431 Class 09

Thomas E. Love

2017-09-26

Today's Agenda

1. Discussion of Assignment 2
2. Silver, Chapters 2 and 3
3. Associations, Using Linear Models (Notes: Ch 11)
    - A study of von Hippel-Lindau disease
    - Associations, Correlation and Scatterplots
    - Fitting a Linear Model
4. Setting Up Project Task B Groups

Assignment 2 feedback

https://github.com/THOMASELOVE/431homework/tree/master/HW2 has the (password-protected pdf and non-protected Rmd) answer sketch, the grades, and the grading rubric

- I'm no longer suggesting people use gg_qq, and I've removed it from the Course Notes and answer sketch.

Four Interesting Essays

Tell us about an example in your own field/work/experience where a "surplus" of information made (or makes) it easier for people dealing with a complex system to cherry-pick information that supports their prior positions. What were the implications of your example in terms of lessons that can be learned?

Visit https://goo.gl/5q6Nrw to read excerpts from four of the more interesting responses.

1. On Screening, Hepatitis C and Liver Cancer
2. On Web MD, and "a little knowledge is a dangerous thing"
3. Is self-control like a muscle? The problem of underpowered studies
4. On the polarizing impact of the free flow of "information"

The Signal and The Noise

Chapter 2: Political Predictions

When forecasting political events,

- Pundits and experts usually do no better than chance
- Pundits and experts usually do worse than crude statistical models.

What are the characteristics of experts who **are** substantially more accurate? How can you tell a *fox* from a *hedgehog*?

Chapter 3: Baseball

When you have a whole lot of data, that's one thing. But what if you have a truly **rich** collection of data?

- How can you build a simple model to describe how the performance of a baseball player varies with age?
- Why is age such an important predictor of future performance?

The Signal and The Noise: Coming Up

Read by October 10 for in-class discussion

- Chapter 4: Weather Predictions
- Chapter 5: Earthquake Predictions

Read by October 17

- Chapter 7: Disease Outbreaks
- Chapter 8: Bayes' Theorem

R setup for Today

```
library(forcats); library(tidyverse)

## source("Love-boost.R")
## isn't needed today

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

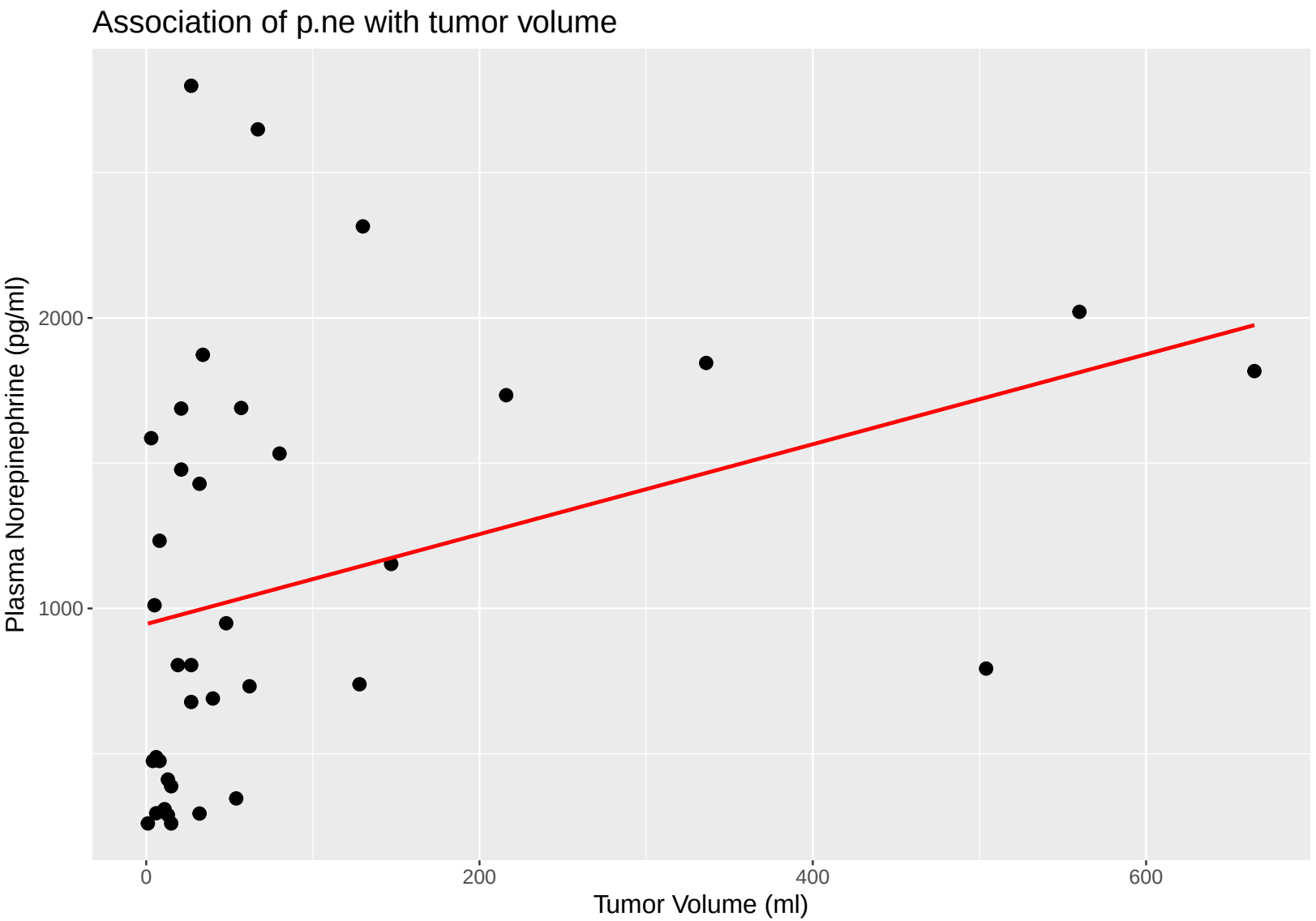
Von Hippel - Lindau study Codebook

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

```
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
```

A Linear Model for the p.ne - volume relationship



The Linear Model

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

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

Coefficients:
(Intercept)     tumorvol
    946.185        1.547
```

The (simple regression / prediction / ordinary least squares) model is

- p.ne = 946.2 + 1.55 * tumorvol.

Using the model to make predictions (PI)

To predict the `p.ne` for a subject with tumor volume 100 ml, we have

- `p.ne` = 946.2 + 1.55 * 100

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
```

Using the model to make predictions (CI)

To predict the `p.ne` for a subject with tumor volume 100 ml, we have

- `p.ne` = 946.2 + 1.55 * 100

A 95% **confidence interval** for the population average of all subjects with volume 100 ml...

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

```
       fit      lwr      upr
1 1100.925 872.0323 1329.818
```

Summary of our Linear (OLS) Model

```
> 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 (1)

```
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 straight line model for these data fitted by ordinary least squares is p.ne = 946 + 1.55 `tumorvol`.
- 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.

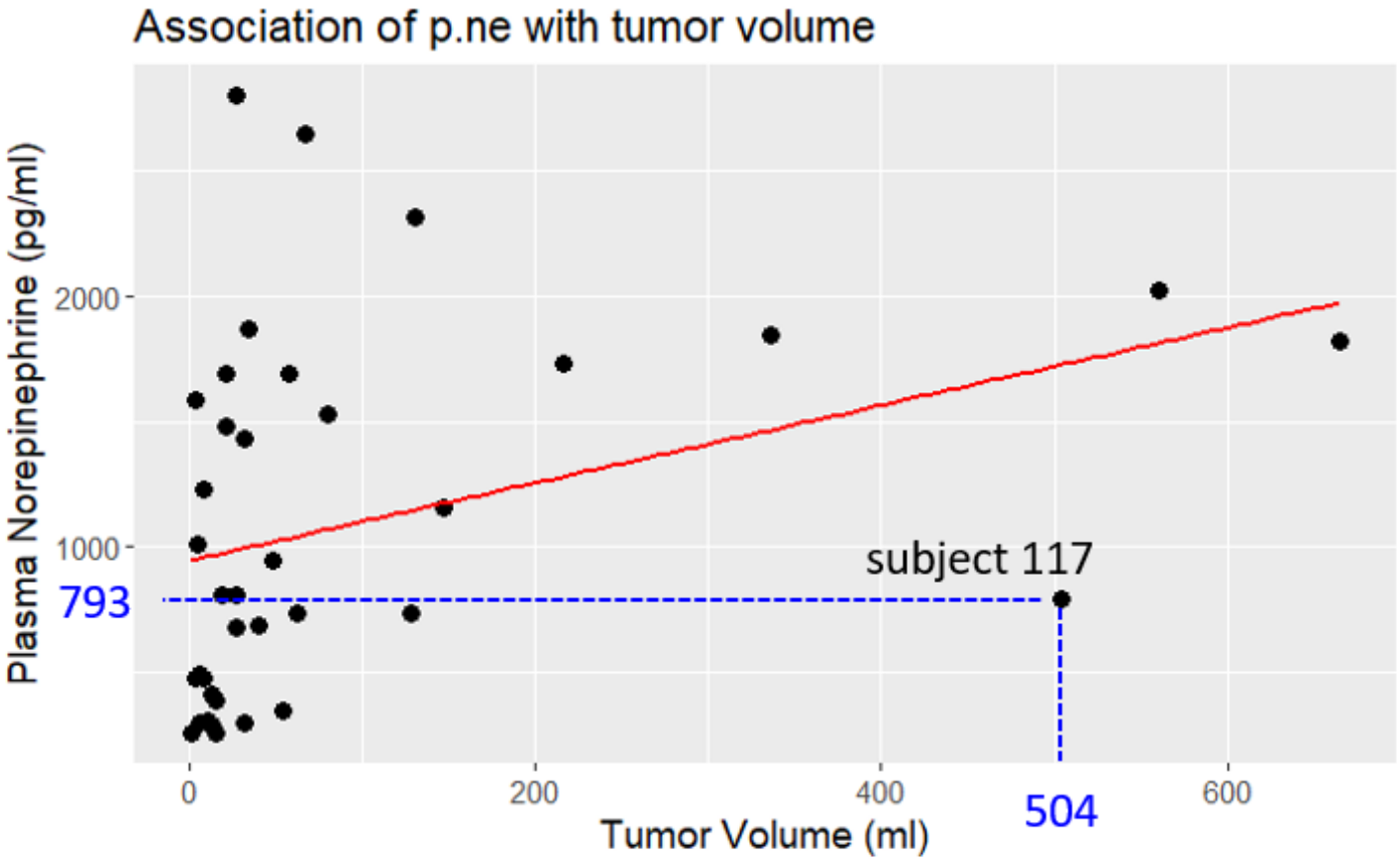
Key Elements of the Summary (2)

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

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

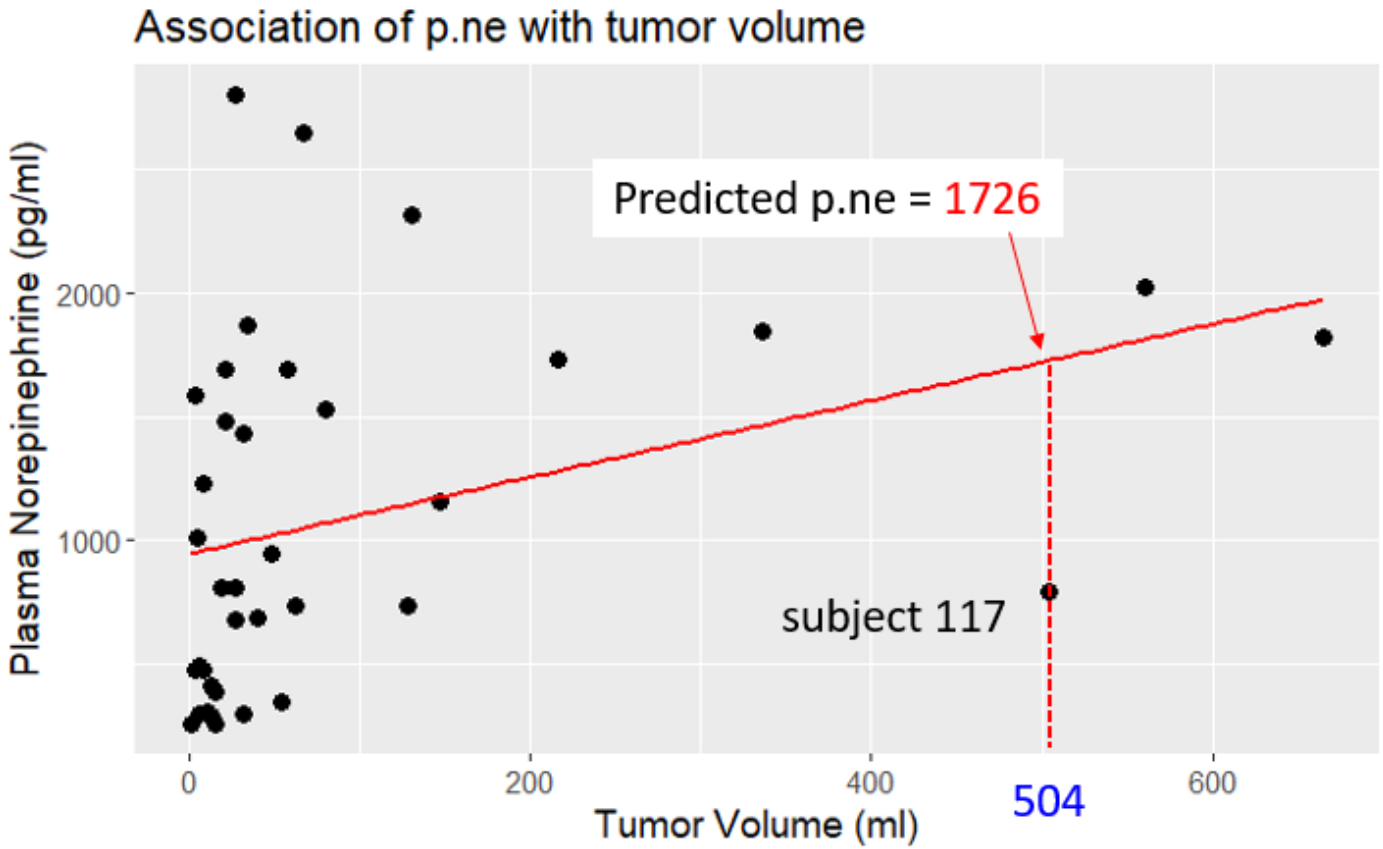
- Here, the **outcome** is `p.ne`, and the **predictor** is `tumorvol`.
- The **residuals** are the observed `p.ne` values minus the model's predicted `p.ne`. The sample residuals are the prediction errors.
- The biggest miss is for a subject whose observed `p.ne` was 1,811 pg/nl higher than the model predicts based on the subject's tumor volume.
- The mean residual will always be zero in an OLS model.

Understanding Regression Residuals (A)



Subject 117 has tumorvol = 504, and observed p.ne = 793 pg/nl.

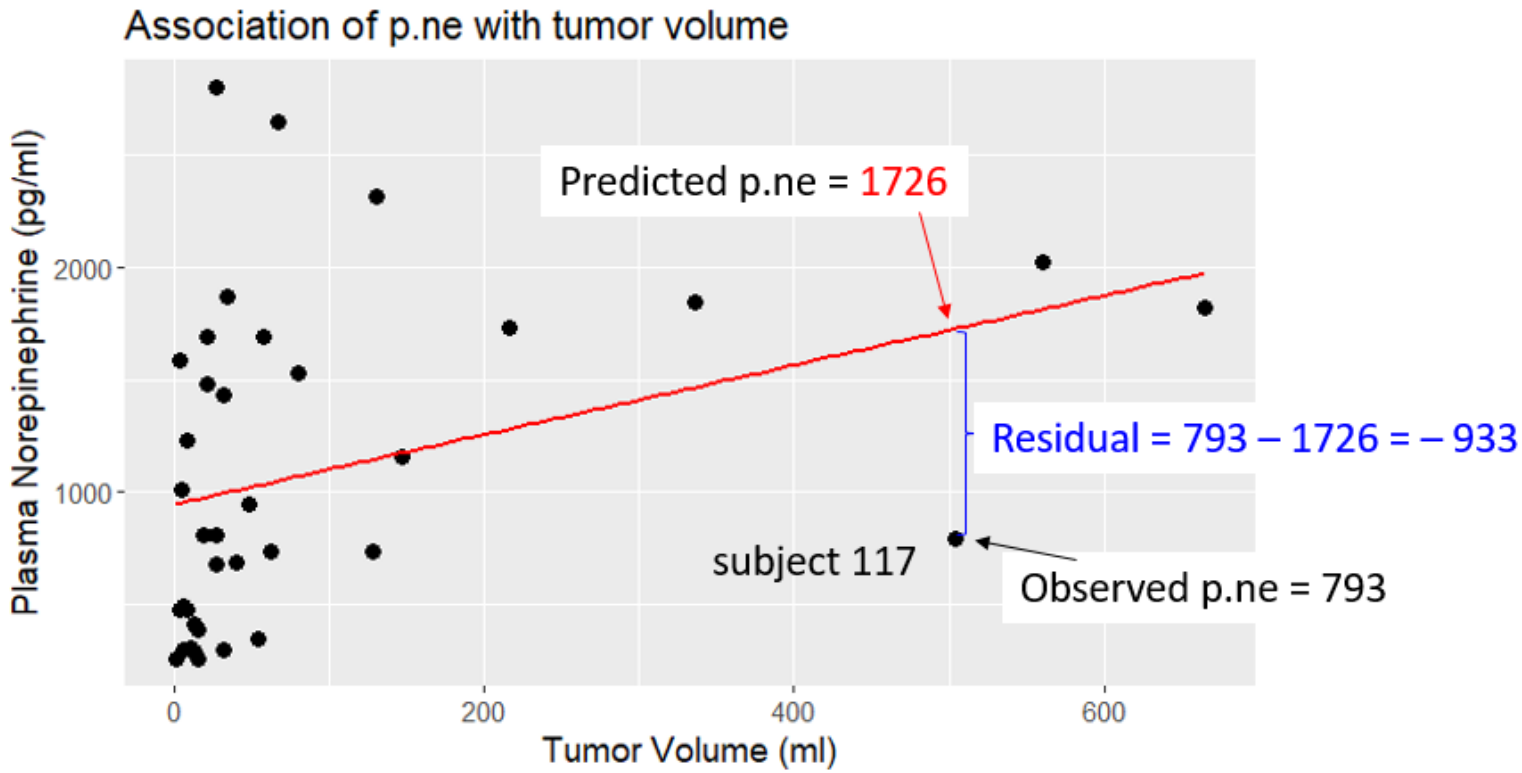
Understanding Regression Residuals (B)



Subject 117 has tumorvol = 504, and observed p.ne = 793 pg/nl.

Model predicts p.ne is 946.2 + 1.55(504) = 1726 pg/nl.

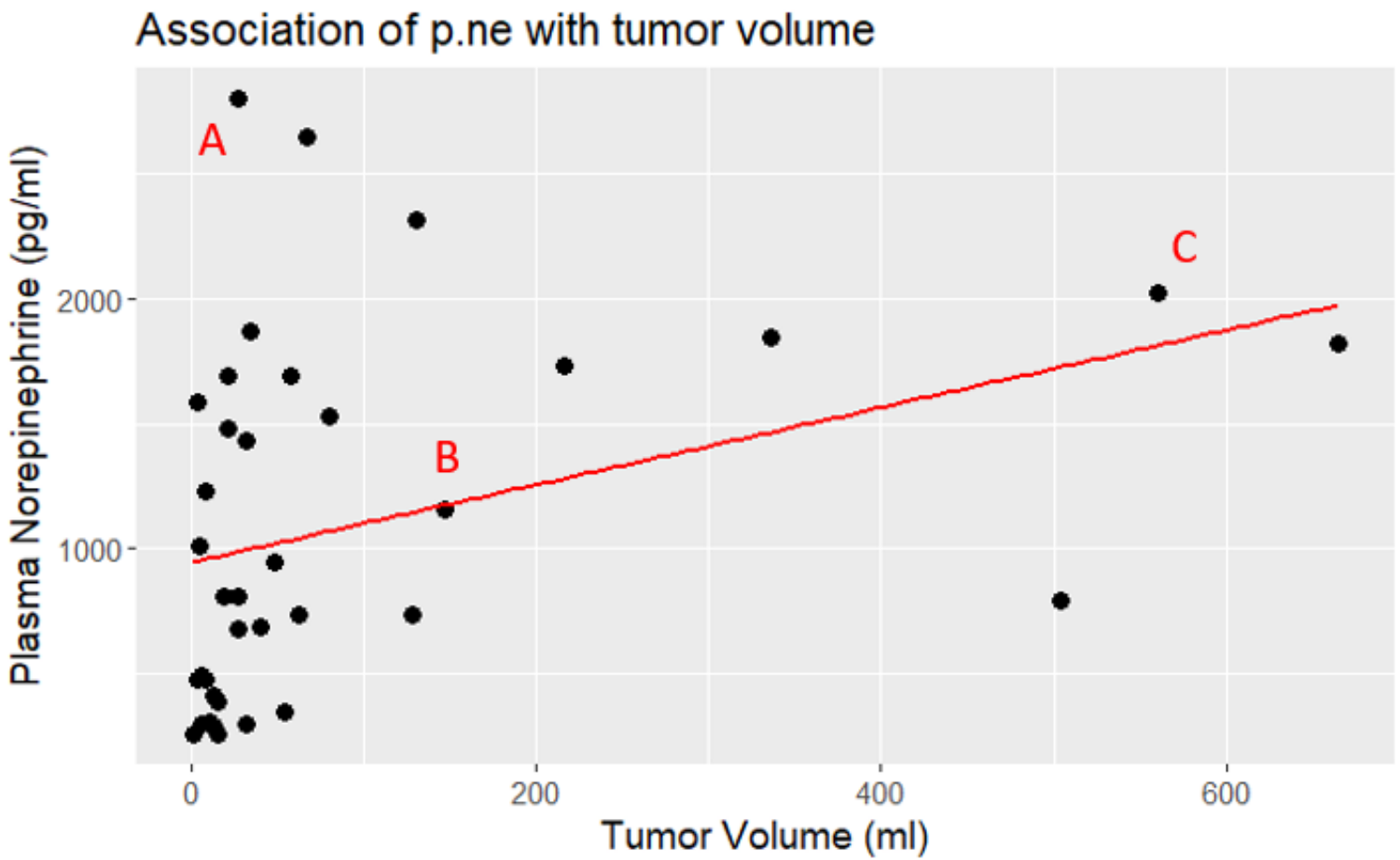
Understanding Regression Residuals (C)



Subject 117 has tumorvol = 504, and observed p.ne = 793 pg/nl.

Model predicts p.ne is 946.2 + 1.55(504) = 1726. So, residual = 793 – 1726 = **-933**

Understanding Regression Residuals (D)



Which point (A, B or C) has the largest positive residual?

Key Elements of the Summary (3)

```
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.

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

```
[1] 0.3465646
```

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.

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

```
[1] 0.3465646
```

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

```
[1] 0.5414319
```

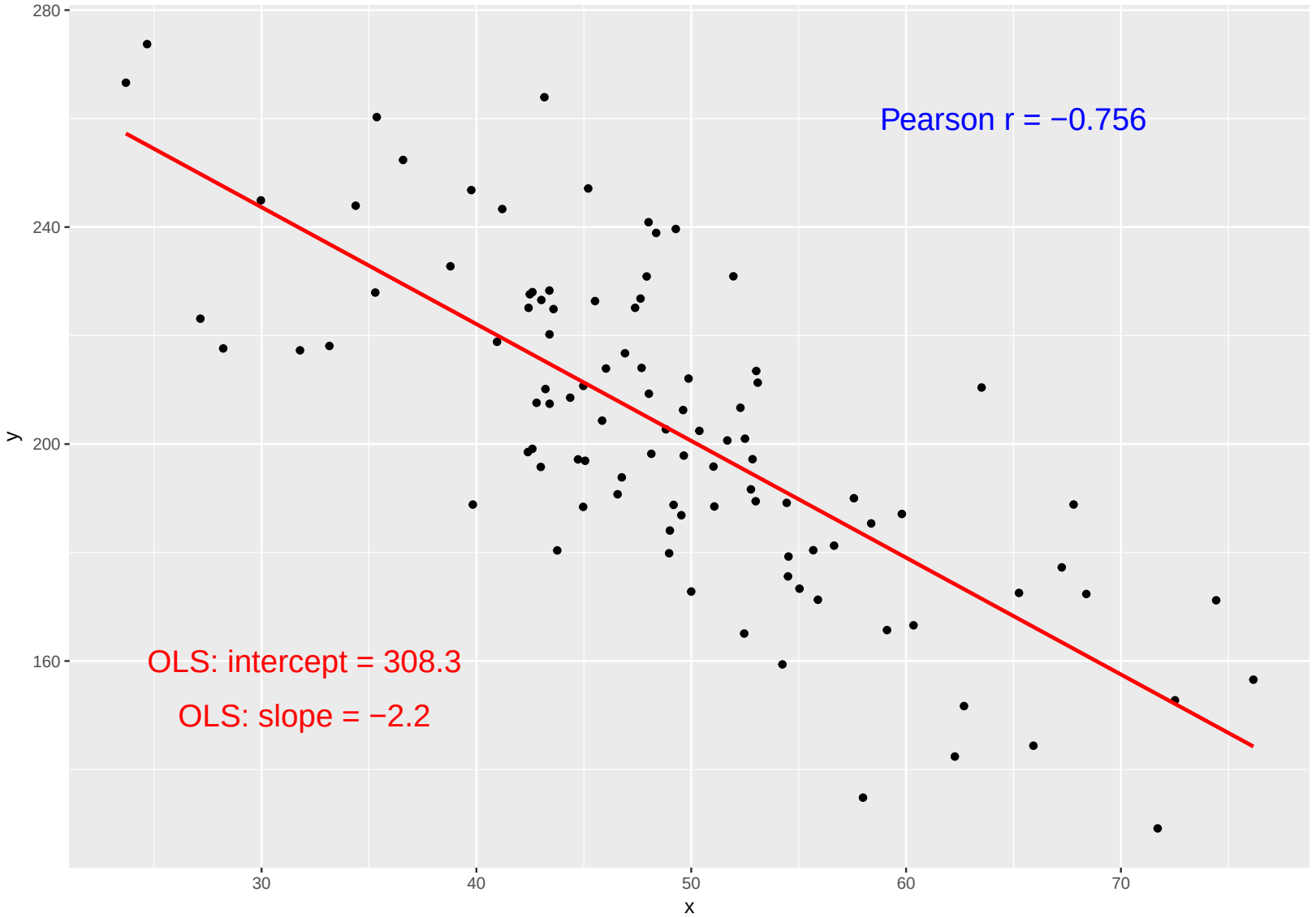
Meaning of Pearson Correlation

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

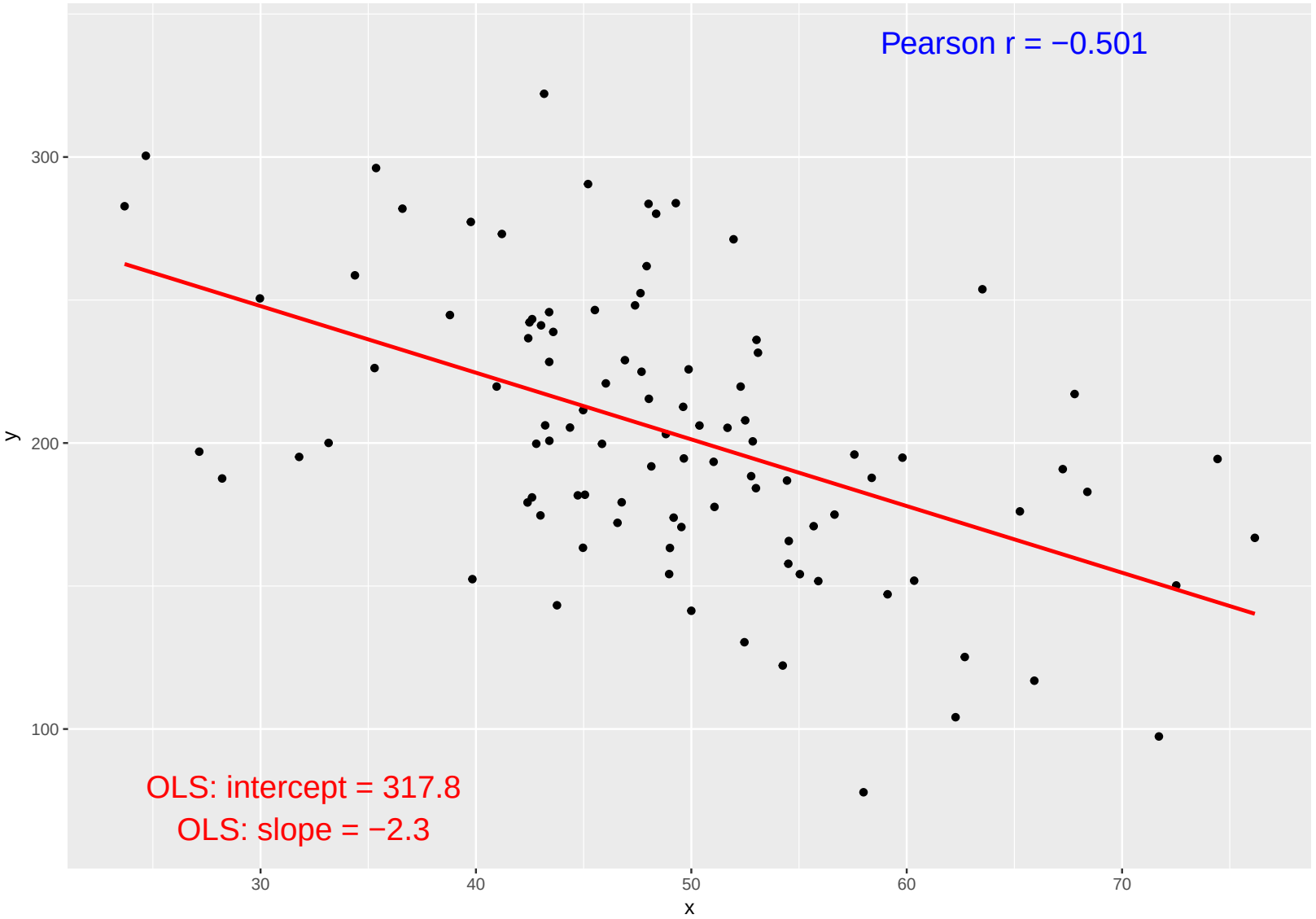
- The Pearson correlation is dimension-free.
- It falls between -1 and +1, with the extremes corresponding to situations where all the points in a scatterplot fall exactly on a straight line with negative and positive slopes, respectively.
- A Pearson correlation of zero corresponds to the situation 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.

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

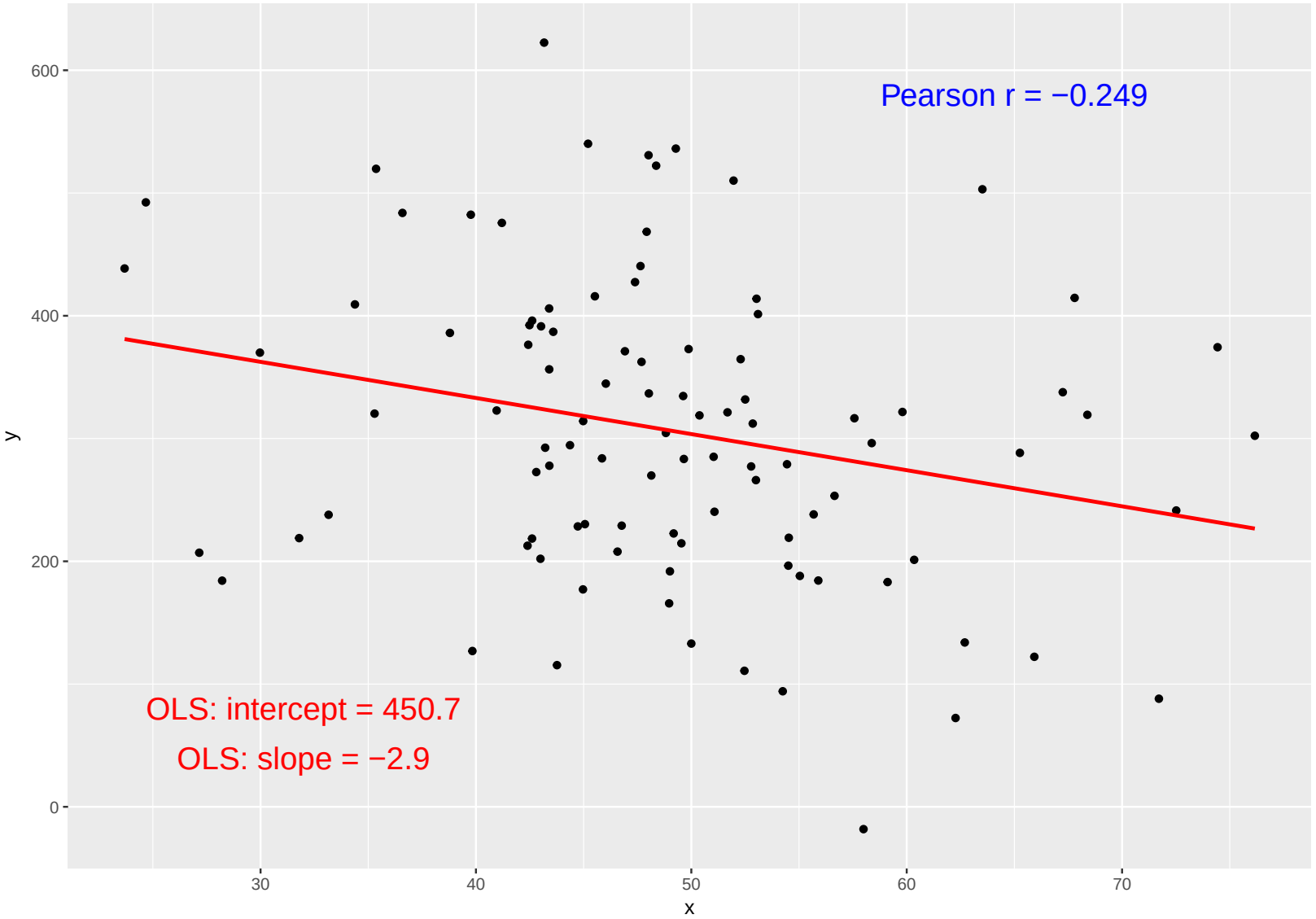
Simulated Example 1



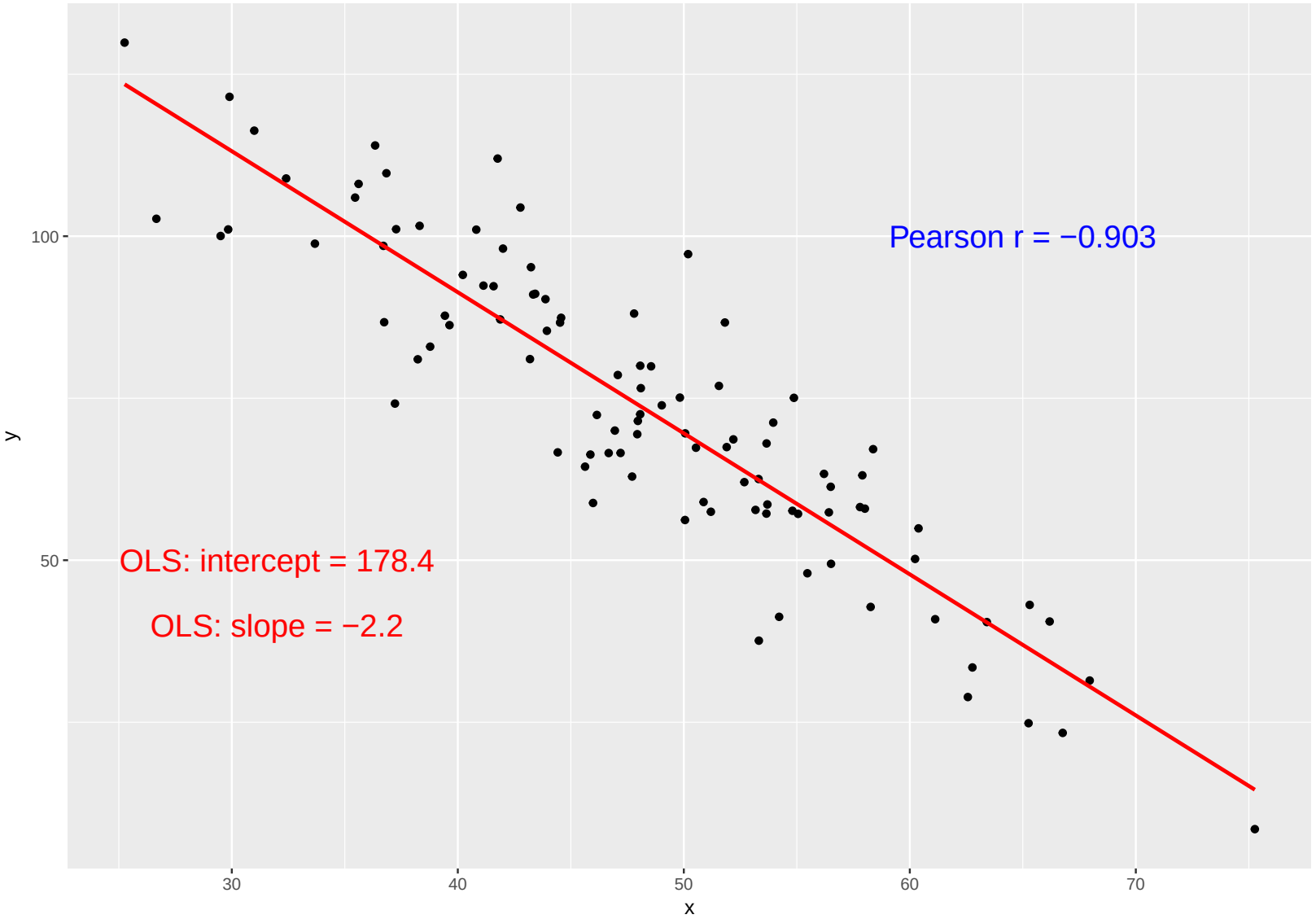
Simulated Example 2



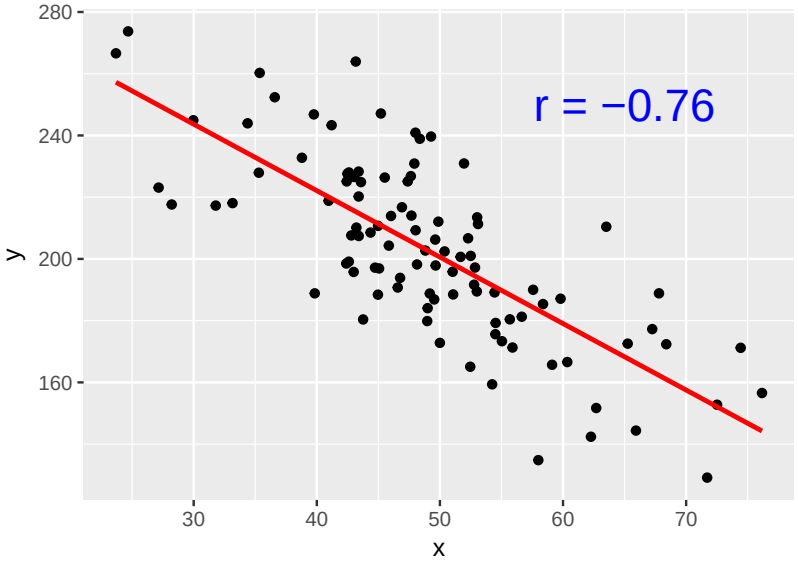
Simulated Example 3

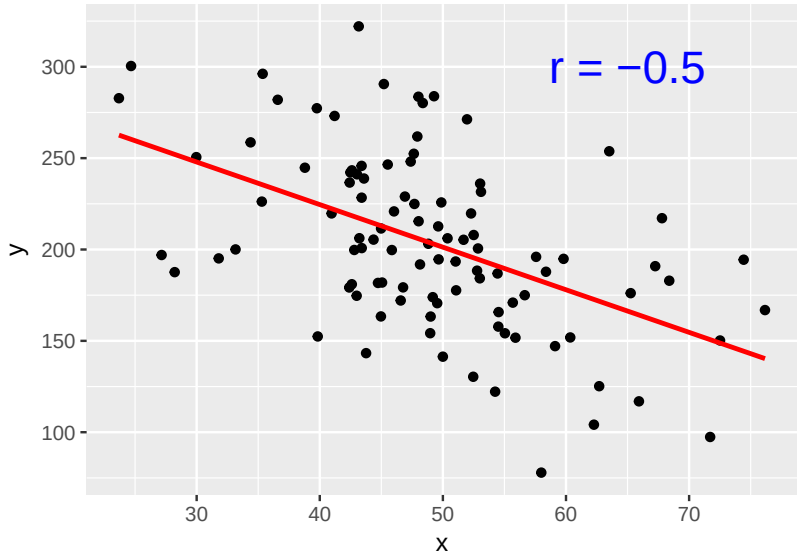


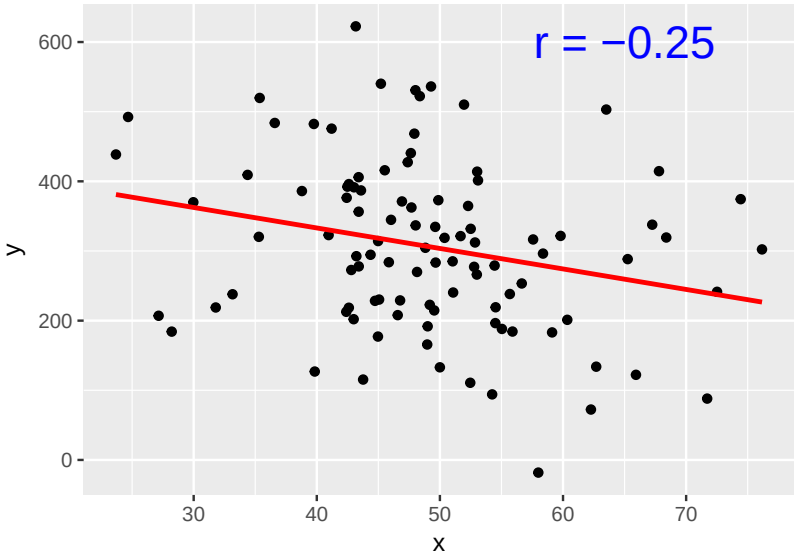
Simulated Example 4

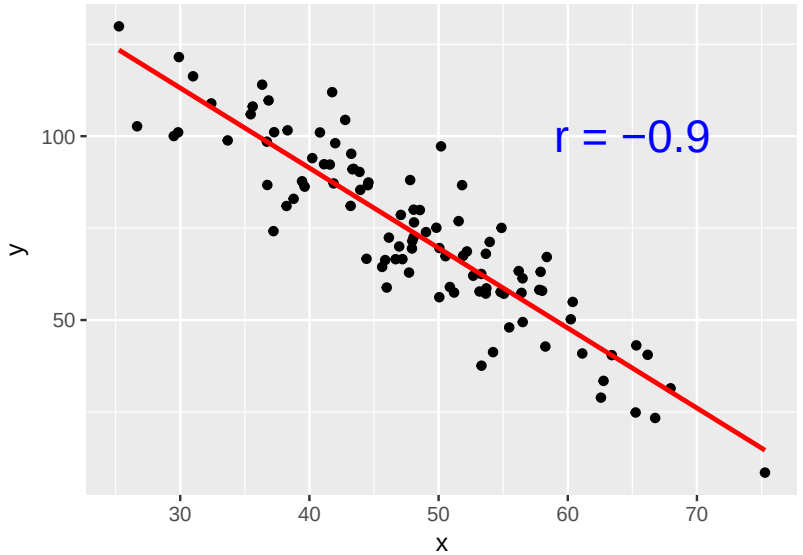


Calibrate Yourself on Correlation Coefficients

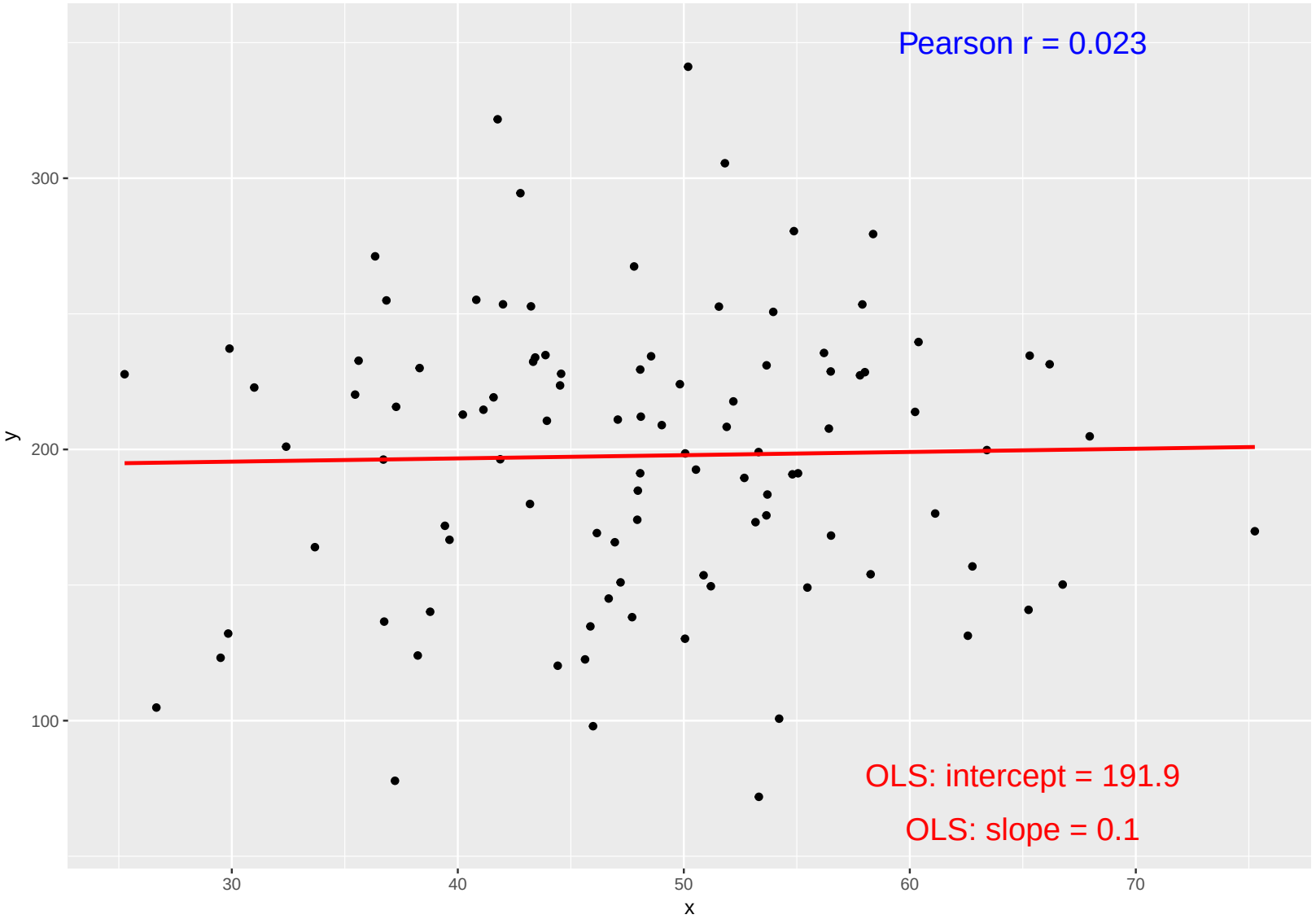




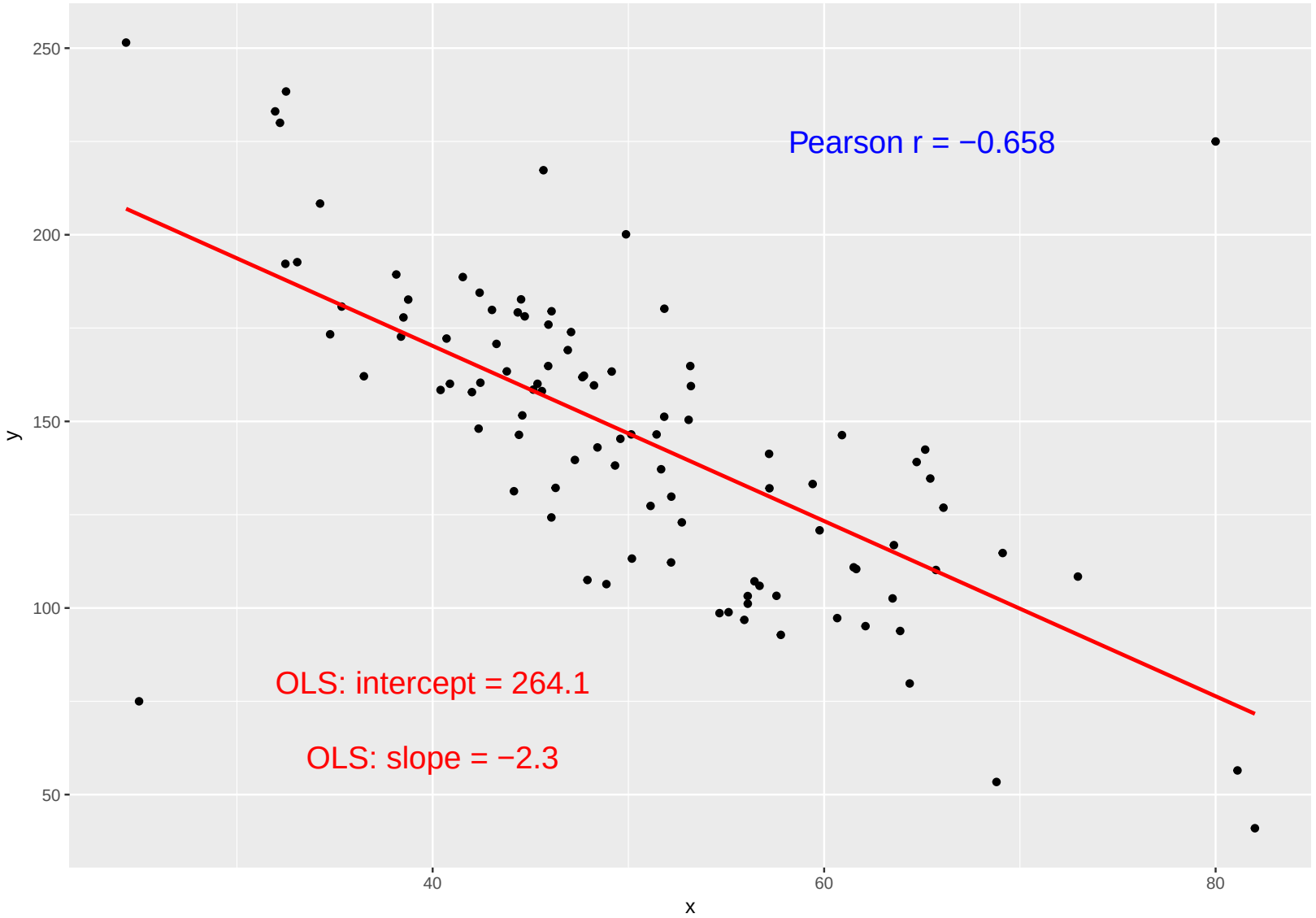




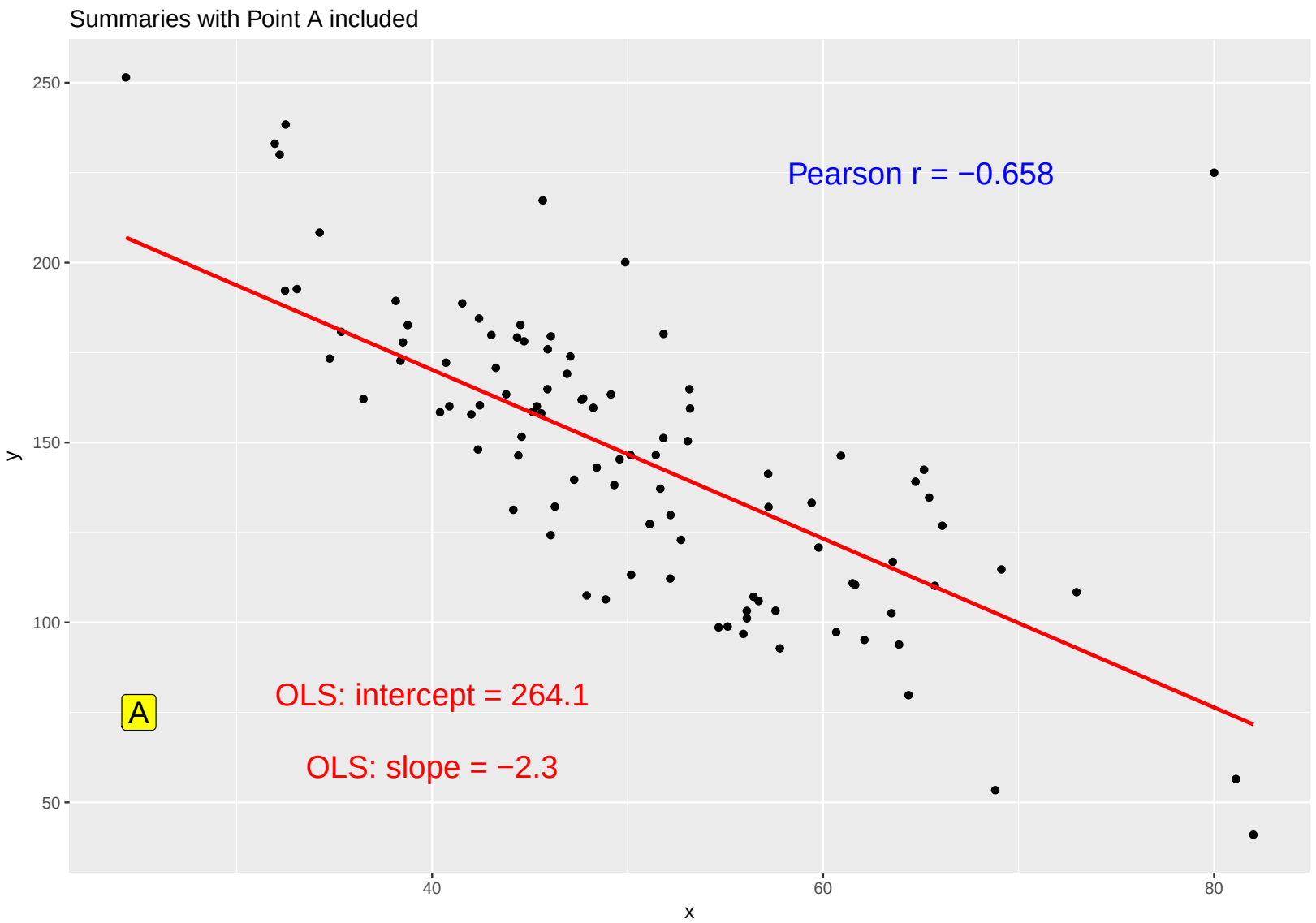
Simulated Example 5



Simulated Example 6

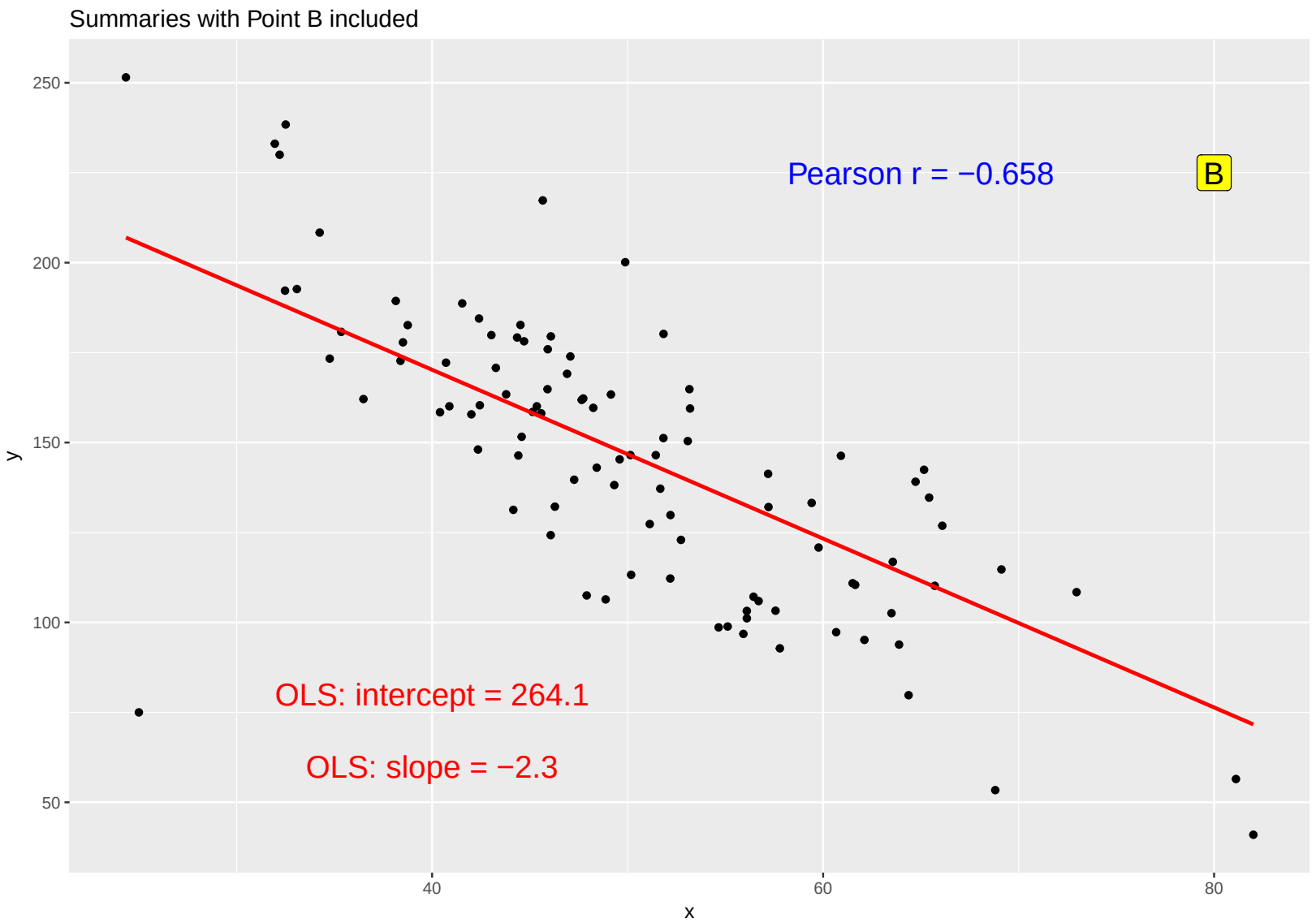


Example 6: What would happen if we omit Point A?

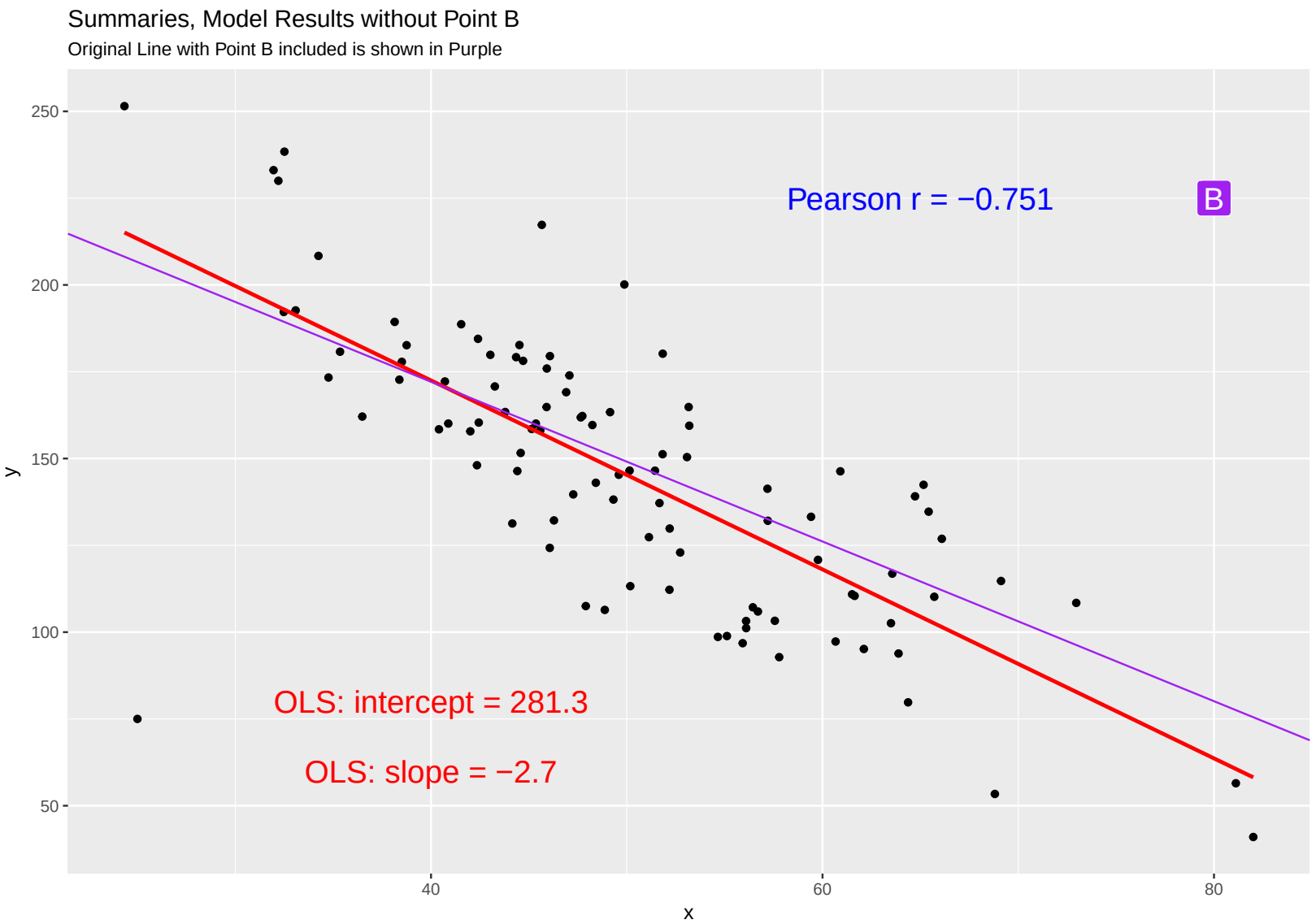


Example 6: Result if we omit Point A

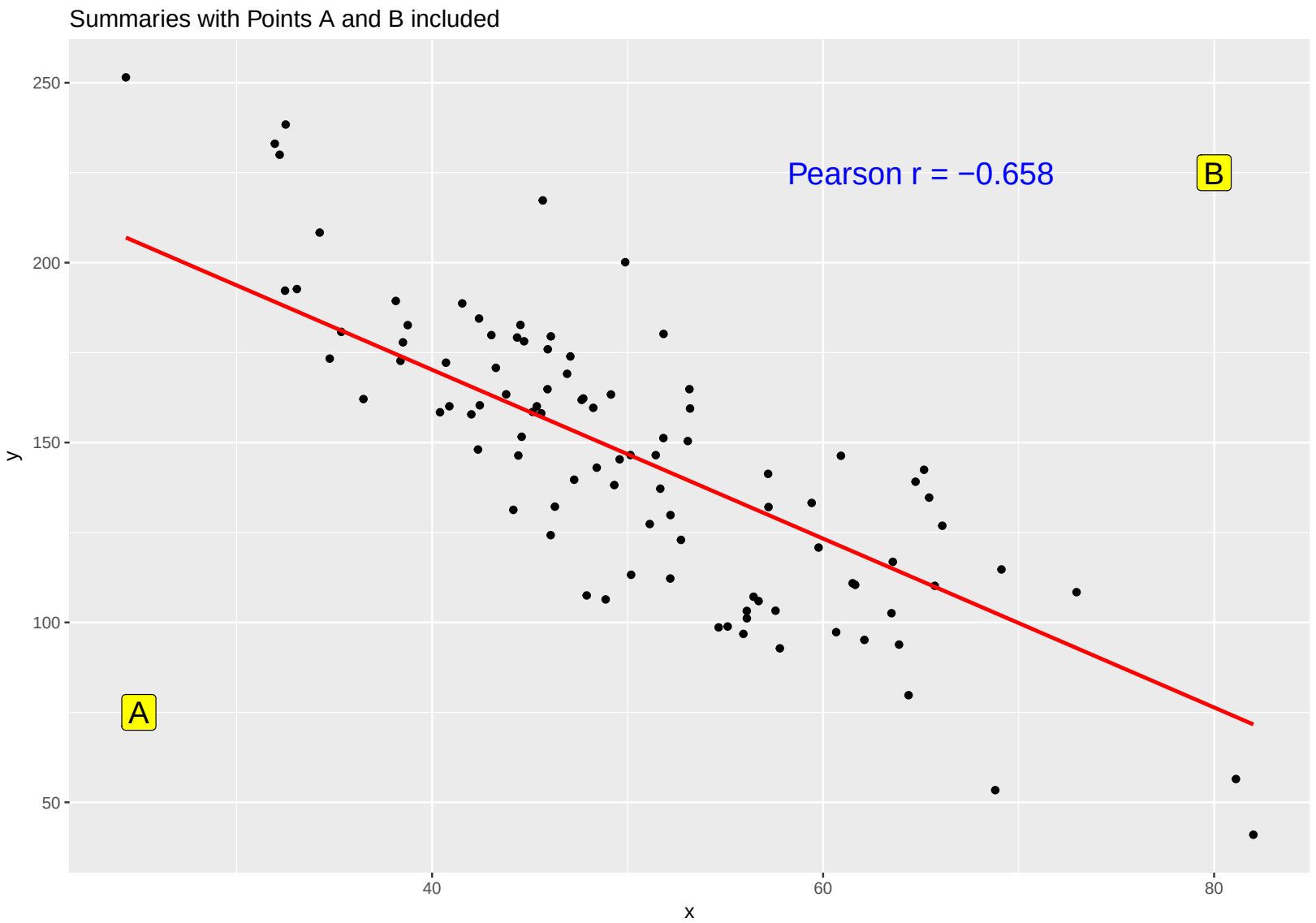
Example 6: What would happen if we omit Point B?



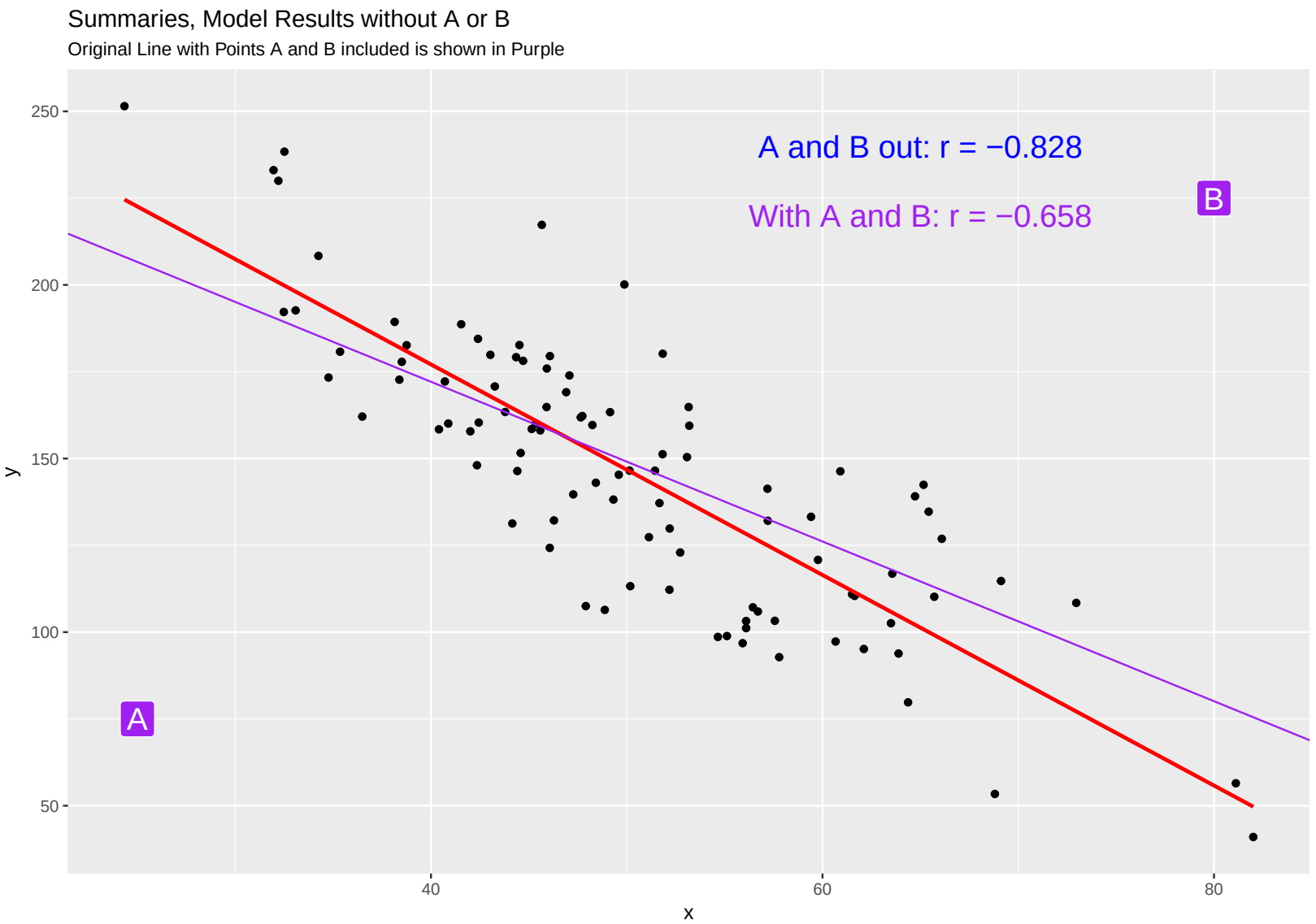
Example 6: Result if we omit Point B



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



Example 6: Result if we omit Points A and B

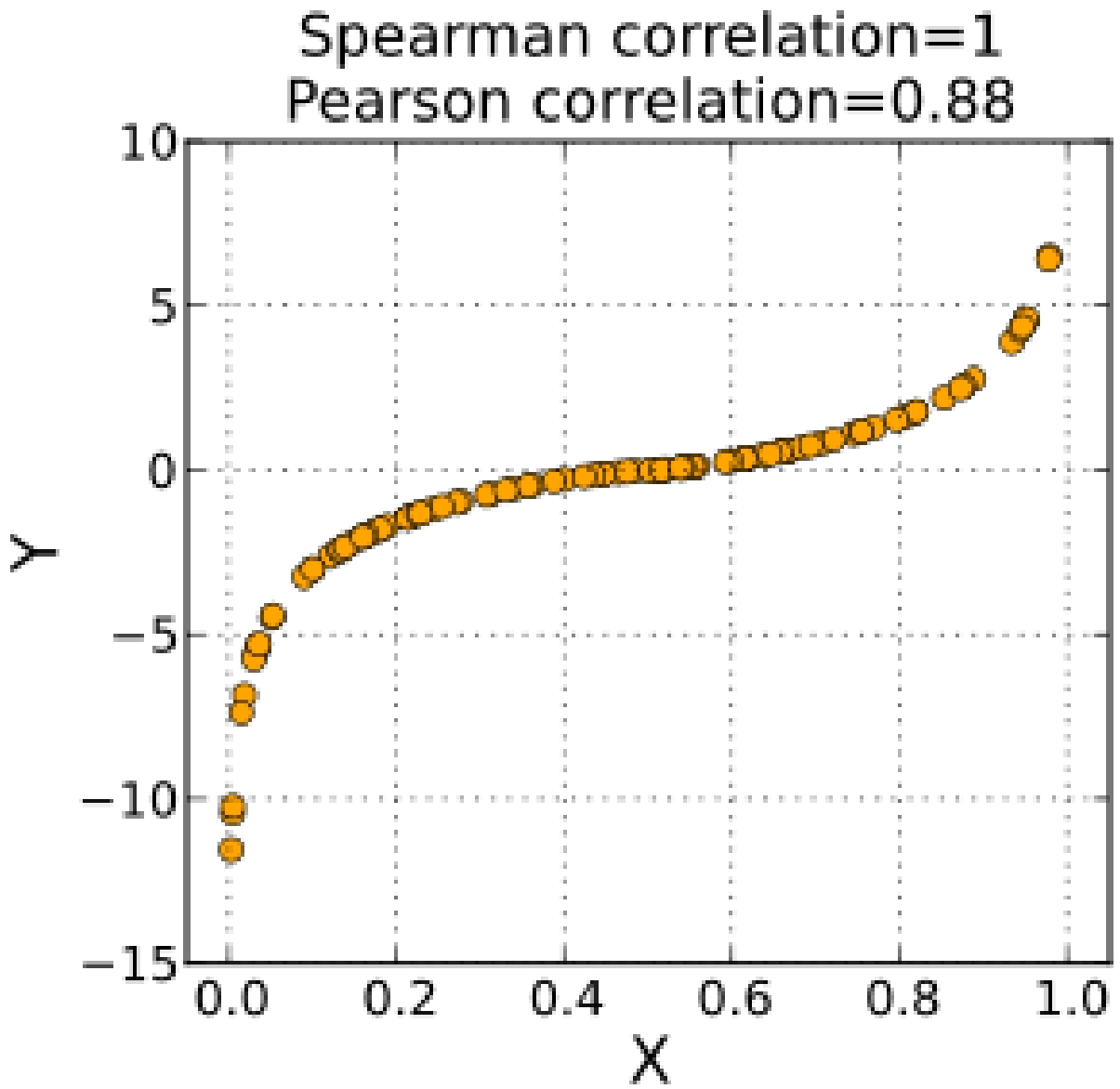


The Spearman Rank Correlation

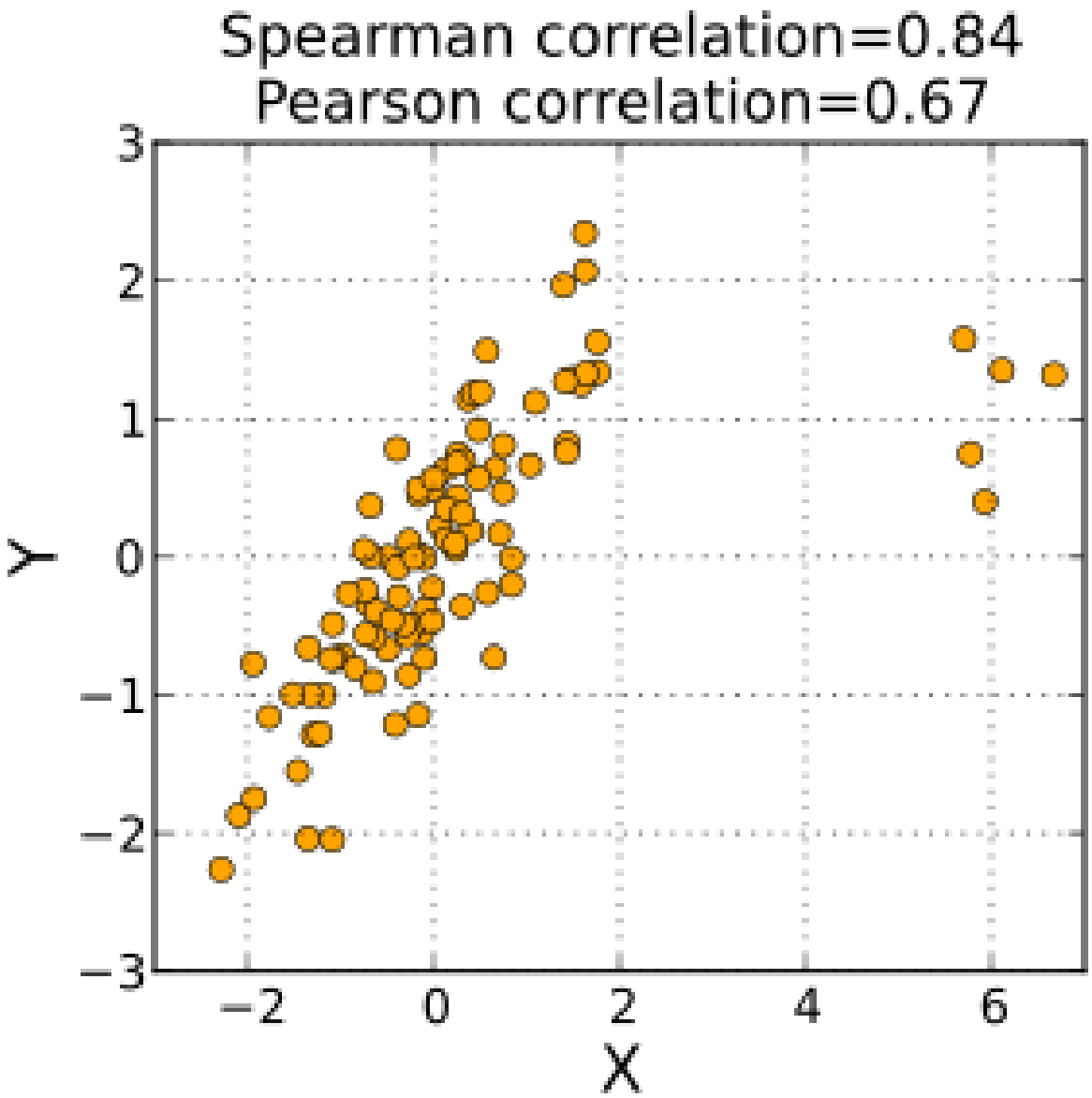
The Spearman rank correlation coefficient assesses how well the association between X and Y can be described using a **monotone function** even if that relationship is not linear.

- 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.

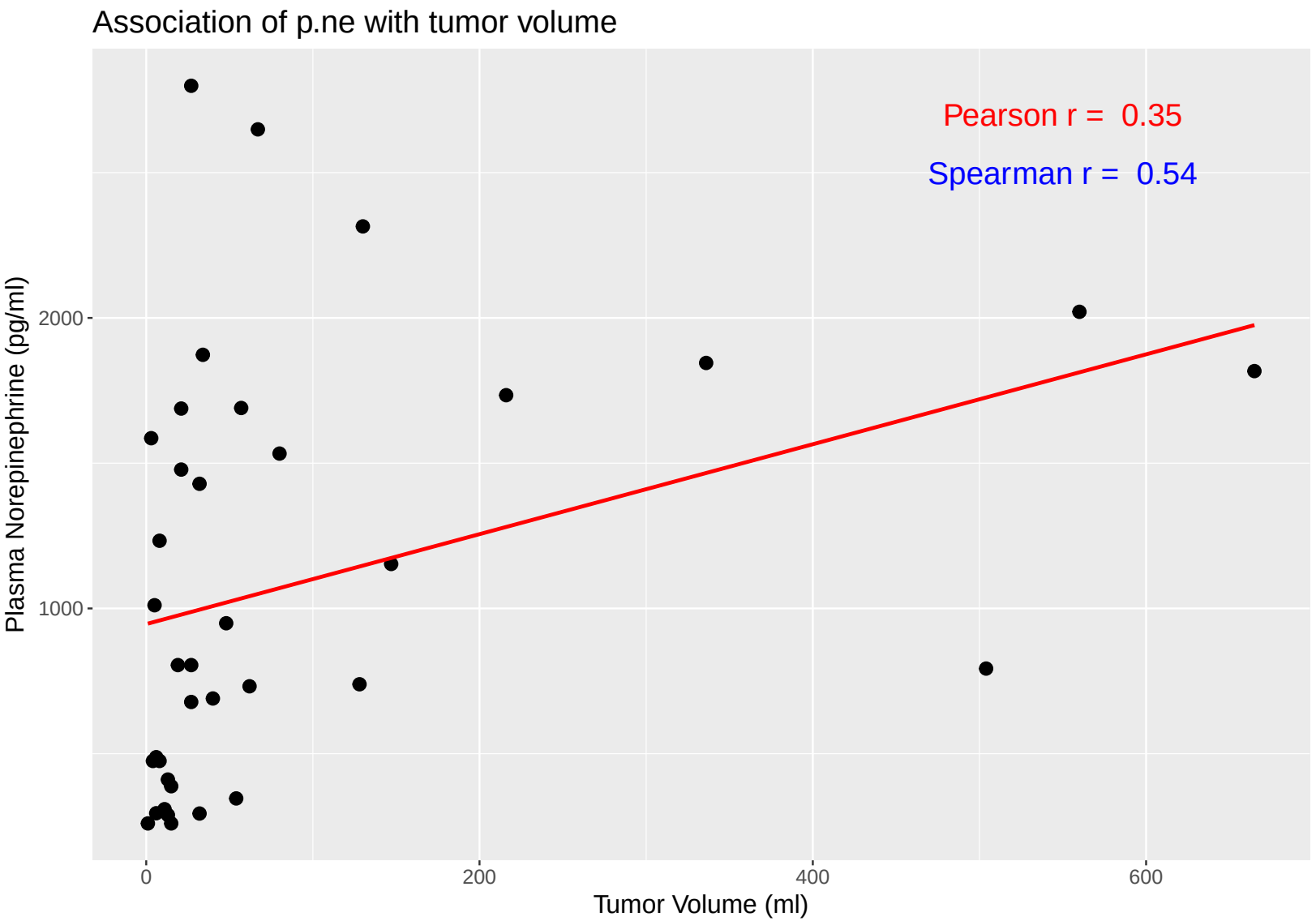
Monotone Association (Source: Wikipedia)



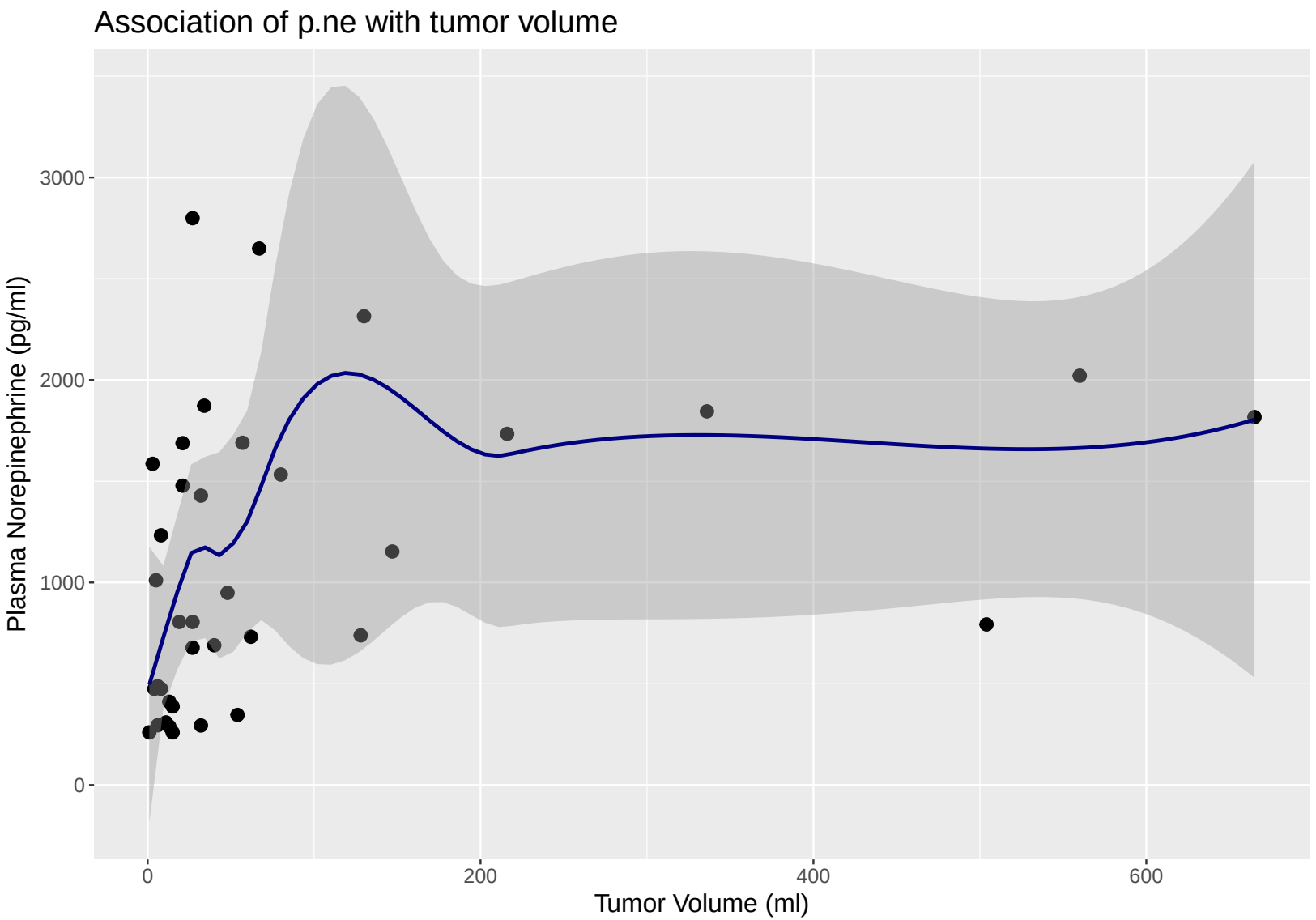
Spearman correlation reacts less to outliers



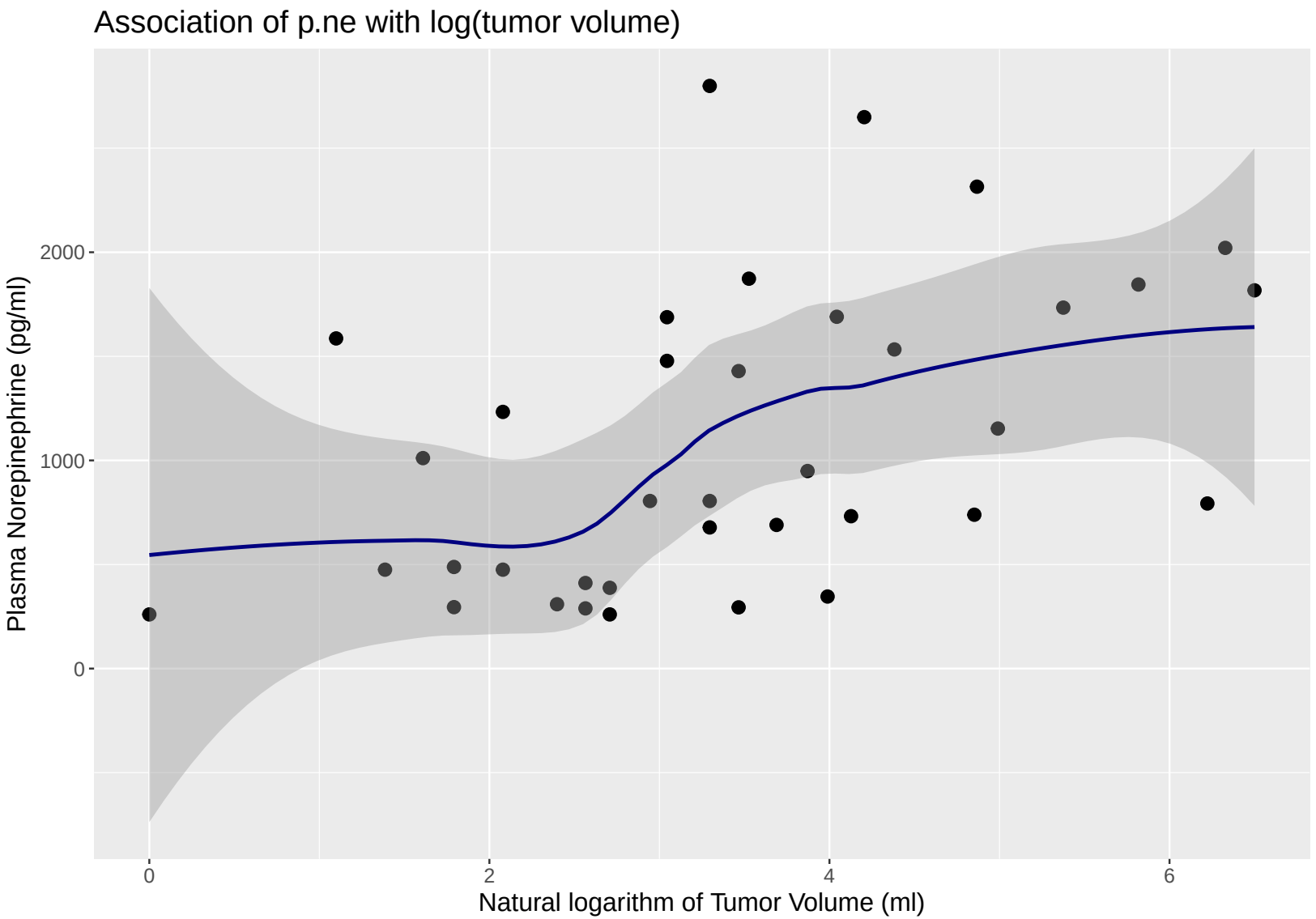
Our Key Scatterplot again



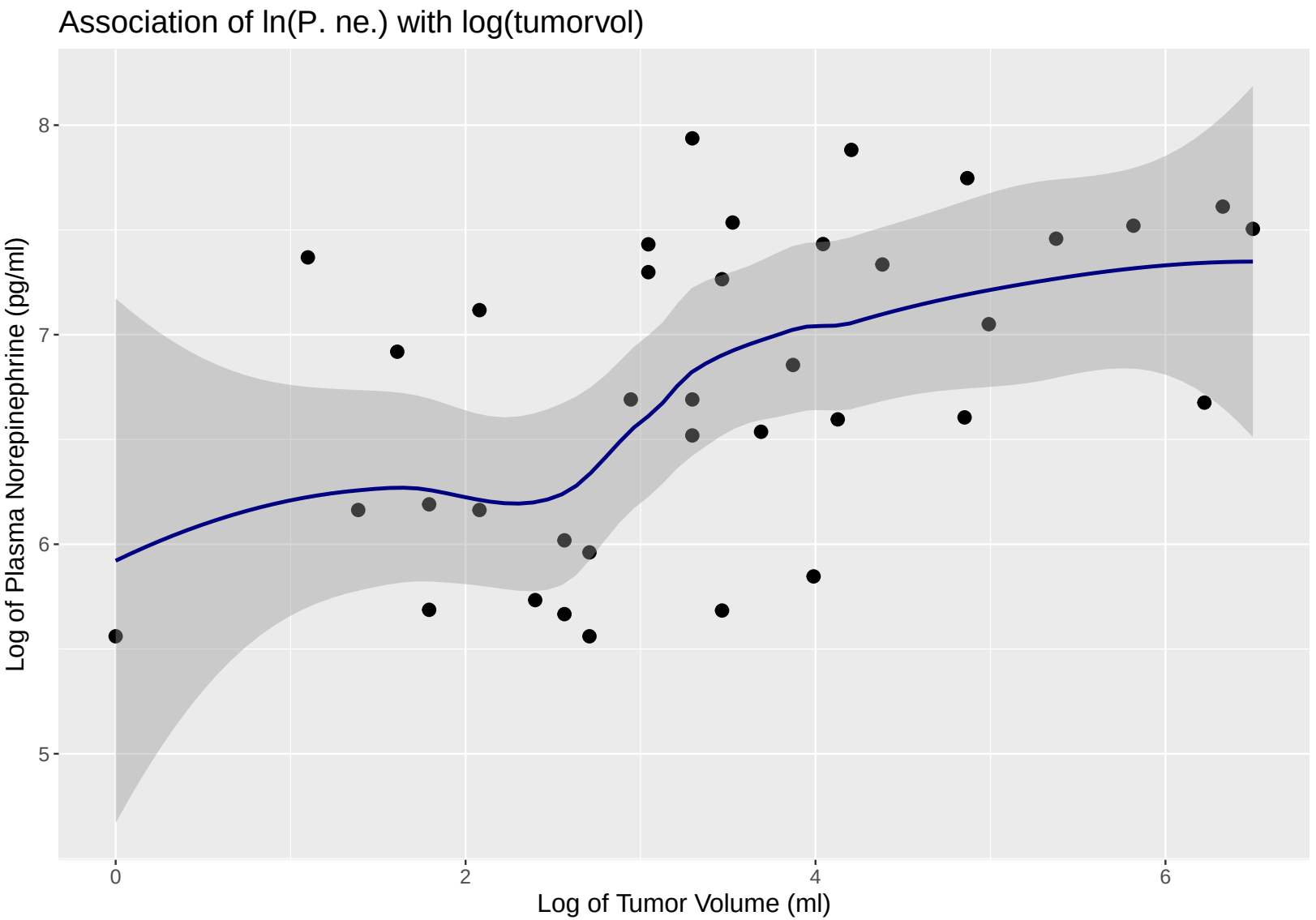
Smoothing using loess, instead



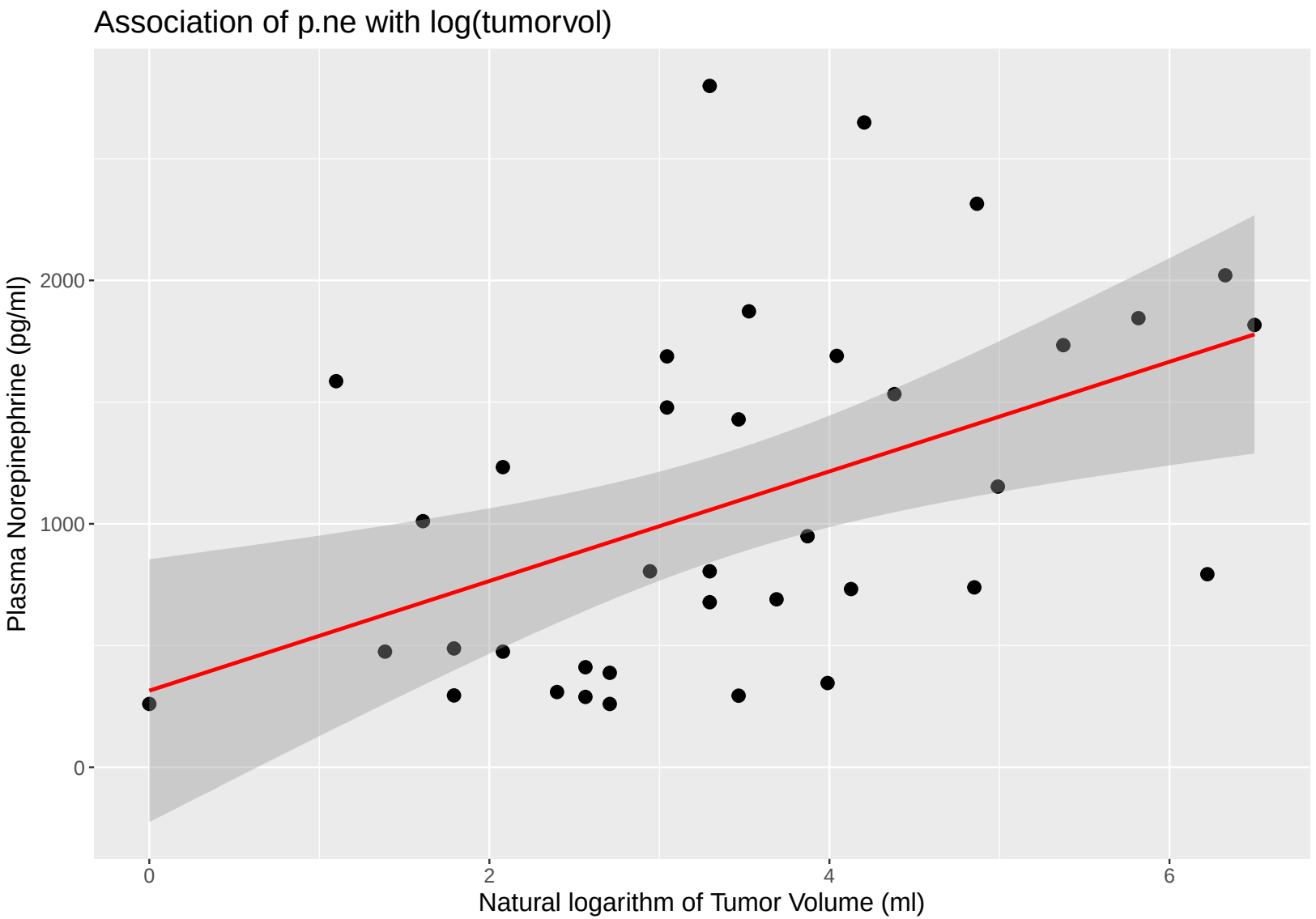
Using the Log transform to spread out the Volumes



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. . .

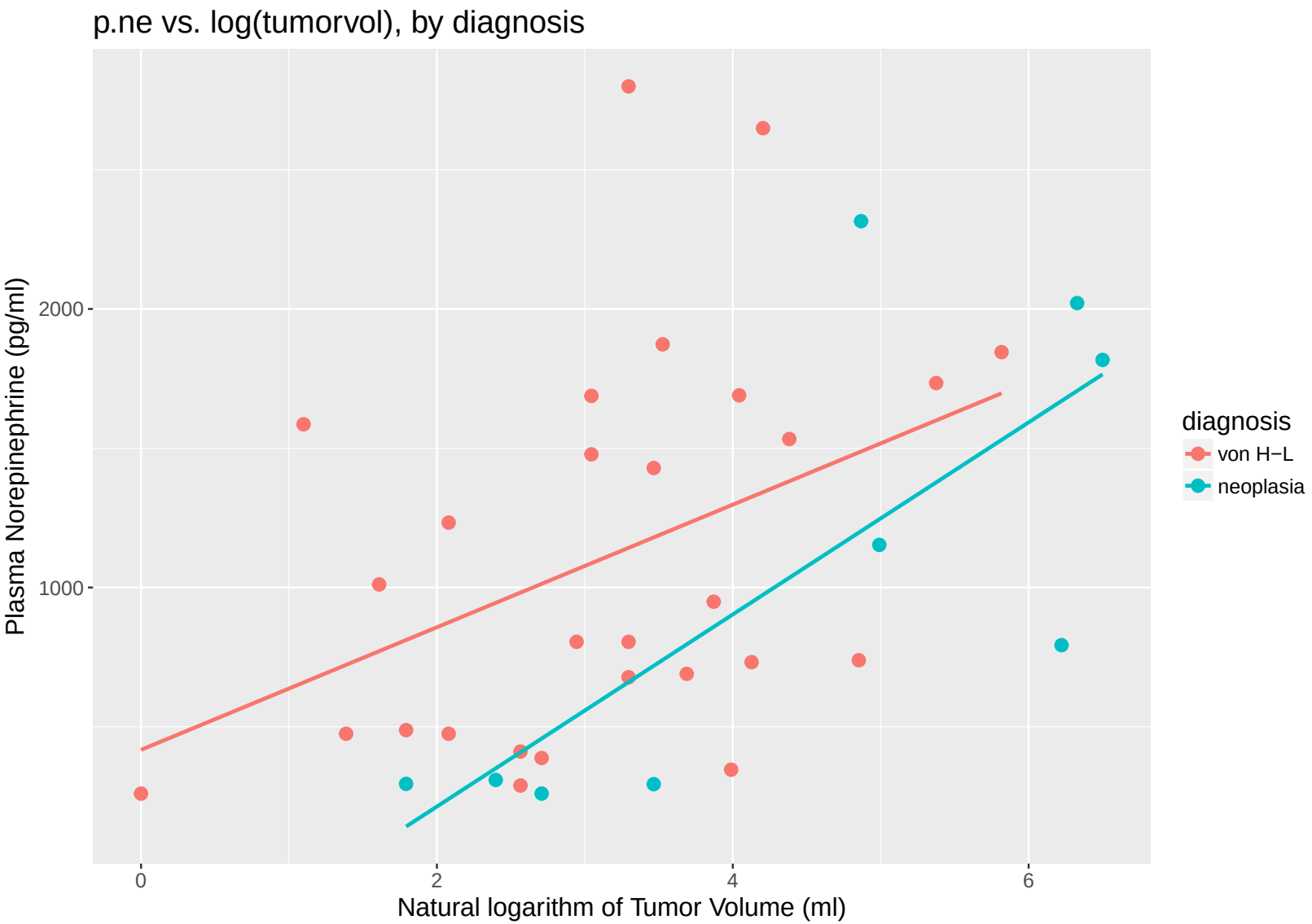
```
VHL <- VHL %>%
  mutate(diagnosis = fct_recode(factor(disease),
                                "neoplasia" = "1",
                                "von H-L" = "0")
  )
```

Now, what does VHL look like?

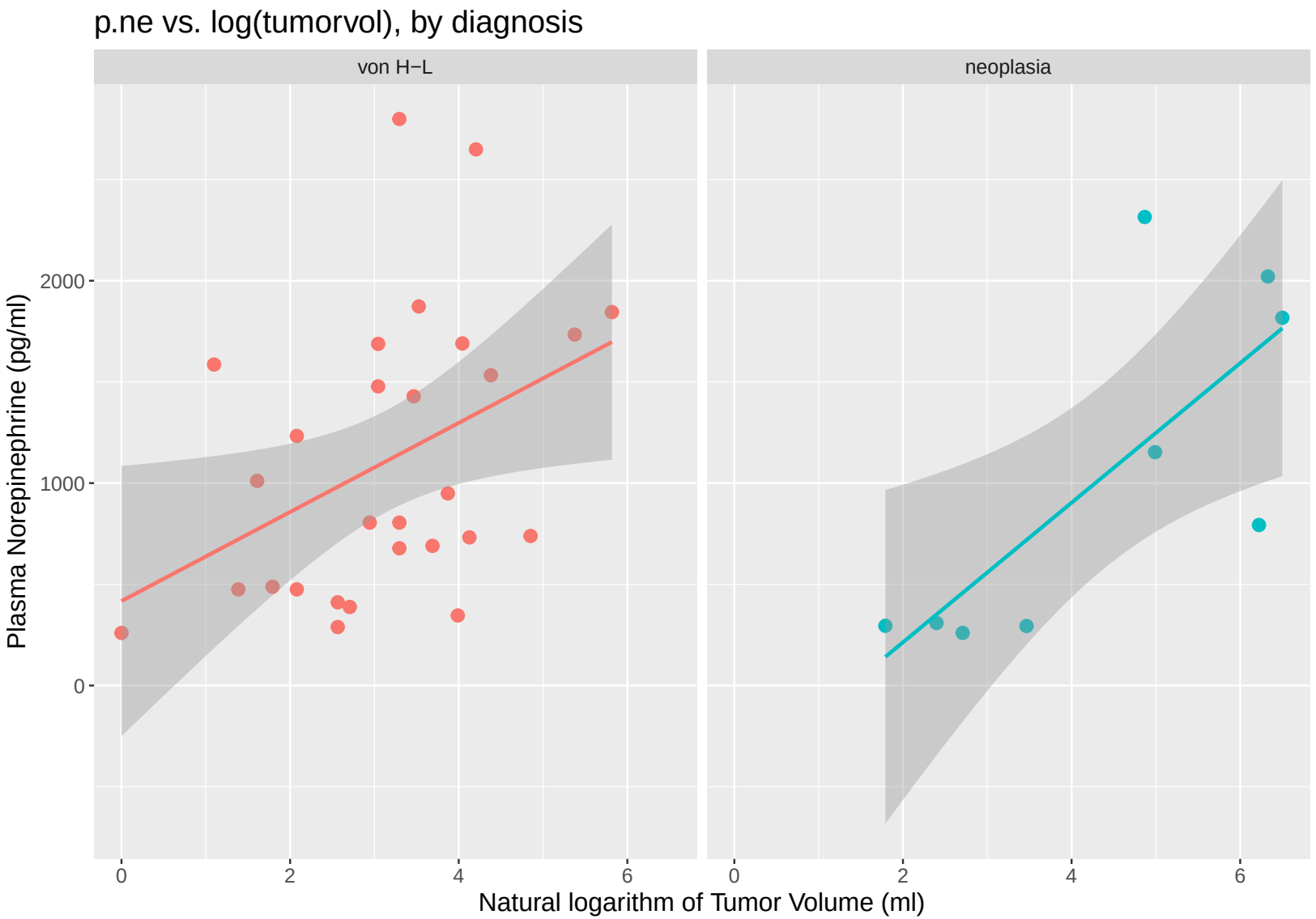
```
VHL
```

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

Setting up the Task B Groups

1. We want ten groups, each with 4-6 people. 5 is ideal.
2. You need the full names of your group members.
3. And their email addresses.
4. Select a group reporter and a group name.
5. Have the reporter fill out the Google Form from the Project Task B instructions.

Google Form for Project Task B groups is linked at https://github.com/thomaselove/431project

The Form's Questions. . .

Fall 2017 Project Task B Groups

This is the form to specify the group name and membership for Task B. Only one person from your group should fill out this form. The Task B groups will be formed in class on 2017-09-26. This form needs to be submitted by noon on 2017-09-27. If you have questions, contact Dr. Love directly.

Your email address (**tel3@case.edu**) will be recorded when you submit this form. Not you? Switch account

* Required

What is the name of your Task B group? *

100 characters or less, please.

Your answer

Please select the names of the members of your group from the list below.

Your group must include 4-6 people, in total. Be sure to check the box for each group member, including yourself.

	In Our Group
Albar, Zainab	☐
Asagba, Oghenerukema	☐