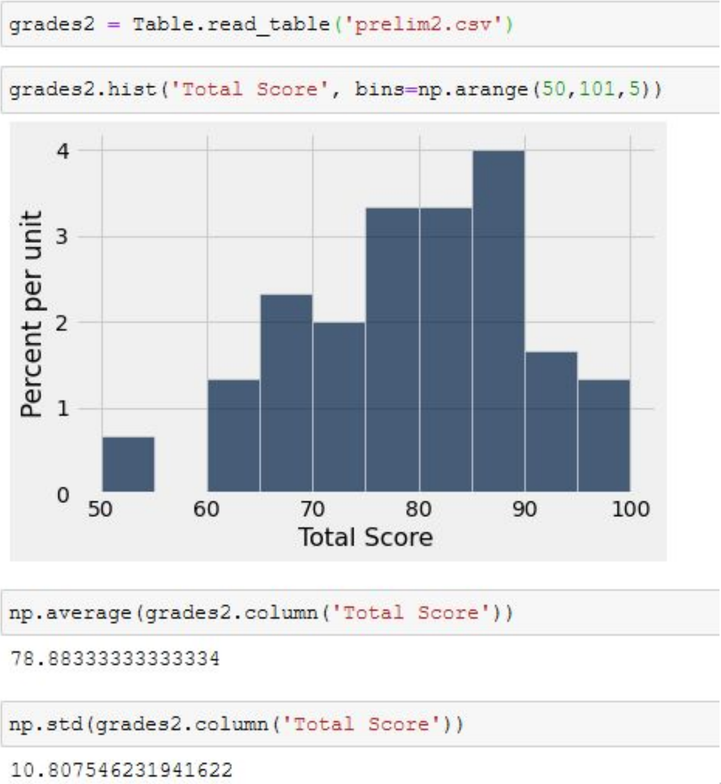
DSFA

Spring 2021

Lecture 31

Regression Inference

Announcements



```
grades2 = Table.read_table('prelim2.csv')
```

```
grades2.hist('Total Score', bins=np.arange(50,101,5))
```

```
np.average(grades2.column('Total Score'))
```

```
78.88333333333334
```

```
np.std(grades2.column('Total Score'))
```

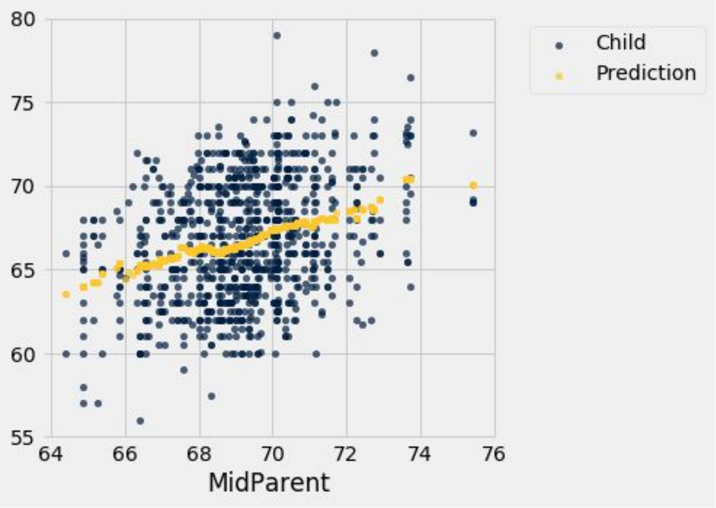
```
10.807546231941622
```

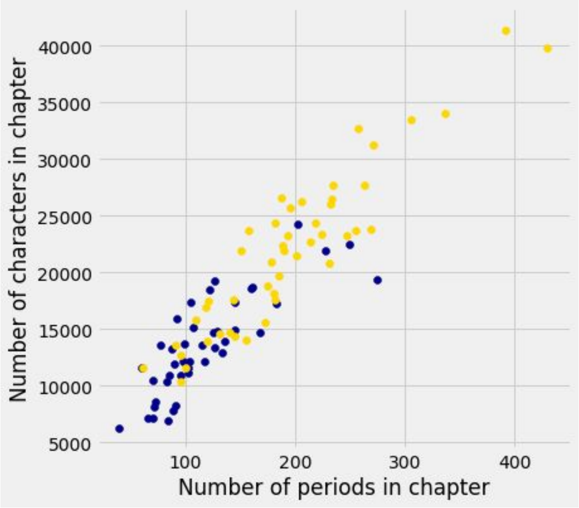
Announcements

- Prelim 2 regrade requests by Monday 5/3, 5PM
- Lab 9 today/Thursday
- HW 5 due this Friday at 5:59PM, usual 1 point bonus for turning in by Thursday midnight.
- Project 3 out Friday, Part 1 due 5/7, Part 2 due 5/14.
- Labs the week of 5/3 and 5/10 will be dedicated to Project 3.
- Final May 22, 1:30PM

Prediction

If we have a line describing the relation between two variables, we can make predictions





Regression Line Equation

In original units, the regression line has this equation:

$$\underbrace{\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y}}_{\text{y in standard units}} = r \times \underbrace{\frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}}_{\text{x in standard units}}$$

$$y = \text{slope} \times x + \text{intercept}$$

$$\textbf{slope of the regression line} \ = \ r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} \ = \ \text{average of } y \ - \ \text{slope} \cdot \text{average of } x$$

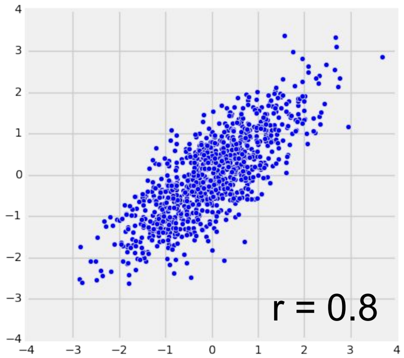
Errors and Predictions

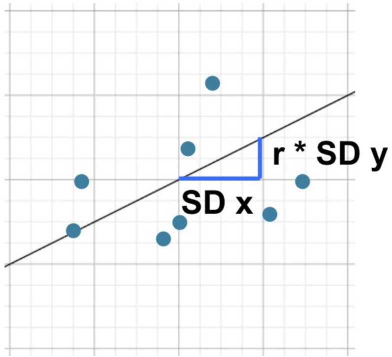
- **error = actual value − prediction**
- RMSE = root mean square error
- Regression line has the minimum RMSE of all lines

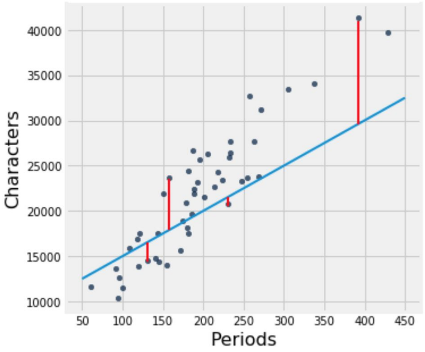
- Names:
  - Regression line
  - Least squares line
  - “Best fit” line

Summary: What we can learn from *r*

- How clustered points are around a line
- How *y* depends on *x*
- How accurate linear regression predictions will be







Prediction from a Sample

Prediction from a Sample

- We’ve been treating dataset as though it were population
- What if we had to make predictions from samples?

(Demo)

Confidence Interval for Prediction

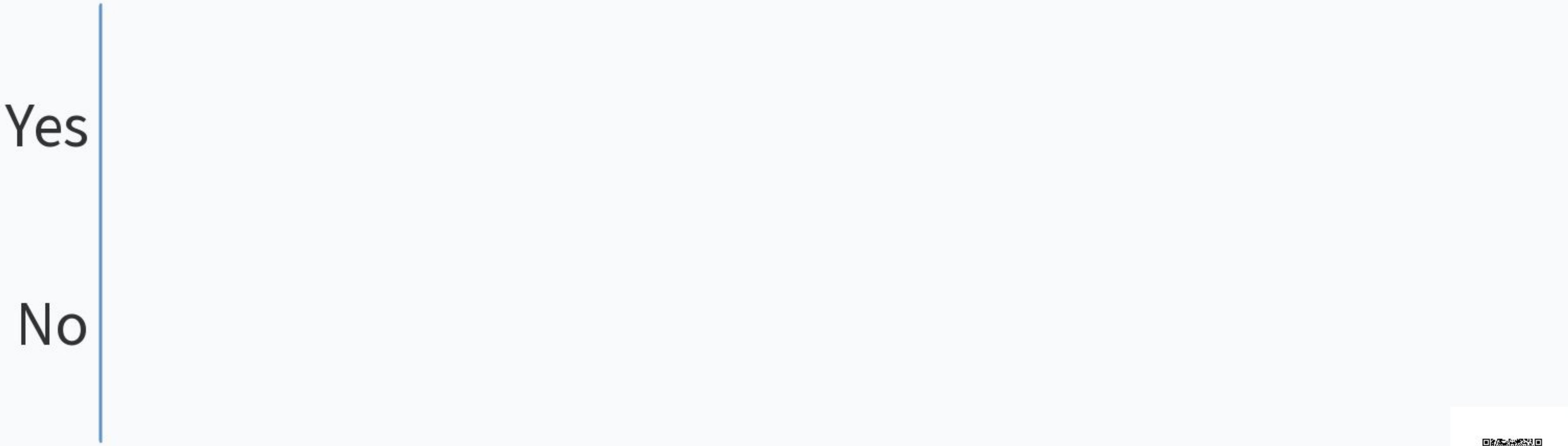
- **Bootstrap:**
  - **Resample the data**
  - **Get a prediction for *y* using the regression line that goes through the resampled data**
  - **Repeat the above two steps, many times**
- Draw the empirical histogram of all the predictions
- Get the "middle 95%" interval
- That's an approximate 95% confidence interval for the predicted value of *y*

(Demo x 2)

When poll is active, respond at **pollev.com/dsfa**

Text **DSFA** to **22333** once to join

Is there a 95% chance that the birth weight of a baby born at 288 gestational days is about 122-125?



Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Test Whether Variables are Correlated

- **Null hypothesis:** The correlation is 0
- **Alternative hypothesis:** It’s not
- **Method:**
  - Construct a 95% confidence interval for the correlation using the bootstrap
  - Check if 0 is in the interval

(Demo)