Introduction to Linear Regression

Jacob M. Schauer

Sample Class
Presented to the University of Twente

November 2019

Objectives

By the end of this lesson, we should be able to:

- Visualize and describe the relationship between two numerical variables.
- Define and describe a linear regression model.
- Fit and interpret a linear regression.

Motivation

Previous grades in statistics course.

```r
library(tidyverse); library(skimr)
grades <- read_csv("grades_2018.csv")
skim(grades)
```

- `id`
- `stat_important`: did student indicate statistics was important to them?
- `stat_major`
- `pretest`
- `grade`

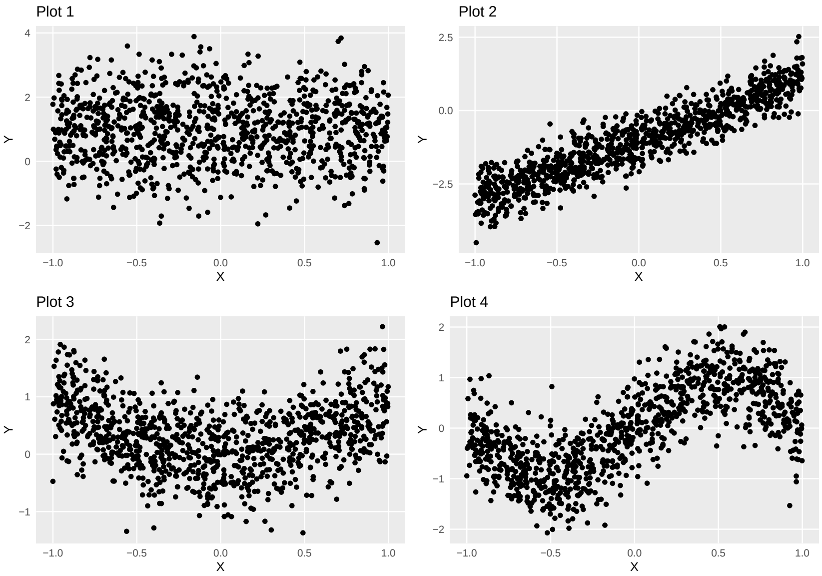
Motivation

- How is a student's pre-test score associated with their final grade?
- If so, how can we use pre-test scores to predict final grades?
  - Identify students at the beginning of the semester in order to give them additional help!

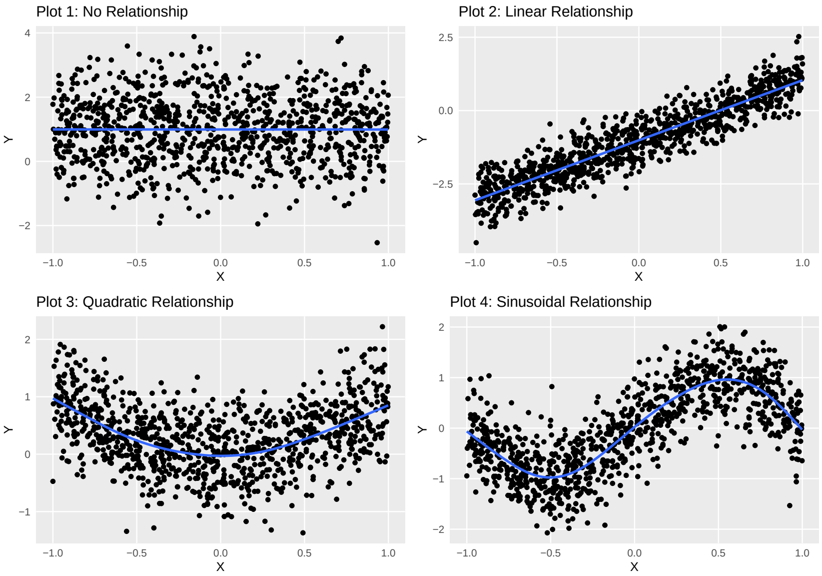
Review

- Data are typically stored in tables (`data.frame`)
  - Rows = observations
  - Columns = measurements on each observation
  - Example: Rows = students, Columns = test scores, etc.
- Paired numerical data: (X_i, Y_i)
  - Natural for a scatterplot
  - X and Y refer to columns in the table
  - i refers to a row
  - Means: $\bar{X}$, $\bar{Y}$
  - Standard deviations: S_X, S_Y

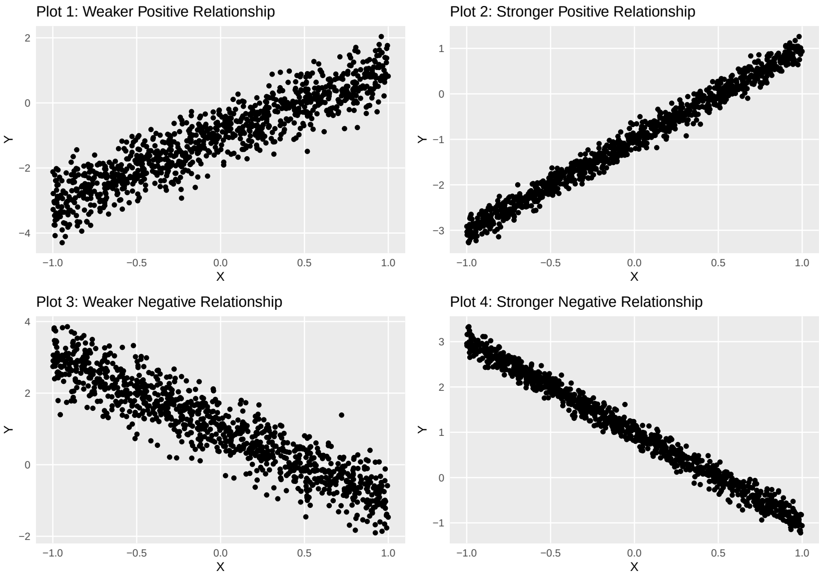
What do we mean that X and Y are "related"?



Shape/trajectory of relationships

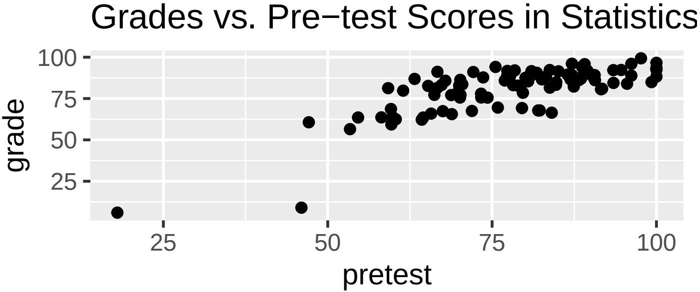


Strength of relationships



Visualizing relationships

```
ggplot(grades) +
  geom_point(aes(pretest, grade)) +
  labs(title = "Grades vs. Pre-test Scores in Statistics")
```



- Trajectory (which way does the line go?)
- Strength (how close are the points to the line?)

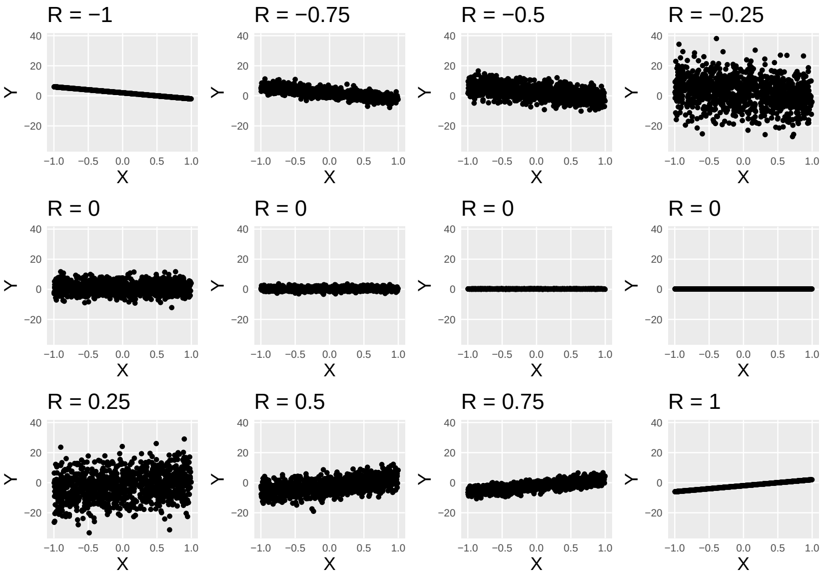
Correlation: the numerical summary of linear relationships

Correlation coefficient $R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$

In `R`, we can compute R as `cor(X, Y)`

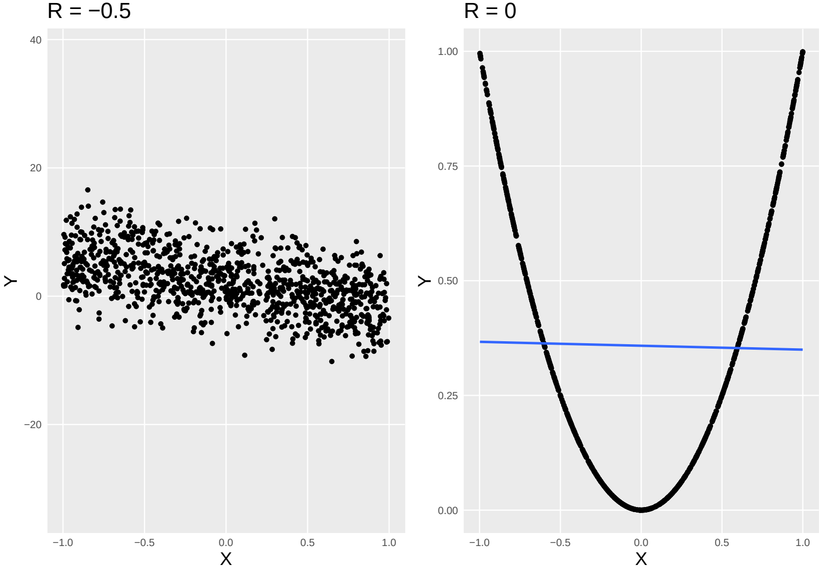
- R is between -1 and 1.
- $R = 0$ means that X and Y are not *linearly* related.
- $R > 0$ means that as X increases, Y increases
- $R < 0$ means that as X increases, Y *decreases*
- $|R| = 1$ means that (X, Y) all lie on a single line
- Often, people will use R^2 as a way to describe the strength of the relationship.
  - R^2 close to 1 indicates a *strong* linear relationship
  - R^2 close to 0 indicates a weak (or no) relationship

Correlations



Nonlinearity?

Correlation really measures the strength of the **linear** relationship



Guess the correlation

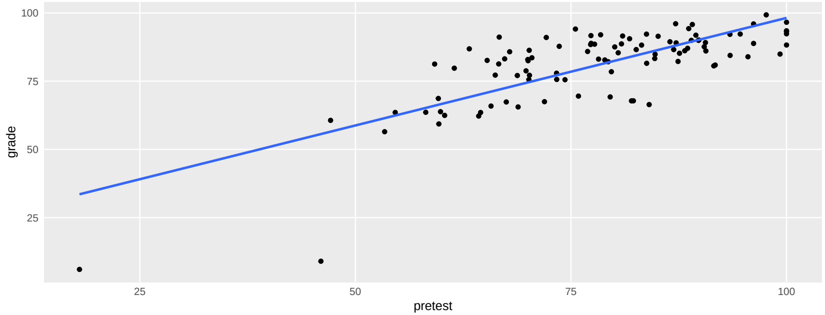
- It is not always easy to just "tell" what the correlation is from a plot.
- It is often difficult to say there is a relationship between variables from their correlation alone.
  - If the relationship is **linear** then the correlation can be really useful!
- **More informative to use plot + correlation**

Pre-test vs. grade

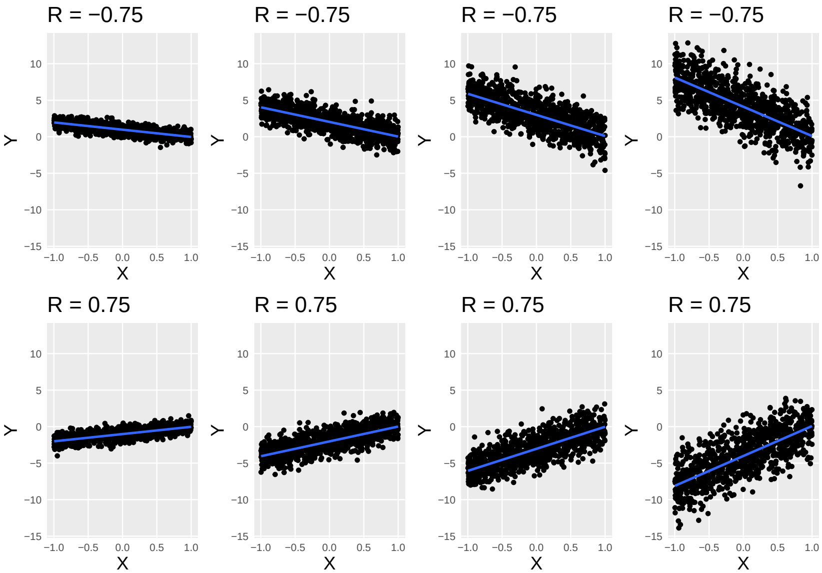
```r
cor(grades$pretest, grades$grade)
```

```
## [1] 0.7560622
```

```r
ggplot(grades) +
  geom_point(aes(pretest, grade)) +
  geom_smooth(aes(pretest, grade),
              method = "lm", se = FALSE)
```



Correlations



What about those lines?

Intercept

Slope

Remember, the formula for a line is $Y = \beta_0 + \beta_1 X$

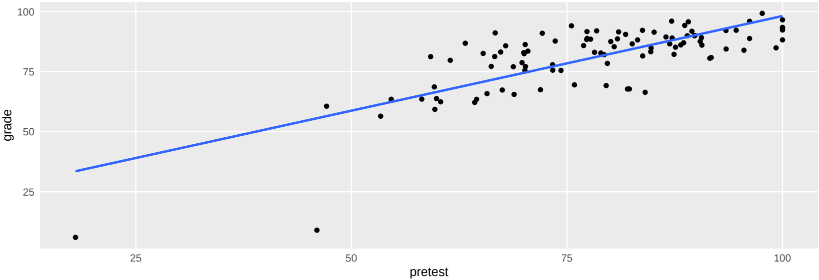
- β_0 tells us what Y is when $X = 0$
- β_1 tells us how much Y changes when X changes.

Example: $Y = 3 + 4 \times X$

- When $X = 0$, $Y = 3$
- When X changes by 1, then Y should change by 4

What about lines with data?

Not every data point will lie directly on the line



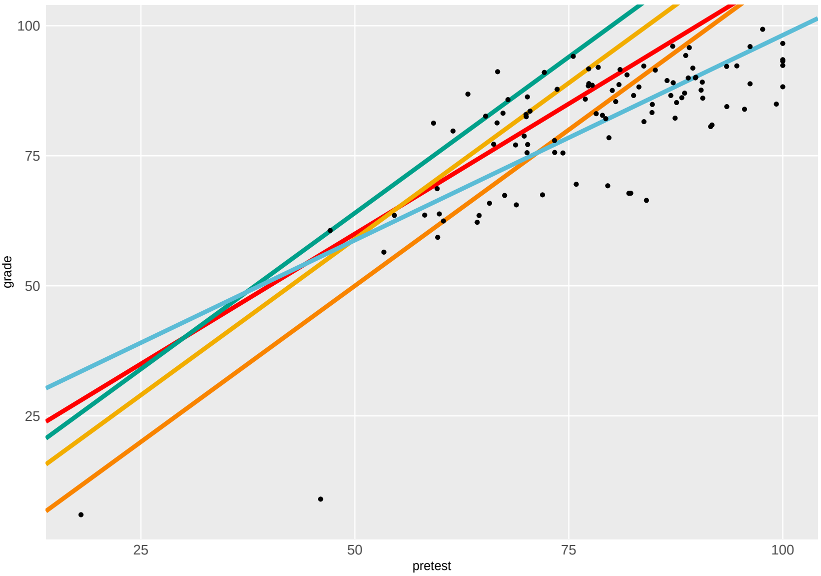
- $\hat{Y}_i = \beta_0 + \beta_1 X_i$
- $Y_i = \beta_0 + \beta_1 X_i + e_i = \hat{Y}_i + e_i$
  - Each value $\hat{Y}$ is like a "prediction" of what Y should be (on average) for a given value of X
  - e_i is like a "prediction error," and is called the **residual**

Linear regression

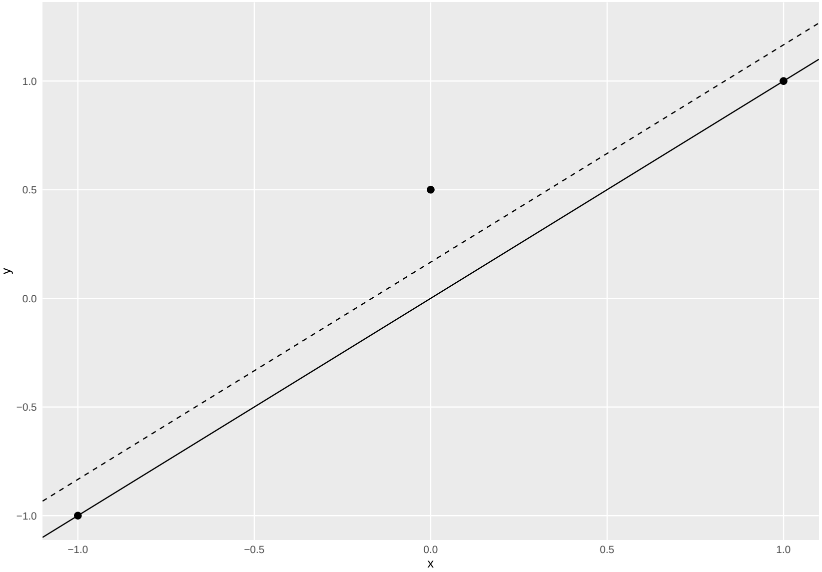
Regression equation: $Y_i = \beta_0 + \beta_1 X_i + e_i$

- β_0 tells us what Y is *on average* when $X = 0$.
- β_1 tells us how much we would *expect* Y to change when X changes.
- e_i is the residual.
- We are also interested in the *residual standard error* S_e.
    - $S_e = \sum_{i=1}^{n} \frac{e_i^2}{n-2} = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n-2}$
    - S_e is like the standard deviation of the e_i describes how big our prediction errors are (or how much the Y_i vary around the regression line).

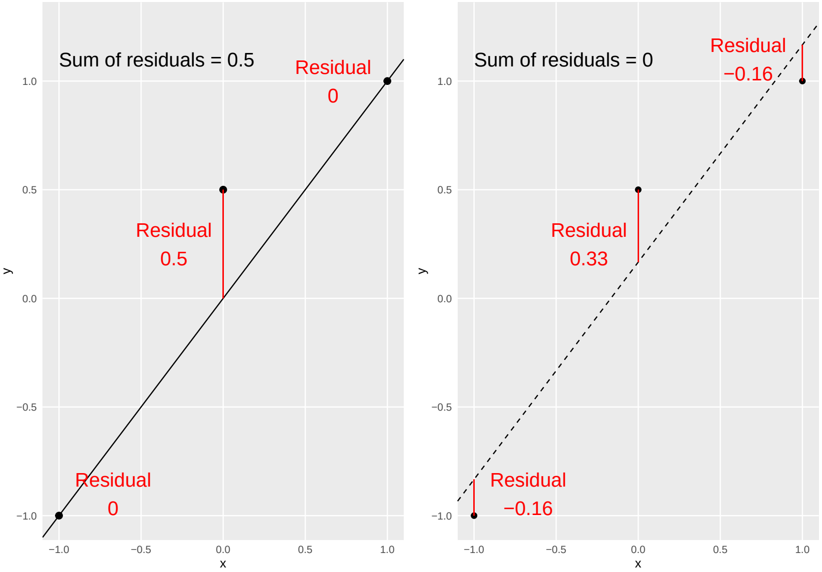
How do we choose β_0 and β_1?



Least squares

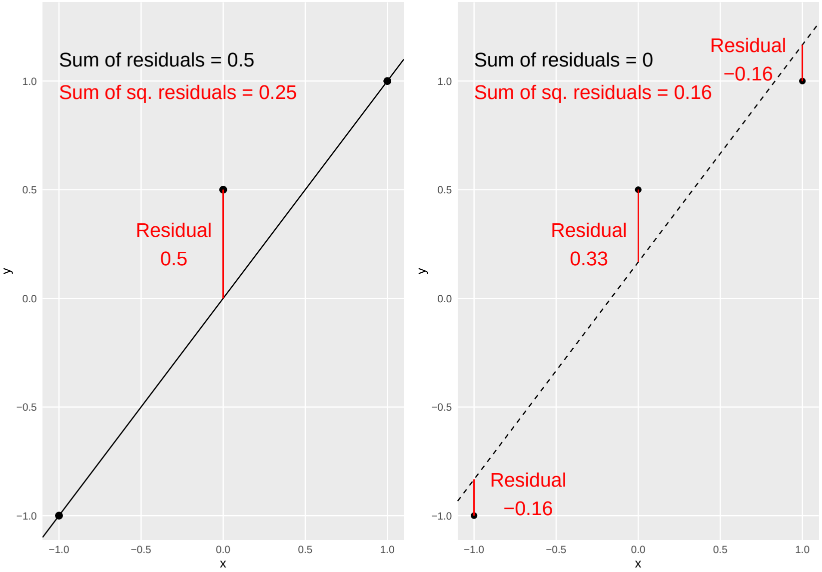


Residuals



Sum of residuals = 0.5
Residual
0
Residual
0.5
Residual
0
Sum of residuals = 0
Residual
−0.16
Residual
0.33
Residual
−0.16
1.0
0.5
0.0
−0.5
−1.0
y
−1.0
−0.5
0.0
0.5
1.0
x

Squared residuals



Least squares

Choose β_0, β_1 such that $\sum_{i=1}^{n} e_i^2$ is smallest

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- This is the same as choosing β_0, β_1 such that S_e is the smallest!

Answer:

$$\begin{aligned} \hat{\beta}_1 &= R\frac{S_y}{S_x} = R\frac{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Linear regression output

```r
lmod <- lm(grade ~ pretest, grades)
summary(lmod)
```

```
Call:
lm(formula = grade ~ pretest, data = grades)

Residuals:
    Min      1Q  Median      3Q     Max
-46.623  -4.913   0.613   6.121  19.233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.35467    5.51650   3.509 0.000688 ***
pretest      0.78843    0.06966  11.318  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.637 on 96 degrees of freedom
Multiple R-squared:  0.5716,    Adjusted R-squared:  0.5672
F-statistic: 128.1 on 1 and 96 DF,  p-value: < 2.2e-16
```

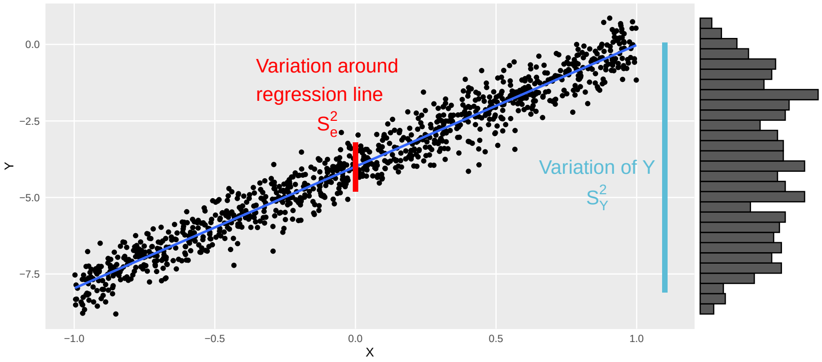
Linear regression output

```
Call:
lm(formula = grade ~ pretest, data = grades)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.623  -4.913   0.613   6.121  19.233 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 19.35467    5.51650   3.509 0.000688 ***
pretest      0.78843    0.06966  11.318  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.637 on 96 degrees of freedom
Multiple R-squared:  0.5716,	Adjusted R-squared:  0.5672 
F-statistic: 128.1 on 1 and 96 DF,  p-value: < 2.2e-16
```

- $\hat{\beta}_1 = 0.79$. This means that for an increase of one point in pre-test score, we would expect someone's final grade to increase by 0.79 points.
  - **Question:** What if pre-test score increases by 10 points?
  - We would expect final grade to increase by $10\hat{\beta}_1 = 7.9$ points.
- $\hat{\beta}_0 = 19.35$. This means that for someone who scores a 0 on the pre-test, we would expect their final grade to be 19.35.
- $S_e = 9.64$. This means that on average, our predictions are off by about 9.64 points in the final grade.

Strength of relationship (revisited)

- R^2 describes the proportion of the variation in Y that is explained by X.
- $R^2 = 1 - \frac{\text{Variation around regression line}}{\text{Variation of } Y} = \frac{(n-1)S_Y^2 - (n-2)S_e^2}{(n-1)S_Y^2}$



Putting it all together

We want to know how pre-test scores are related to final grades.

- Correlation R
  - Direction, size, meaning (*linear relationship*)
- Scatterplot
  - Shape of relationship (if any) and strength
- Linear regression equation
  - Regression coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$
  - For a 1-unit increase in X, we would expect a $\hat{\beta}_1$ increase in Y
  - Residual standard error S_e tells us how close to the regression line our Y_i are on average.
- R^2
  - How much variation in Y is explained by X.

Putting it all together

```r
# correlation
cor(grades$pretest, grades$grade)

# scatterplot
ggplot(grades) +
  geom_point(aes(pretest, grades))

# regression equation
lm(grades ~ pretest, grades)

# scatterplot with regression equation and correlation
r <- cor(grades$pretest, grades$grade)
ggplot(grades) +
  geom_point(aes(pretest, grades)) +
  geom_smooth(aes(pretest, grades), method = "lm")
```

Activity

- Work together to answer the questions in the lab.

Key points

When examining relationships between numerical variables:

- Plot your data
- Compute R or R^2
- **Interpret** your plot and correlation coefficient

When examining *linear* relationships

- Run a regression model
- **Interpret the intercept, slope, and residual standard error!**
- **Interpret the R^2 value!**

For next class

- **Read** *OpenIntro Stats* pages 328-340
- **Watch** Outliers, Inference for regression
- **Think** about what we think of as *random* in a regression model and how that is related to the residuals.
- **Find** an article online that uses linear regression. How do they use it? Did they accurately interpret the statistics involved?
  - You can usually find something at FiveThirtyEight