Week 3: Basic regression

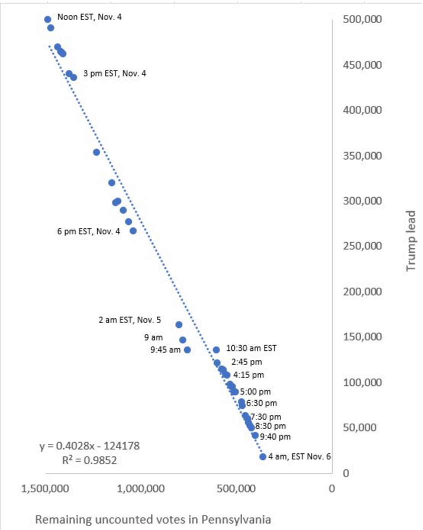
4. How useful is a linear model

Stat 140 - 04

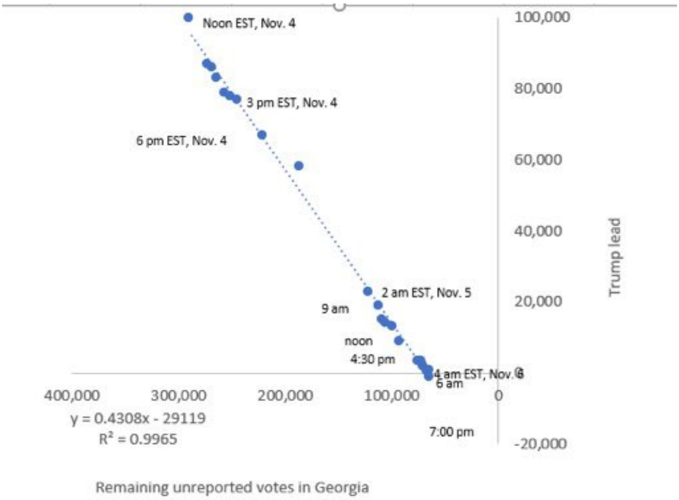
Mount Holyoke College

Dr. Shan Shan

Slides posted at *http://sshanshans.github.io/stat140*



* Source: https://www.nbcnews.com/politics/2020-elections/pennsylvania-president-results?icid=election_usmap



Noon EST, Nov. 4
3 pm EST, Nov. 4
6 pm EST, Nov. 4
2 am EST, Nov. 5
9 am
noon
4:30 pm
4 am EST, Nov. 6
6 am
7:00 pm
100,000
80,000
60,000
40,000
20,000
0
-20,000
Trump lead
400,000
300,000
200,000
100,000
y = 0.4308x - 29119
R² = 0.9965
Remaining unreported votes in Georgia

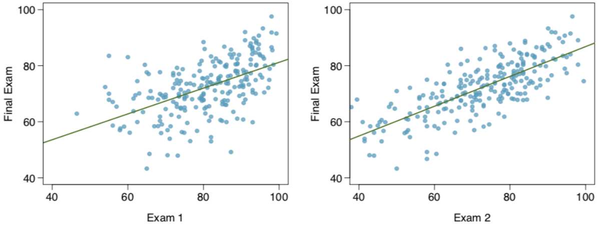
What does the intercept mean here?
Is it useful?

What does the intercept mean here?
Is it useful?

What is R^2?

The two scatterplots below show the relationship between final and mid-semester exam grades recorded during several years for a Statistics course at a university.

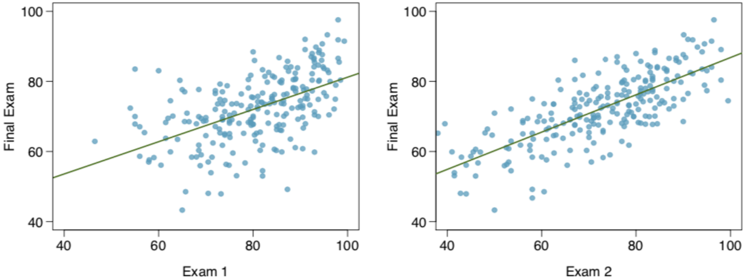
- **Final exam** the final
- **Exam 1** first midterm
- **Exam 2** second midterm



Poll question

Which of these models would you prefer to use for predicting sales?

a. Exam 1
b. Exam 2

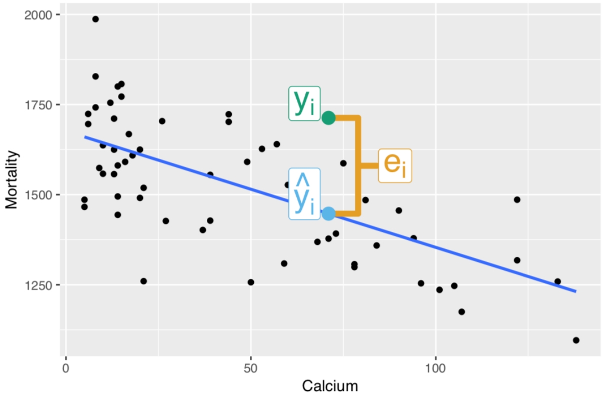


Being as specific and concrete as possible, write down a rule for selecting your preferred model

1. based only on **visual characteristics** of the plot.
2. based only on **a quantitative summary** of the data. You can describe how you would calculate your numeric summary of the data in a general sense; if you'd like you can write down a formula.

Residuals:

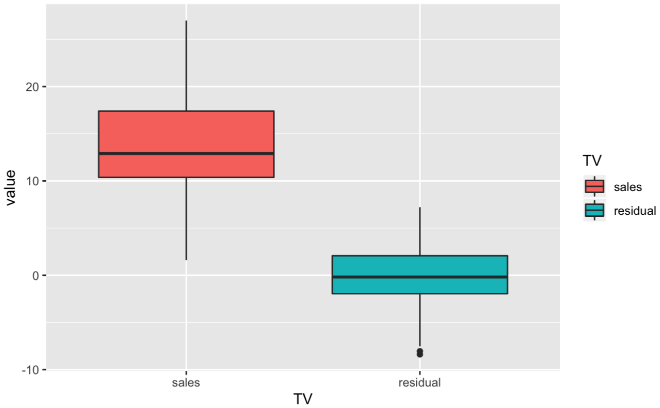
- $e_i = y_i - \hat{y}_i$ (vertical distance between point and line)
- Smaller residuals mean the predictions were better.
- The key is to measure the spread of residuals.



Measure spread of residuals with the standard deviation. We call this the **residual standard error**, s_{RES}.

- Exam 1: 4.28
- Exam 2: 3.26

The variability in the residuals describes how much variation remains after using the model



Let's compute the reduction in variation.

$$\frac{s^2_{\text{sales}} - s^2_{\text{RES}}}{s^2_{\text{sales}}} = 0.61$$

This number describes the amount of variation in the y-variable that is explained by the least squares line.

An value of 61% indicates that 61% of the variation in final exam grades can be accounted for by Exam 1 grades.

Variation accounted by the model

- Exam 1: 0.61
- Exam 2: 0.73

meaning,

- 61% of the variation in final exam grades can be accounted for by Exam 1 grades;
- 73% of the variation in final exam grades can be accounted for by Exam 2 grades

Statisticians found the variation accounted by the model can be computed by R^2, the **square of correlation**.

Square of the correlation coefficient R: between 0 and 1, closer to 1 is better.

R^2 describes the amount of variation in the y-variable that is explained by the least squares line.

Compute R^2 from R

```r
linear_fit <- lm(Mortality ~ Calcium, data = mortality_water)
summary(linear_fit)
```

```
##
## Call:
## lm(formula = Mortality ~ Calcium, data = mortality_water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -348.61 -114.52    -7.09   111.52   336.45
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1676.3556    29.2981  57.217  < 2e-16 ***
## Calcium       -3.2261     0.4847  -6.656 1.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143 on 59 degrees of freedom
## Multiple R-squared:  0.4288, Adjusted R-squared:  0.4191
## F-statistic:  44.3 on 1 and 59 DF,  p-value: 1.033e-08
```

b0 intercept

b1 slope

Useful later

R squared