STAT 305: Chapter 4

Part I

Amin Shirazi

ashirazist.github.io/stat305_s2020.github.io

Chapter 4, Section 1

Linear Relationships Between Variables

Describing Relationships

Idea

Describing Relationships between variables

This chapter provides methods that address a more involved problem of describing relationships between variables and require more computation. We start with relationships between two variables and move on to more.

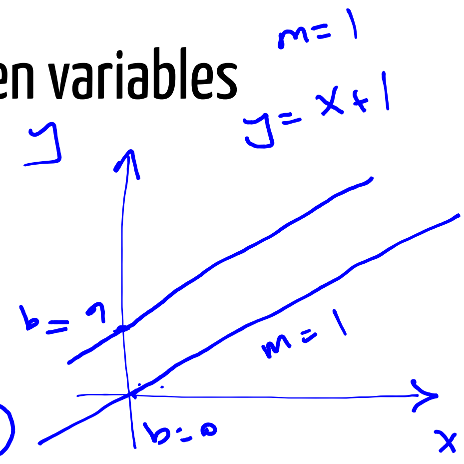
Fitting a line by least squares

Goal: Notice a relationship between two quantitative variables.

> We would like to use an equation to describe how a dependent (response) variable, y, changes in response to a change in one or more independent (experimental) variable(s), x.

Describing Relationships between variables

Line review



Recall a linear equation of the form

$$y = mx + b$$

Where m is the slope and b is the intercept of the line.

In statistics, we use the notation $y = \beta_0 + \beta_1 x + \epsilon$ where we assume β_0 and β_1 are unknown parameters and ϵ is some error.

The goal is to find estimates b_0 (intercept) and b_1 (slope) for the parameters.

$\hat{\beta}_0 = b_0$, $\hat{\beta}_1 = b_1$ estimates

β_0, β_1 are unknown (fixed) parameters

Idea

Describing Relationships

We have a standard idea of how our experiment works:

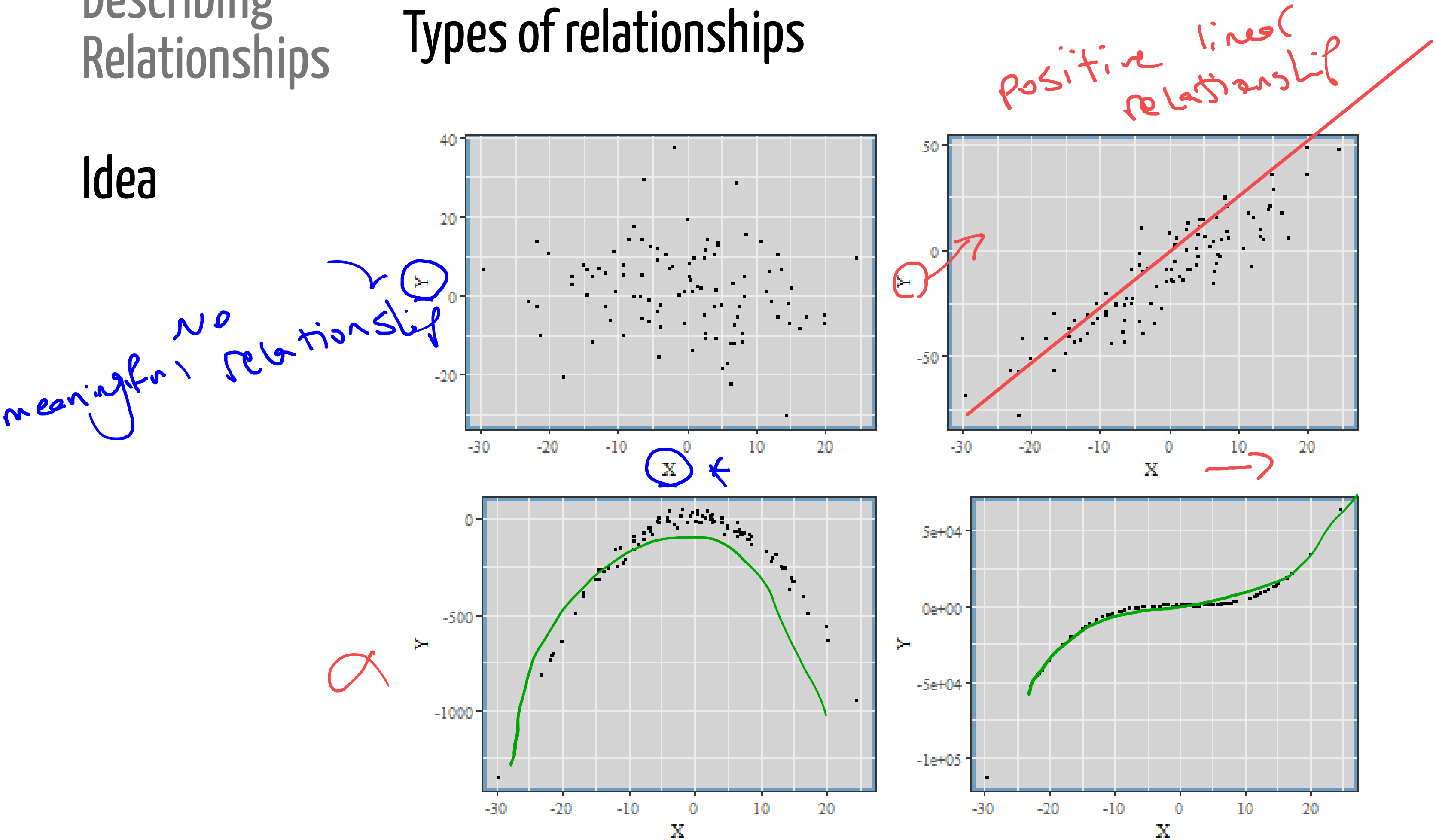
- Bivariate data oftern arise because a quantitative experimental variable *x* has been varied between several different setting (treatment).
- It is helpful to have an equation relating *y* (the response) to *x* when the purposes are summarization, interpolation, limited extrapolation, and/or process 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…

Types of relationships

Idea



The Underlying Idea

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

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

true relationship

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.

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 (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

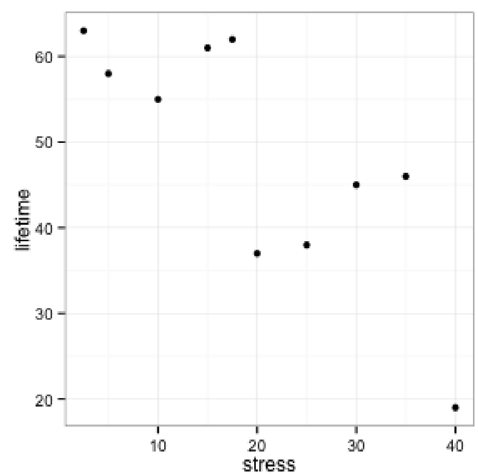
A good first place to investigate the relationship between our experimental variables (in this case, stress) and the response (in this case, lifetime) is to use a scatterplot and look to see if there might be any basic mathematical function that could describe the relationship between the variables.

Example: Stress on Bars (continued)

Our data:

stress (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



Describing Relationships

Idea

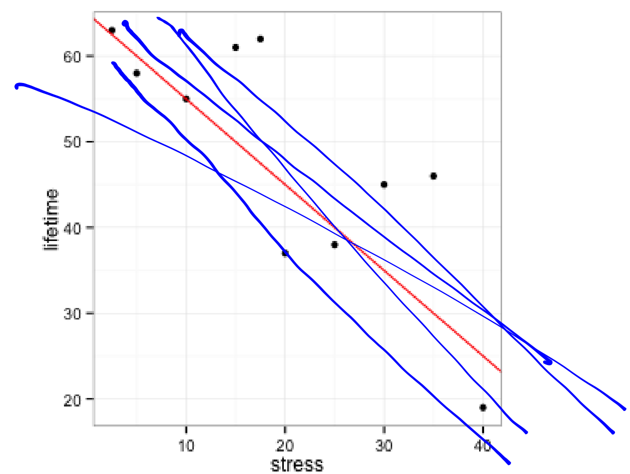
Ex: Bar Stress

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.



Describing Relationships

Idea

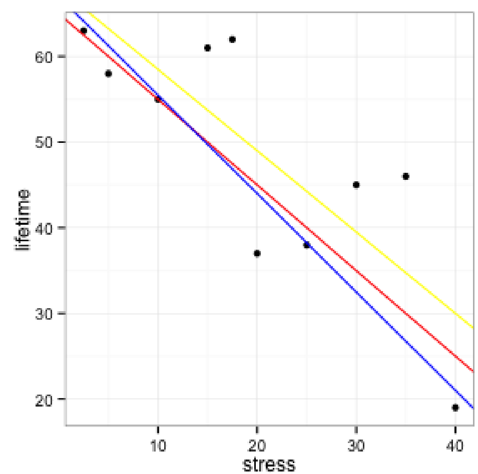
Ex: Bar Stress

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?

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

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

Describing Relationships

Idea

Ex: Bars

Fitting Lines

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.

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

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

observed value

fitted values

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

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

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

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:

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

Deriving the Least Squares Estimates (continued)

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 $\mathbf{b_1}$:

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

Describing Relationships

Idea

Ex: Bars

Fitting Lines

Best Estimate

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

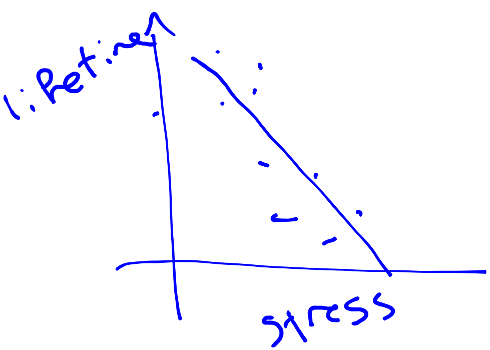
$$b_0 = \bar{y} - b_1\bar{x}$$

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

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

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 them the obvious best choice as we continue the course.



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

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 $(\mathrm{kg/mm^2})$	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

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

Using this we can estimate b_1:

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

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

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

$$\approx -0.9009$$

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:

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

$$= \left(\frac{484}{10}\right) - b_1 \left(\frac{200}{10}\right)$$

$$= 48.4 - \left(\frac{-1272.5}{1412.5}\right) 20.0$$

$$= 66.4177$$

true relationship:

$* \; y = \beta_0 + \beta_1 X + \varepsilon$

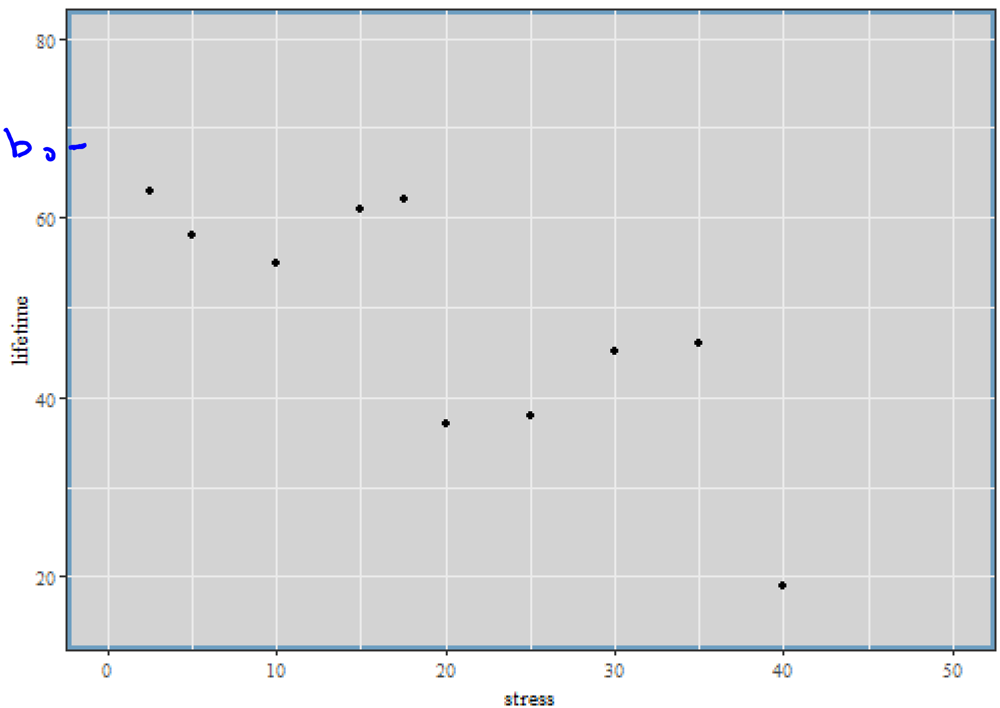
Which gives us the **Fitted Relationship**:

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

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

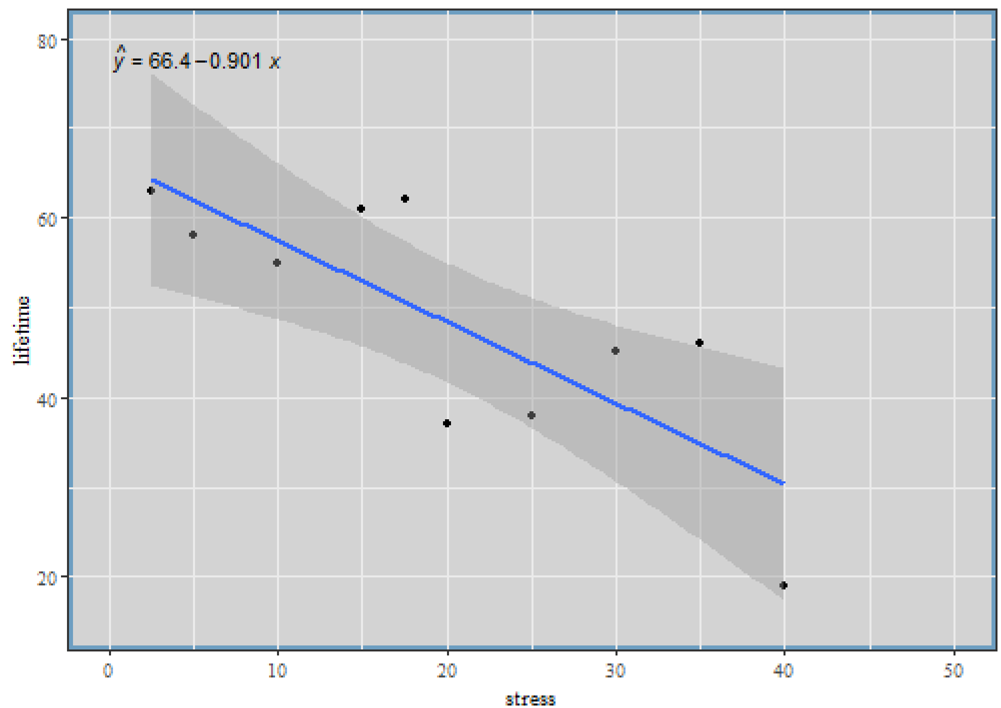
$$\hat{y} = 66.4177 - 0.9009x$$



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



Describing Relationships

Idea

Ex: Bars

Fitting Lines

Best Estimate

When making predictions, don't *extrapolate*.

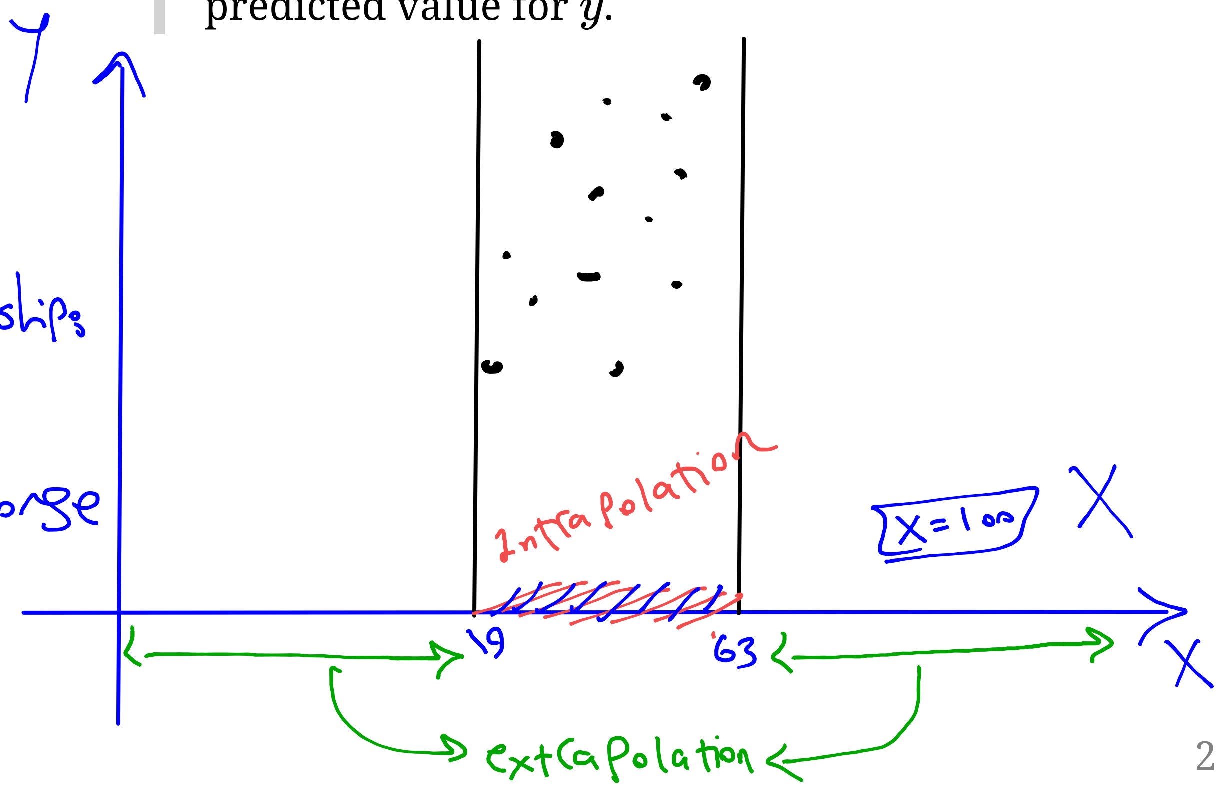
> **Extrapolation** is when a value of x beyond the range of our actual observations is used to find a predicted value for y. We don't know the behavior of the line beyond our collected data.
>
> **Interpolation** is when a value of x within the range of our observations is used to find a predicted value for y.

Fitted relationships:

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

Predict new response values:

$\hat{y} = b_0 + b_1(20)$



Good Fit

Describing Relationships

Idea

Ex: Bars

Fitting Lines

Best Estimate

Good Fit

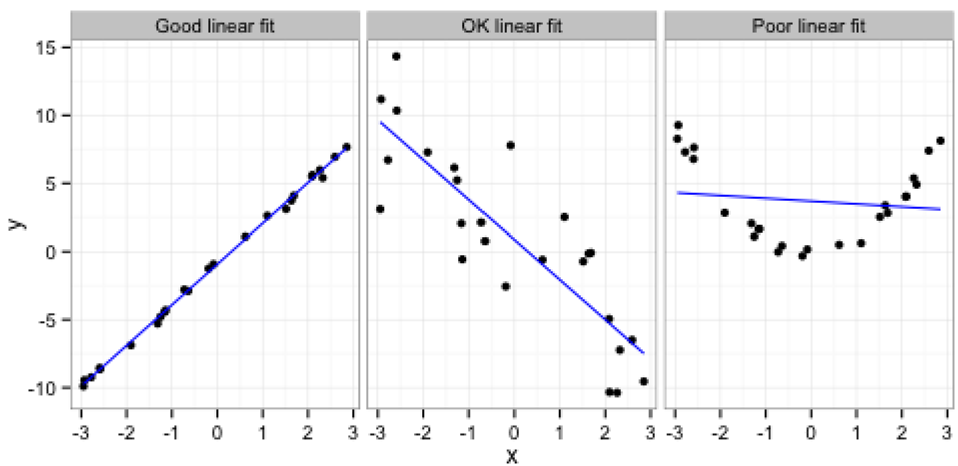
Knowing when a relationship fits the data well

So far we have been fitting lines to describe our data. A first question to ask may be something like:

- **Q**: What kind of situations can a linear fit be used to describe the relationship between an expreimental variable and a response?

- **A**: Any time both the experimental variable and the response variable are numeric.

However all fits are not created the same:



Correlation

Correlation

Visually we can assess if a fitted line does a good job of fitting the data using a scatterplot. However, it is also helpful to have methods of quantifying the quality of that fit.

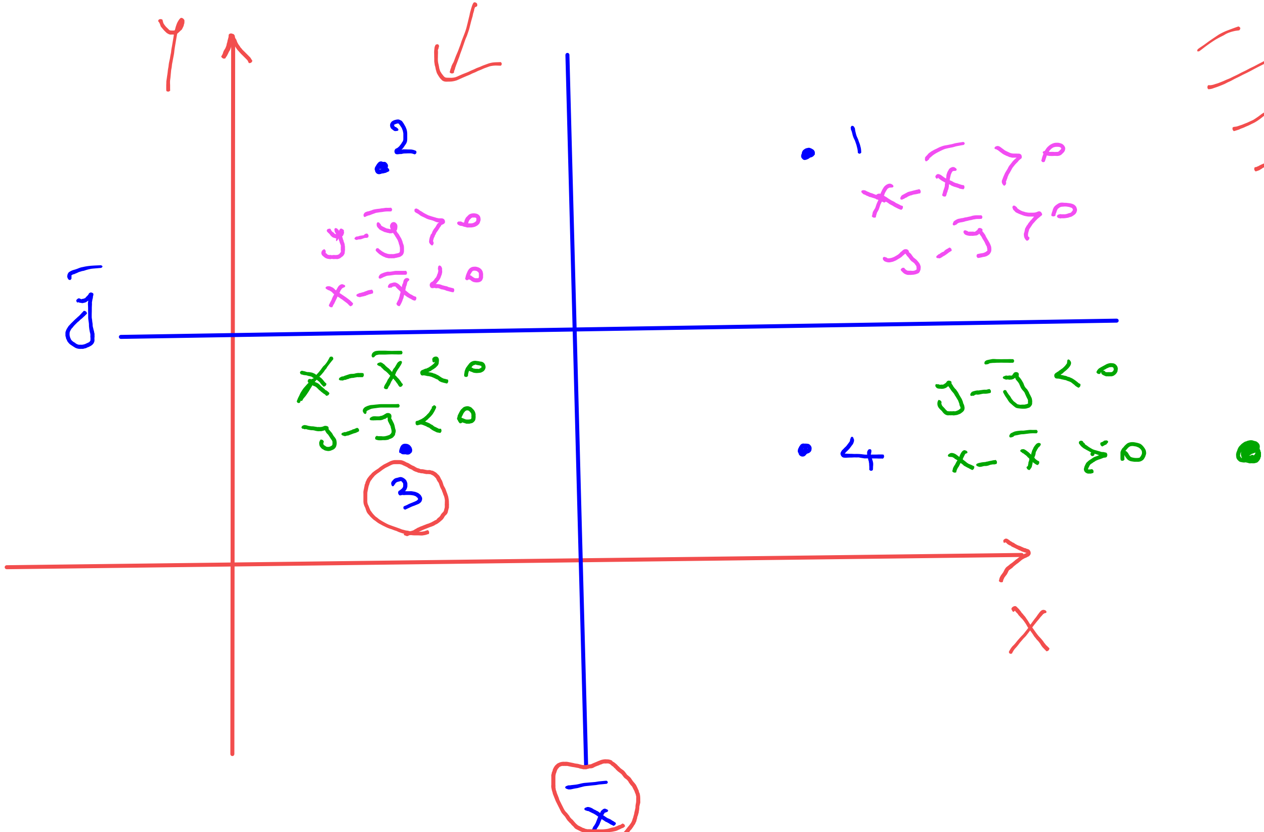
> **Correlation** gives the strength and direction of the linear relationship between two variables.

For a sample consisting of data pairs (x_1, y_1), (x_2, y_2), (x_3, y_3), ... (x_n, y_n), the sample linear correlation, r, is defined by

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

which can also be written as

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



$\Rightarrow x_i > \bar{x}$ $\forall i = 1, \dots, n$

$y_i > \bar{y}$

correlation : $r = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$ always ≥ 0

So, what determines the sign of sample correlation

is

$$\sum (x_i - \bar{x})(y_i - \bar{y})$$

Contribution to r

area ① $\sum (+)(+)$ ⟹ ⊕ ✓

area ② $\sum (-)(+)$ ⟹ ⊖ ✓

area ③ $\sum (-)(-)$ ⟹ ⊕ ✓

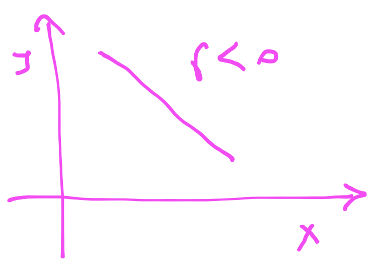
area ④ $\sum (+)(-)$ ⟹ ⊖ ✓

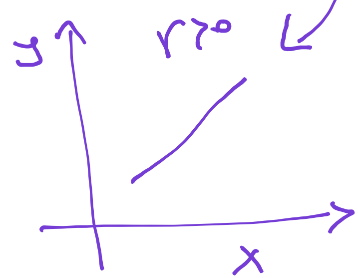
Correlation

1. Sample correlation (aka, sample linear correlation)

The value of r is always between -1 and +1. (or $|r| \leq 1$)

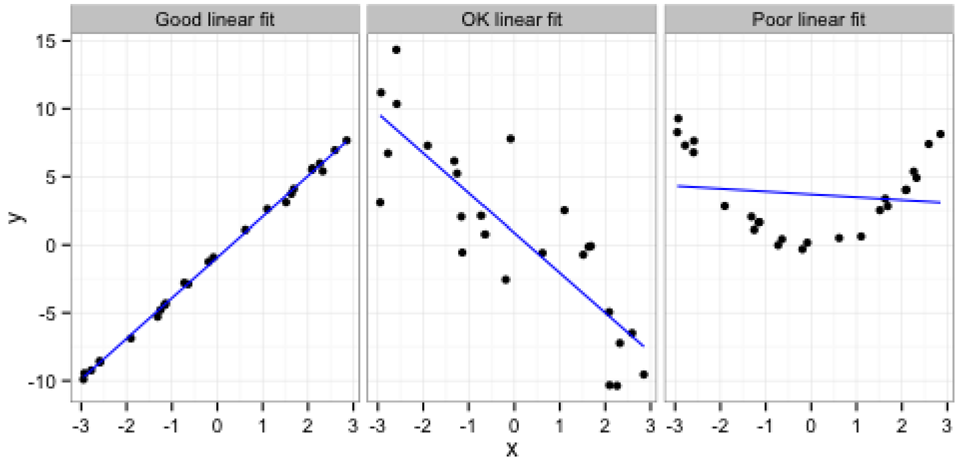
- The closer the value is to -1 or +1 the stronger the linear relationship.
- Negative values of r indicate a negative relationship (as x increases, y decreases).
- Positive values of r indicate a positive relationship (as x increases, y increases).





- One possible rule of thumb:

Range of r	Strength	Direction
0.9 to 1.0	Very Strong	Positive
0.7 to 0.9	Strong	Positive
0.5 to 0.7	Moderate	Positive
0.3 to 0.5	Weak	Positive
-0.3 to 0.3	Very Weak/No Relationship	
-0.5 to -0.3	Weak	Negative
-0.7 to -0.5	Moderate	Negative
-0.9 to -0.7	Strong	Negative
-1.0 to -0.9	Very Strong	Negative



The values of r from left to right are in the plot above are:

```
r=0.9998782        r=-0.8523543    r=-0.1347395
```

- In there first case the linear relationship is almost perfect, and we would happily refer to this as a **very strong**, **positive** relationship between x and y.
- In there second case the linear relationship is seems appropriate - we could safely call it a **strong**, **negative** linear relationship between x and y.
- In there third case the value of r indicates that there is **no linear relationship** between the value of x and the value of y.

1. Sample correlation (aka, sample linear correlation)

Example: Stress and Lifetime of Bars

We can use it to calculate the following values:

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

$$\sum_{i=1}^{10} y_i = 484, \sum_{i=1}^{10} y_i^2 = 25238, \sum_{i=1}^{10} x_i y_i = 8407.5,$$

and we can write:

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

$$= \frac{8407.5 - 10(20)(48.5)}{\sqrt{(5412.5 - 10(20)^2)(25238 - 10(48.4)^2)}}$$

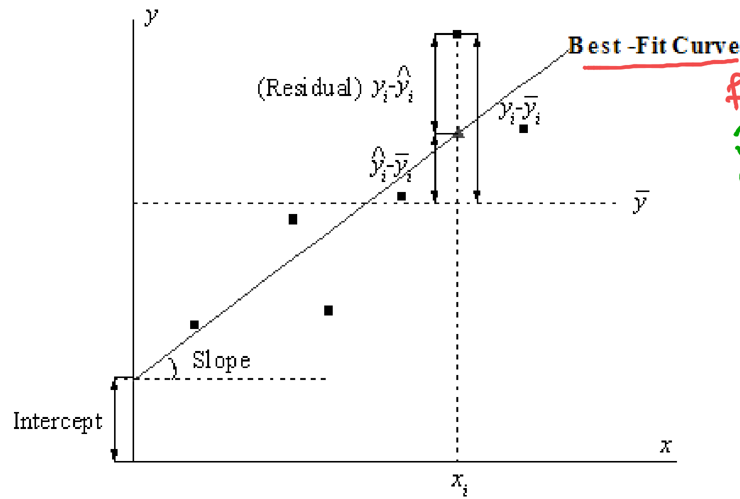
$$= -0.795$$

So we would say that stress applied and lifetime of the bar have a **strong, negative, linear relationship**.

Residuals

Residuals

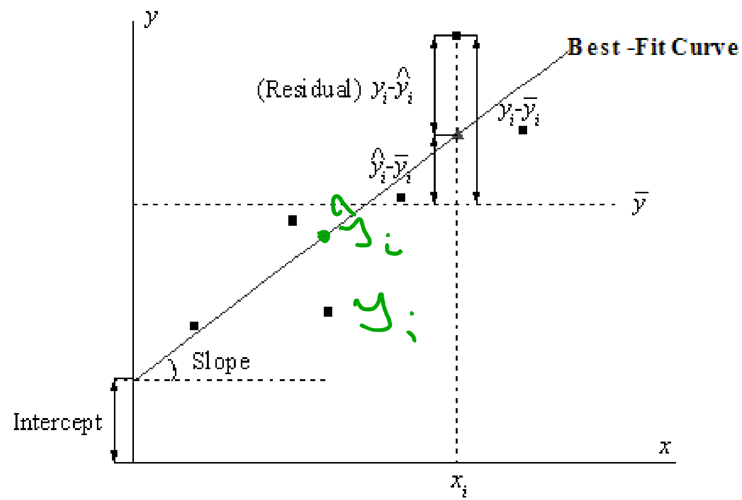
- The "residue" left over from fitting a line



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

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

Residuals



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

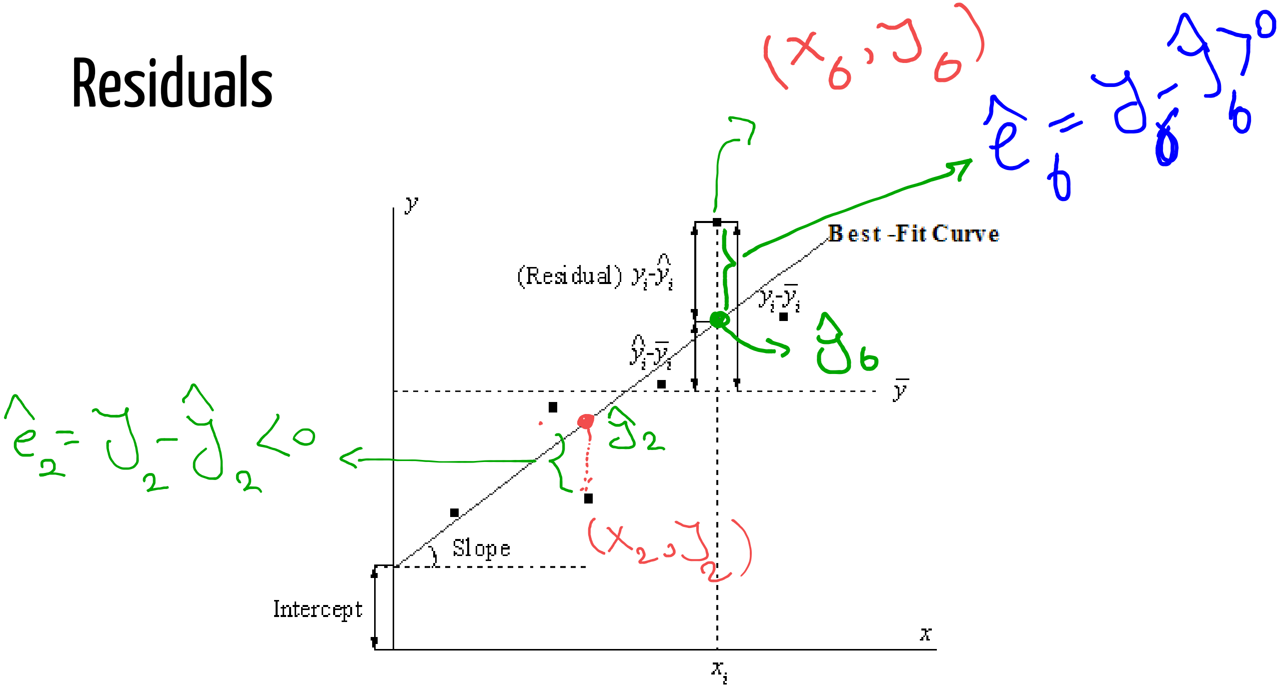
$$e_i = y_i - \hat{y}_i$$

- **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"

obviously, we'd like our residuals to be small

Assessing Models

Assessing models

When modeling, it's important to assess the (1) **validity** and (2) **usefulness** of your model.

To assess the validity of the model, we will look to the residuals. If the fitted equation is the good one, the residuals will be:

- Ptternless (cloud like, random scatter)
- Centered at zero
- Bell shaped distribution

To check if these three things hold, we will use two plotting methods.

> A **residual plot** is a plot of the residuals, $e = y - \hat{y}$ vs. x (or $\hat{y}$ in the case of multiple regression, Section 4.2).

Describing Relationships

Idea

Fitting Lines

Best Estimate

Good Fit

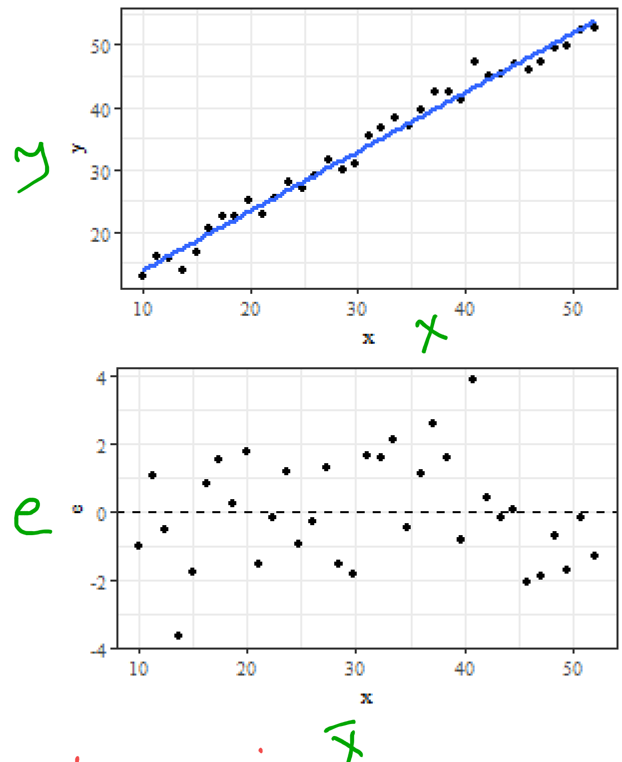
Correlation

Residuals

Assessment

Assessing models

Residual plot



* Patternless residuals → good fit
* centered at zero

Describing Relationships

Idea

Fitting Lines

Best Estimate

Good Fit

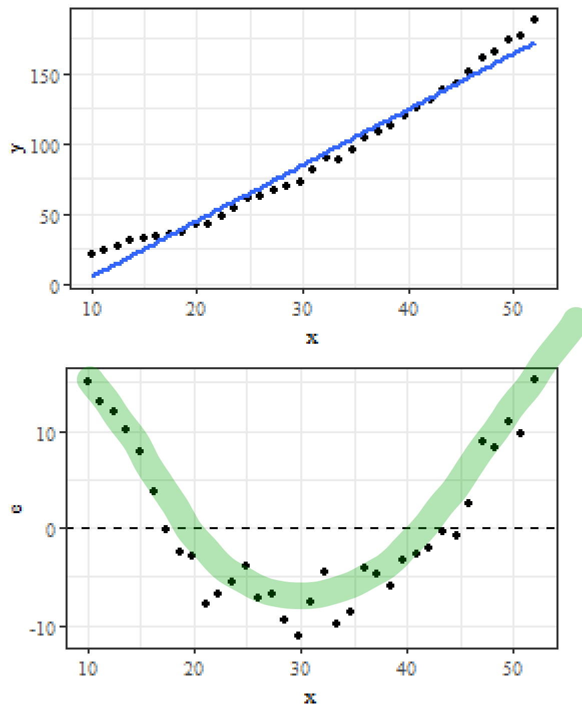
Correlation

Residuals

Assessment

Assessing models

Residual plot



Not a good fit. <=

* Pattern in residual plot
* Not centered at zero

Assessing models

Residual plot