CompSci 116: Lecture 10: Prediction - Regression

Jeff Forbes

March 28, 2019

Project

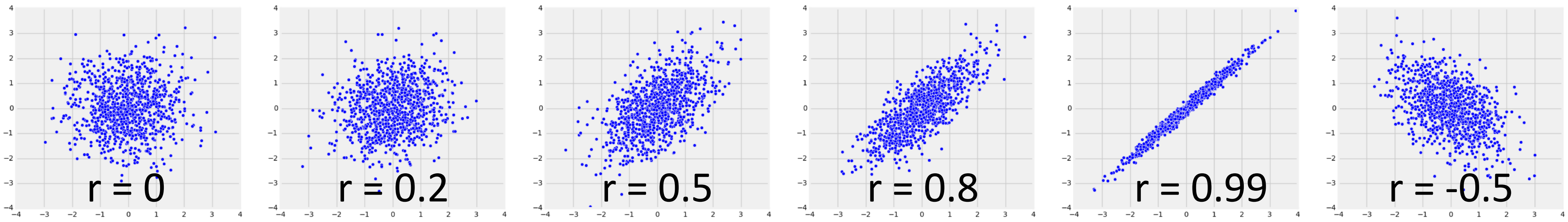
- Exploration of question(s) of interest to you
- Today
  - Who is on your project team?
  - What generally do you hope to address?

- By April 4
  1. Data source
  2. What question will you address?
  3. How will you address those questions?

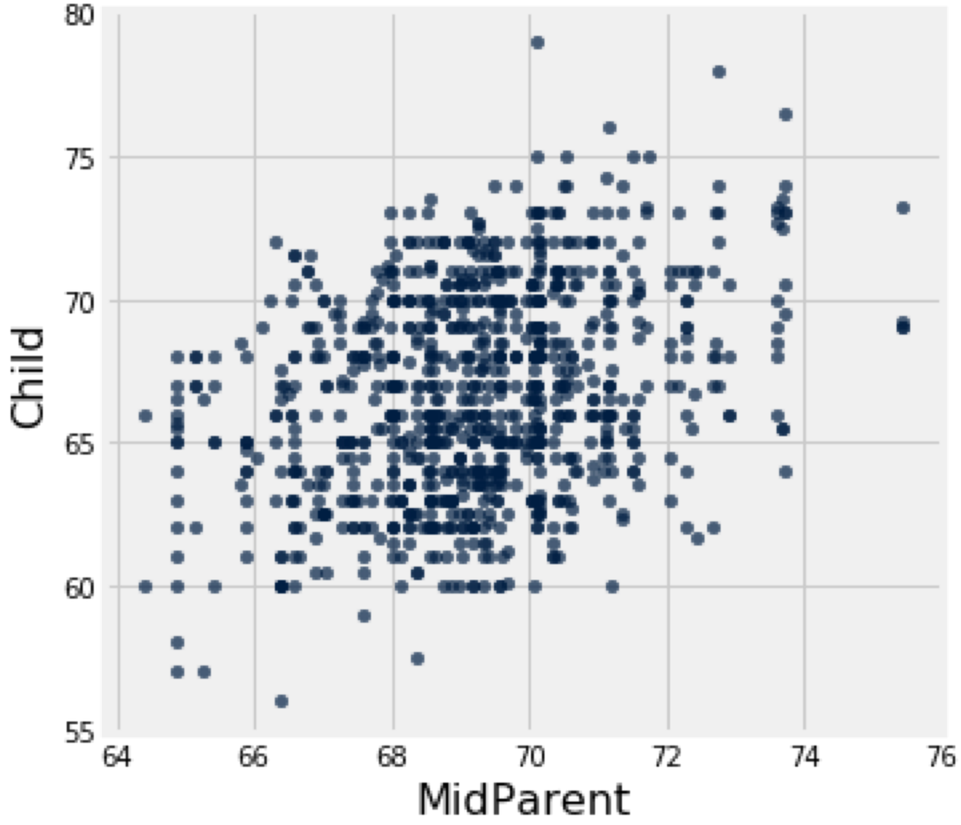
The Correlation Coefficient *r*

Correlation Coefficient (r) $= \text{average}(x_{su} * y_{su})$

```python
def standard_units(any_numbers):
    """Convert any array of numbers to standard units."""
    return

def correlation(t, x, y):
    """Return the correlation coefficient (r) of two variables."""
    return
```

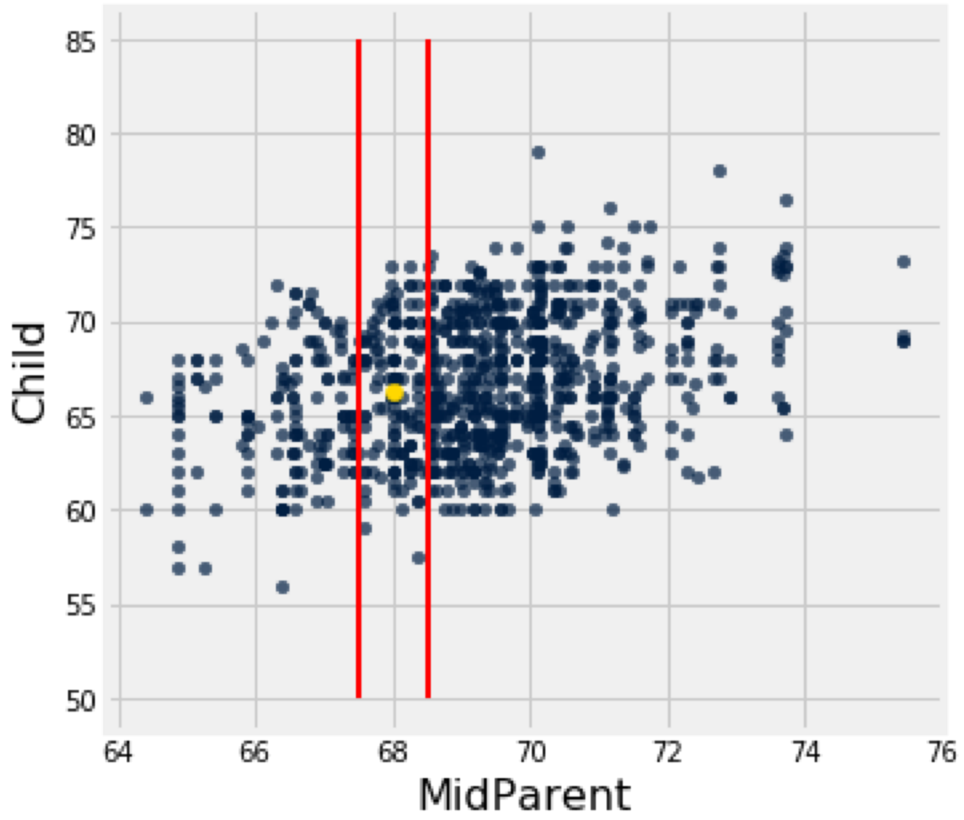


Galton's Heights



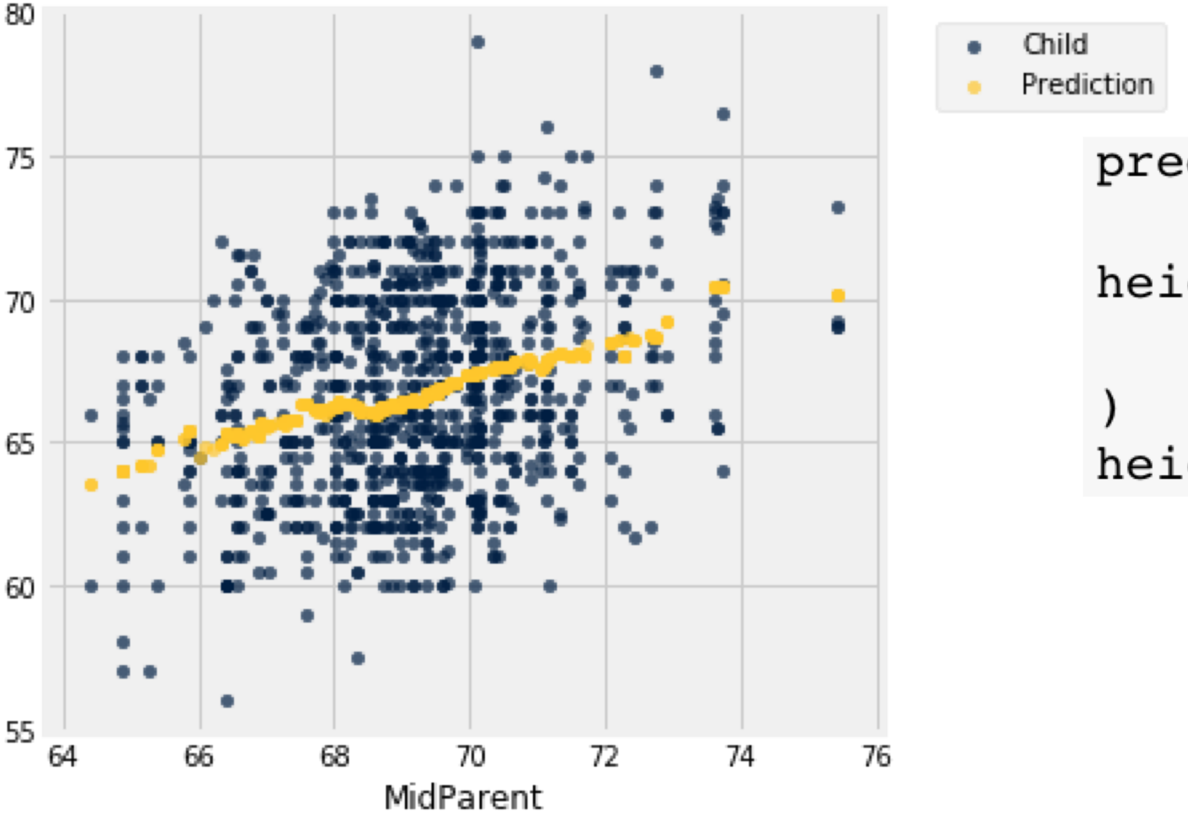

```
galton = Table.read_table('galton.csv')

heights = Table().with_column(
    'MidParent', galton.column('midparentHeight'),
    'Child', galton.column('childHeight')
    )
heights.scatter('MidParent')
```

Galton's Heights



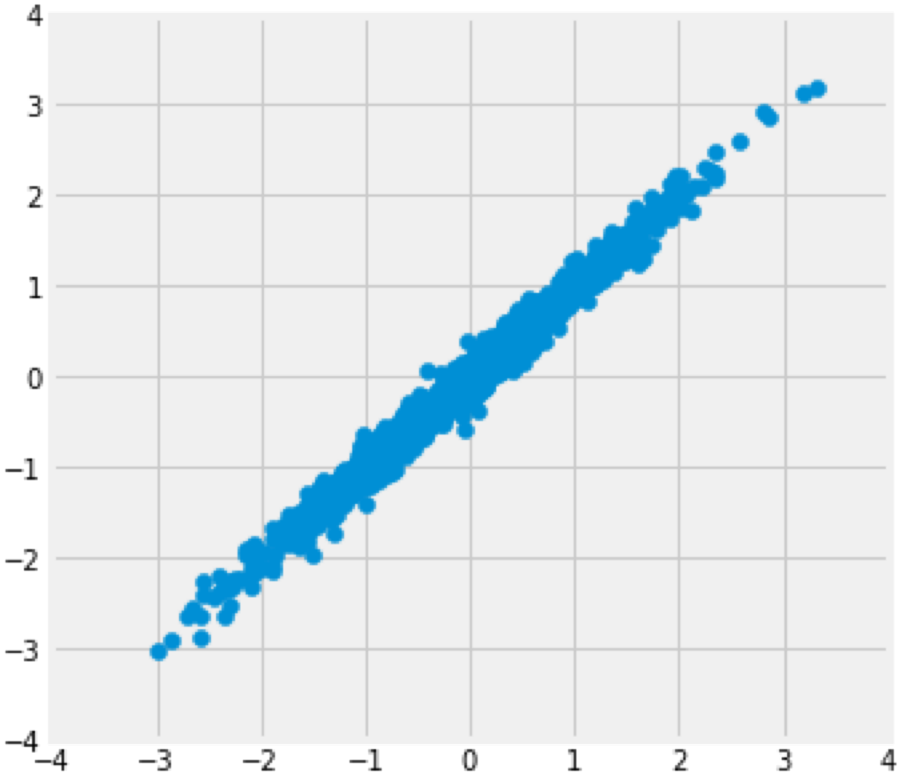

```python
def predict_child(x):
    chosen = heights.where('MidParent',
                           are.between(x - 0.5,
                                       x + 0.5))
    return np.average(chosen.column('Child'))

yp = predict_child(68)
heights.scatter('MidParent')
plots.scatter(68, yp, s=30, color='yellow')
```

Galton's Heights



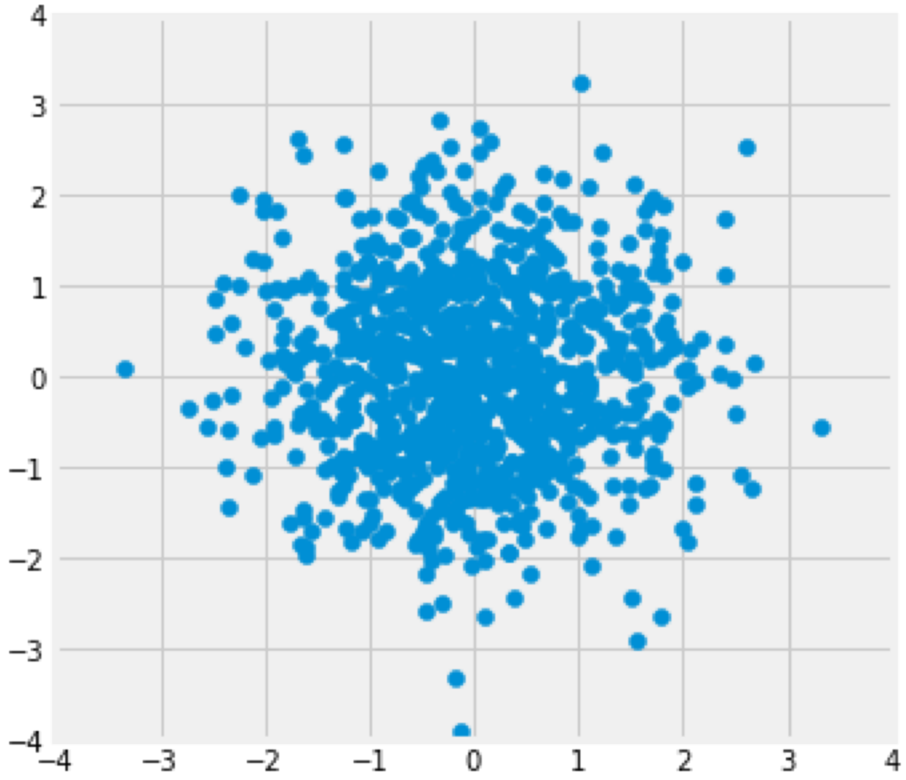

```
predictions = heights.apply(predict_child,
                            'MidParent')
heights = heights.with_column(
    'Prediction', predictions
)
heights.scatter('MidParent')
```

Where is the prediction line?



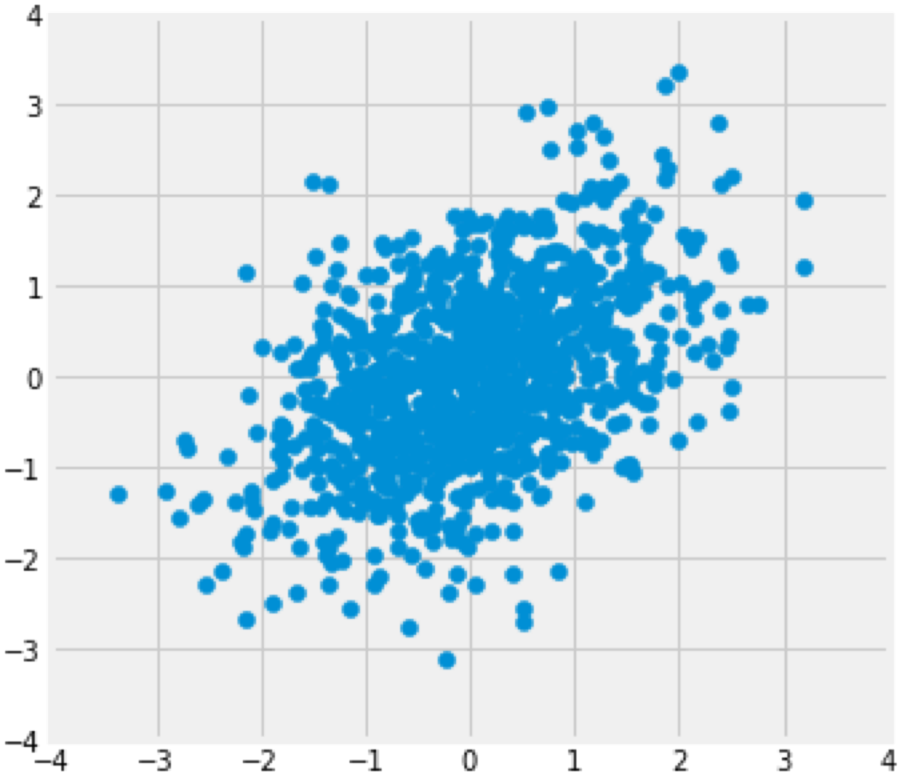
r = 0.99

Where is the prediction line?



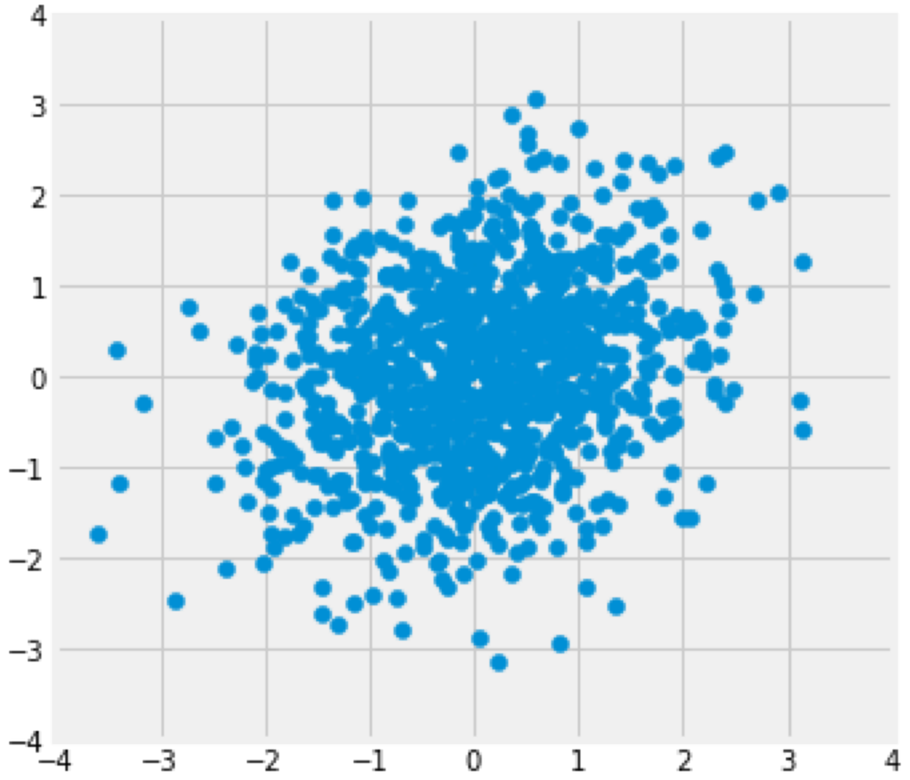
r = 0.0

Where is the prediction line?



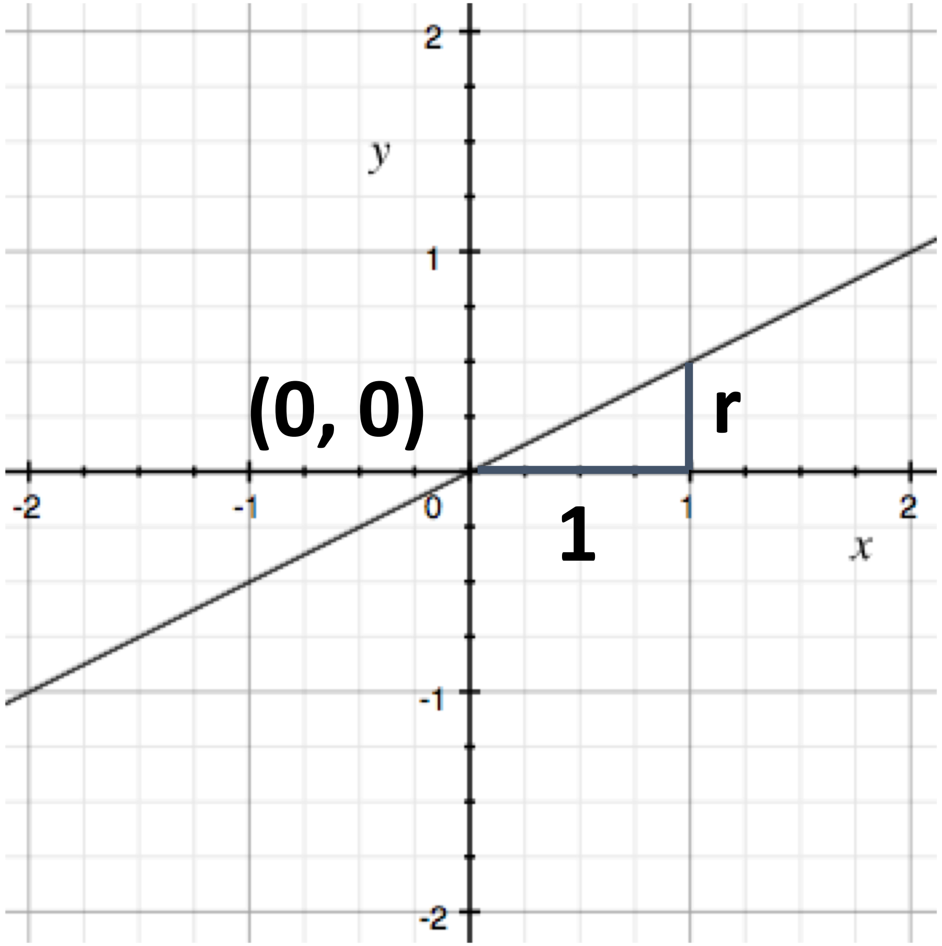
r = 0.5

Where is the prediction line?



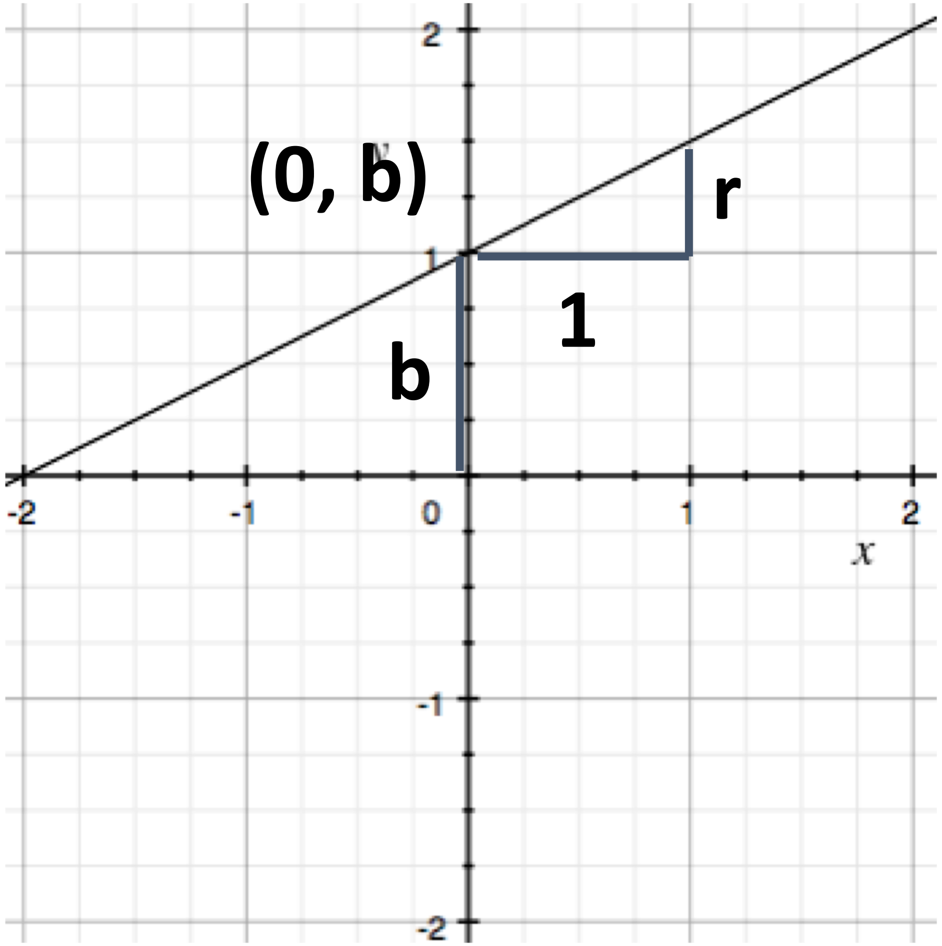
r = 0.2

Algebra! Equation of a Line



$$y = r \times x$$

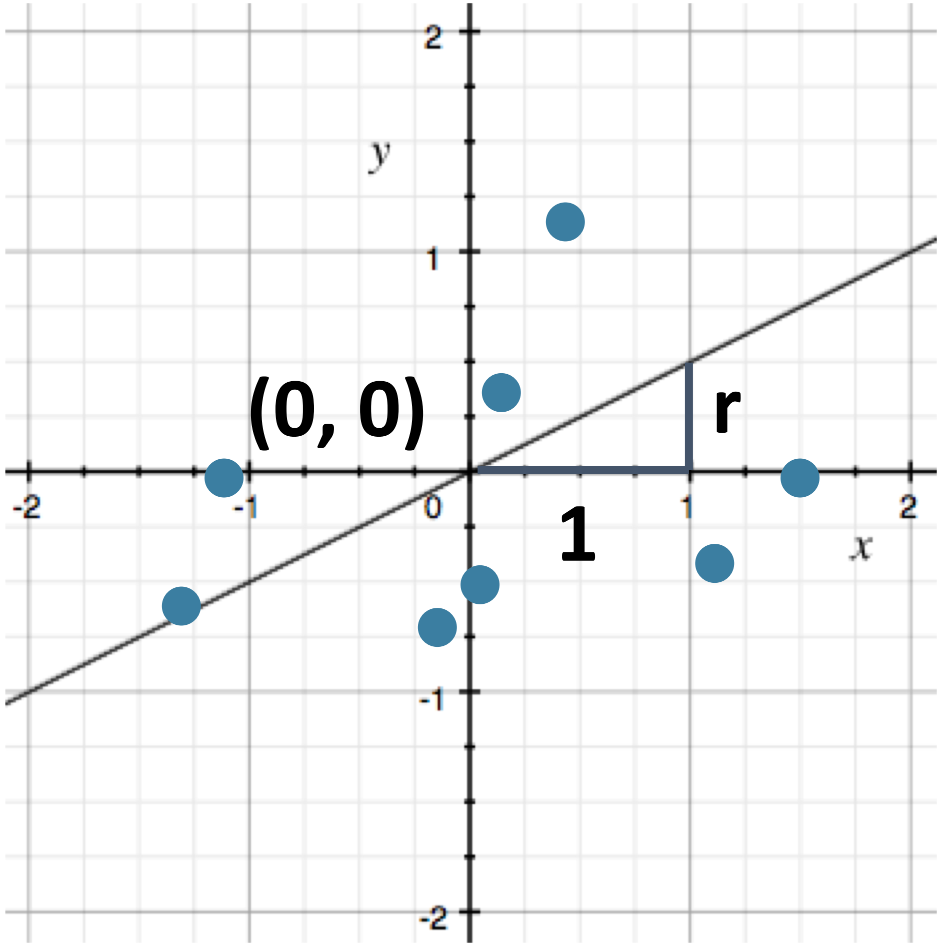
Algebra! Equation of a Line



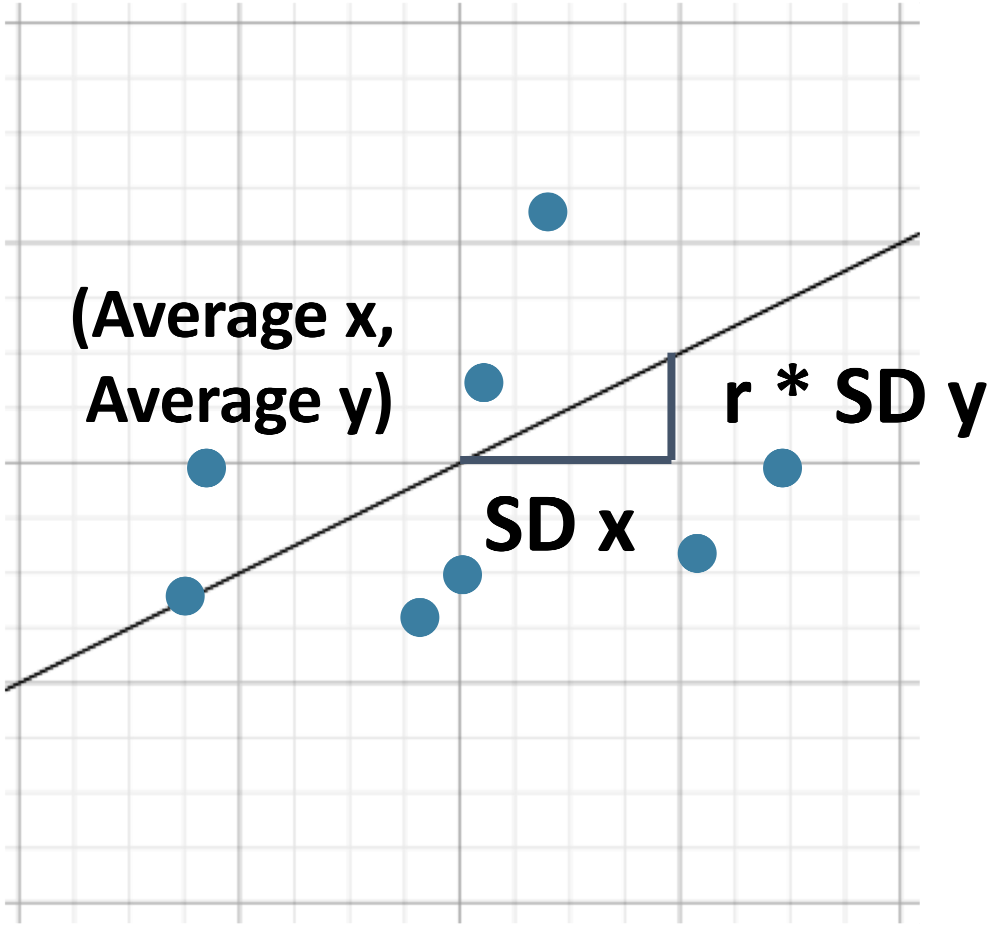
$$y = r \times x + b$$

Regression Line

Standard Units

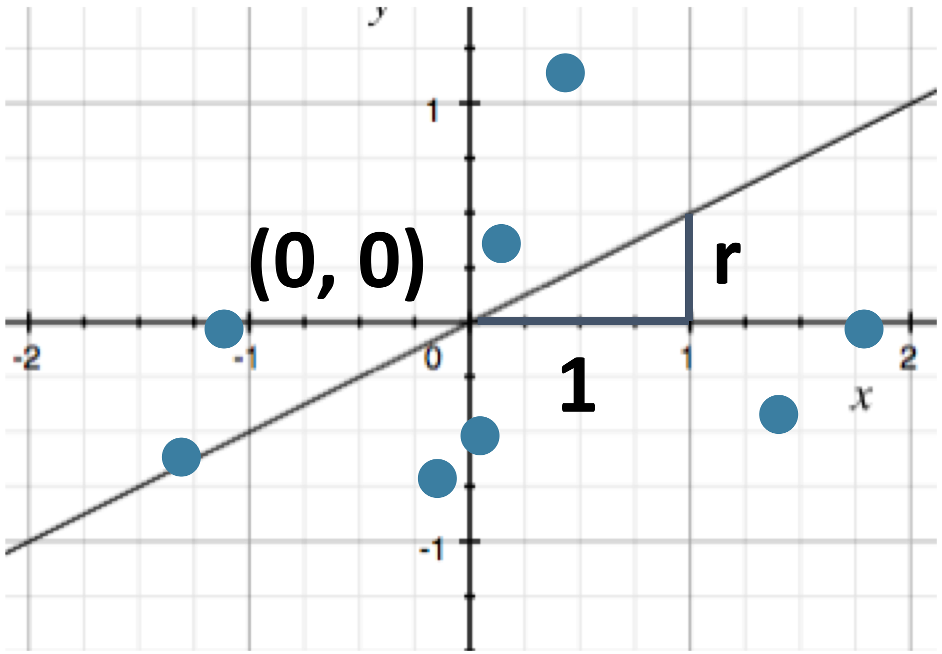


Original Units



Regression Line Equation

In standard units, the equation of the regression line is:



Fitted value

Observed value

$$y_{(\text{su})} = r \times x_{(\text{su})}$$

Correlation coefficient

Regression Line Equation

In original units, the regression line has this equation:

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

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

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

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

Discussion Question

A course has a midterm (average 70; standard deviation 10) and a really hard final (average 50; standard deviation 12)

If the scatter diagram comparing midterm & final scores for students has a typical oval shape with correlation 0.75, then...

What do you expect the average final score would be for students who scored 90 on the midterm?

How about 60 on the midterm?

http://bit.ly/FoDS-s19-0328-1

Error in Estimation

- error = actual value – estimate
- Typically, some errors are positive and some negative
- To measure the rough size of the errors
  - square the errors to eliminate cancellation
  - take the mean of the squared errors
  - take the square root to fix the units
  - Root Mean Square Error (rmse)

(Demo)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
    - “Best fit” line
    - Least squares line
    - Regression line

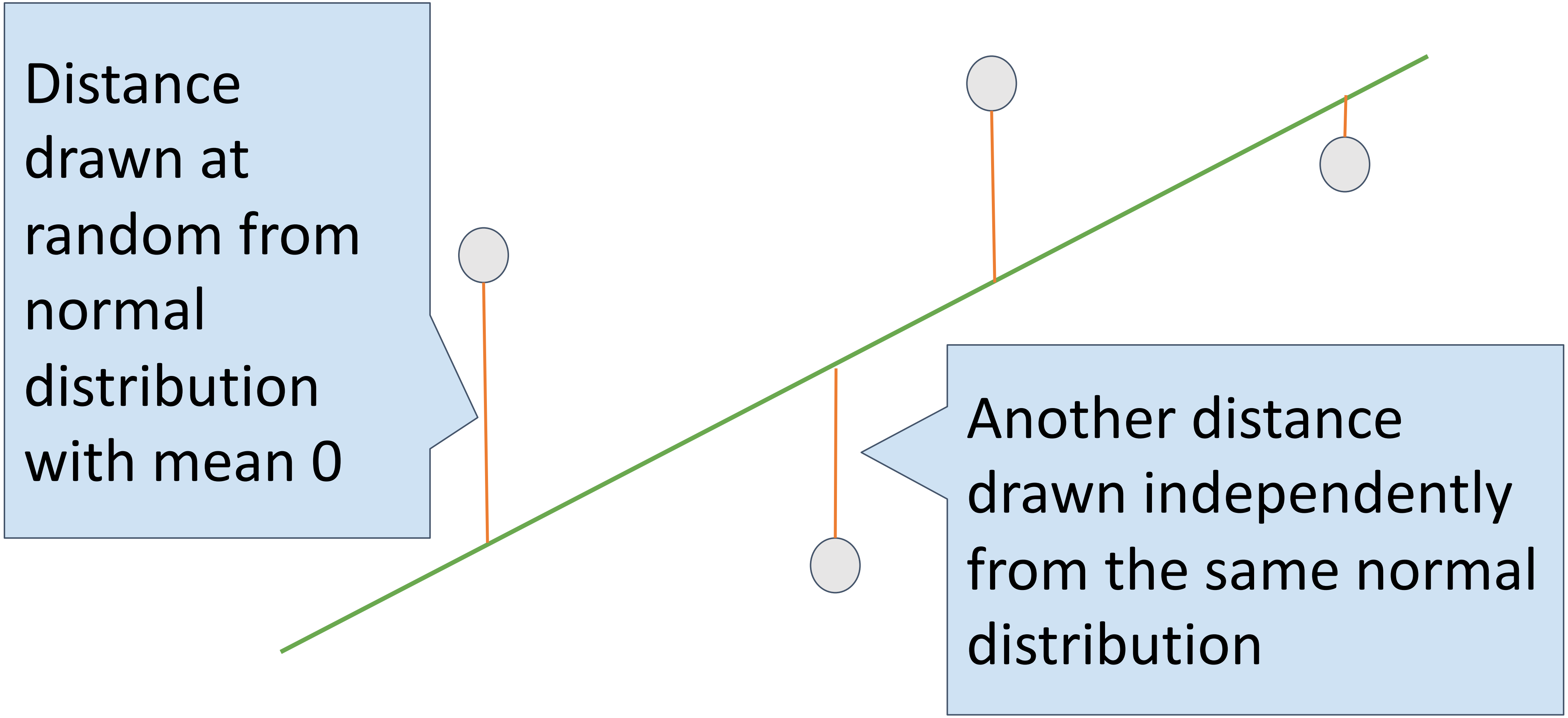
(Demo)

Numerical Optimization

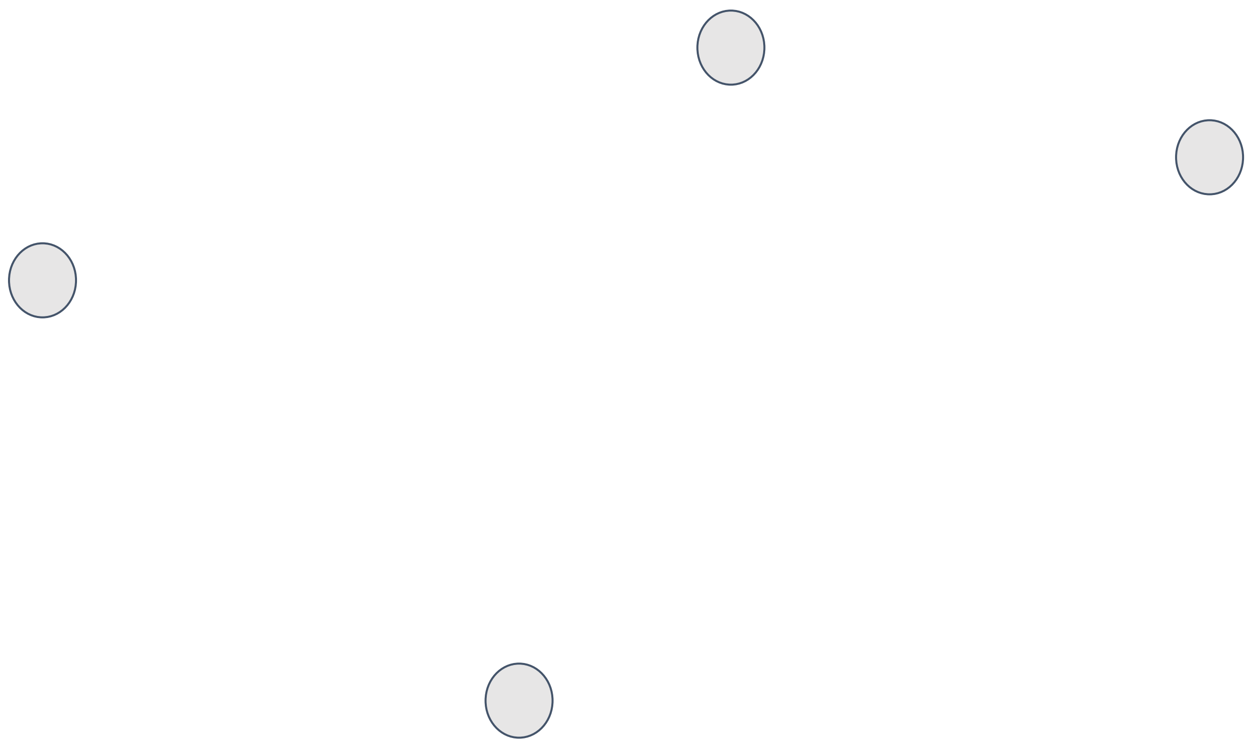
- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- If the function `mse(a, b)` returns the mse of estimation using the line "estimate = ax + b",
    - then `minimize(mse)` returns array $[a_0, b_0]$
    - a_0 is the slope and b_0 the intercept of the line that minimizes the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

(Demo)

A “Model”: Signal + Noise



What We Get to See



Regression Prediction

If the data come from the regression model,

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

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 predicted value of *y*.

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 interval also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.

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.

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, reject the null hypothesis.
  - If the interval does contain 0, there isn't enough evidence to reject the null hypothesis.