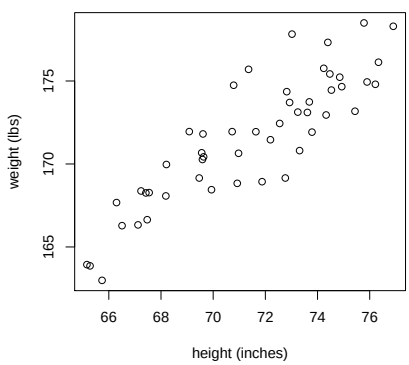
Lecture 24: Linear Regression Part I

Chapter 7.1-7.2

Questions for Today

Say we have the height/weight of 50 individuals and we display the scatterplot/bivariate plot of the seemingly linear relationship:



Questions:

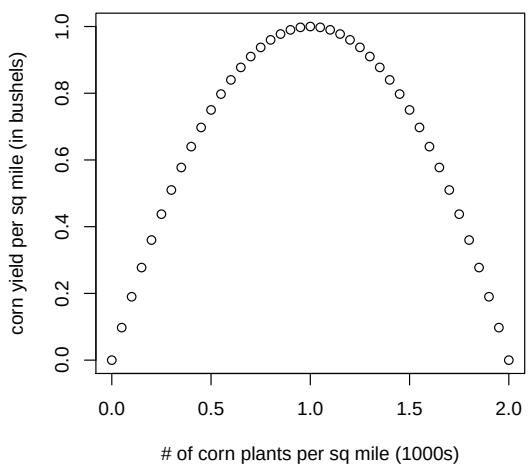
- What is the "best" fitting line through these points?
- What do we mean by "best"?

Regression

There are many types of regression, all in order to estimate the relationship between variables.

Example of Non-Linear Relationship

At first as you plant more corn plants, you have higher yield, but past a certain point plants fight for limited resources and they die.

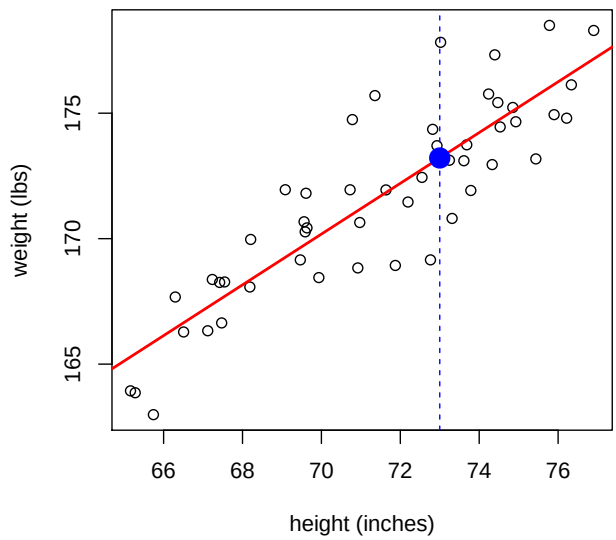


Modeling x and y Linearly

Framework

Fitted Value

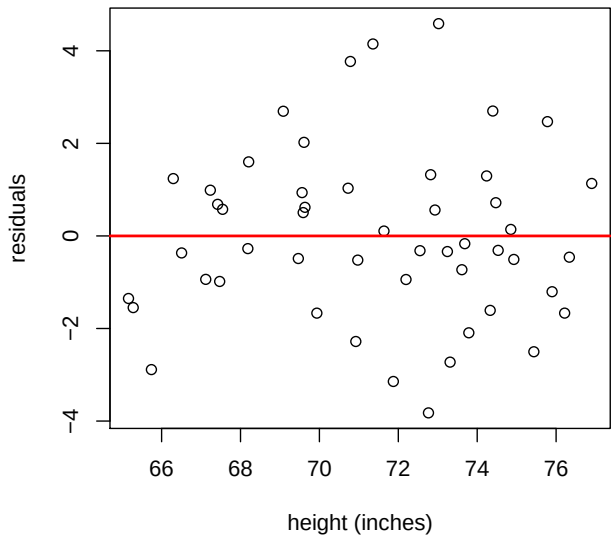
Here $\widehat{y} = 100 + 0.99x$. Thus for $x = 73$, $\widehat{y} = 173.22$:



Residuals

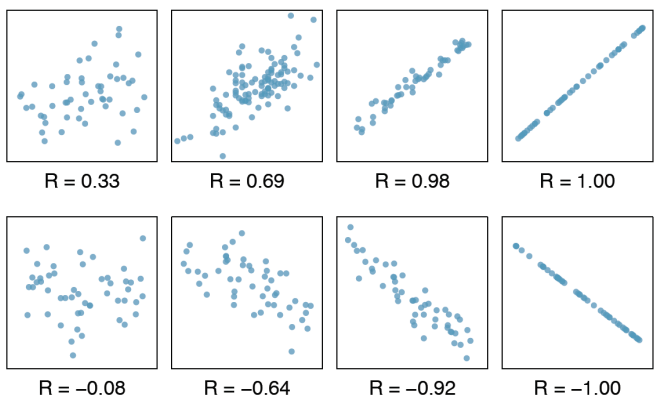
Residual Plot

Residual plots: take previous plot and flatten the red line by subtracting $\widehat{y}$ from y.



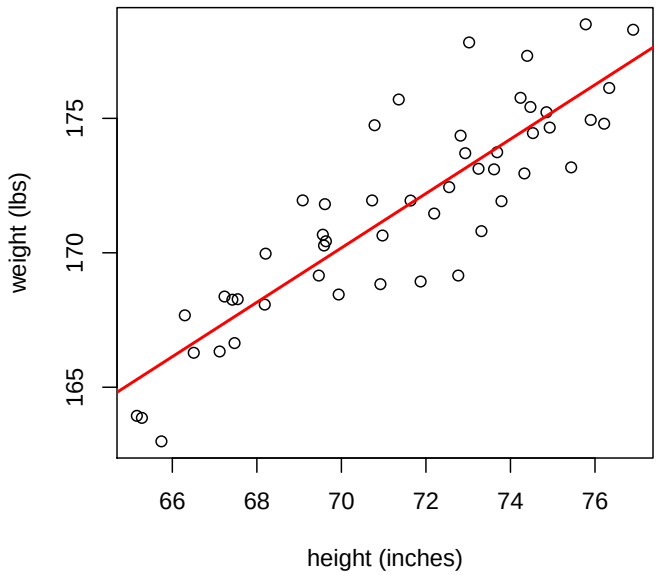
Correlation Coefficient

The correlation coefficient R is a value between $[-1, 1]$ that measures the strength of the linear relationship between x and y.



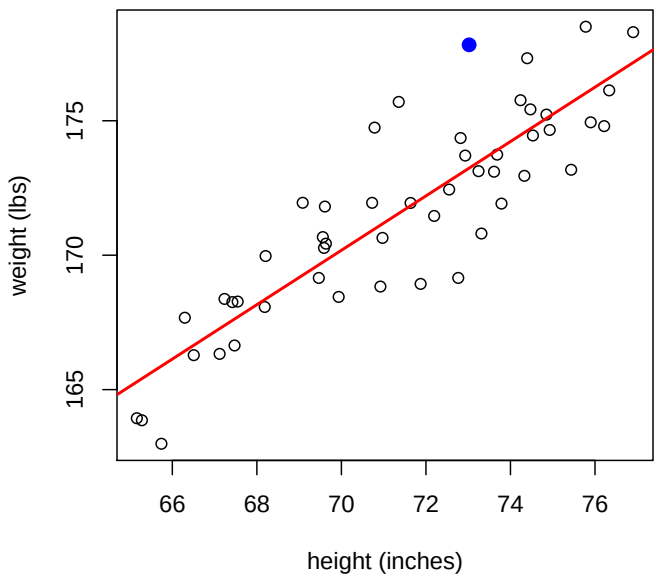
Best Fitting Line

What does "best fitting line" mean?



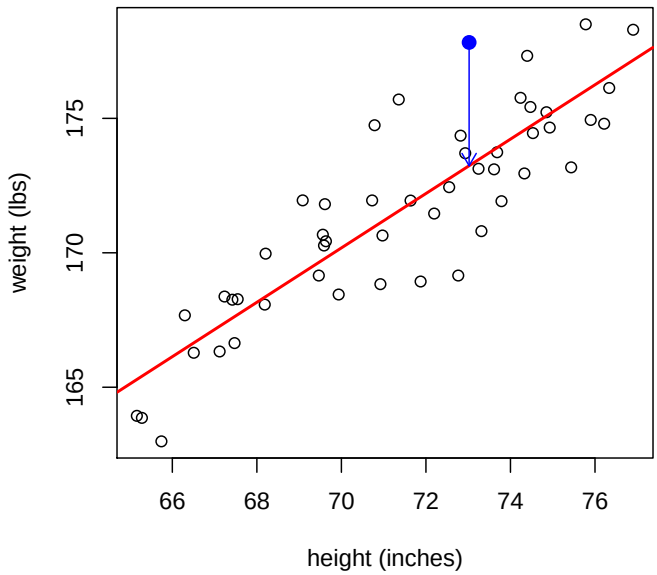
Best Fitting Line

Consider ANY point x_i for $i = 1, \ldots, 50$ (in blue).



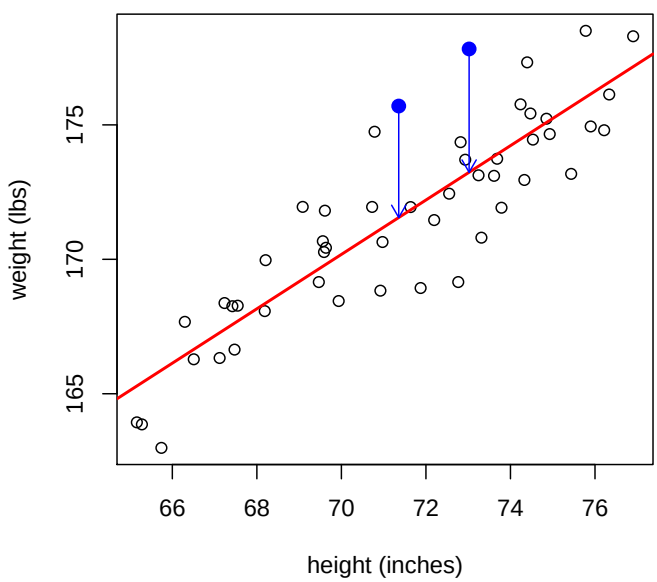
Best Fitting Line

Now consider this point's deviation from the regression line



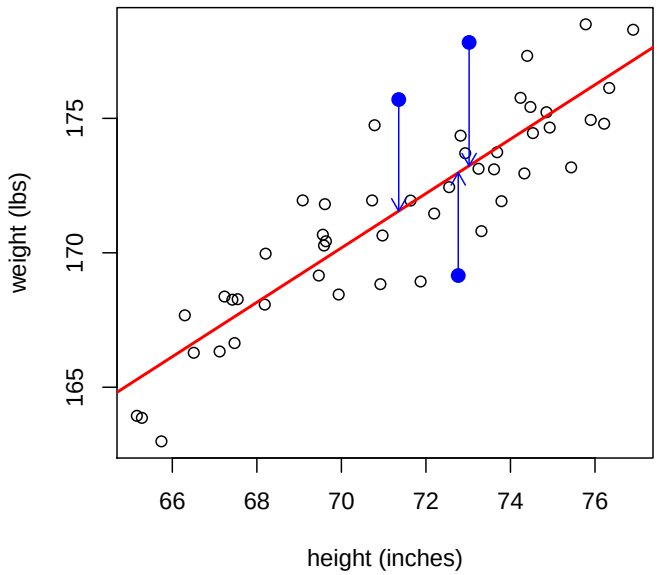
Best Fitting Line

Do this for another point x_i...



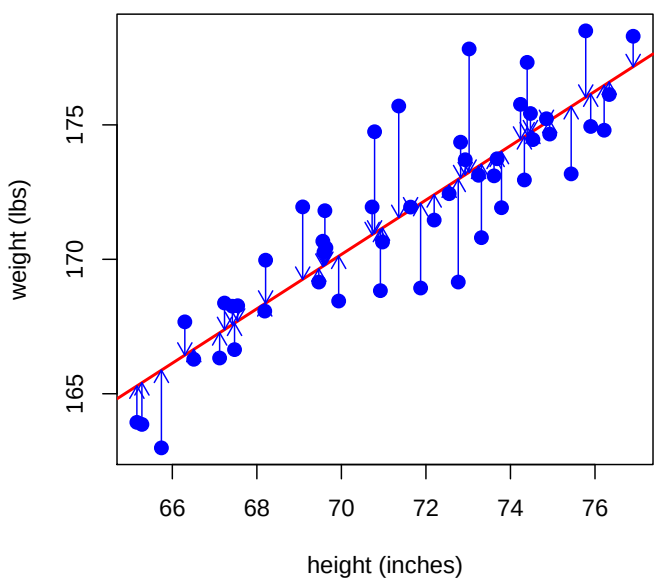
Best Fitting Line

Do this for another point x_i...



Best Fitting Line

The regression line minimizes the sum of the squared arrow lengths.

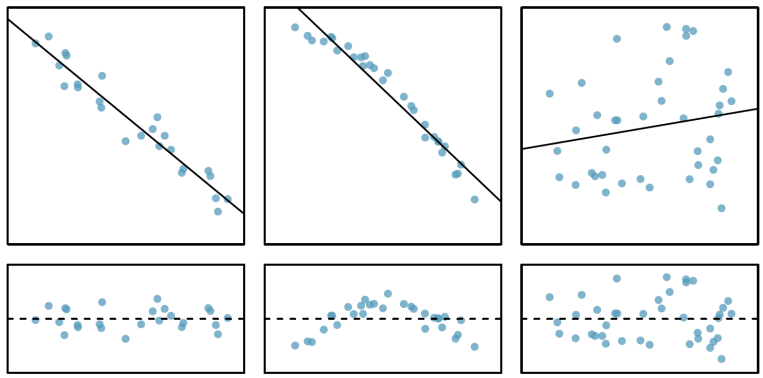


Least Squares

Conditions for Simple Linear Regression

Behavior of Residuals: 3 Examples

Sample data + regression on top, residual plots on bottom.



- Plots 1 and 3 are roughly linear.
- Plots 1 and 3 have roughly constant variability, but the 3rd plot has higher variability

Finding the Least Squares Line

Next Time

- How to interpret regression line parameter estimates
- Categorical Variable for x: male vs female, new vs used, etc.
- Inference for linear regression