DATA 8
Spring 2020

Lecture 33

Regression Inference

Regression roadmap

- Last Monday:
  - Association and correlation
- Last Wednesday
  - Prediction, scatterplots and lines
- Last Friday:
  - Least squares: finding the “best” line for a dataset
- Monday:
  - Residuals: analyzing mistakes and errors
- **Today**
  - **Regression inference: understanding uncertainty**

Residuals

Review: Residuals

- Error in regression estimate
- One residual corresponding to each point (*x*, *y*)
- **residual**
  = observed *y* - regression estimate of *y*
- In other words:
  - **observed y = regression estimate + residual**

Review: Residual Plots

A scatter diagram of residuals

- Should look like an unassociated blob for linear relations
- But will show patterns for non-linear relations
- Used to check whether linear regression is appropriate
- Look for curves, trends, changes in spread, outliers, or any other patterns

Properties of residuals

- Residuals from a linear regression **always** have
  - **Zero** mean
    - (so **rmse = SD of residuals**)
  - **Zero** correlation with x
  - **Zero** correlation with the fitted values

- These are all true **no matter what the data look like**
  - Just like deviations from mean are zero on average

(Demo)

Discussion Questions

How would we adjust our regression line…

- if the average residual were 10?
- if the residuals were positively correlated with x?
- if the residuals were above 0 in the middle and below 0 on the left and right?

A Measure of Clustering

Correlation, Revisited

- Last week, we said "The correlation coefficient measures how clustered the points are around a straight line."

- We can now quantify this statement.

(Demo)

SD of Fitted Values

- $$\frac{\text{SD of fitted values}}{\text{SD of } y} = |r|$$

- SD of fitted values $= |r| * (\text{SD of } y)$

Variance of Fitted Values

- Variance = Square of the SD
  = Mean Square of the Deviations
- Variance has weird units, but good math properties

- $$\frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2$$

A Variance Decomposition

By definition,

$$y = \text{fitted values} + \text{residuals}$$

Tempting (but wrong) to think that:

$$\cancel{\mathrm{SD}(y) = \mathrm{SD}(\text{fitted values}) + \mathrm{SD}(\text{residuals})}$$

But it **is** true that:

$$\mathrm{Var}(y) = \mathrm{Var}(\text{fitted values}) + \mathrm{Var}(\text{residuals})$$

(a result of the **Pythagorean theorem!**)

A Variance Decomposition

Var(y) = Var(fitted values) + Var(residuals)

- $$\frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2$$

- $$\frac{\text{Variance of residuals}}{\text{Variance of } y} = 1 - r^2$$

A Variance Decomposition

Var(y) = Var(fitted values) + Var(residuals)

- $$\frac{\text{SD of fitted values}}{\text{SD of } y} = |r|$$

- $$\frac{\text{SD of residuals}}{\text{SD of } y} = \sqrt{1 - r^2}$$

Residual Average and SD

- The average of residuals is always 0
- SD of residuals = $\sqrt{(1 - r^2)}$ * SD of y
- SD of predictions = $|r|$ * SD of y

(Demo)

Discussion Question

Midterm: Average 70, SD 10

Final: Average 60, SD 15

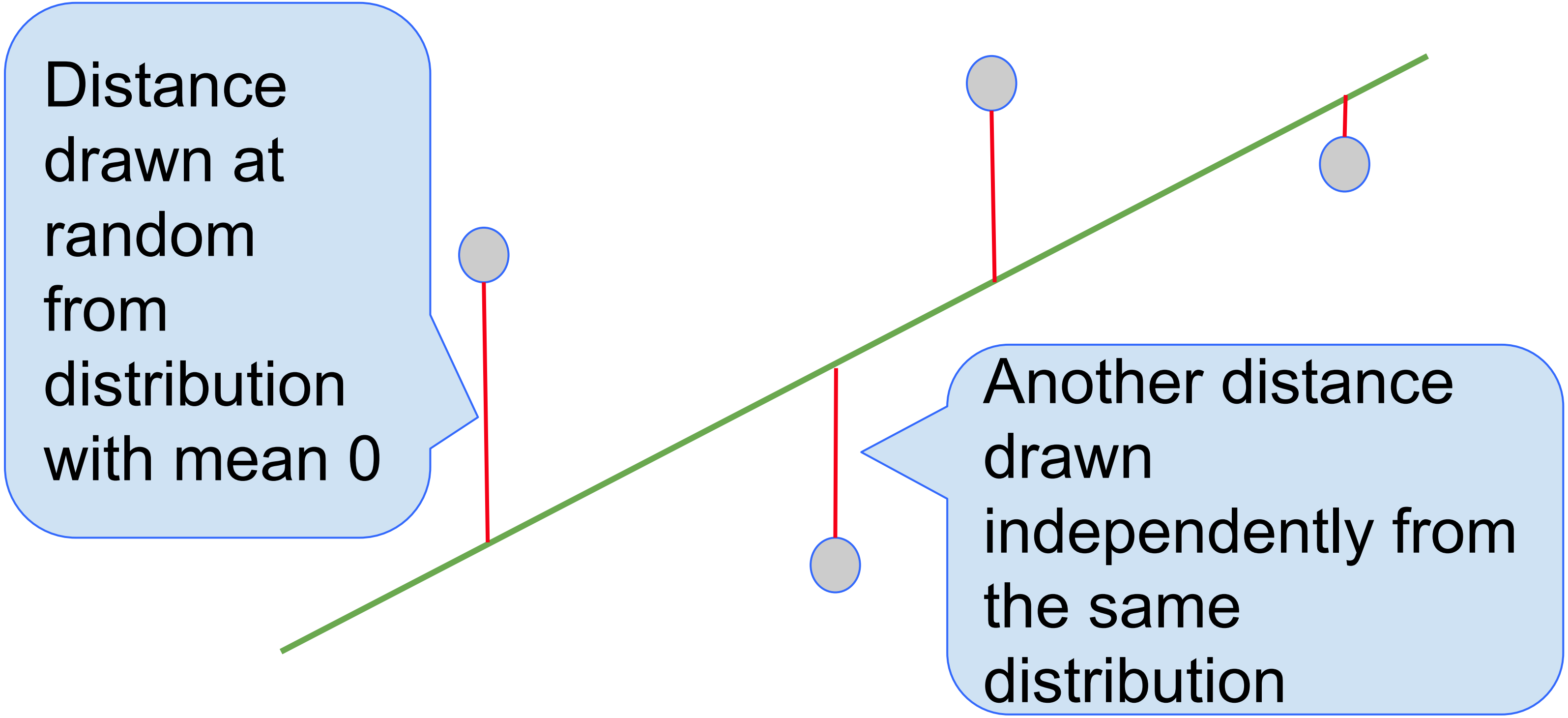
$r = 0.6$

Fill in the blank:

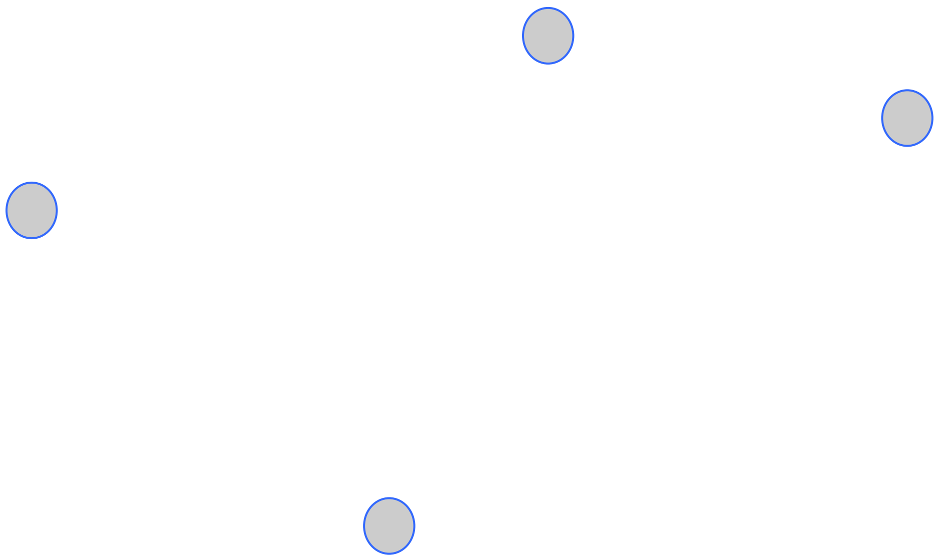
For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within ____________ points.

Regression Model

A “Model”: Signal + Noise



What We Get to See



(Demo)

Prediction Variability

Regression Prediction

- **If the data come from the regression model,**
- **and if the sample is large, then:**

- The regression line is close to the true line
- Given a new value of *x*, predict *y* by finding the point on the regression line at that *x*

(Demo)

Confidence Interval for Prediction

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the height of the true line at *y*.

(Demo)

Predictions at Different Values of *x*

- Since *y* is correlated with *x*, the predicted values of *y* depend on the value of *x*.

- The width of the prediction’s CI also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.

(Demo)

The True Slope

Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat.
- Draw the empirical histogram of all the generated slopes.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the slope of the true line.

(Demo)

Rain on the Regression Parade

We observed a slope based on our sample of points.

But what if the sample scatter plot got its slope just by chance?

What if the true line is actually FLAT?

Test Whether There Really is a Slope

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval doesn't contain 0, the data are more consistent with the alternative
  - If the interval does contain 0, the data are more consistent with the null

(Demo)