STAT 305: Chapter 4

Part I

Amin Shirazi

Course page:
ashirazist.github.io/stat305_s2020.github.io

Chapter 4, Section 1

Linear Relationships Between Variables

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.

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).

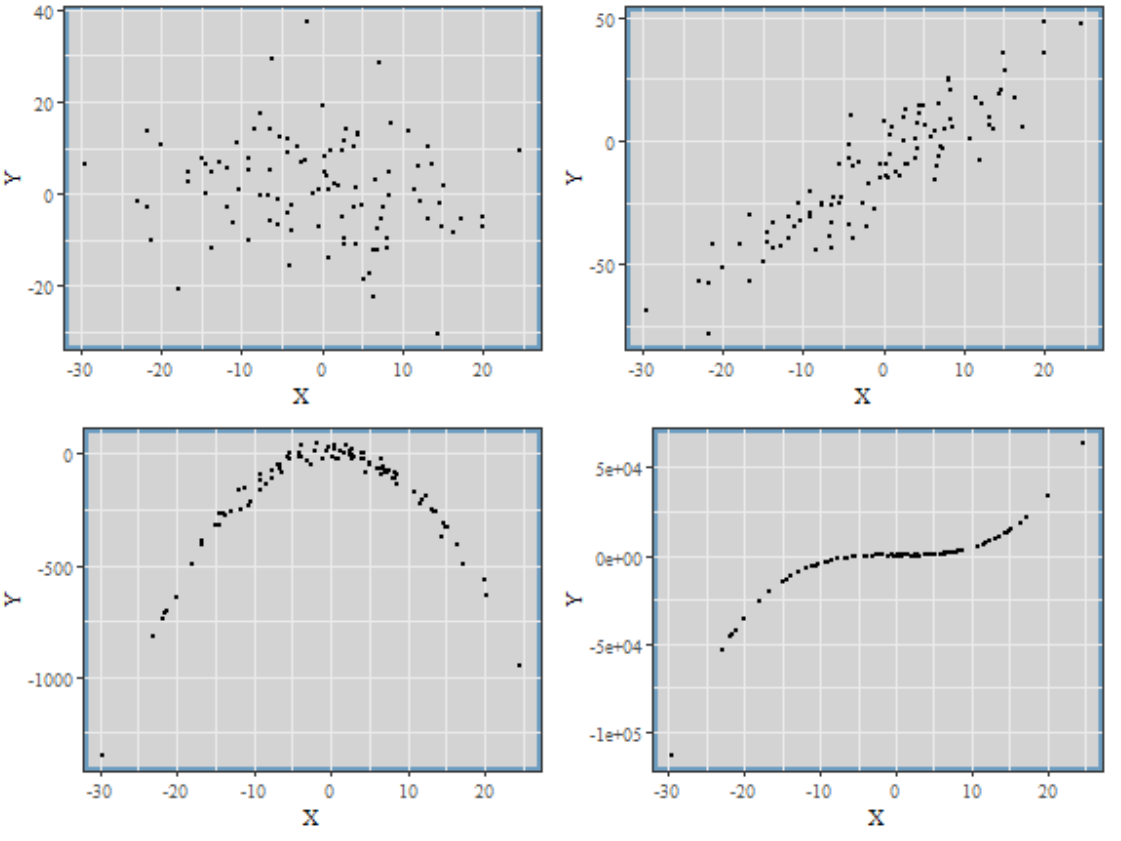
It is helpful to have an equation relating y (the response) to x when the purposes are summarization, interpolation, limited extrapolation, and/or process optimization/ adjusment.

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:

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

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

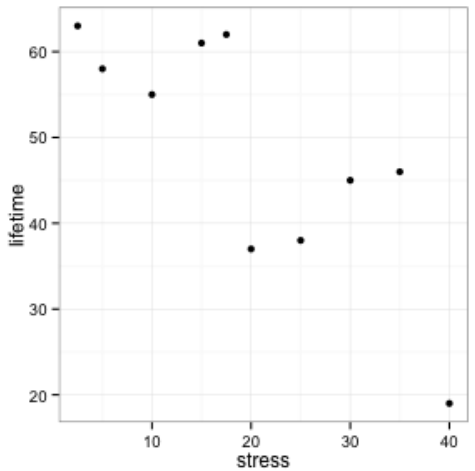
A good first place to investigate the relationship between our experimental variables (in this case, stress) and the response (in this case, lifetime) is to use a scatterplot and look to see if there might be any basic mathematical function that could describe the relationship between the variables.

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

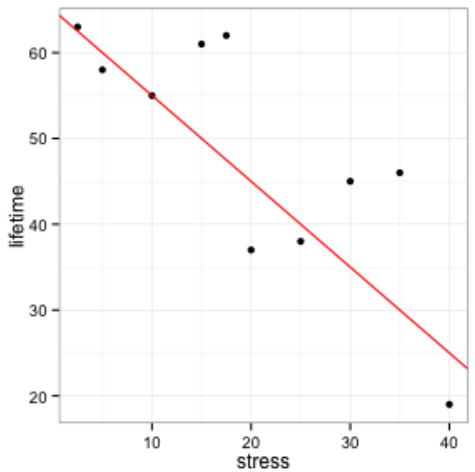


Example: Stress on Bars (continued)

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.

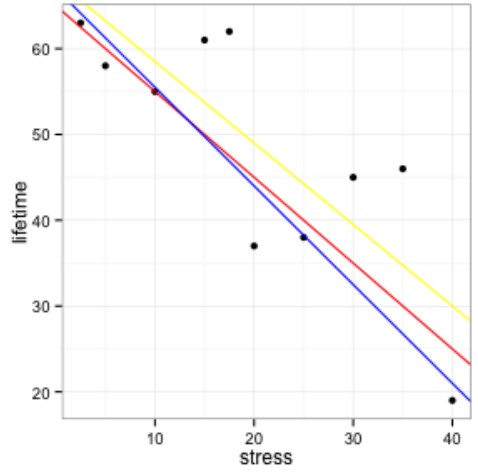


Example: Stress on Bars (continued)

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?

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.

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**

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

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

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

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)

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 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 (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:

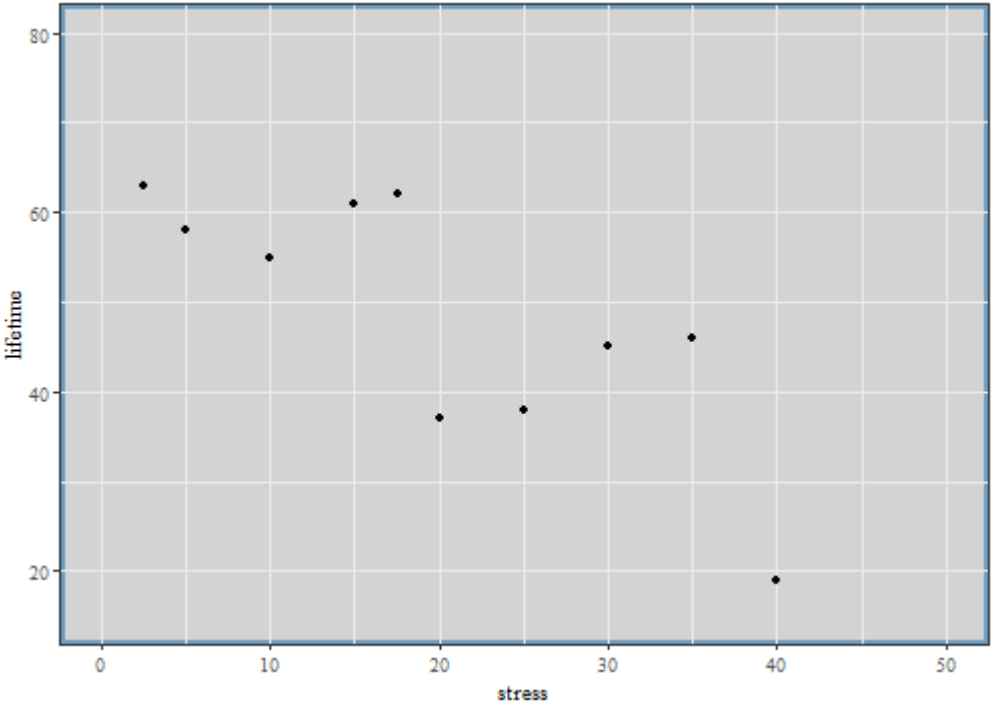
$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} \\ &= \left(\frac{484}{10}\right) - b_1 \left(\frac{200}{10}\right) \\ &= 48.4 - \left(\frac{-1272.5}{1412.5}\right) 20.0 \\ &= 66.4177 \end{aligned}$$

Which gives us the **Fitted Relationship**:

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

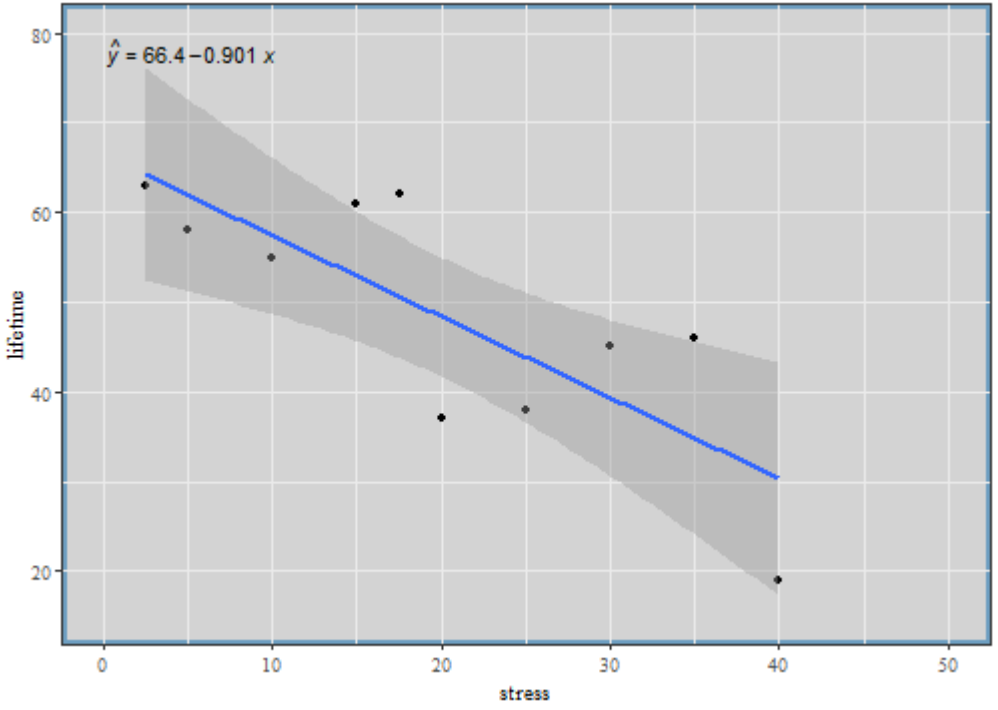
$$\hat{y} = 66.4177 - 0.9009x$$



Example: Stress on Bars

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

Fitted line



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.

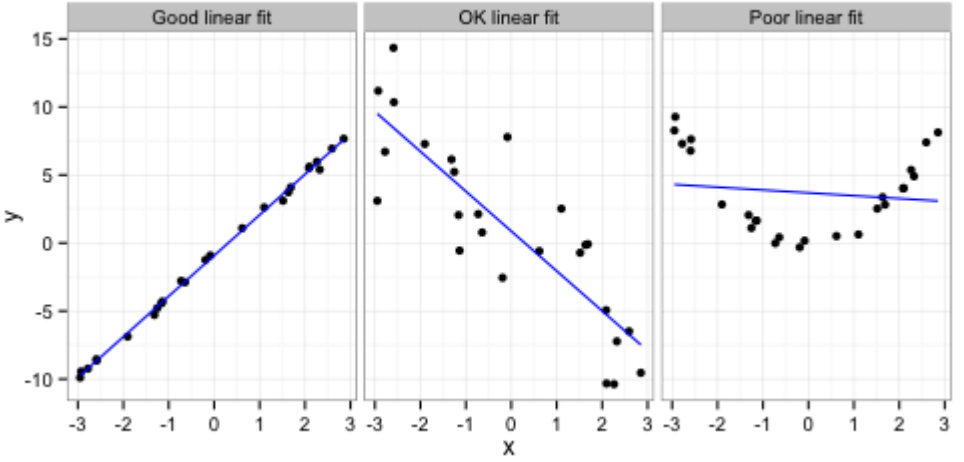
Good Fit

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

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

Correlation

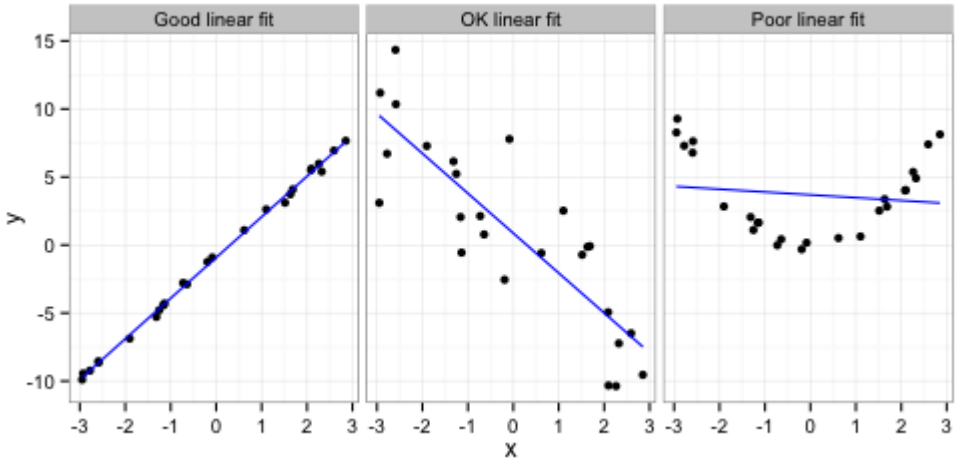
1. Sample correlation (aka, sample linear correlation)

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



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:

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

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

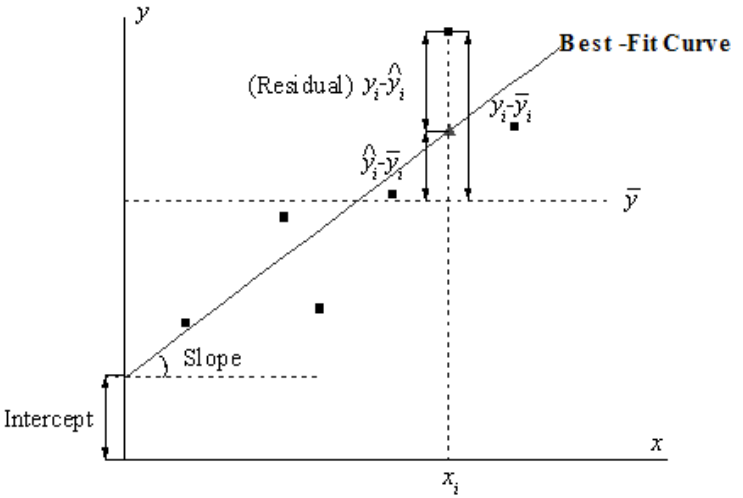
$$= -0.795$$

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

Residuals

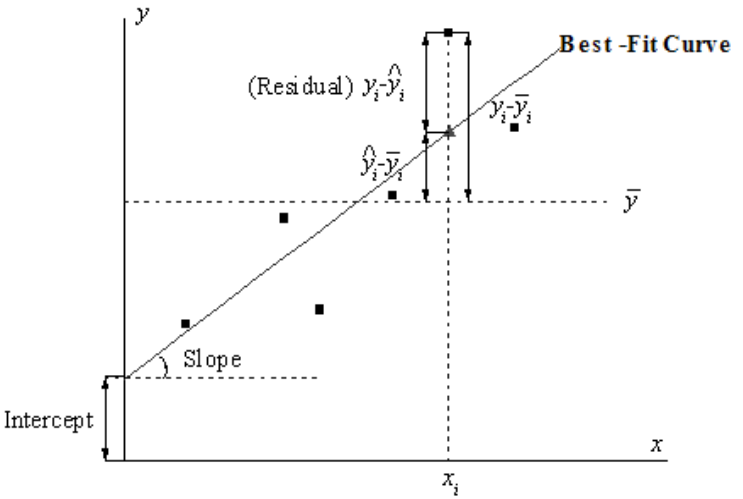
Residuals

- The "residue" left over from fitting a line



- 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:

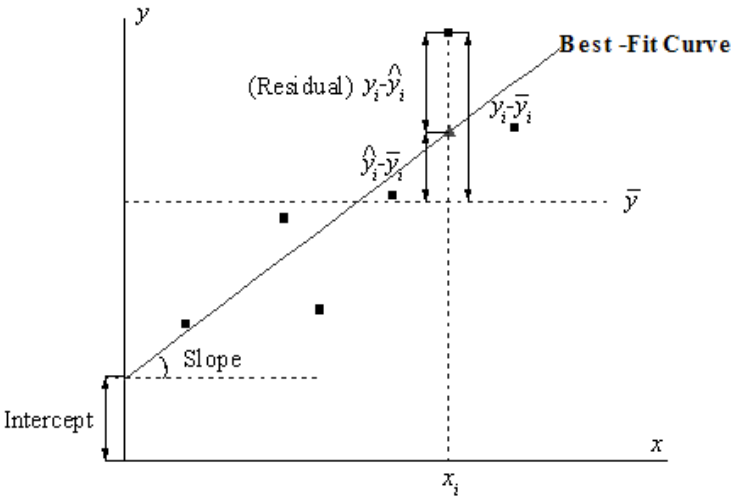
$$e_i = y_i - \hat{y}_i$$

- **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"

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:

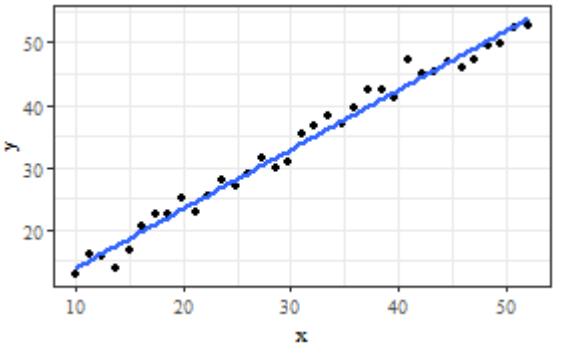
- Ptternless (cloud like, random scatter)
- Centered at zero
- Bell shaped distribution

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).

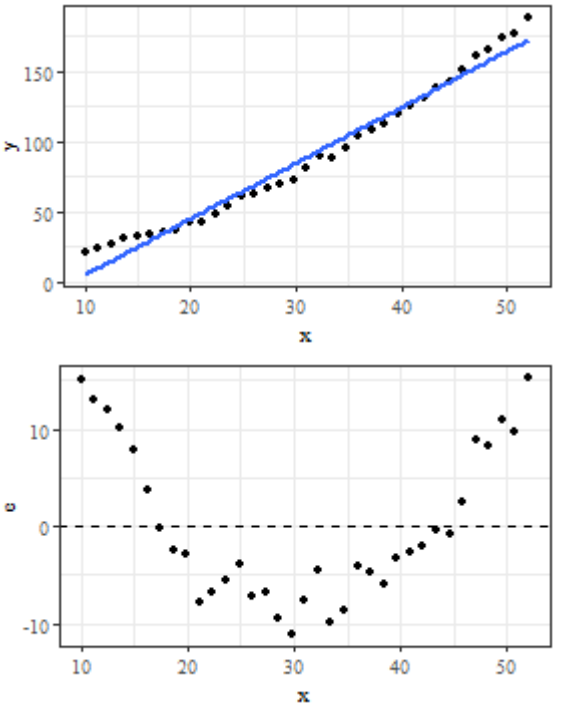
Assessing models

Residual plot



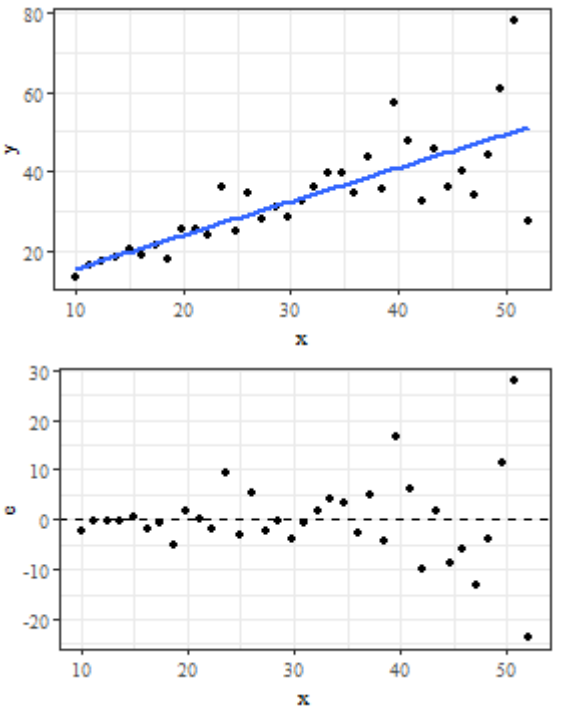
Assessing models

Residual plot



Assessing models

Residual plot

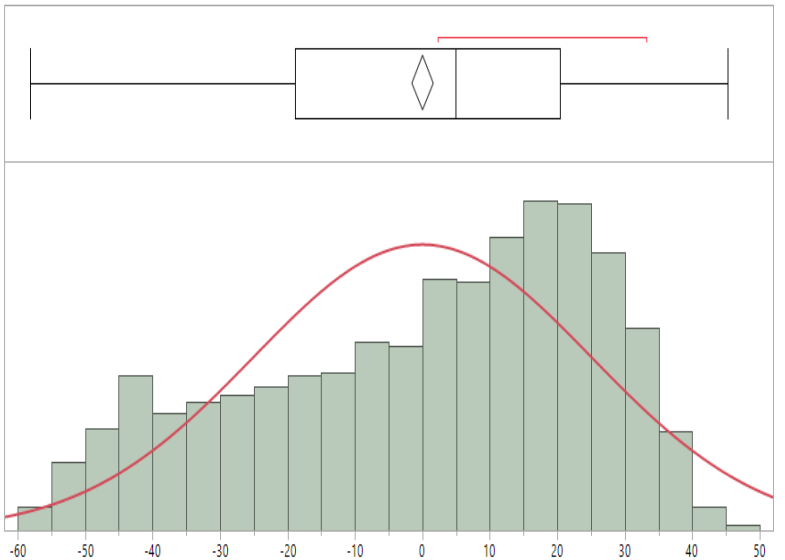


Normality of residuals

- In addition to the residual versus predicted plot, there are other residual plots we can use to check regression assumptions.
- A **histogram of residuals** and a **normal probability plot (QQ-plot)** of residuals can be used to evaluate whether our residuals are approximately normally distributed.
    - However, unless the residuals are far from normal or have an obvious pattern, we generally don't need to be overly concerned about normality.
- Note that we check the residuals for normality. We don't need to check for normality of the raw data. Our response and predictor variables do not need to be normally distributed in order to fit a linear regression model.

Normality of residuals

Draw a histogram of the residuals (review the JMP toturial for histograms)

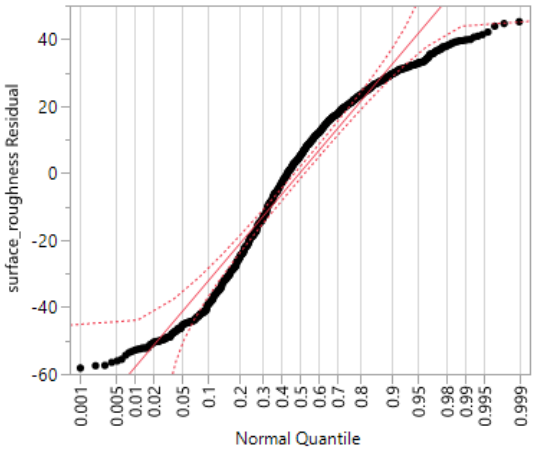


It seems the residuals are not normaly distributed in this example.The residuals have a left skewed distirbution.

Normality of residuals

As the instructions on the JMP toturials (and also HW #3), you can draw **Normal QQ-plot** to evaluate if the residuals meet the assumptions of normaly distributed.

Plotting Normal QQ-plot of the same example



- Again, the QQ-plot also confirms that the assumption of Normal distribution of residuals is violated to some extend in this example.
- More examination is required to fix the issue or to find the problem.

Coefficient of Determination

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.

Calculating Coeffecient of Determination (R^2

Method a. Using the data and our fitted relationship:

For an experiment with response values $y_1, y_2, \ldots, y_n$ and fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ 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 ($\hat{y}_i$)

Calculating Coeffecient of Determination (R^2

Method b. Using Sums of Squares

For an experiment with response values $y_1, y_2, \ldots, y_n$ and fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ 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

$$SSR = SSTO - SSE$$

Calculating Coeffecient of Determination (R^2

Method b. Using Sums of Squares

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.

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.

Calculating Coeffecient of Determination (R^2

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

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 linear relationship between the stress the bars were placed under and the lifetime.

Precautions

Precautions about Simple Linear Regression (SLR)

- r only measures linear relationships
- R^2 and r can be drastically affected by a few unusual data points.

Using a computer

You can use JMP (or `R`) to fit a linear model. See BlackBoard for videos on fitting a model using JMP.

Section 4.2

Fitting Curves and Surfaces by Least Squares

Multiple Linear Regression

Linear Relationships

- The idea of simple linear regression can be generalized to produce a powerful engineering tool: **Multiple Linear Regression** (MLR).
- SLR is associated with **line fitting**
- MLR is associated with **curve fitting and surface fitting**
- What we mean by multiple **linear** relationship is that the relation between the variables and the response is linear **in their parameters**.
  - **Multiple linear regression in general:** when there are more than one experimental variable in the experiment

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

  - **polynomial equation of order k:**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + + \beta_3 x^3 + \cdots + \beta_k x^k$$

Non-Linear Relationships

- And there are also **non-linear relationship** where the relationship between the variables and the response is non-linear **in their parameters**.

$$y = \beta_0 + e^{\beta_1} x$$

$$y = \frac{\beta_0}{\beta_1 + \beta_2 x}$$

An issue

- The point is that fitting curves and surfaces by the least square method needs a lot of matrix algebra concepts and it is difficult to be done by hand.
- We need software to fit surfaces and curves.

Example

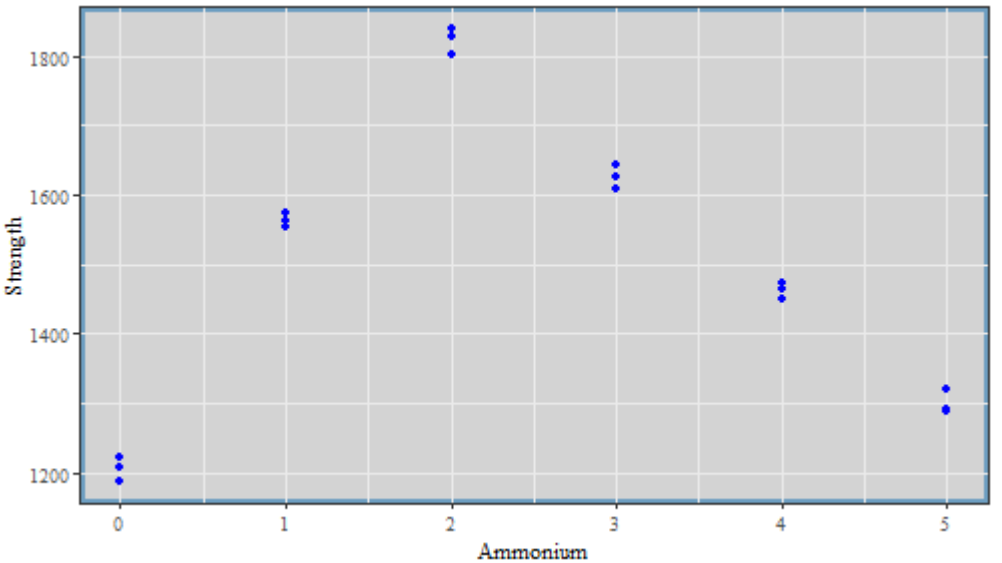
Example:

Compressive Strength of Fly Ash Cylinders as a Function of Amount of Ammonium Phoshate Additive

Ammonium Phosphate(%)	Compressive Strength (psi)	Ammonium Phosphate(%)	Compressive Strength (psi)
0	1221	3	1609
0	1207	3	1627
0	1187	3	1642
1	1555	4	1451
1	1562	4	1472
1	1575	4	1465
2	1827	5	1321
2	1839	5	1289
2	1802	3	1292

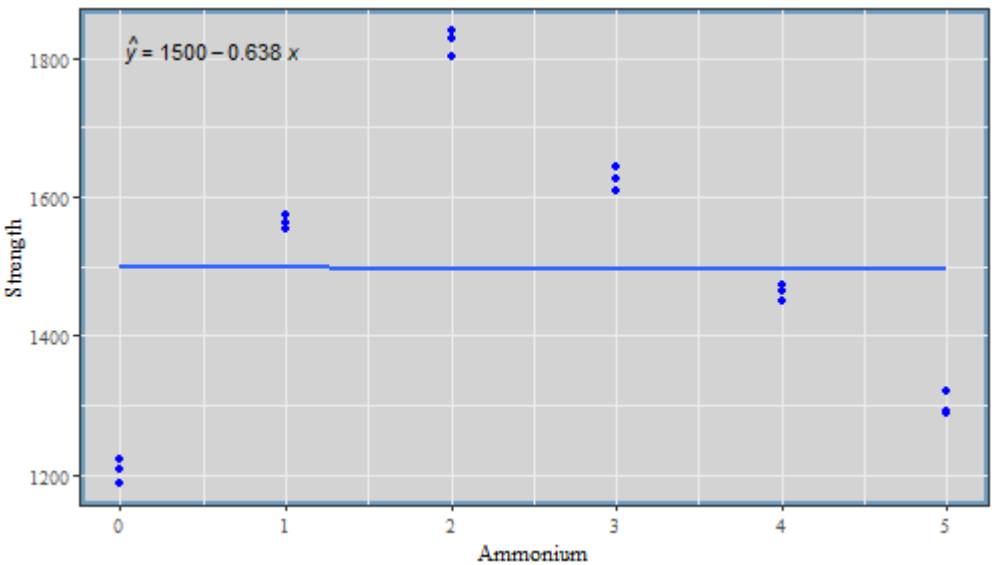
Example:

Compressive Strength of Fly Ash Cylinders as a Function of Amount of Ammonium Phoshate Additive



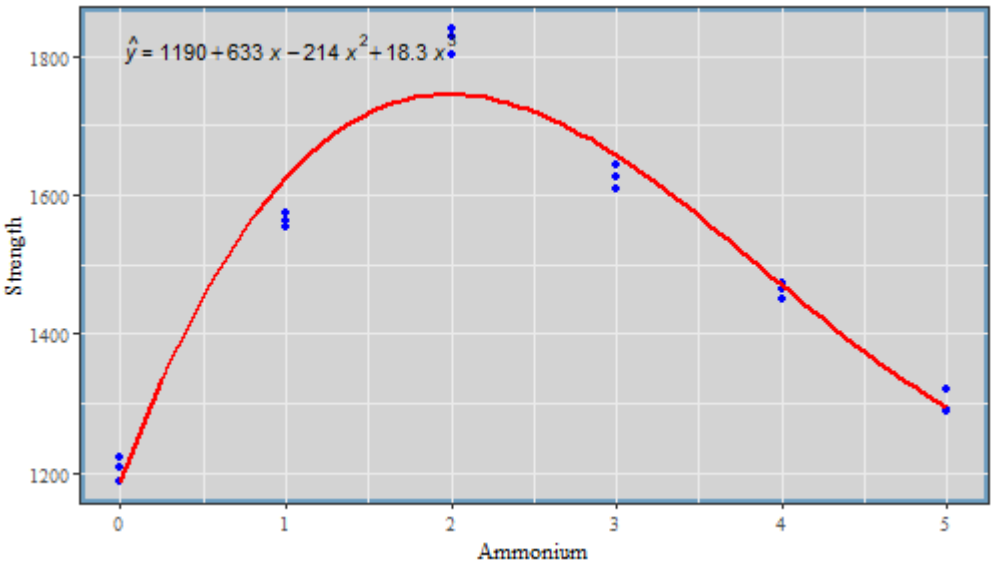
Example:

Compressive Strength of Fly Ash Cylinders as a Function of Amount of Ammonium Phoshate Additive



Example:

Compressive Strength of Fly Ash Cylinders as a Function of Amount of Ammonium Phoshate Additive



One More Example in Fitting Surface and Curves

Example: Hardness of Alloy

A group of researchers are studying influences on the hardness of a metal alloy. The researchers varied the percent copper and tempering temperature, measuring the hardness on the Rockwell scale.

The goal is to describe a relationship between our response, Hardness, and our two experimental variables, the percent copper (x_1) and tempering temperature (x_2).

Example: Hardness of Alloy

Percent Copper	Temperature	Hardness
0.02	1000	78.9
	1100	65.1
	1200	55.2
	1300	56.4
0.10	1000	80.9
	1100	69.7
	1200	57.4
	1300	55.4
0.18	1000	85.3
	1100	71.8
	1200	60.7
	1300	58.9

Example: Hardness of Alloy

Theoretical Relationship:

We start by writing down a theoretical relationship. With one experimental variable, we may start with a line. Extending that idea for two variables, we start with a plane:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Observed Relationship:

In our data, the true relationship will be shrouded in error.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{errors}$$

$$= [\quad \text{signal} \quad] + [\text{noise}]$$

Example: Hardness of Alloy

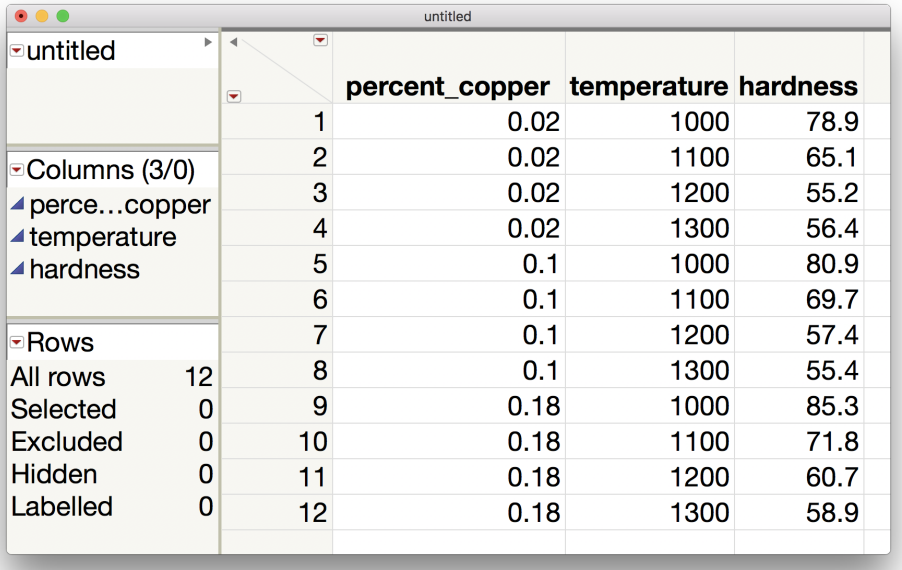
Fitted Relationship:

If we are right about our theoretical relationship, though, and the signal-to-noise ratio is small, we might be able to estimate the relationship:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Example: Hardness of Alloy

Enter the data in JMP



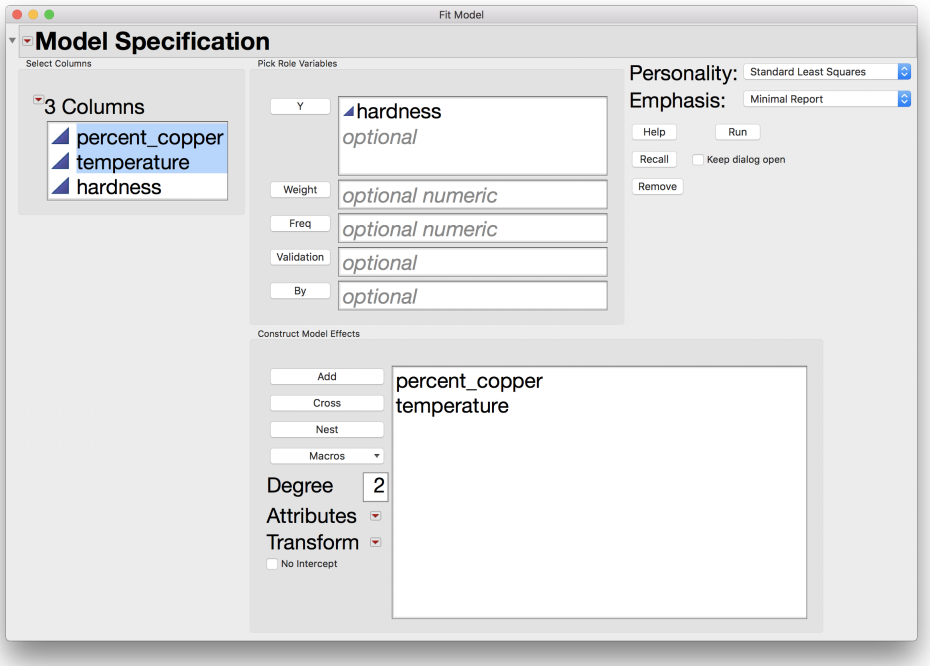
	percent_copper	temperature	hardness
1	0.02	1000	78.9
2	0.02	1100	65.1
3	0.02	1200	55.2
4	0.02	1300	56.4
5	0.1	1000	80.9
6	0.1	1100	69.7
7	0.1	1200	57.4
8	0.1	1300	55.4
9	0.18	1000	85.3
10	0.18	1100	71.8
11	0.18	1200	60.7
12	0.18	1300	58.9

Example: Hardness of Alloy

In JMP, go to

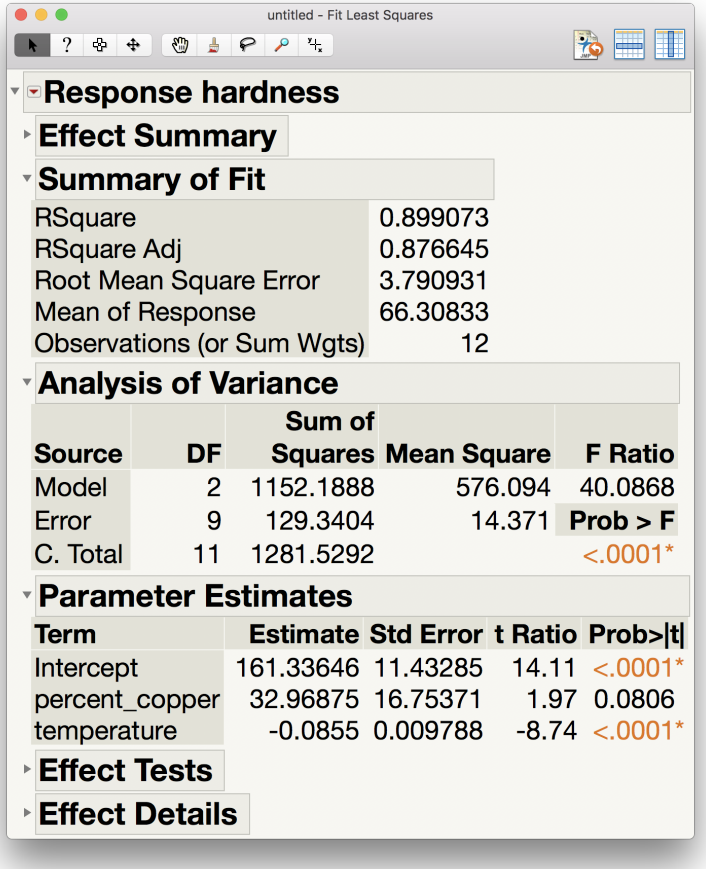
> `Analyze > Fit Model`

to define the model you are fitting:



Example: Hardness of Alloy

After clicking Run we get the following model fit results:



Example: Hardness of Alloy

From this output, we can get the value of R^2, the coeffecient of determination:

Summary of Fit	
RSquare	0.899073
RSquare Adj	0.876645
Root Mean Square Error	3.790931
Mean of Response	66.30833
Observations (or Sum Wgts)	12

Since $R^2 = 0.899073$, we can say

> 89.9074% of the variability in the hardness we observed can be explained by its relationship with temperature and percent copper.

Example: Hardness of Alloy

From this output, we can get the sum of squares.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	1152.1888	576.094	40.0868
Error	9	129.3404	14.371	Prob > F
C. Total	11	1281.5292		<.0001*

This "Analysis of Variance" table has the same format across almost all textbooks, journals, software, etc. In our notation,

- $SSR = 1152.1888$
- $SSE = 129.3404$
- $SSTO = 1281.5292$

We can use these for lots of purposes. In this class, we have seen that we can get R^2:

$$R^2 = 1 - \frac{SSE}{SSTO} = 1 - \frac{129.3404}{1281.5292} = 0.8990734$$

Example: Hardness of Alloy

The parameter estimates give us the fitted values used in our model:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	161.33646	11.43285	14.11	<.0001*
percent_copper	32.96875	16.75371	1.97	0.0806
temperature	-0.0855	0.009788	-8.74	<.0001*

Since we defined percent copper as x_1 earlier and temperature as x_2 then we can write:

$$\hat{y} = 161.33646 + 32.96875 \cdot x_1 - 0.0855 \cdot x_2$$

We can use this to get fitted values. If we use temperature of 1000 degrees and percent copper of 0.10 then we would predict a hardness of

$$\hat{y} = 161.33646 + 32.96875 \cdot (0.10) - 0.0855 \cdot (1000)$$

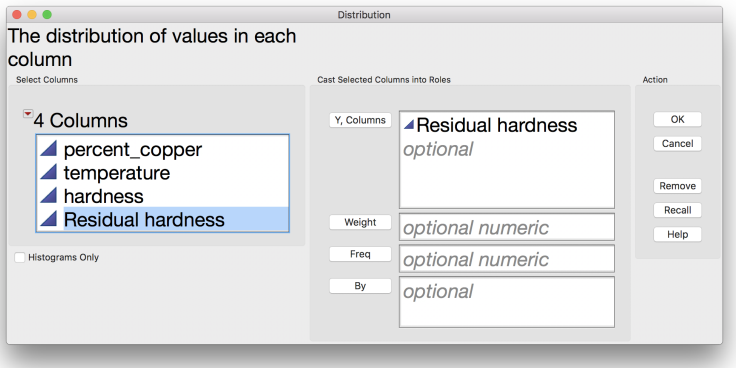
$$= 161.33646 + 3.296875 - 85.5$$

$$= 79.13333$$

Example: Hardness of Alloy

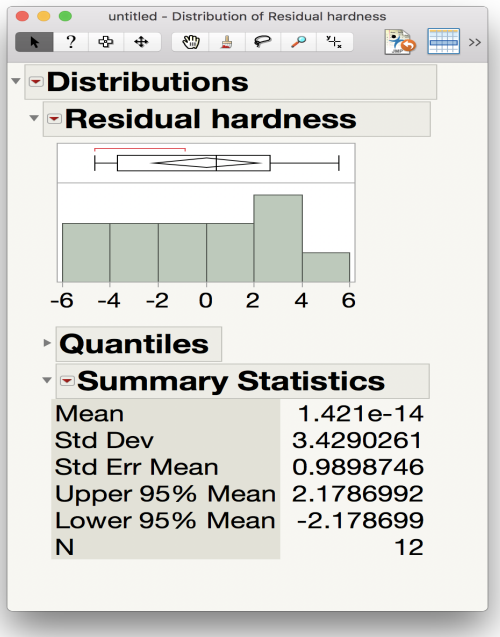
While our model looks pretty good, we still need to check a few things involving residuals. We can save our residuals from the model fit drop down and analyze them.

From Analyze > Distribution:



Example: Hardness of Alloy

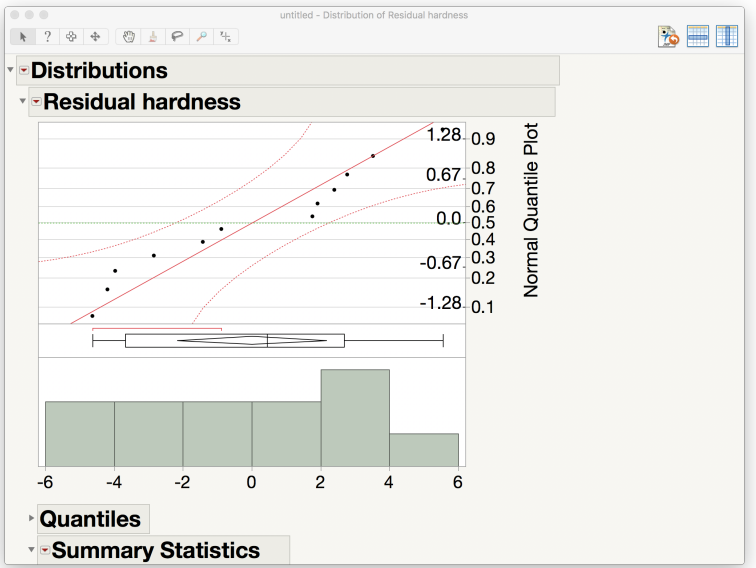
There aren't many residuals here (just 12) but we would like to make sure that the histogram has rough bell-shape (normal residuals are good). I would call this one inconclusive.



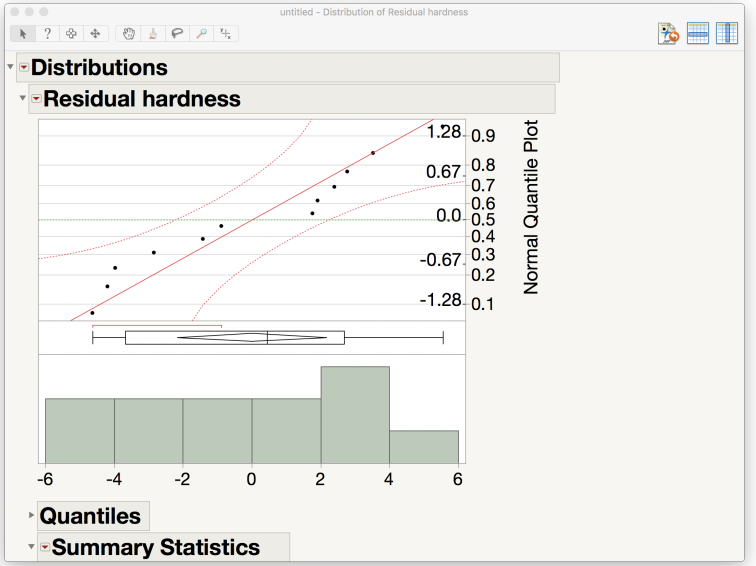
Example: Hardness of Alloy

Another way to check if the residuals are approximately normal is to compare the quantiles of our residuals to the theoretical quantiles of the true normal distribution.

From the dropdown menu, choose Normal Quantile Plot to get:



Example: Hardness of Alloy

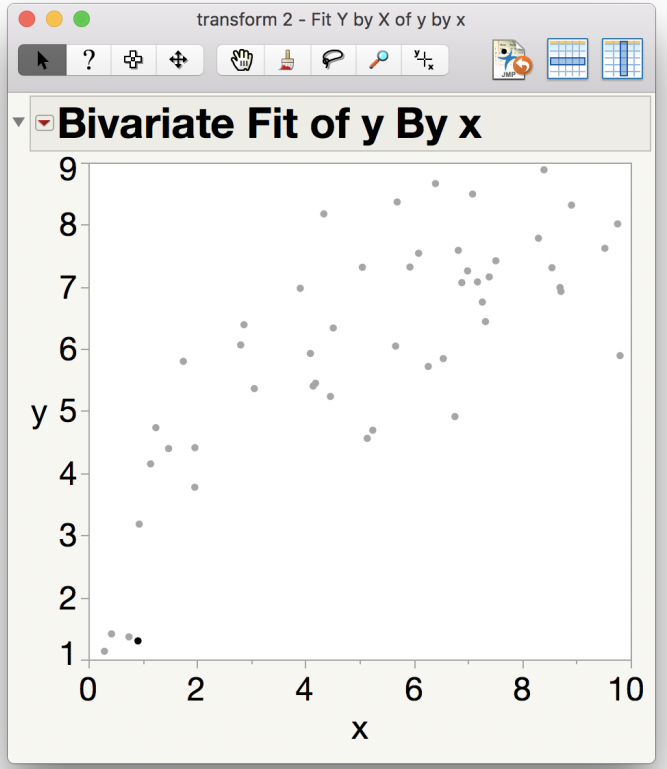


- If the points all fall on the line, then the residuals have the same spread as the normal distribution (i.e., the residuals follow a bell-shape, which is what we want).
- If they stay within the curves, then we can say the residuals follow a rough bell shape (which is good).
- If points fall outside the curves, our model has problems (which is bad).

Transformations

Transformations: Fitting complicated relationships

Consider the simulated dataset 'transform.csv' in the lecture module. Here's the scatterplot:



Transformations: Fitting complicated relationships

Consider the residual plot you would get by trying to fit a line. What would that look like?

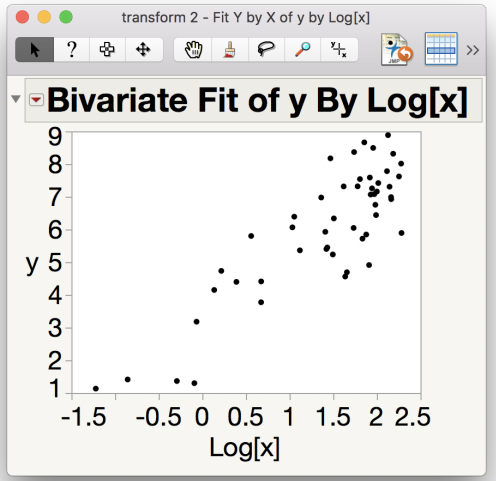
Now consider the residual plot you would get by trying to fit a quadratic. What would that look like?

What can we do about the size of the residuals??

We need a function that can both adjust the scale our responses and account for the curve!!

Transformations: Fitting complicated relationships

One possible function that could do that: $ln(x)$.



Transforming our variables can allow us to get better fits, but you need to be careful about the meaning of the relationship. For instance, the slope now means "the change in the response when *the natural log of x is increased by 1* - the relationship to x itself is not always easy to translate back.

Dangers in Fits

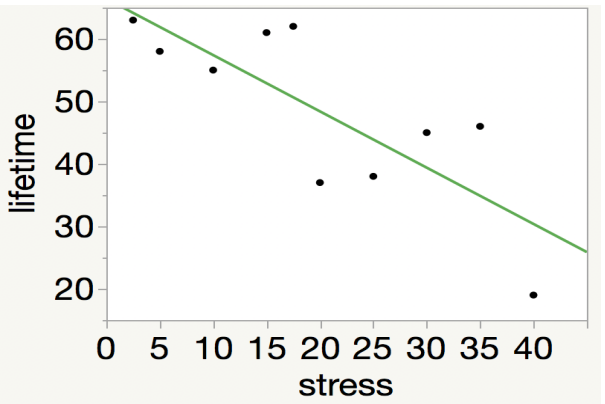
Dangers in Fitting Relationships

Example: Stress and Lifetime of Bars

Consider the bars example again

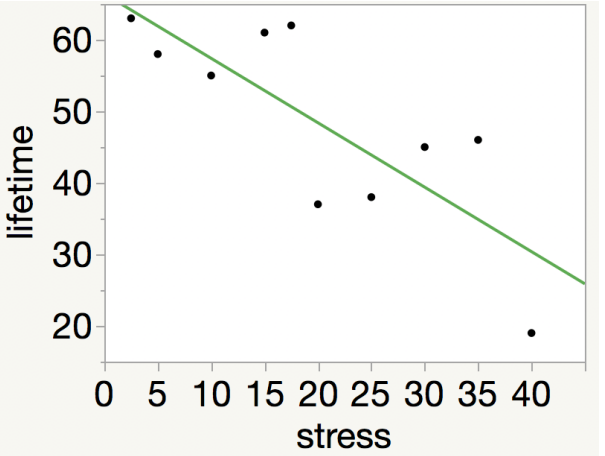
stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

Here's the linear fit:



Dangers in Fitting Relationships

Example: Stress and Lifetime of Bars



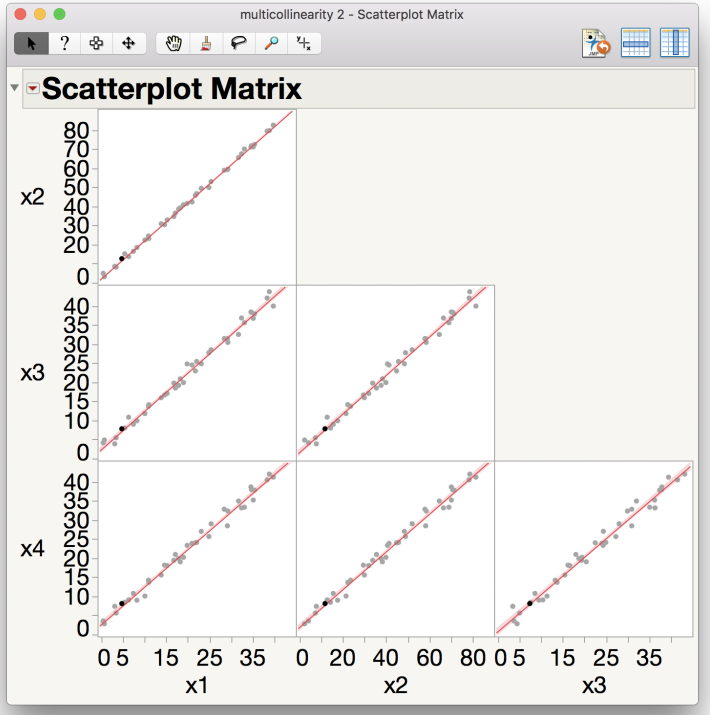
The fitted line doesn't touch all the points, but we can push our relationship further by adding $(stress)^2$, $(stress)^3$, $(stress)^4$, and so on.

Everytime we add a new term to the polynomial, we give the fitted relationship the ability to make one more turn.

This leads to a problem called **overfitting**: our model is just following *the data*, including the errors, instead of uncovering *the true relationship*.

Multicollinearity

Multicollinearity occurs when you have strongly correlated experimental variables.



Multicollinearity

Multicollinearity can lead to several problems:

- Since the variables are all related to each other, the impact each variable has in the relationship to the response becomes difficult to determine
- Since the disentangling the relationships is difficult, the estimates of the slopes for each variable become very sensitive (different samples lead to very different estimates)
- Since the correlated experimental variables will have similar relationships to the response, most of them are not needed. Including them leads to an overfit.

Ultimately while it may look like a good fit on paper, the model will be inaccurate.

Finding the Best Fit

- Again, we can use the **Least Squares** principle to find the best estimates, b_0, b_1, and b_2.
- The calculations are fairly advanced now that we have three values to estimate,
- so these calculations are usually done in statistical software (like JMP).

Judging The Fit

- Not all Theoretical Relationships we may imagine are real!
- Perhaps a better relationship could be found using

$$y = \beta_0 + \beta_1 x_1 + \beta_2 \ln(x_2)$$

- We determine which relationships to try by examining plots of the data, fit statistics (like R^2), and plots of residuals.
- Be careful of overfitting and multicollinearity (when the experimental variables are correlated).