Lecture 24: Linear Regression Part I

Chapter 7.1-7.2

Quiz 9

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Question 1: What is p-hacking?

Quiz 9

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Question 1: What is p-hacking?
Answer 1: Data-dredging AKA “trying multiple things until you get the desired result”
http://simplystatistics.org/2013/08/26/statistics-meme-sad-p-value-bear/

Quiz 9

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Question 2: Say a scientist obtains a p-value of 0.01. An incorrect interpretation of this is that it is the probability of a "false alarm" (type I error)... If one wants to make a statement about this being a false alarm, what additional piece of information is required?

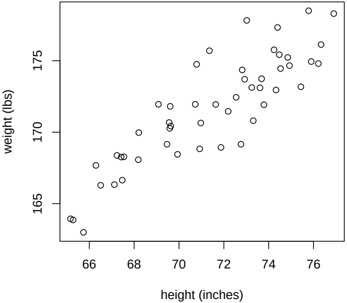
Quiz 9

http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Question 2: Say a scientist obtains a p-value of 0.01. An incorrect interpretation of this is that it is the probability of a "false alarm" (type I error)... If one wants to make a statement about this being a false alarm, what additional piece of information is required?
Answer 2: The plausibility of the hypothesis being tested for.

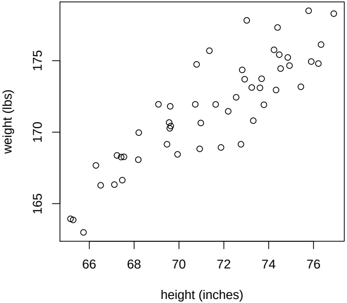
Questions for Today

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



Questions 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

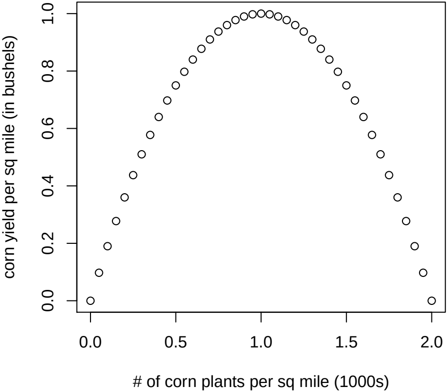
There are many types of regression, all in order to estimate the relationship between variables.

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.

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.

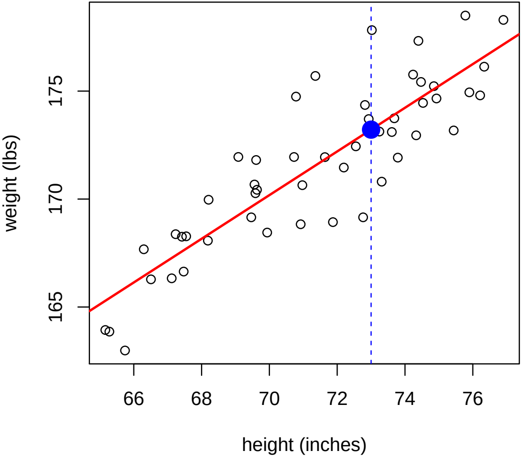


Modeling x and y Linearly

Procedure

Fitted Value

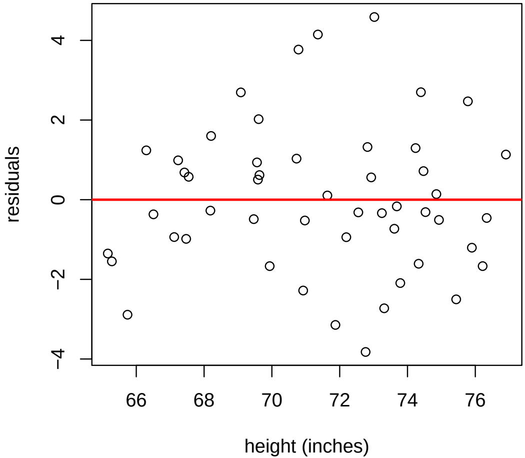
Here $\widehat{y} = 100 + 0.99x$. Thus for $x = 73$, $\widehat{y} = 173.22$:



Residuals

Residual Plot

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

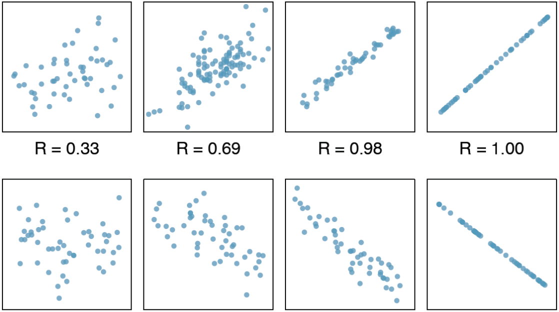


Correlation Coefficient

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

Correlation Coefficient

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



R = 0.33

R = 0.69

R = 0.98

R = 1.00

R = –0.08

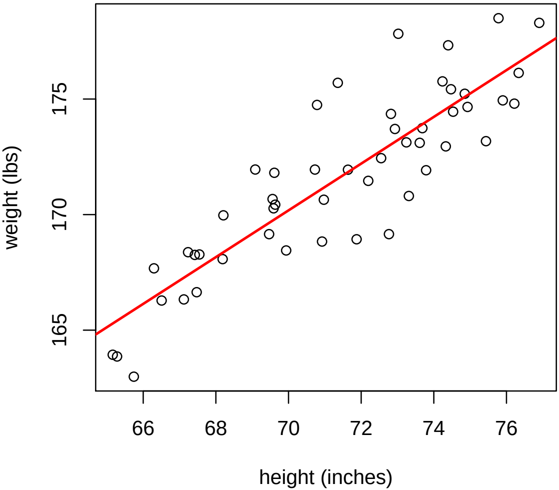
R = –0.64

R = –0.92

R = –1.00

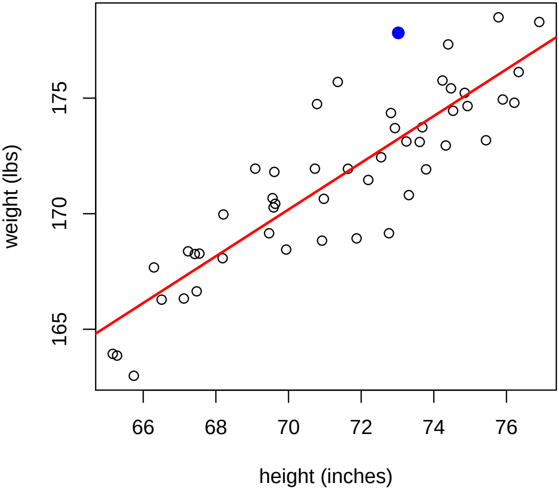
Best Fitting Line

What does "best fitting line" mean?



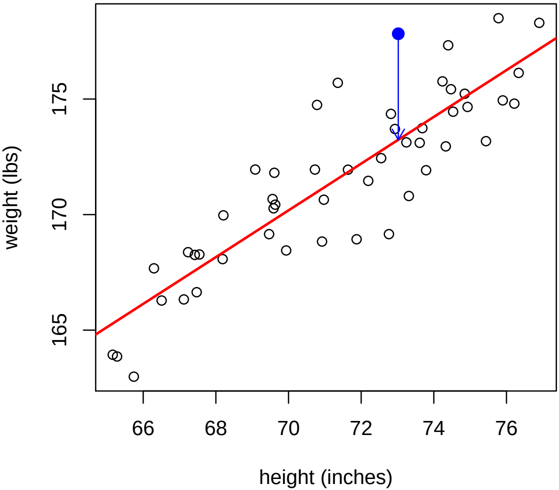
Best Fitting Line

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



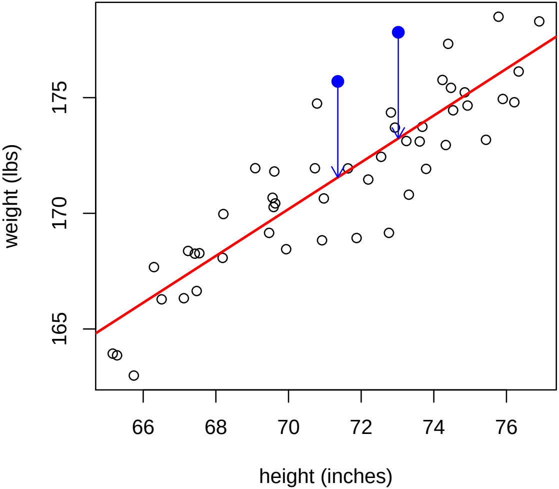
Best Fitting Line

Now consider this point's deviation from the regression line



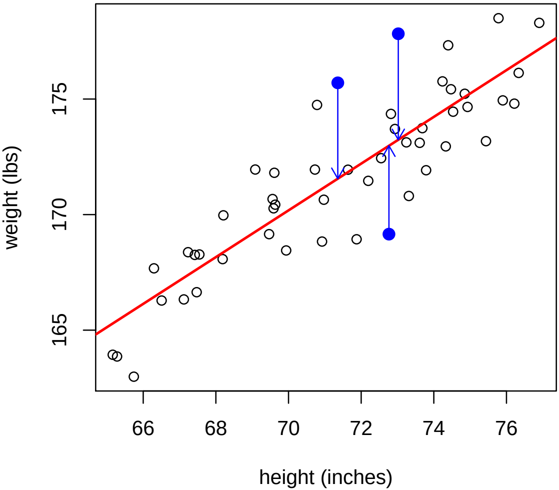
Best Fitting Line

Do this for another point x_i...



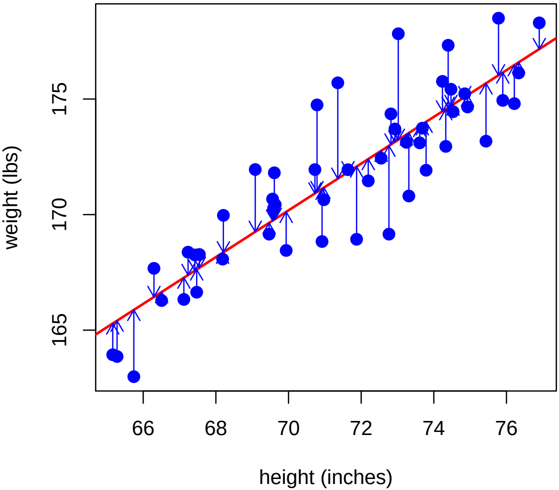
Best Fitting Line

Do this for another point x_i...



Best Fitting Line

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

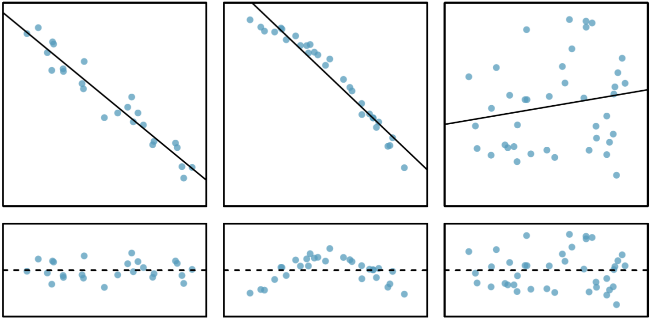


Least Squares

Conditions for Simple Linear Regression

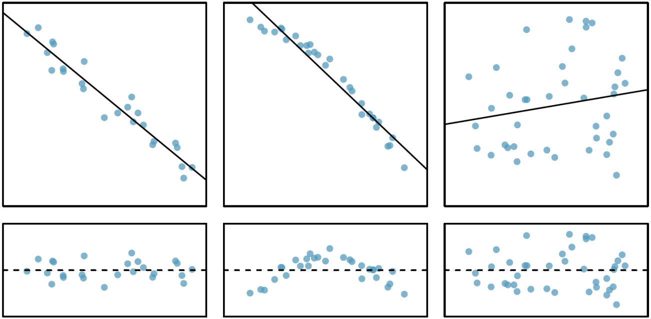
Behavior of Residuals: 3 Examples

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



Behavior of Residuals: 3 Examples

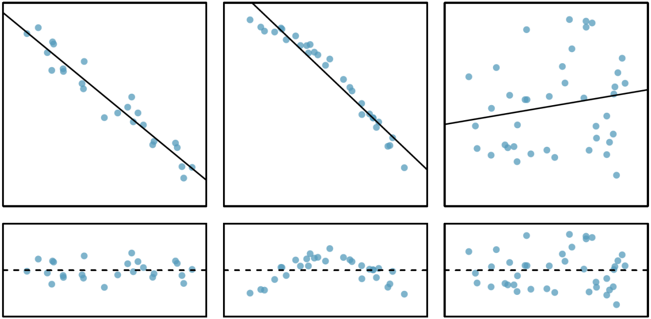
Sample data + regression on top, residual plots on bottom.



- Plots 1 and 3 are roughly linear.

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

Finding the Point Estimate of the Intercept b_0

Measuring the Strength of a Fit

If $R = -1$ or $R = 1$ we have a perfect linear fit between x and y, if $R = 0$ then there is no fit.

Measuring the Strength of a Fit

If $R = -1$ or $R = 1$ we have a perfect linear fit between x and y, if $R = 0$ then there is no fit.

However R^2 is a more commonly used measure of the strength of fit. For SLR, it is correlation coefficient squared, but not for other kinds of regression.

Measuring the Strength of a Fit

If $R = -1$ or $R = 1$ we have a perfect linear fit between x and y, if $R = 0$ then there is no fit.

However R^2 is a more commonly used measure of the strength of fit. For SLR, it is correlation coefficient squared, but not for other kinds of regression.

R^2 of a linear model describes the proportion of the total variation in y that is explained by 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