MIS2502:
Data Analytics
Linear Regression

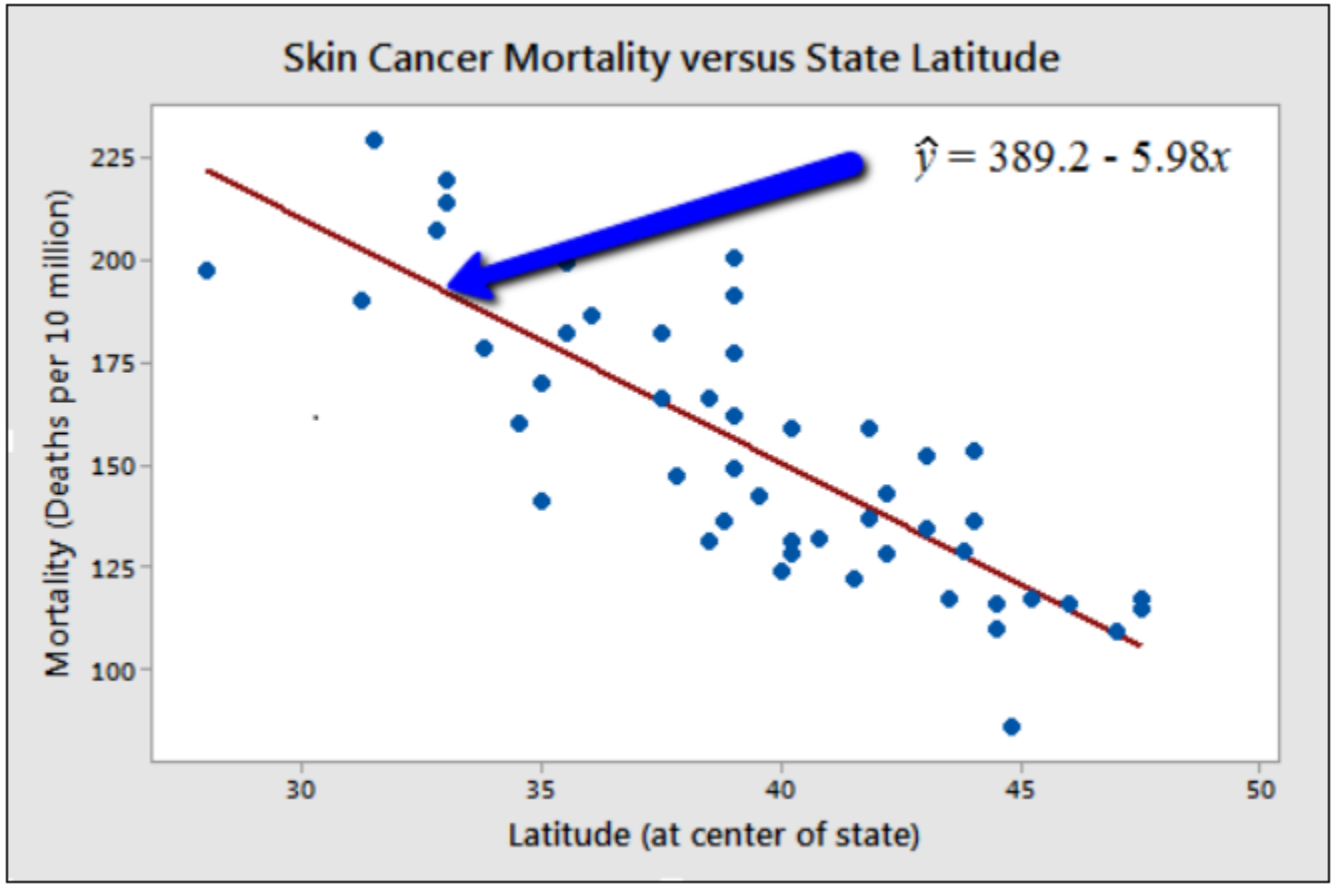
Jing Gong
gong@temple.edu
http://community.mis.temple.edu/gong

Notes adapted from http://www.math.utah.edu/~hughes/Chapter_05.pdf

Simple Linear Regression

- Objective: To summarize and study relationships between two continuous (quantitative) variables
  - One variable, denoted *x*, is regarded as the **predictor**, **explanatory**, or **independent** variable.
  - The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable.

Example

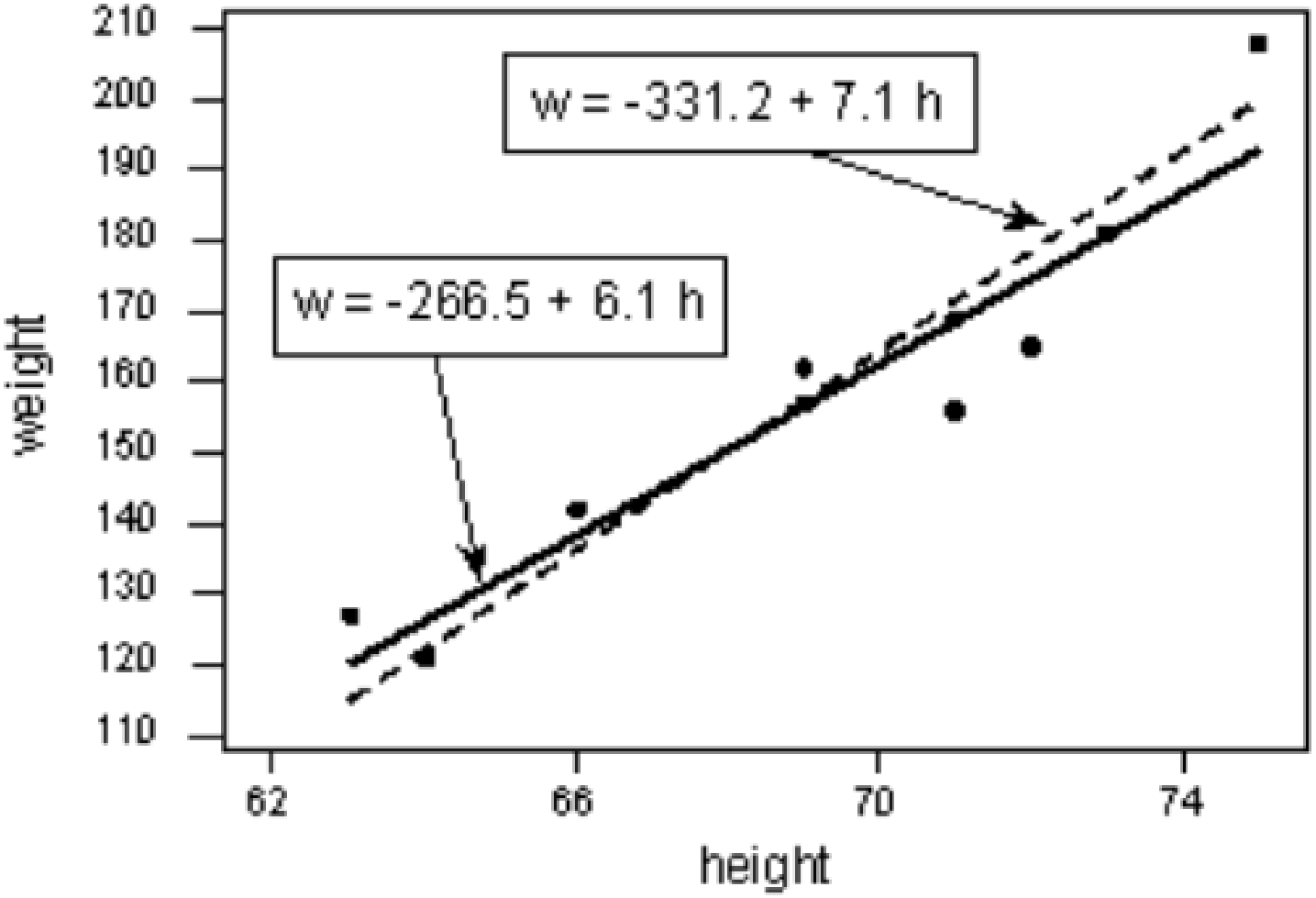


Outcome variable (y): mortality due to skin cancer

Predictor variable (x): State Latitude

https://onlinecourses.science.psu.edu/stat501/node/251

Which Line is Better?



https://onlinecourses.science.psu.edu/stat501/node/252

Regression equation

- Regression equation for the best fitting line: $\hat{y} = b_0 + b_1 x$
  - x is the predictor value
  - $\hat{y}$ is the predicted response (or fitted value)
  - b_0 is the intercept of the fitting line
  - b_1 is the slope of the fitting line

How to Determine the "best" line

- We want the line to be as close as possible to the data points

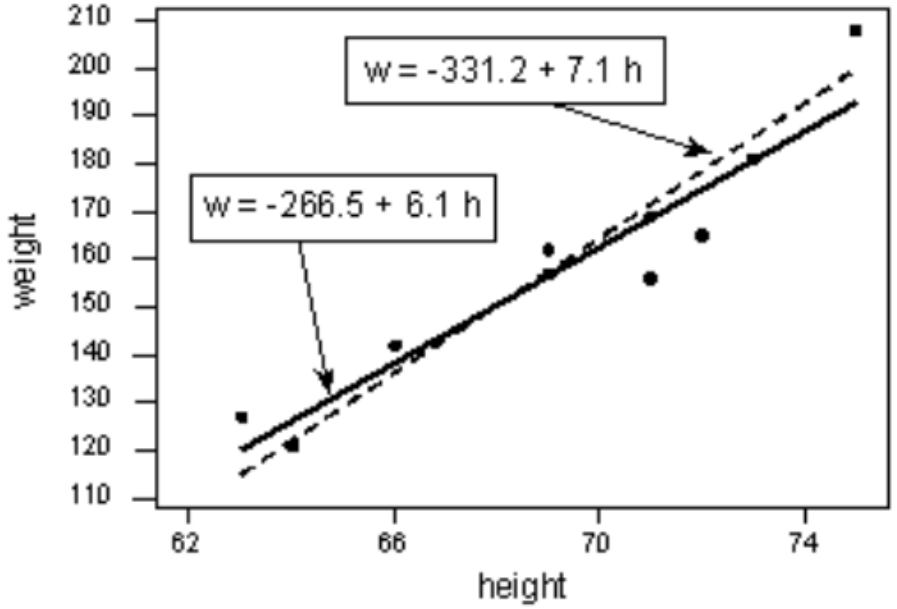
- **Least Squares**: we need to find the values b_0 and b_1 that minimize the sum of the squared prediction errors:

$$Sum\ of\ squared\ errors\ (SSE)$$

$$= \sum [y - \hat{y}]^2 = \sum [y - (b_0 + b_1 x\ )]^2$$

  - $y - \hat{y}$ is the prediction error
  - $[y - \hat{y}]^2$ is the squared prediction error
  - The symbol Σ tells us to add up the squared prediction errors for all data points

Which Line is Better?



Dash Line: SSE = 766.5
Solid Line: SSE = 597.4

Therefore, of the two lines, the solid line, $w = -266.53 + 6.1376h$, best summarizes the data.

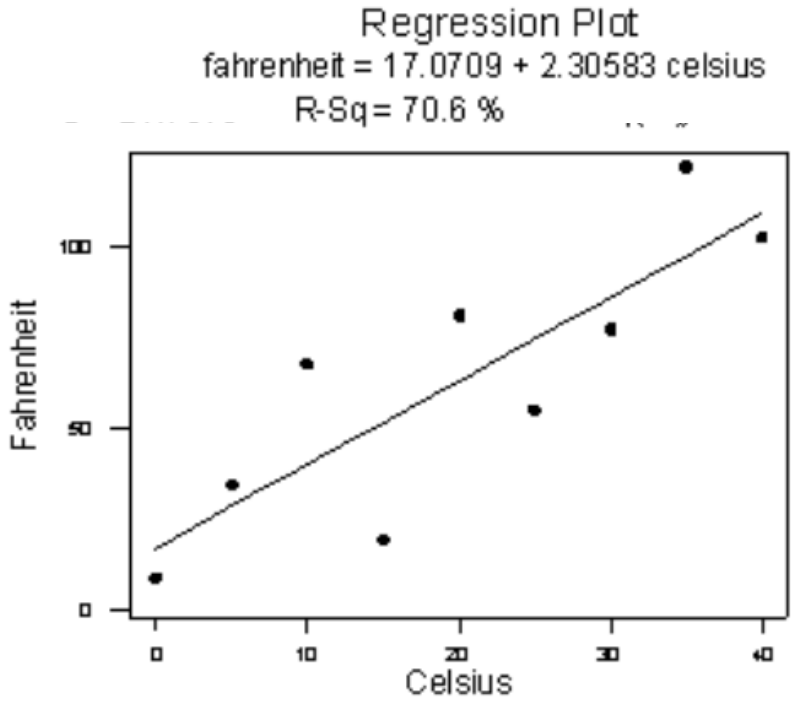
https://onlinecourses.science.psu.edu/stat501/node/252

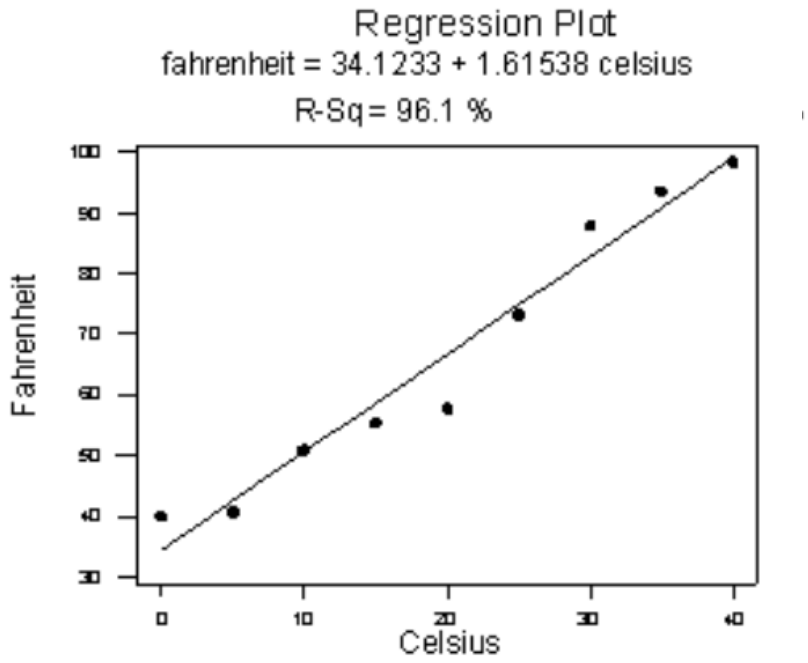
Coefficient of Determination (R^2)

- Measures usefulness of regression prediction

- R^2 measures what fraction of the variation in the values of the response variable (y) is explained by the regression line
  - R^2=1: regression line explains all (100%) of the variation in y
  - R^2=.50: regression line explains half (50%) of the variation in y

Comparing R^2





The Regression plot on the right gives higher R-sq (R^2), therefore better prediction

Extension: Multiple Linear Regression

- What if we have two or more predictor variables (x_1, x_2, …)

- The good news: everything you learned about the simple linear regression model extends

Multiple Linear Regression (Example)

- Are a person's brain size and body size predictive of his or her intelligence?
  - Outcome (y): Performance IQ scores (**PIQ**) as measure of the individual's intelligence.
  - Predictor (x_1): **Brain size** based on the count obtained from MRI scans
  - Predictor (x_2): **Height**.
  - Predictor (x_3): **Weight**.

Fitting Multiple Linear Regression

- Regression equation:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

- Again, we need to find the values b_0, b_1, b_2, b_3 that minimize the sum of the squared prediction errors:

$$Sum\ of\ squared\ errors\ (SSE)$$

$$= \sum [y - \hat{y}]^2$$

$$= \sum [y - (b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3)]^2$$

Fitting and Interpreting Linear Regression Models in R

- R makes it easy to fit a linear regression to your data.

- The centerpiece for linear regression in R is the lm( ) function.