STAT 305: Lecture 9

Chapter 4: Describing Relationships Between Variables

Course page: imouzon.github.io/stat305

Examples

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"

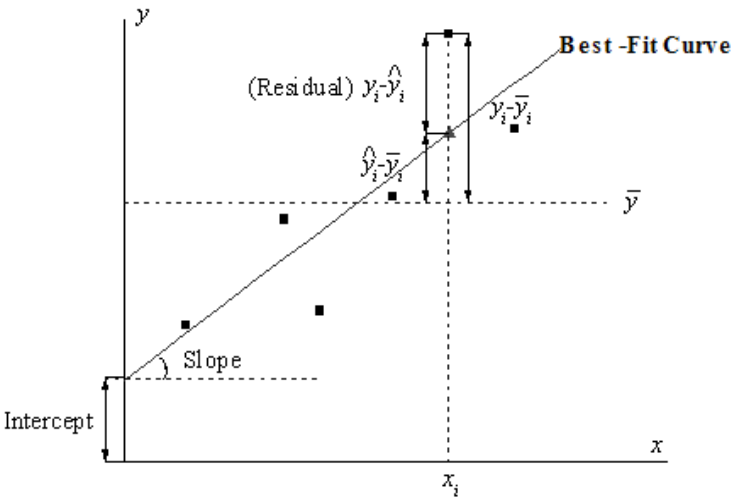
Experiment 4: Changing categorical speed changing applied weight

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"

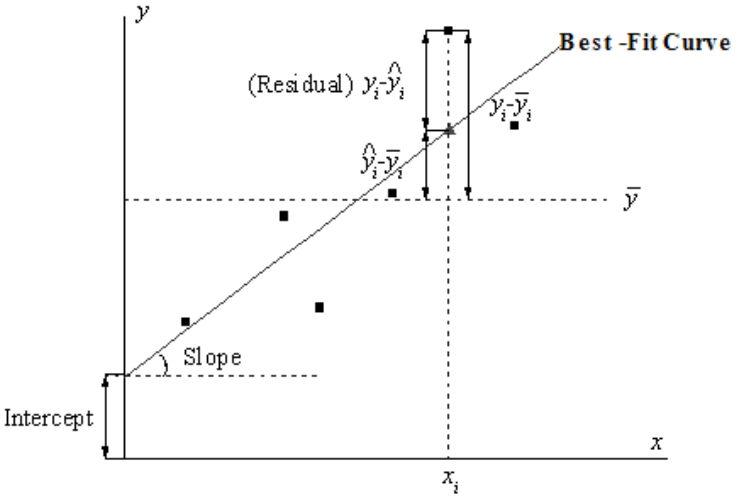
Residuals

- 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:

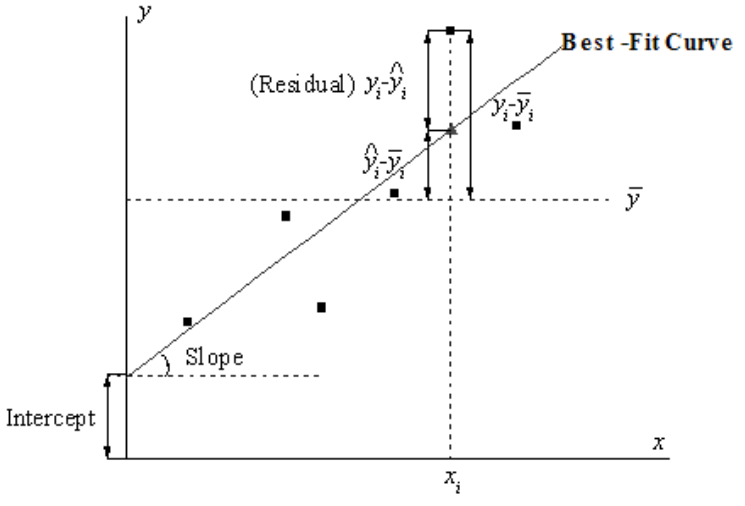
$$e_i = y_i - \hat{y}_i$$

- **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"

Obviously, we would like our residuals to be small compared to the size of response values.

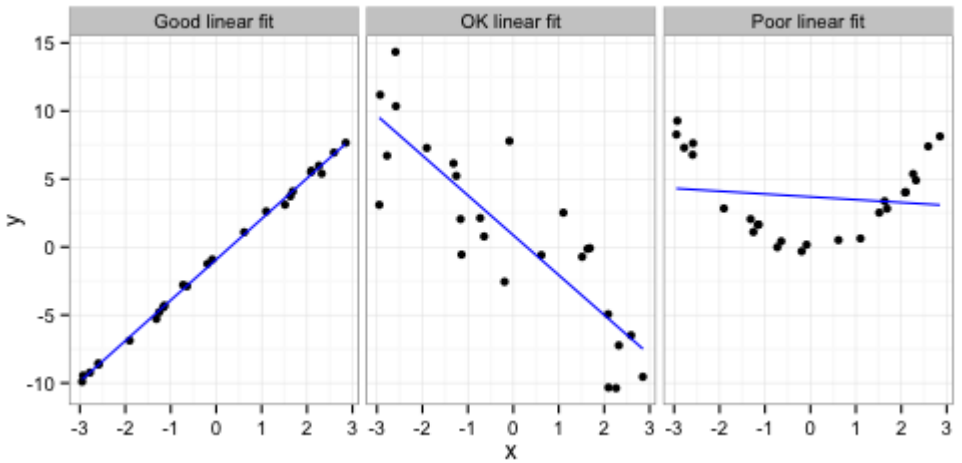
Knowing when a relationship fits the data well

Good Fit

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:



Describing Fit Numerically

1. Sample correlation (aka, sample linear correlation)

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

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

1. Sample correlation (aka, sample linear correlation)

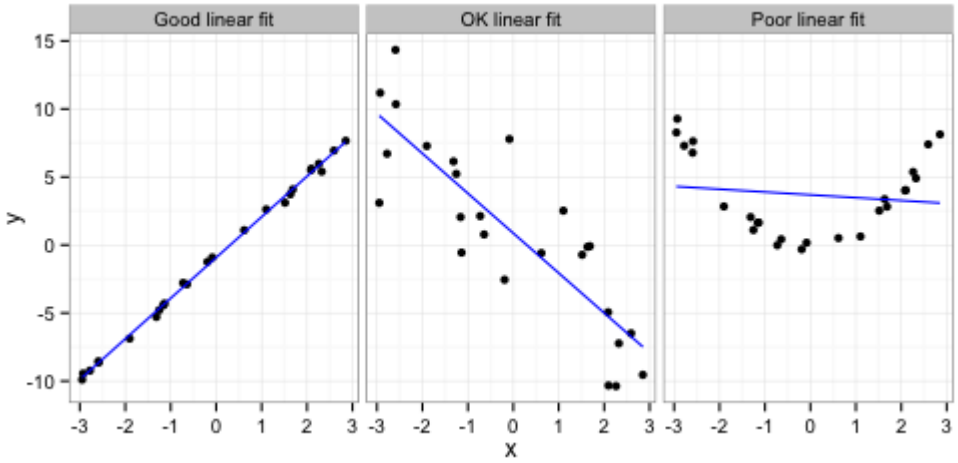
The value of r is always between -1 and +1.

- 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

Good Fit

Numeric Desc.



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

The data can be found in Lecture 9. 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:

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

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

$$= -0.795$$

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

2. Coeffecient of Determination (R^2)

We know that our responses have variability - they are not always the same. We hope that the relationship between our response and our explanatory variables explains some of the variability in our responses.

R^2 is the fraction of the total variability in the response (y) accounted for by the fitted relationship.

- When R^2 is close to 1 we have explained **almost all** of the variability in our response using the fitted relationship (i.e., the fitted relationship is good).

- When R^2 is close to 0 we have explained **almost none** of the variability in our response using the fitted relationship (i.e., the fitted relationship is bad).

There are a number of ways we can calculate R^2. Some require you to know more than others or do more work by hand.

2. Calculating Coeffecient of Determination (R^2)

Method a. Using the data and our fitted relationship:

For an experiment with response values $y_1, y_2, \ldots, y_3$ and fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_3$ we calcuate the following:

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

- This is the longest way to calculate R^2 by hand.
- It requires you to know every response value in the data (y_i) and every fitted value ($\hat{y}_i$)

2. Calculating Coeffecient of Determination (R^2)

Method b. Using Sums of Squares

For an experiment with response values $y_1, y_2, \ldots, y_3$ and fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_3$ we calcuate the following:

- Total Sum of Squares (SSTO): a baseline for the variability in our response.

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

- Error Sum of Squares (SSE): The variability in the data after fitting the line

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

- Regression Sum of Squares (SSR): The variability in the data accounted for by the fitted relationship

$$SSR = SSTO - SSE$$

2. Calculating Coeffecient of Determination (R^2)

Method b. Using Sums of Squares, continued

We can write the R^2 using these sums of squares:

$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- **Q**: What's the advantage of using the sums of squares?
- **A**: The values of SSTO, SSE, and SSR are used in many statistical calculations. Because of this, they are commonly reported by statistical software. For instance, fitting a model in JMP produces these as part of the output.

2. Calculating Coeffecient of Determination (R^2)

Method c. A special case when the relationship is linear

If the relationship we fit between y and x is linear, then we can use the sample correlation, r to get:

$$R^2 = (r)^2$$

NOTE: Please, please, please, understand that this is only true for linear relationships.

Example: Stress and Lifetime of Bars

The data can be found in Lecture 9.

Earlier, we found $r = -0.795$.

Since we are describing the relationship using a line, then we can use the special case:

$$R^2 = (r)^2 = (-0.795)^2 = 0.633$$

In other words, 63.3% of the variability in the lifetime of the bars can be explained by the stress the bars were placed under.