

Lecture 24: Linear Regression Part I

Chapter 7.1-7.2

1 / 25

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/>

2 / 25

Quiz 9

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

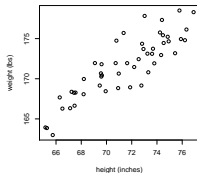
Question 2: Say a scientist obtains a p-value of 0.01. An incorrect interpretation of this is that it is the probability of a “false alarm” (type I error)... If one wants to make a statement about this being a false alarm, what additional piece of information is required?

Answer 2: The plausibility of the hypothesis being tested for.

3 / 25

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”?

4 / 25

Regression

There are many types of **regression**, all in order to estimate the relationship between variables. We start by considering **simple**

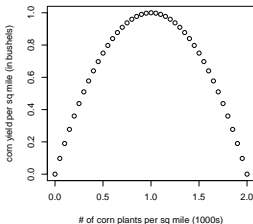
linear regression (SLR):

- ▶ a single **explanatory variable / independent variable / predictor variable** x
- ▶ an **outcome variable / dependent variable** y
- ▶ a presumed linear relationship between them

5 / 25

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.



6 / 25

Modeling x and y Linearly

The **SLR model** assumes that the relationship between x and y can be modeled by a line:

$$y = \beta_0 + \beta_1 x$$

where

- ▶ β_0 is the unknown **intercept parameter**
- ▶ β_1 is the unknown **slope parameter**

7 / 25

Procedure

Based on n pairs of observations (x_i, y_i)

1. Compute **point estimates**
 - ▶ b_0 of parameter β_0
 - ▶ b_1 of parameter β_1
2. Associate standard errors SE_{b_0} and SE_{b_1}
3. For both the intercept and slope
 - ▶ Build confidence intervals
 - ▶ Do hypothesis test

$$H_0 : \beta = 0$$

vs

$$H_A : \beta \neq 0$$

The equation

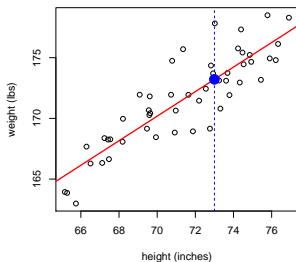
$$\hat{y} = b_0 + b_1 x$$

is called the **least squares line** where \hat{y} is the **fitted/predicted value**.

8 / 25

Fitted Value

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



9 / 25

Residuals

Residuals are what's leftover: leftover variation in the data unexplained by the model:

$$\text{Residual} = \text{Data} - \text{Fit}$$

$$e_i = y_i - \hat{y}_i$$

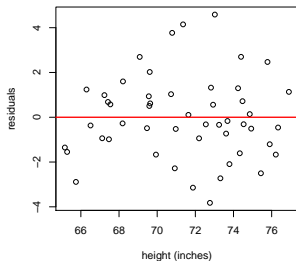
where e_i is the **residual** of the i^{th} observation (x_i, y_i) .

We can think of the e_i 's as **deviations** from the model. The smaller the deviations, the better the fit.

10 / 25

Residual Plot

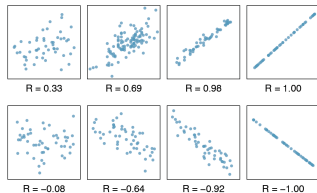
Residual plots: take previous plot and flatten the red line by subtracting \hat{y} from y .



11 / 25

Correlation Coefficient

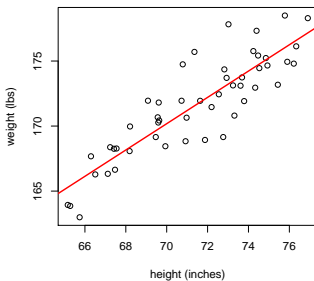
The correlation coefficient R is a value between $[-1, 1]$ that measures the strength of the linear relationship between x and y .



12 / 25

Best Fitting Line

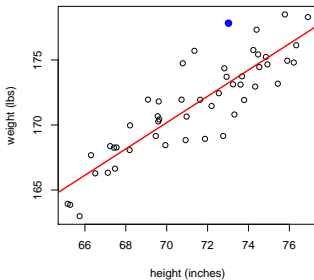
What does "best fitting line" mean?



13 / 25

Best Fitting Line

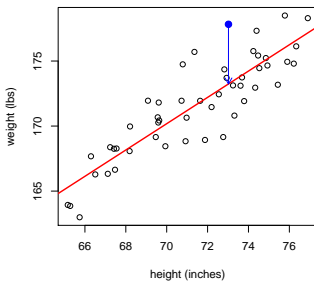
Consider ANY point x_i for $i = 1, \dots, 50$ (in blue).



14 / 25

Best Fitting Line

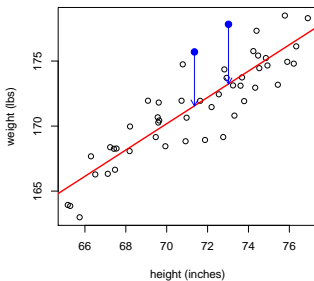
Now consider this point's deviation from the **regression line**



15 / 25

Best Fitting Line

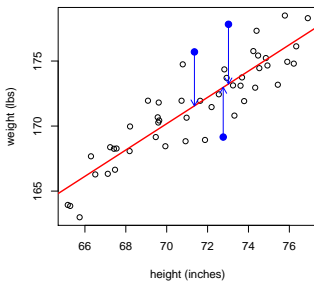
Do this for another point x_j ...



16 / 25

Best Fitting Line

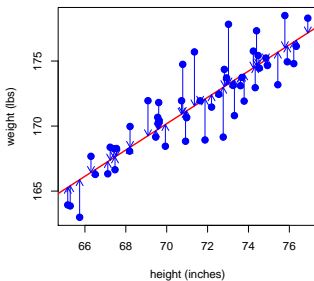
Do this for another point x_j ...



17 / 25

Best Fitting Line

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



18 / 25

Least Squares

i.e. the regression line minimizes:

$$e_1^2 + e_2^2 + \dots + e_n^2$$

This is called **minimizing the least squares criterion**.

Why not minimize

$$|e_1| + |e_2| + \dots + |e_n| ?$$

It's easier to do calculus on x^2 than $|x|$

19 / 25

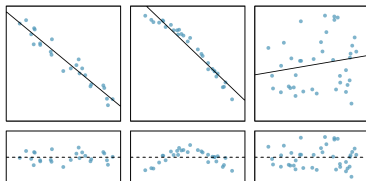
Conditions for Simple Linear Regression

- ▶ **Linearity:** The data should show a linear trend.
- ▶ **Independence:** The residuals should be independent
- ▶ **Nearly normal residuals:** The residuals e_i must be nearly normal (verify with QQ-plot) with mean 0.
- ▶ **Constant variability:** The variability of points around the least squares line remains roughly constant (i.e. for all values of x).

20 / 25

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

21 / 25

Finding the Least Squares Line

To find the least squares line we need to find the point estimates:

- ▶ The point estimate b_1 of the slope β_1 is

$$b_1 = \frac{s_y}{s_x} R$$

- ▶ The regression line **always** goes through (\bar{x}, \bar{y}) . We use this fact to find the point estimate of b_0 of the intercept β_0 .

22 / 25

Finding the Point Estimate of the Intercept b_0

Given the slope and a point on the line (x_0, y_0) , the equation for the line can be written as

$$\begin{aligned}\text{slope} &= \frac{\text{rise}}{\text{run}} = \frac{y - y_0}{x - x_0} \\ y - y_0 &= \text{slope} \times (x - x_0)\end{aligned}$$

So

$$\begin{aligned}y - \bar{y} &= b_1(x - \bar{x}) \\ \text{so } y &= (\bar{y} - b_1\bar{x}) + b_1x \\ \text{so } b_0 &= \bar{y} - b_1\bar{x}\end{aligned}$$

23 / 25

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.**

24 / 25

Next Time

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