STAT 305: Chapter 4

Part I

Amin Shirazi

Course page:
ashirazist.github.io/stat305.github.io

Recap on Numerical Summaries (Chapter 3)

Summaries of Variablity (Measures of Spread)

Chapter 4: Describing Relationships Between Variables

Introduction to Models

Recap: Summaries of Variablity (Measures of Spread)

Motivated by asking what kind of *variability is seen in the data* or *how spread out* the data is.

Range: The difference between the highest and lowest values (Range = max - min)

IQR: The Interquartile Range, how spread out is the middle 50% (IQR = Q3 - Q1)

Variance/Standard Deviation: Uses squared distance from the mean.

	Variance	Standard Deviation
Population	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$
Sample	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Summarizing Data Numerically

Example: Taking a sample of size 5 from a population we record the following values:

58, 60, 61, 68, 56

Find the variance and standard deviation of this sample.

Example: Finding the Variance

Since we are told it is a sample, we need to use **sample variance**. The mean of 58, 60, 61, 68, 56 is 60.6

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{5}(x_i - \bar{x})^2$$

$$= \frac{1}{n-1}\left((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + (x_4 - \bar{x})^2 + (x_5 - \bar{x})^2\right)$$

$$= \frac{1}{5-1}\left((58 - 60.6)^2 + (60 - 60.6)^2 + (61 - 60.6)^2 + (68 - 60.6)^2 + (56 - 60.6)^2\right)$$

$$= \frac{1}{4}\left((-2.6)^2 + (-0.6)^2 + (0.4)^2 + (7.4)^2 + (-4.6)^2\right)$$

$$= \frac{1}{4}(6.76 + 0.36 + 0.16 + 54.76 + 21.16)$$

$$= 20.8$$

Example: Finding the Standard Deviation

With s^2 known, finding s is simple:

$$\begin{aligned} s &= \sqrt{s^2} \\ &= \sqrt{20.8} \\ &= 4.5607017 \end{aligned}$$

Chapter 4, Section 1

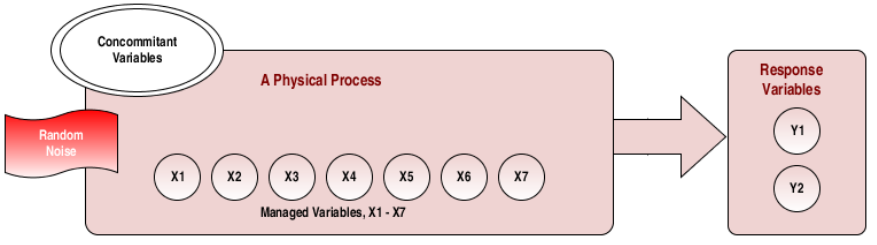
Linear Relationships Between Variables

Describing Relationships

Idea

Describing Relationships

We have a standard idea of how our experiment works:



Bivariate data oftern arise because a quantitative experimental variable *x* has been varied between several different setting (treatment).

It is helpful to have an equation relating *y* (the response) to *x* when the purposes are summarization, interpolation, limited extrapolation, and/or process optimization/ adjusment.

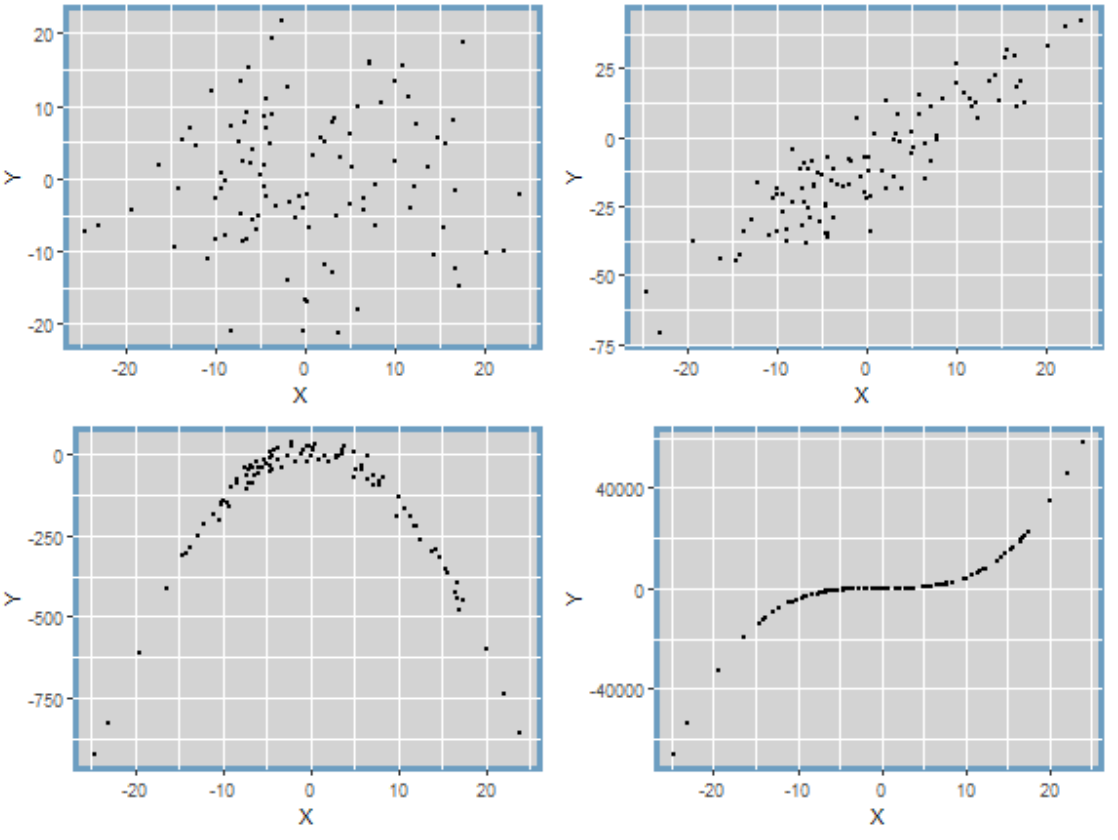
and we know that with an valid experiment, we can say that the changes in our experimental variables actually *cause* changes in our response.

But how do we describe those response when we know that random error would make each result different...

Describing Relationships

Types of relationships

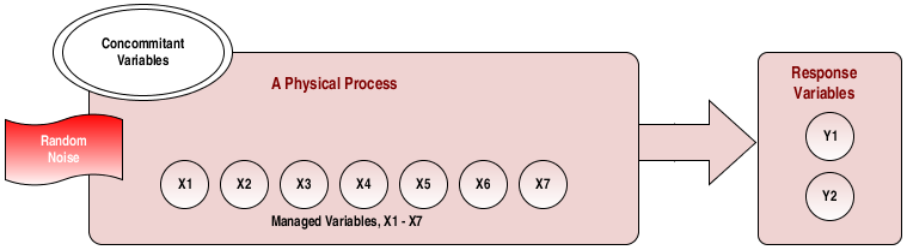
Idea



Describing Relationships

The Underlying Idea

Idea



We start with a valid mathematical model, for instance a line:

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

In this case,

- β_0 is the intercept - when $x = 0$, $y = \beta_0$.
- β_1 is the slope - when x increase by one unit, y increases by β_1 units.

Describing Relationships

Idea

Ex: Bar Stress

Example: Stress on Bars

An experiment examining the effects of **stress** on **time until fracture** is performed by taking a sample of 10 stainless steel rods immersed in 40% CaCl solution at 100 degrees Celsius and applying different amounts of uniaxial stress.

The results are recorded below:

stress $(\mathrm{kg/mm^2})$	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

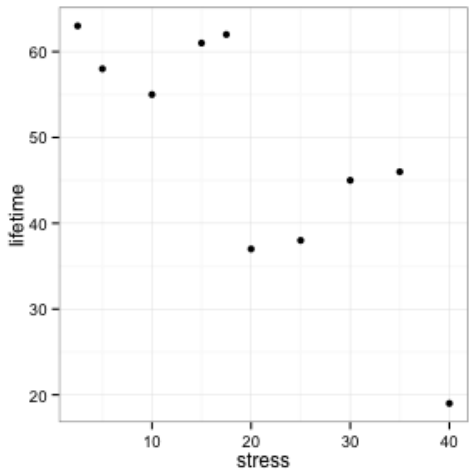
A good first place to investigate the relationship between our experimental variables (in this case, stress) and the response (in this case, lifetime) is to use a scatterplot and look to see if there might be any basic mathematical function that could describe the relationship between the variables.

Example: Stress on Bars (continued)

Our data:

stress $(\mathrm{kg/mm^2})$	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

- Plotting stress along the x-axis and plotting lifetime along the y-axis we get

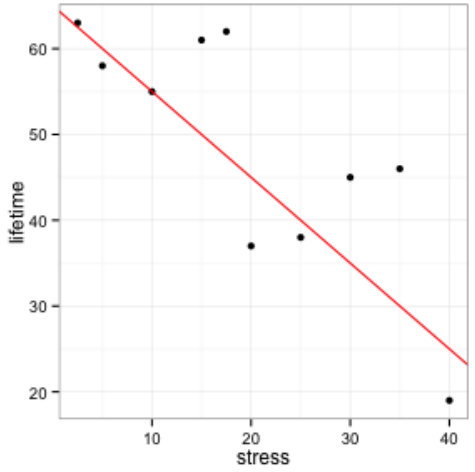


Example: Stress on Bars (continued)

Our data:

stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

- Examining the plot, we might determine that there could be a linear relationship between the two. The red line looks like it fits the data pretty well.

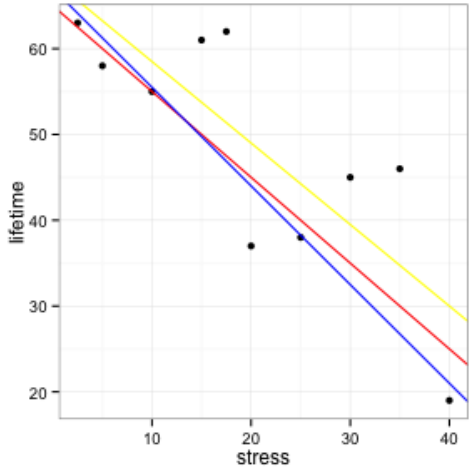


Example: Stress on Bars (continued)

Our data:

stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

- But there are several other lines that fit the data pretty well, too.



- How do we decide which is best?

Describing Relationships

Idea

Ex: Bars

Fitting Lines

Where the line comes from

When we are trying to find a line that fits our data what we are *really* doing is saying that there is a true physical relationship between our experimental variable x is related to our response y that has the following form:

Theoretical Relationship

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

However, the response we observe is also effected by random noise:

Observed Relationship

$$y = \beta_0 + \beta_1 \cdot x + \text{errors}$$

$$= \text{signal} + \text{noise}$$

If we did a good job, hopefully we will have small enough errors so that we can say

$$y \approx \beta_0 + \beta_1 \cdot x$$

Where the line comes from

So, if things have gone well, we are attempting to estimate the value of β_0 and β_1 from our observed relationship

$$y \approx \beta_0 + \beta_1 \cdot x$$

Using the following notation:

- b_0 is the estimated value of β_0 and
- b_1 is the estimated value of β_1
- $\hat{y}$ is the estimated response

We can write a **fitted relationship**:

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

The key here is that we are going from the underlying *true, theoretical* relationship to an *estimated* relationship.

In other words, we will never get the true values β_0 and β_1 but we can estimate them.

However, this doesn't tell us *how* to estimate them.

The principle of Least Squares

A good estimte should be based on the data.

Idea

Suppose that we have observed responses $y_1, y_2, \ldots, y_n$ for experimental variables set at $x_1, x_2, \ldots, x_n$.

Ex: Bars

Then the **Principle of Least Squares** says that the best estimate of β_0 and β_1 are values that **minimize**

Fitting Lines

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Best Estimate

In our case, since $\hat{y}_i = b_0 + b_1 \cdot x_i$ we need to choose values for b_0 and b_1 that minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 \cdot x_i))^2$$

In other words, we need to minimize something with respect to two values we get to choose - we can do this by taking derivatives.

Deriving the Least Squares Estimates(Optional reading)

We can rewrite the target we want to minimize so that the variables are less tangled together:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - (b_0 + b_1 x_i)\right)^2$$

$$= \sum_{i=1}^{n}\left(y_i^2 - 2y_i(b_0 + b_1 x_i) + (b_0 + b_1 x_i)^2\right)$$

$$= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} 2y_i(b_0 + b_1 x_i) + \sum_{i=1}^{n}(b_0 + b_1 x_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n}(2y_i b_0 + 2y_i b_1 x_i) + \sum_{i=1}^{n}\left(b_0^2 + 2b_0 b_1 x_i + (b_1 x_i)^2\right)$$

$$= \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} 2y_i b_0 - \sum_{i=1}^{n} 2y_i b_1 x_i + \sum_{i=1}^{n} b_0^2 + \sum_{i=1}^{n} 2b_0 b_1 x_i + \sum_{i=1}^{n} b_1^2 x_i^2$$

$$= \sum_{i=1}^{n} y_i^2 - 2b_0\sum_{i=1}^{n} y_i - 2b_1\sum_{i=1}^{n} y_i x_i + nb_0^2 + 2b_0 b_1\sum_{i=1}^{n} x_i + b_1^2\sum_{i=1}^{n} x_i^2$$

Deriving the Least Squares Estimates (continued)

Idea
Ex: Bars
Fitting Lines
Best Estimate

How do we minimize it?

- Since we have two "variables" we need to take derivates with respect to both.
- Remember we have our data so we know every value of x_i and y_i and can treat those parts as constants.

The derivative with respect to $\mathbf{b_0}$:

$$-2\sum_{i=1}^{n} y_i + 2nb_0 + 2b_1\sum_{i=1}^{n} x_i$$

The derivative with respect to $\mathbf{b_1}$:

$$-2\sum_{i=1}^{n} y_i x_i + 2b_0\sum_{i=1}^{n} x_i + 2b_1\sum_{i=1}^{n} x_i^2$$

Deriving the Least Squares Estimates (continued)

We set both equal to 0 and solve them at the same time:

$$-2\sum_{i=1}^{n} y_i + 2nb_0 + 2b_1 \sum_{i=1}^{n} x_i = 0$$

$$-2\sum_{i=1}^{n} y_i x_i + 2b_0 \sum_{i=1}^{n} x_i + 2b_1 \sum_{i=1}^{n} x_i^2 = 0$$

We can rewrite the first equation as:

$$b_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - b_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \bar{y} - b_1 \bar{x}$$

and then replace all b_0 in the second equation (there is some algebra type stuff along the way, of course)

Deriving the Least Squares Estimates (continued)

After a little simplification we arrive at our estimates:

Least Squares Estimates for Linear Fit

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Wrap Up

- Don't try to memorize the derivation. I will never ask you to do that on an exam.
- Try to understand the simplification steps - the ones that moved constants out of summations for example.
- This is one rule - there are others, but **Least Squares Estimates** have some useful properties that will make

Describing Relationships

Idea

Ex: Bars

Fitting Lines

Best Estimate

Example: Stress on Bars

stress $(\mathrm{kg/mm}^2)$	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

Estimating the best slope and intercept using least squares:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

In our case we have the following:

Example: Stress on Bars

stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

$$\sum_{i=1}^{10} y_i = 484, \sum_{i=1}^{10} x_i = 200, \sum_{i=1}^{10} x_i y_i = 8407.5, \sum_{i=1}^{10} x_i^2 = 5412.5,$$

Using this we can estimate b_1:

$$b_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{8407.5 - 10\left(\frac{200}{10}\right)\left(\frac{484}{10}\right)}{5412.5 - 10\left(\frac{200}{10}\right)^2}$$

$$= \frac{-1272.5}{1412.5}$$

$$\approx -0.9009$$

Describing Relationships

Idea

Ex: Bars

Fitting Lines

Best Estimate

Example: Stress on Bars

stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

$$\sum_{i=1}^{10} y_i = 484, \sum_{i=1}^{10} x_i = 200, \sum_{i=1}^{10} x_i y_i = 8407.5, \sum_{i=1}^{10} x_i^2 = 5412.5,$$

And using b_1 we can estimate b_0:

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= \left(\frac{484}{10}\right) - b_1 \left(\frac{200}{10}\right) \\ &= 48.4 - \left(\frac{-1272.5}{1412.5}\right) 20.0 \\ &= 66.4177 \end{aligned}$$

Which gives us the **Fitted Relationship**:

Describing Relationships

Idea

Ex: Bars

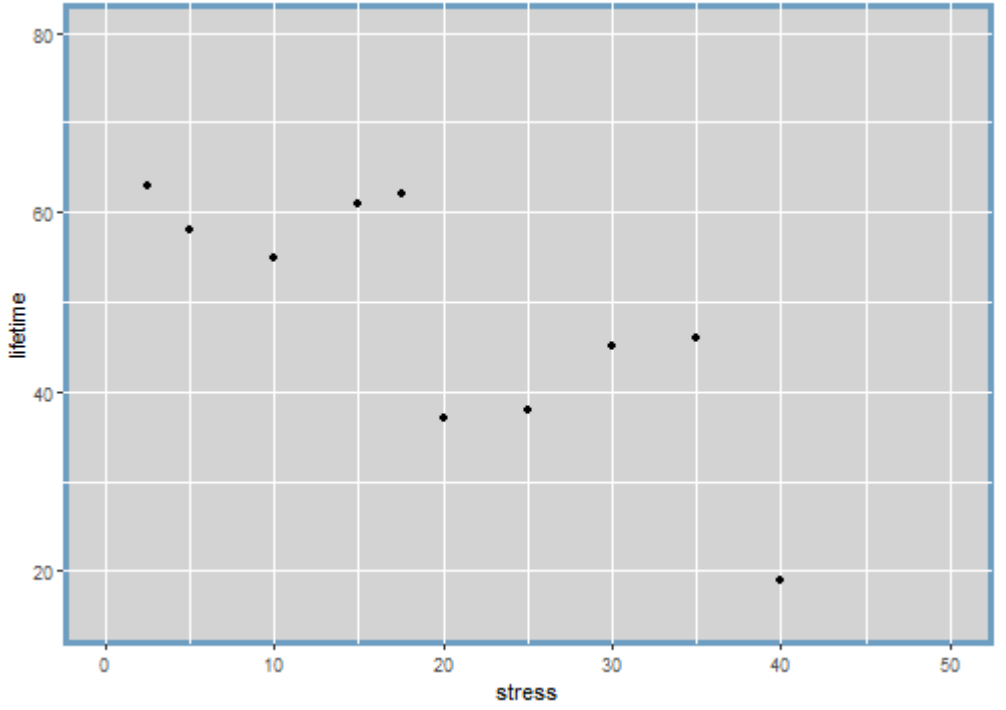
Fitting Lines

Best Estimate

Example: Stress on Bars

stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

$$\hat{y} = 66.4177 - 0.9009x$$



Describing Relationships

Idea

Ex: Bars

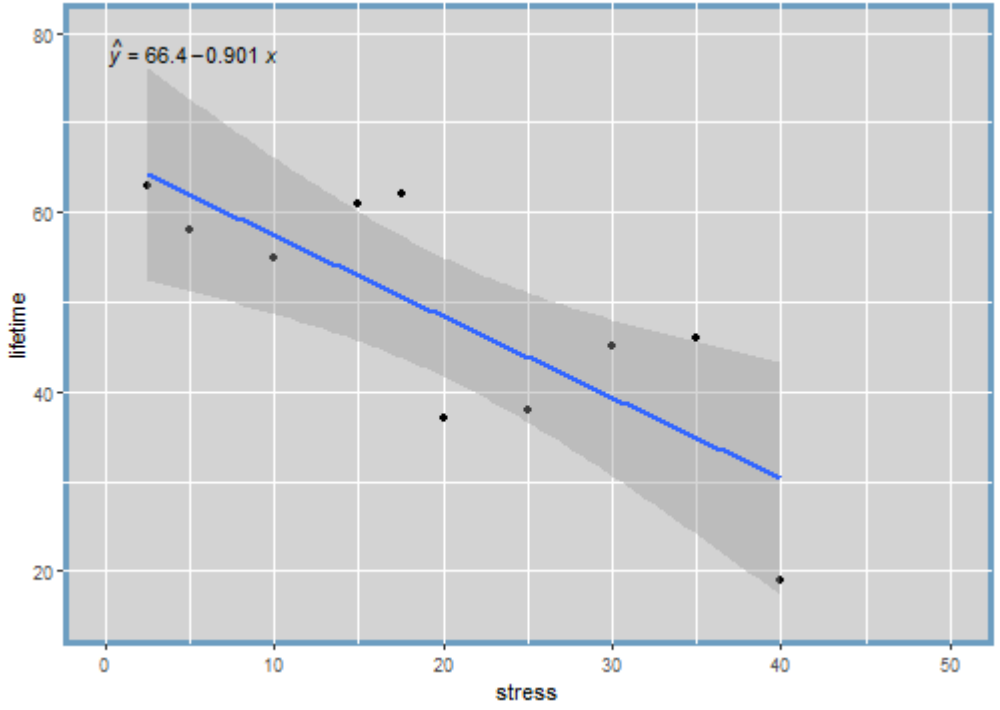
Fitting Lines

Best Estimate

Example: Stress on Bars

stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

Fitted line



JMP

Describing Relationships

Using JMP

Topics to be covered in JMP

- Fitting linear relationships
- Describing quality of fit (correlation, R^2)
- Fitting relationships using multiple variables
- Fitting non-linear relationships

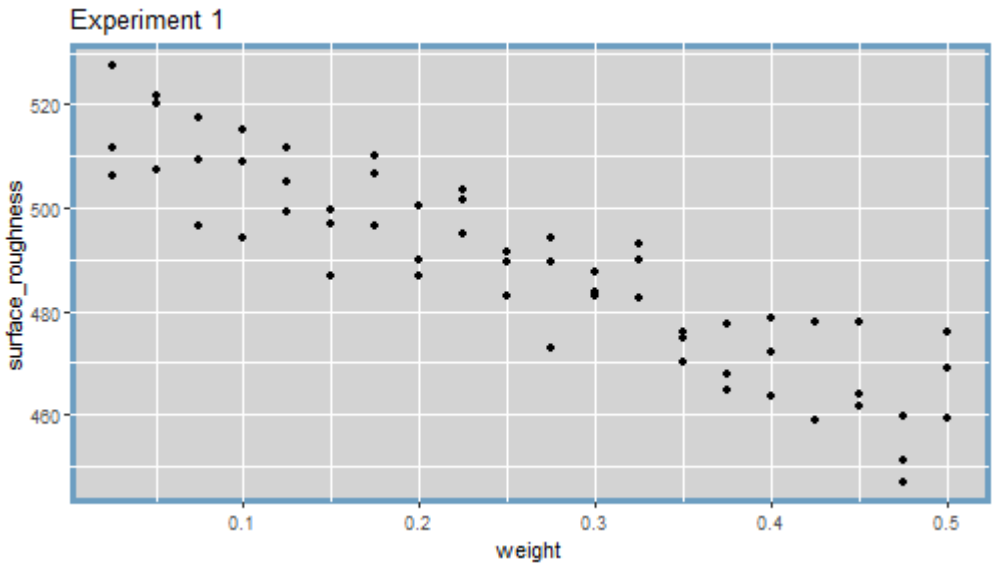
An example

Example: Manufacturing Ball Bearings

Controlling surface roughness is an important part of the manufacture of bearing balls. The key step in this smoothing the balls involves the use of a spinning weighted disc. Two important aspects of this are the rotation speed of the disc and the weight applied to the disc. Since higher weights and higher rotation speed are all known to cause shorter lifetimes for the discs (which requires halts in production, costs of new discs, and so on), a team of engineers are attempting to better understand the relationship between the rotation speed, the weight, and the resulting surface roughness of the balls produced.

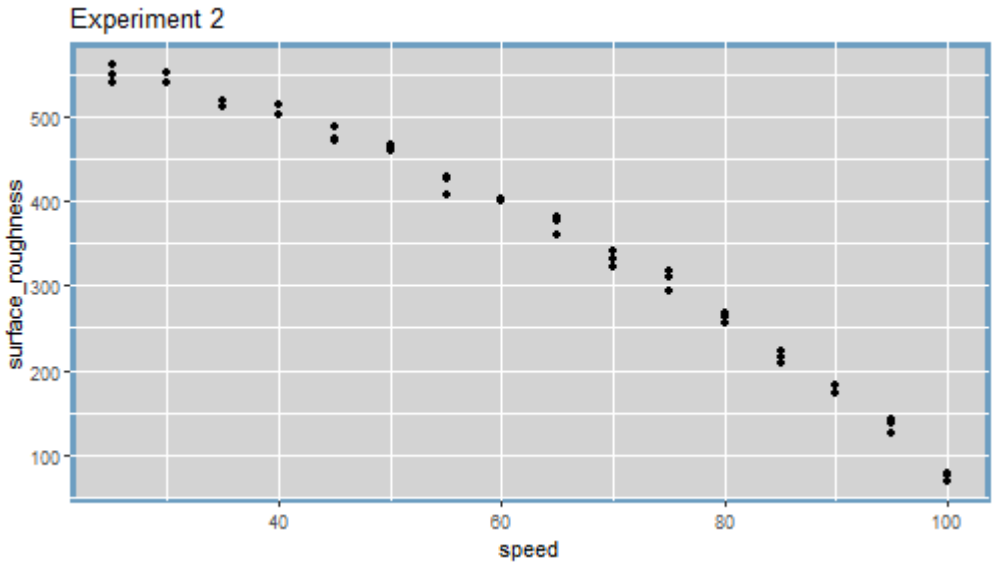
Experiment 1: Constant speed, changing applied weight

With the disc rotation speed locked at 50 rotations/second, the team of engineers created 60 batches of balls using differently weighted discs (0.025 g, 0.050 g, 0.075 g, 0.100 g, ..., 0.500 g) and randomly selected one ball from each batch. The results are recorded in the dataset "balls-001.csv" on the course page.



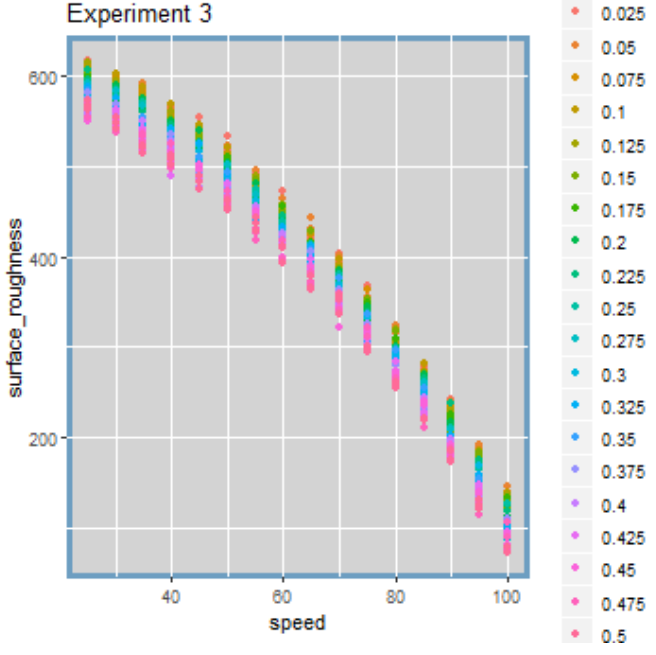
Experiment 2: Changing speed, constant applied weight

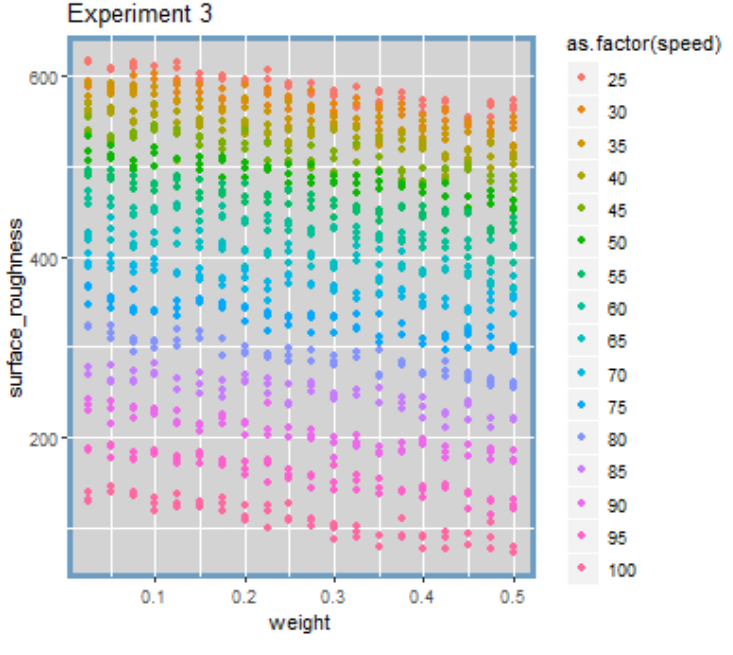
With an better understanding of the relationship between weight and surface roughness, the team turned their attention to rotation speed. This time the produced 3 batches for each of 15 rotation speeds (25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, and 95 rotations per second). The results are recorded in the dataset "balls-002.csv" on the course page.



Experiment 3: Changing speed changing applied weight

With a better understanding of the relationship between weight and surface roughness, the team turned their attention to rotation speed. This time the produced 3 batches for each combination of 20 weights (0.025 g, 0.050 g, 0.075 g, 0.100 g, ..., 0.500 g) and 15 rotation speeds (25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, and 95 rotations per second). The results are recorded in the dataset "balls-003.csv"



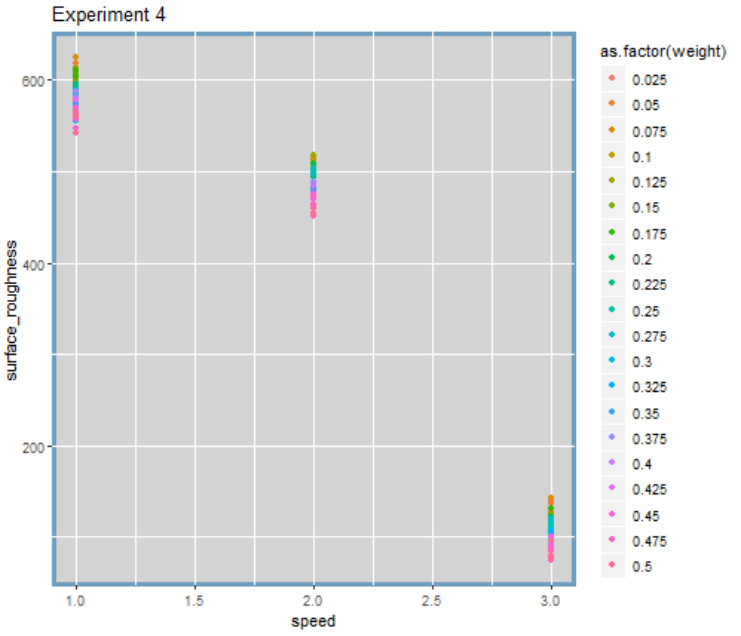


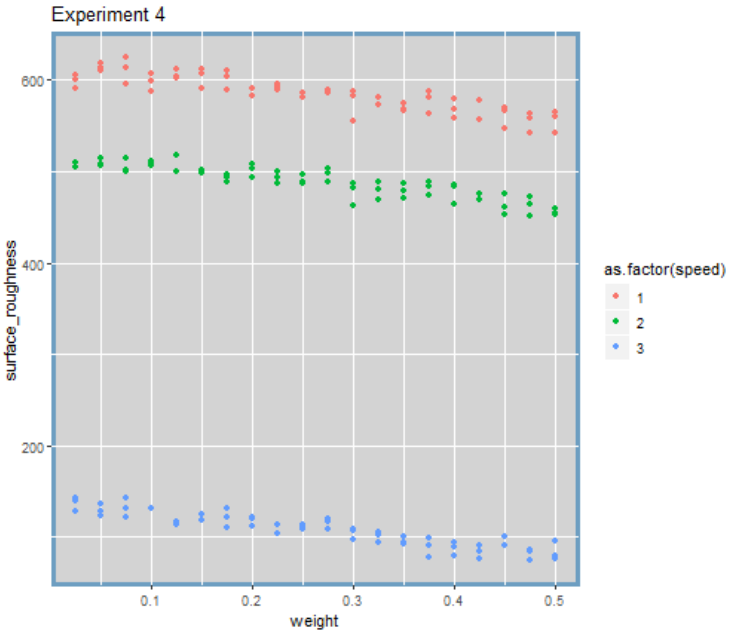
Experiment 4: Changing categorical speed changing applied weight

Now that they have a complete model, what if they had attempted this experiment with a machine in which rotation speed only consisted of "low, medium, and high"?

Again, time the produced 3 batches for each combination of 20 weights (0.025 g, 0.050 g, 0.075 g, 0.100 g, ..., 0.500 g) and three rotation speeds: low (encoded as 1), medium (encoded as 2), high (encoded as 3). The results are recorded in the dataset "balls-004.csv"

Experiment 4: Changing categorical speed changing applied weight

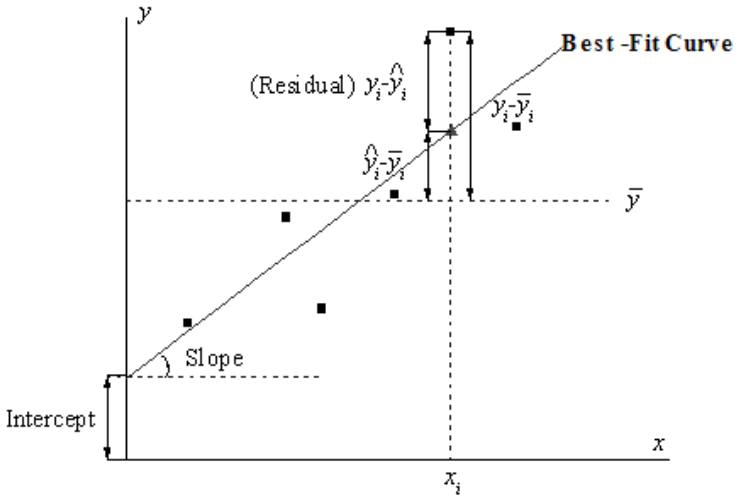




Residuals

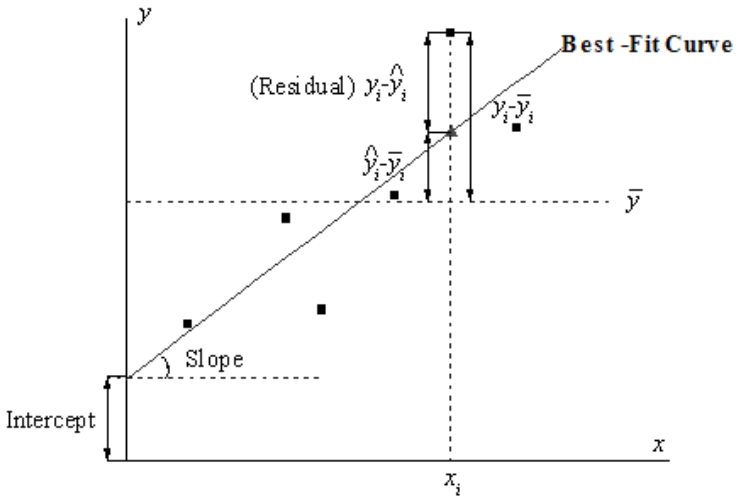
Residuals

- The "residue" left over from fitting a line



- Each point represents some (x_i, y_i) pair from our data
- We use the Least Squares approach to find the best fit line, $\hat{y} = b_0 + b_1 x$
- For any value x_i in our data set, we can get a fitted (or predicted) value $\hat{y}_i = b_0 + b_1 x_i$

Residuals



- The residual is the difference between the observed data point and the fitted prediction:

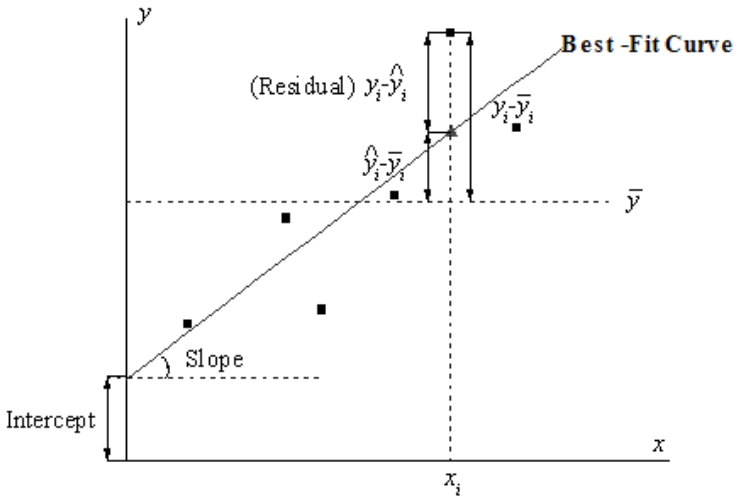
$$e_i = y_i - \hat{y}_i$$

- **In the linear case**, using $\hat{y} = b_0 + b_1 x$, we can also write

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

for each pair (x_i, y_i).

Residuals



ROPe: **R**esiduals = **O**bserved - **P**redicted (using symbol e_i)

- If $e_i > 0$ then $y_i - \hat{y}_i > 0$ and $y_i > \hat{y}_i$ meaning the observed is larger than the predicted - we are "underpredicting"
- If $e_i < 0$ then $y_i - \hat{y}_i < 0$ and $y_i < \hat{y}_i$ meaning the observed is smaller than the predicted - we are "overpredicting"