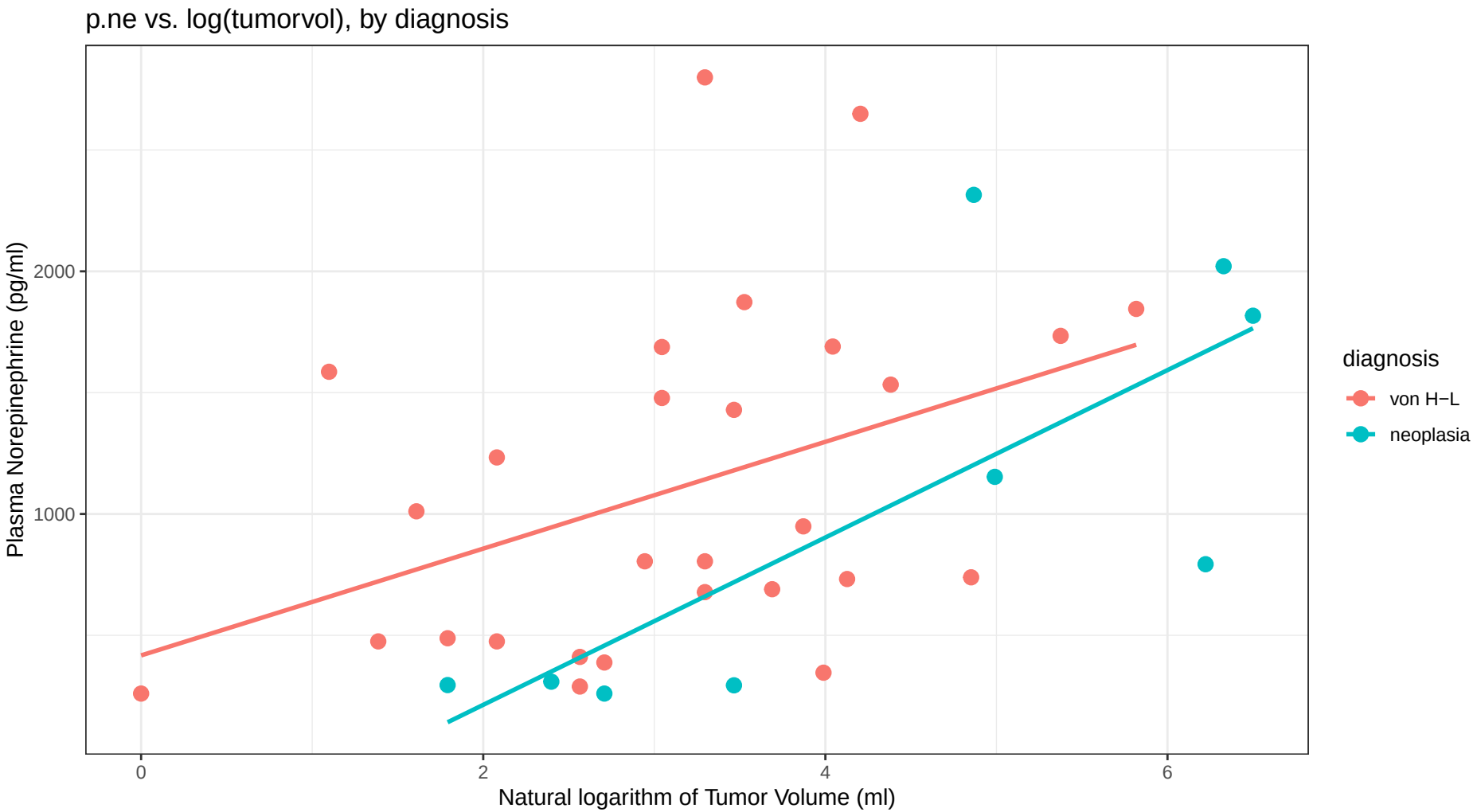
```
VHL <- VHL %>%
  mutate(diagnosis =
           fct_recode(factor(disease),
                      "neoplasia" = "1",
                      "von H-L" = "0")
  )
```

Now, what does VHL look like?

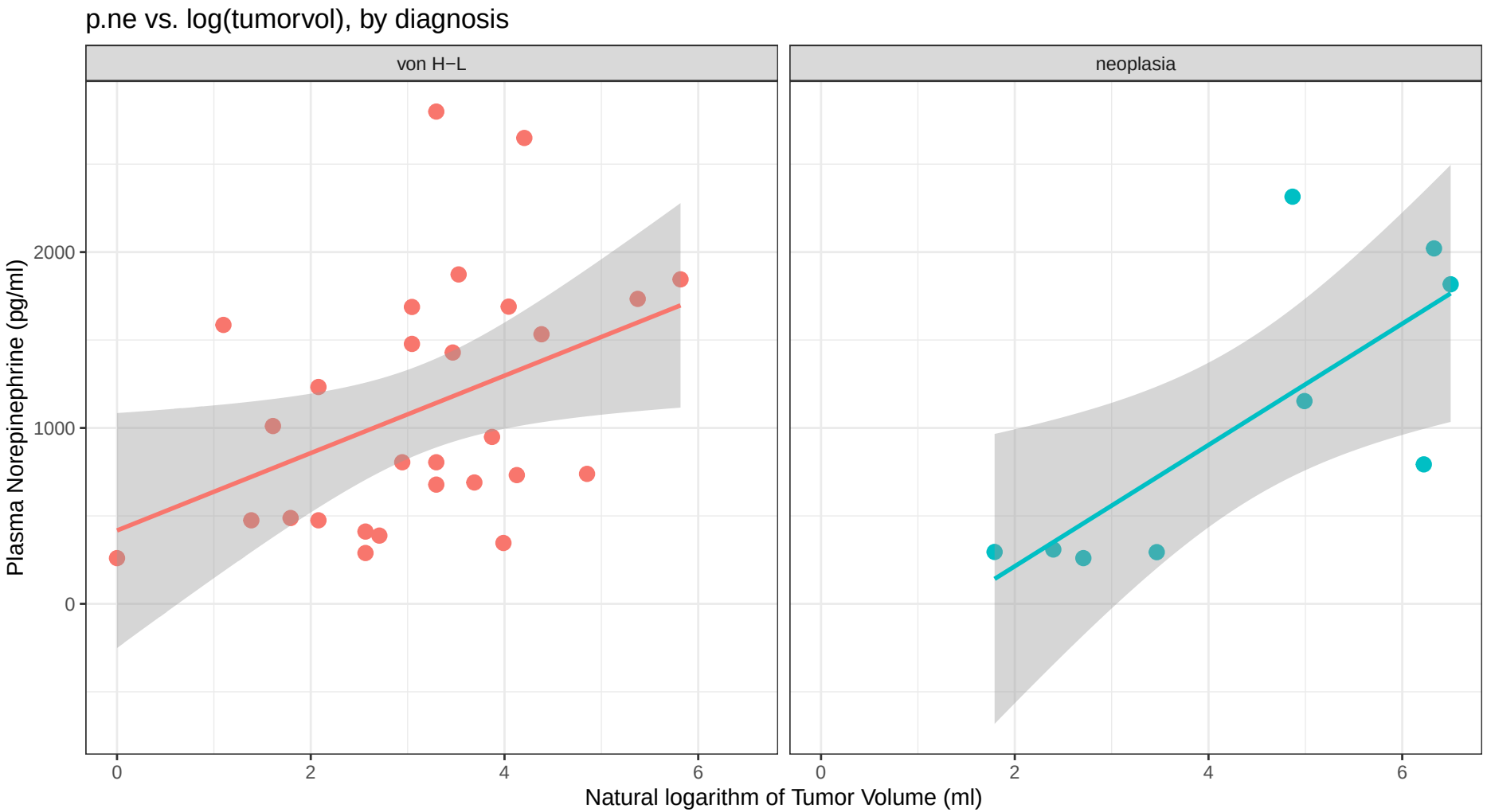
```
VHL
```

```
# A tibble: 37 x 5
      id disease  p.ne tumorvol diagnosis
   <dbl>   <dbl> <dbl>    <dbl> <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



Faceted Scatterplots by diagnosis



Separate Models by Diagnosis?

```
model2_vhl <- lm(p.ne ~ log(tumorvol),
                 data = filter(VHL, diagnosis == "von H-L"))

coef(model2_vhl)
```

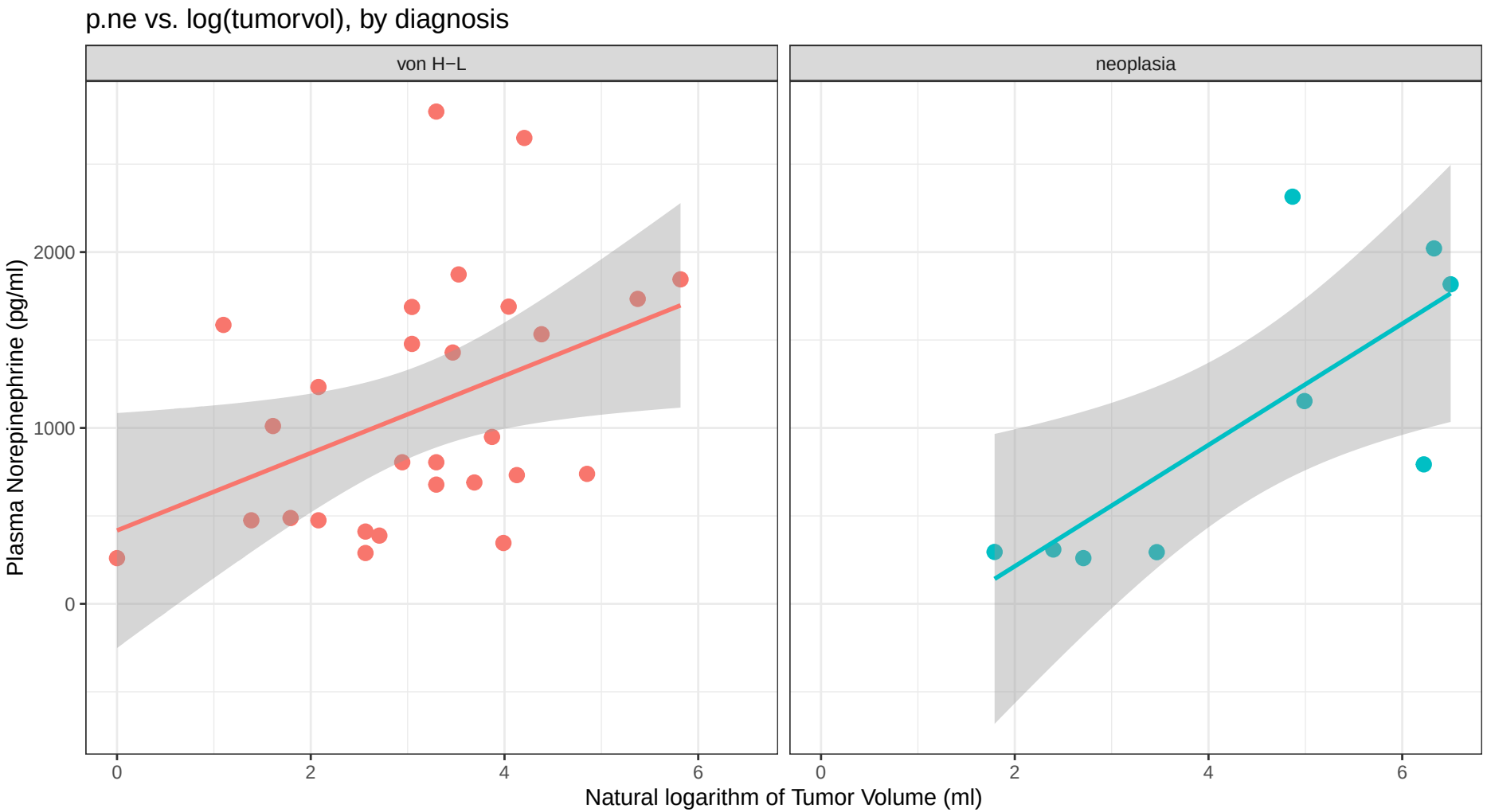
```
  (Intercept) log(tumorvol)
     417.2040      220.0463
```

```
model2_neo <- lm(p.ne ~ log(tumorvol),
                 data = filter(VHL, diagnosis == "neoplasia"))

coef(model2_neo)
```

```
  (Intercept) log(tumorvol)
    -476.0978      344.8253
```

Does this match our plot?

Faceted Scatterplots by diagnosis, again



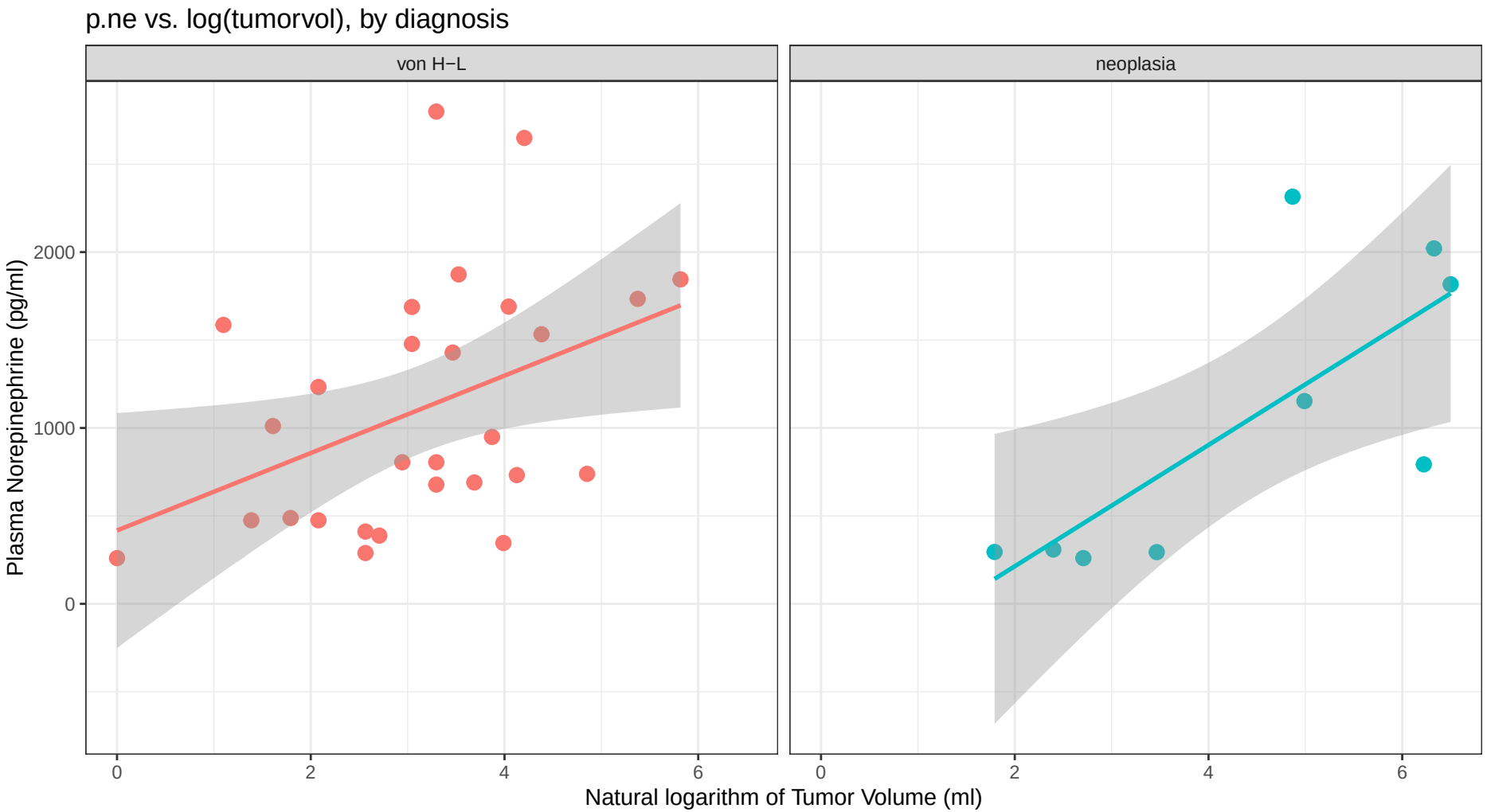
Correlation Coefficients

```
VHL %>%
  group_by(diagnosis) %>%
  summarize(Correlation = cor(log(tumorvol), p.ne),
            Rsquare = (cor(log(tumorvol), p.ne)^2) )
```

```
# A tibble: 2 x 3
  diagnosis Correlation Rsquare
  <fct>           <dbl>   <dbl>
1 von H-L         0.412   0.169
2 neoplasia       0.756   0.572
```

Does this match our plot?

Faceted Scatterplots by diagnosis, one more time



What do we predict if log(tumorvol) = 3?

`log(tumorvol)` = 3 implies `tumorvol` = `exp(3)` = 20.0855369 ml.

From our `model2_vhl`, we'd predict:

- 417 + 220 (3) = 1,077 pg/nl of `p.ne` for a VHL patient with `tumorvol` = 20.0855369 ml.

From our `model2_neo`, we'd predict:

- -476 + 345 (3) = 559 pg/nl of `p.ne` for a Neoplasia patient with `tumorvol` = 20.0855369 ml.

Model including two predictors

```
model3 <- lm(p.ne ~ log(tumorvol) + diagnosis, data = VHL)
model3
```

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

Coefficients:
       (Intercept)       log(tumorvol)
             273.2               265.8
diagnosisneoplasia
            -404.4
```

But this model only changes the intercept?

```
coef(model3)
```

```
       (Intercept)      log(tumorvol)
          273.1745           265.7977
diagnosisneoplasia
          -404.4333
```

- Model for VHL is `p.ne` = 273 + 266 `log(tumorvol)`.
  - `p.ne` prediction if `log(tumorvol)` = 3 is 1,071 pg/nl.
- Model for neoplasia is `p.ne` = (273 - 404) + 266 `log(tumorvol)`, or -131 + 266 `log(tumorvol)`.
  - `p.ne` prediction if `log(tumorvol)` = 3 is 667 pg/nl.

Is that what we want?

Model accounting for different slopes *and* intercepts

```
model4 <- lm(p.ne ~ log(tumorvol) * diagnosis, data = VHL)
model4
```

```
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
```

model4 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)

These are our initial (separated) models, in this case.

model4 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?

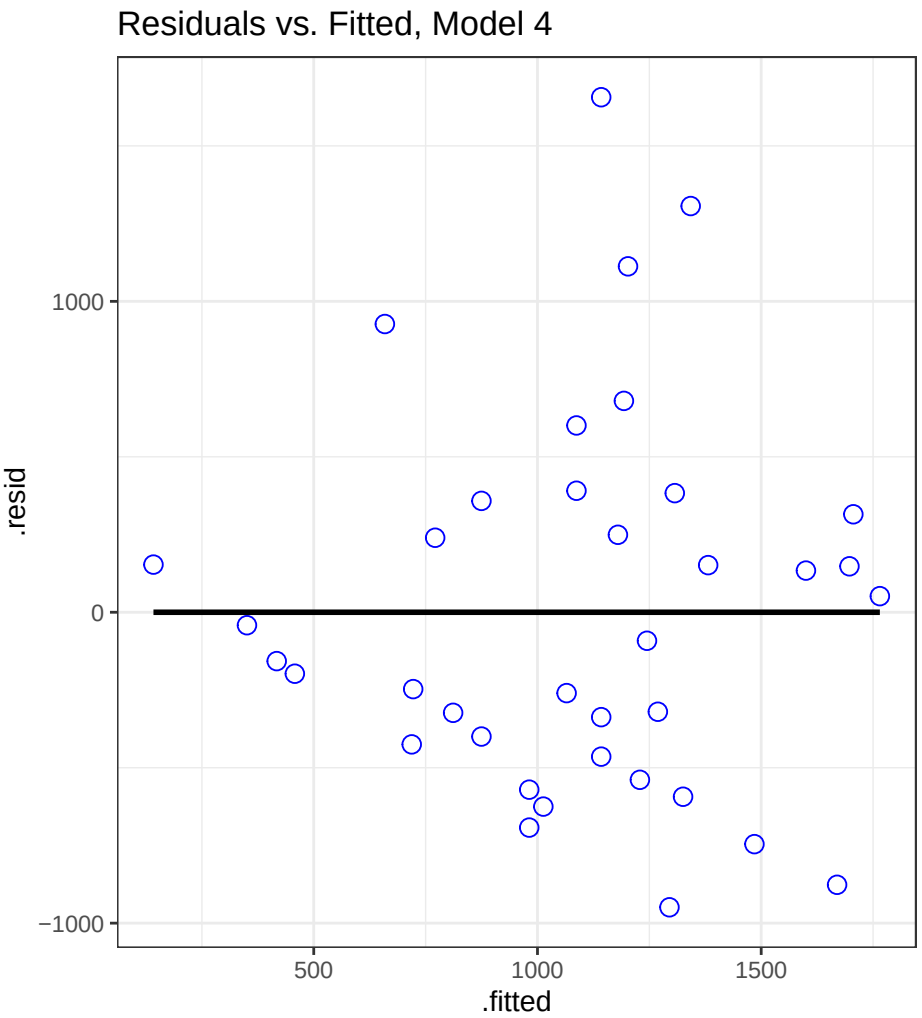
```
predict(model4, newdata = tibble(tumorvol = 200,
        diagnosis = "neoplasia"), interval = "prediction")
```

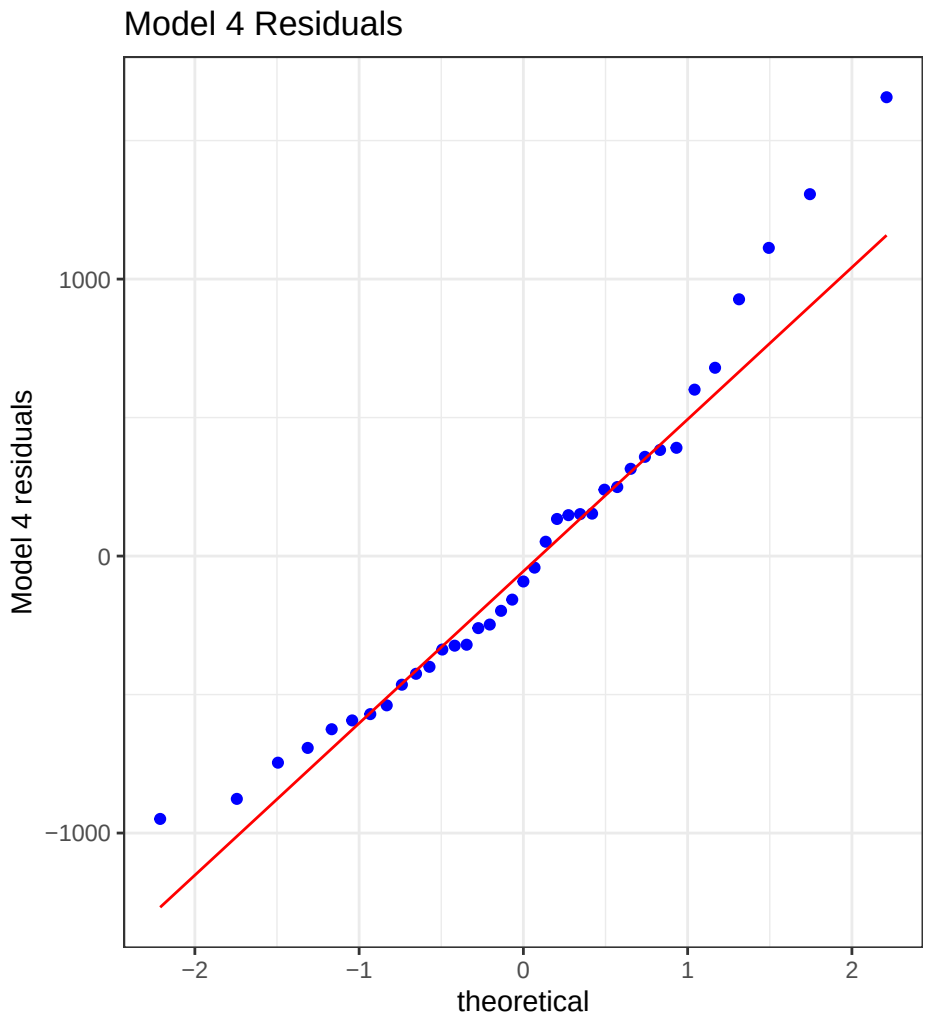
```
       fit      lwr     upr
1 1350.896 -28.0571 2729.85
```

```
predict(model4, newdata = tibble(tumorvol = 200,
        diagnosis = "von H-L"), interval = "prediction")
```

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

How about the Residuals of model4?





Tidying the model4 coefficients, with broom

```
tidy(model4, conf.int = TRUE, conf.level = 0.90) %>%
  select(term, estimate, conf.low, conf.high) %>%
  knitr::kable(digits = 1)
```

term	estimate	conf.low	conf.high
(Intercept)	417.2	-120.9	955.4
log(tumorvol)	220.0	61.7	378.4
diagnosisneoplasia	-893.3	-2007.8	221.2
log(tumorvol):diagnosisneoplasia	124.8	-136.7	386.3

model4, summarized at a glance, with broom

```
glance(model4) %>% select(r.squared, sigma, AIC)
```

```
# A tibble: 1 x 3
  r.squared sigma   AIC
      <dbl> <dbl> <dbl>
1     0.290  634.  588.
```

Compare this to m1log...

```
glance(m1log) %>% select(r.squared, sigma, AIC)
```

```
# A tibble: 1 x 3
  r.squared sigma   AIC
      <dbl> <dbl> <dbl>
1     0.224  643.  587.
```

Conclusions about VHL data

- Model 4, accounting for the interaction of diagnosis with the log of tumor volume, was able to account for about 29% of the variation in the plasma norepinephrine levels.
- m1log, which didn't include diagnosis but just the log of tumor volume, accounts for about 22% of the variation in plasma norepinephrine levels.
- Model 1, our original linear model, which didn't account for diagnosis and didn't fit assumptions well (using raw tumor volume) accounted for about 12% of the variation in plasma norepinephrine levels.

Can we draw a lot more from this yet?

Small Groups!

Group Task: Kidney Cancer Death Rates

The map on the next slide shows U.S. counties.

- The shaded counties are in the bottom 10% of age-standardized rates for death due to cancer of the kidney/ureter for white males, in 1980-1989.

Your Tasks

1. Describe the patterns you see in the map.
2. Speculate as to the cause of these patterns.

Lowest kidney cancer death rates

Don't look ahead, at least not yet.

Highest kidney cancer death rates

So what did we hear about today?

- The central role of linear regression in understanding associations between quantitative variables.
- The interpretation of a regression model as a prediction model.
- Assessment of key regression summaries, including residuals.
- Using `tidy`, `glance` and `augment` from `broom` to summarize the model.
- Measuring association through correlation coefficients.
- How we might think about "adjusting" for the effect of a categorical predictor on a relationship between two quantitative ones.
- How a transformation might help us "linearize" the relationship shown in a scatterplot.
- Thinking about outliers.

Notes on the Kidney Cancer example, 1

I first asked you what you noticed about the map, in the hope that someone would point out the obvious pattern, which is that many of the countries in the Great Plains but relatively few near the coasts are shaded.

- Why might that be? Could these be the counties with more old people? Ah, but these rates are age-adjusted.
- They're mostly in rural areas: could the health care there be worse than in major cities? Or perhaps people living in rural areas have less healthy diets, or are exposed to more harmful chemicals? Maybe, but the confusing fact is that the highest 10% and the lowest 10% each show disproportionately higher rates in those Great Plains counties.

Notes on the Kidney Cancer example, 2

- Consider a county with 100 white males. If it has even one kidney death in the 1980s, its rate is 1 per thousand per year, which is among the highest in the nation. If it has no such deaths, its rate will be 0, which is the lowest in the nation.
- The observed rates for smaller counties are *much* more variable, and hence they are more likely to be shaded, even if nothing special is truly going on.
- If a small county has an observed rate of 1 per thousand per year, it's probably random fluctuation. But if a large county (like Cuyahoga) has a very high rate, it is probably a real phenomenon.

Source

My source for this example was Andrew Gelman and Deborah Nolan's book *Teaching Statistics: a bag of tricks* which is the source of a number of things we'll see in the course, including some of the "age guessing" example we've previously done.