STAT 305: Chapter 4

Part II

Amin Shirazi

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

Good Fit

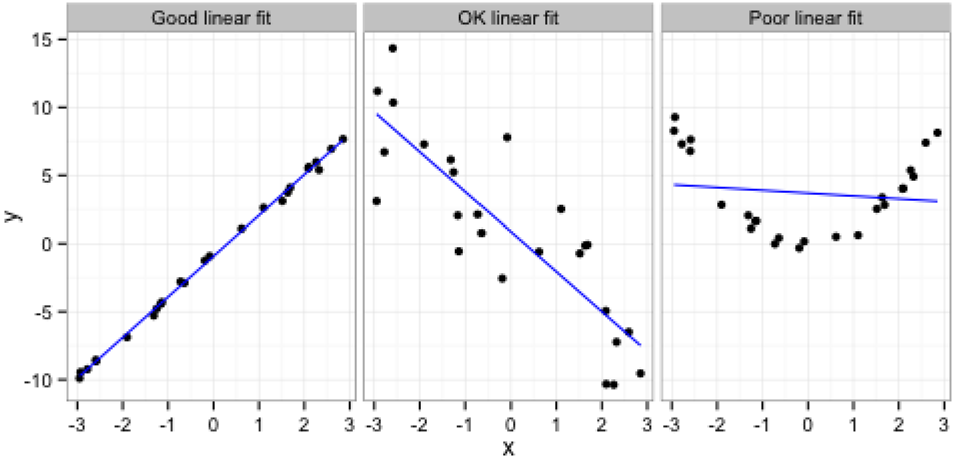
Good Fit

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:



Good Fit

Describing Fit Numerically

Numeric Desc.

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

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

Good Fit

Numeric Desc.

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).

Good Fit

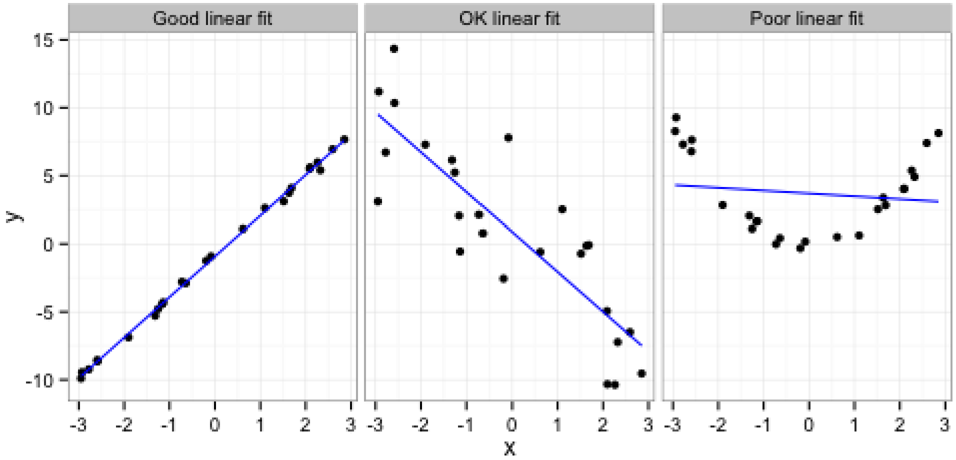
Numeric Desc.

- 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.

Good Fit

Numeric Desc.

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**.

Good Fit

Numeric Desc.

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.

Good Fit

Numeric Desc.

2. Calculating Coeffecient of Determination (R^2)

Method a. Using the data and our fitted relationship:

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

Good Fit

Numeric Desc.

2. Calculating Coeffecient of Determination (R^2)

Method b. Using Sums of Squares

For an experiment with response values $y_1, y_2, \ldots, y_n$ 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

$$SSR = SSTO - SSE$$

Good Fit

Numeric Desc.

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.

Good Fit

Numeric Desc.

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.

Good Fit

Numeric Desc.

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

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.

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

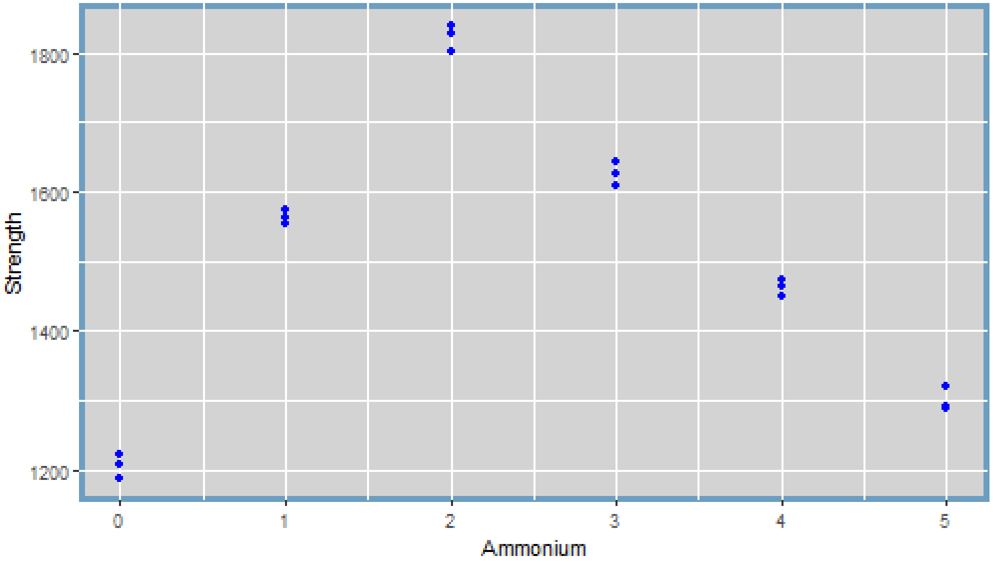
- The point is that fitting curves and surfaces by the least square method needs a lot of matrix algebra concepts and it is difficult to be done by hand.
- We need software to fit surfaces and curves.

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

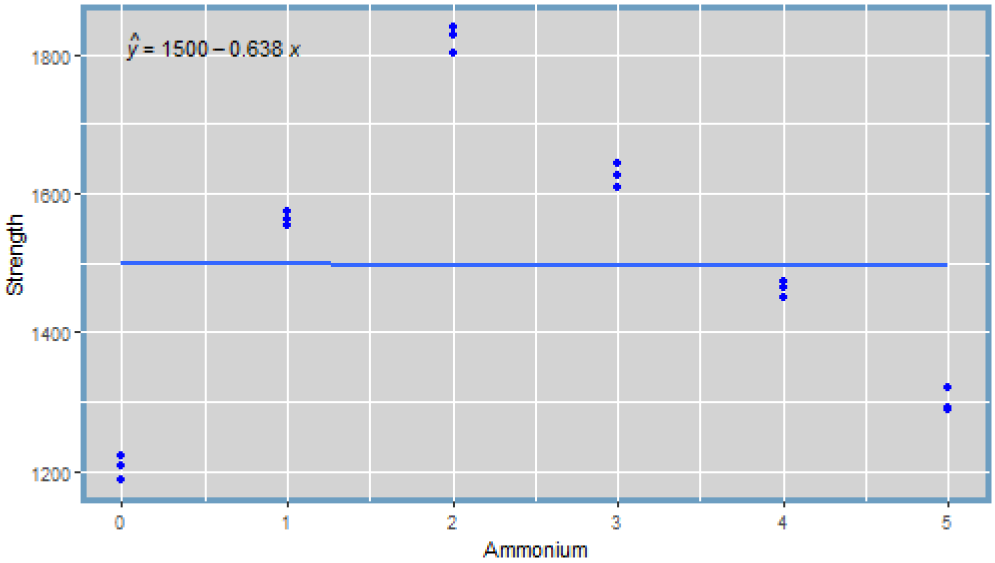


Fitting Curves

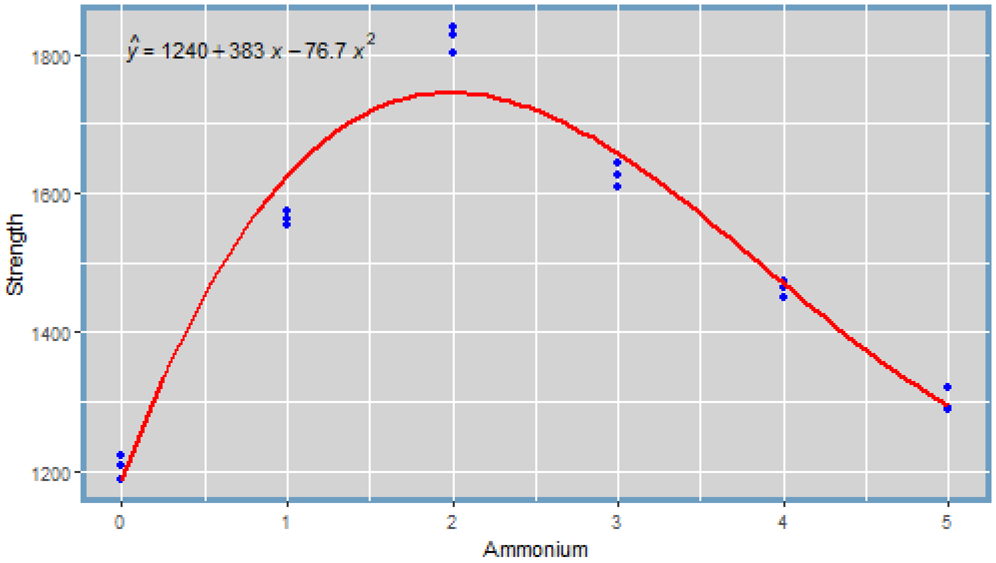
MLR

Example

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).

Fitting Curves

MLR

Ex: Hard Alloy

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.

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

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

Example: Hardness of Alloy

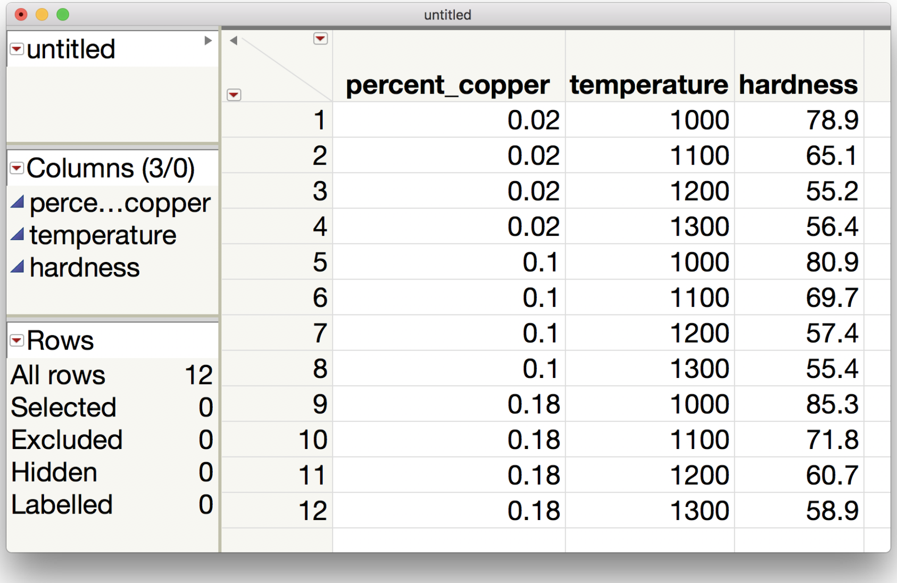
Fitted Relationship:

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



untitled

Columns (3/0)

- perce…copper
- temperature
- hardness

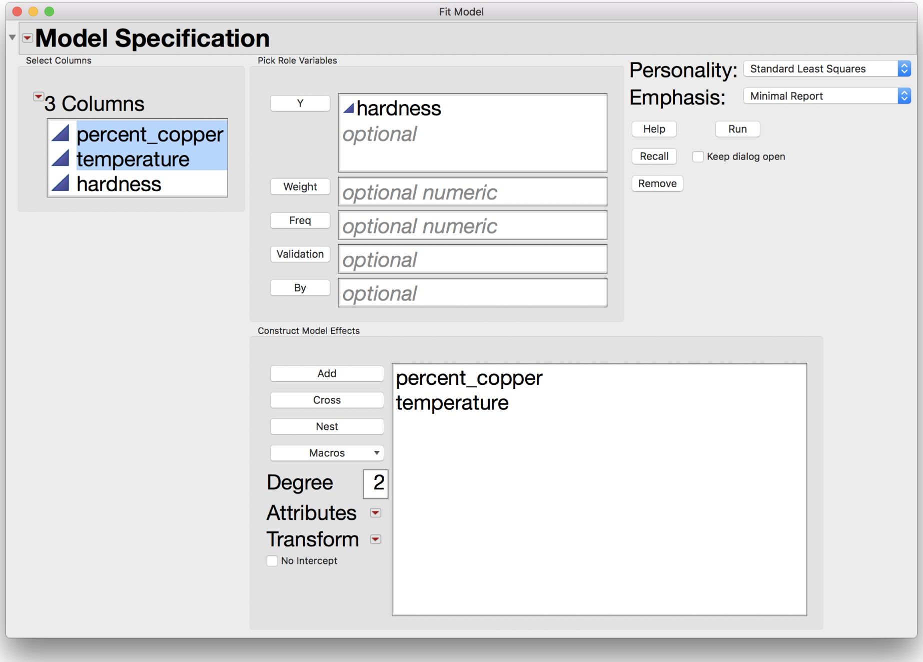
Rows

All rows	12
Selected	0
Excluded	0
Hidden	0
Labelled	0

	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:



Fit Model

Model Specification

Select Columns

3 Columns
- percent_copper
- temperature
- hardness

Pick Role Variables

- Y: hardness
- Weight: optional numeric
- Freq: optional numeric
- Validation: optional
- By: optional

Personality: Standard Least Squares

Emphasis: Minimal Report

Construct Model Effects

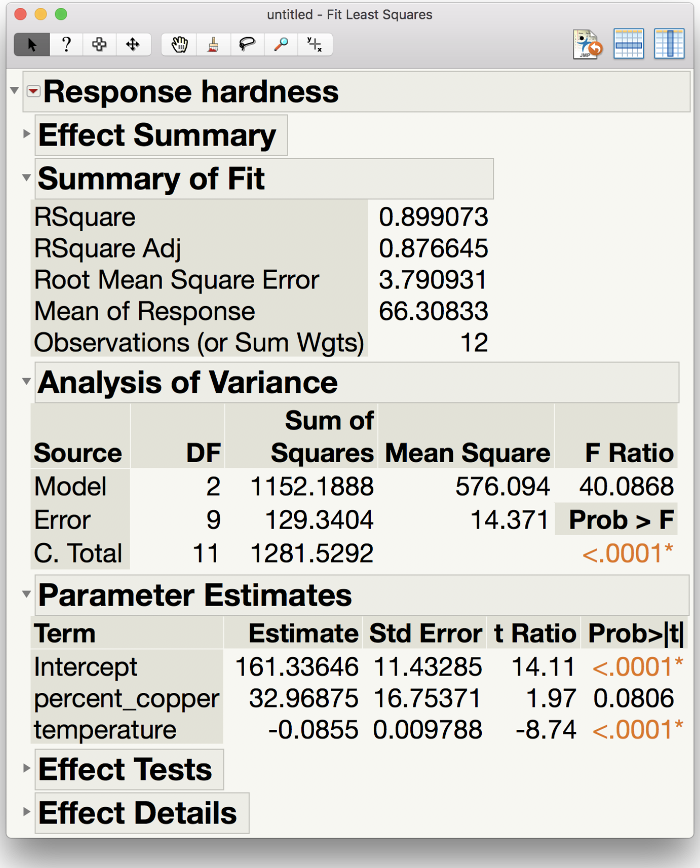
- percent_copper
- temperature

Degree: 2

□ No Intercept

Example: Hardness of Alloy

After clicking Run we get the following model fit results:



Response hardness

Effect Summary

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

Analysis of Variance

		Sum of		
Source	DF	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*

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*

Effect Tests

Effect Details

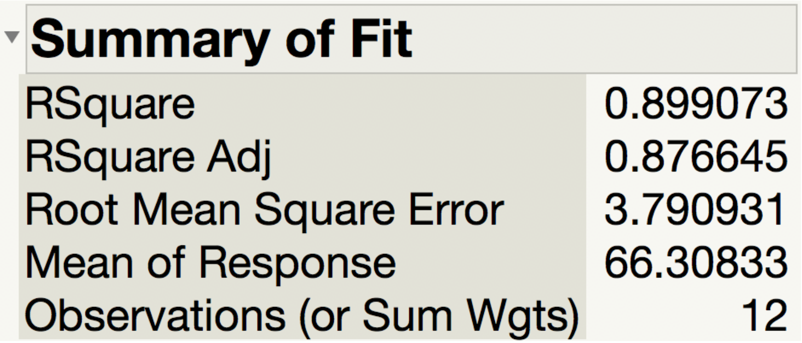
Fitting Curves

MLR

Ex: Hard Alloy

Example: Hardness of Alloy

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



Summary of Fit

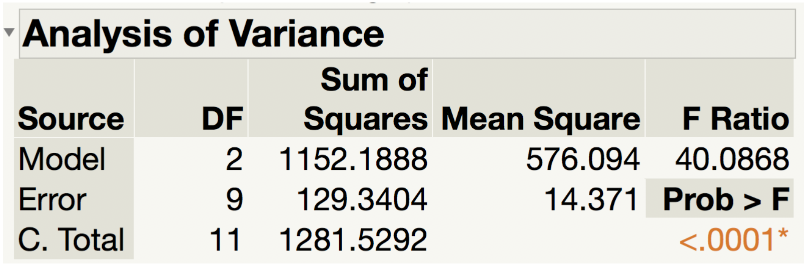
RSquare	0.899073
RSquare Adj	0.876645
Root Mean Square Error	3.790931
Mean of Response	66.30833
Observations (or Sum Wgts)	12

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:

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

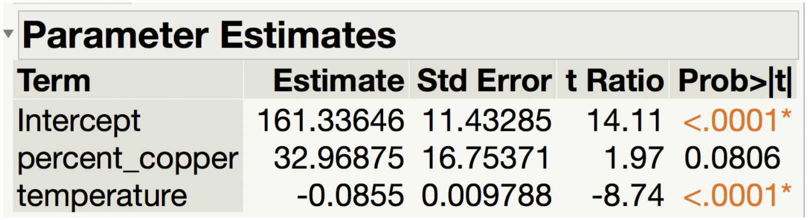
Fitting Curves

MLR

Ex: Hard Alloy

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:

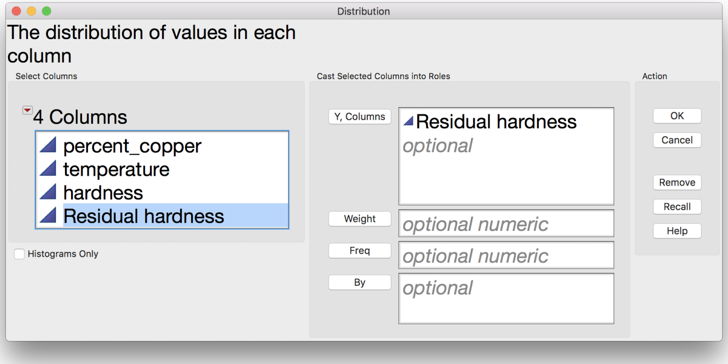
$$\hat{y} = 161.33646 + 32.96875 \cdot x_1 - 0.0855 \cdot x_2$$

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

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:



Distribution

The distribution of values in each column

Select Columns

4 Columns

- percent_copper
- temperature
- hardness
- Residual hardness

□ Histograms Only

Cast Selected Columns into Roles

Y, Columns: Residual hardness

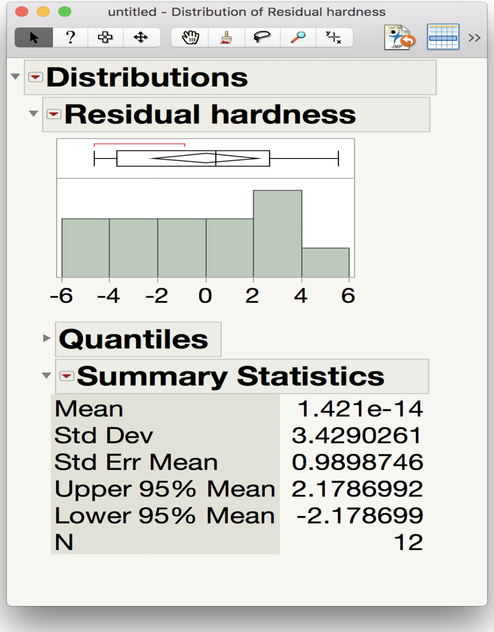
Fitting Curves

MLR

Ex: Hard Alloy

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.



Distributions

Residual hardness

Quantiles

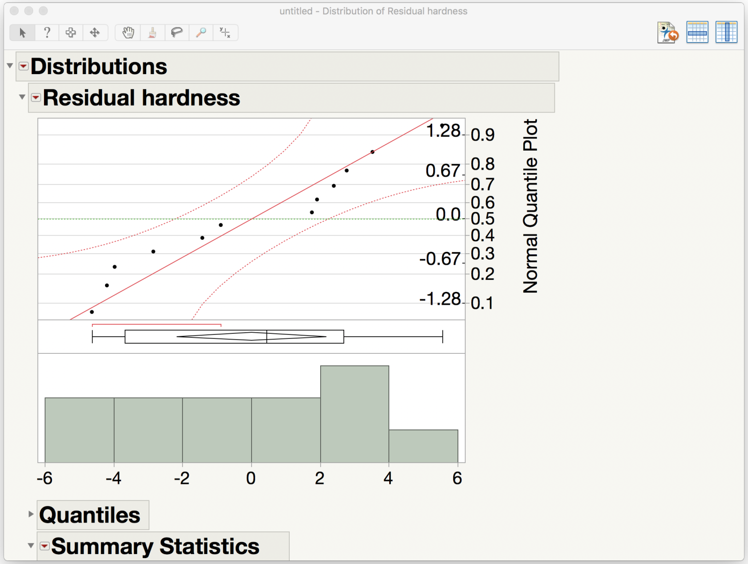
Summary Statistics

Mean	1.421e-14
Std Dev	3.4290261
Std Err Mean	0.9898746
Upper 95% Mean	2.1786992
Lower 95% Mean	-2.178699
N	12

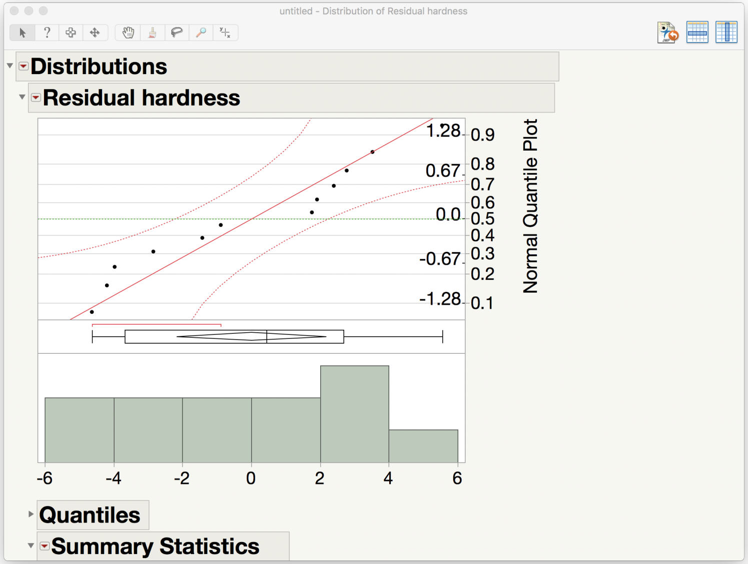
Example: Hardness of Alloy

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

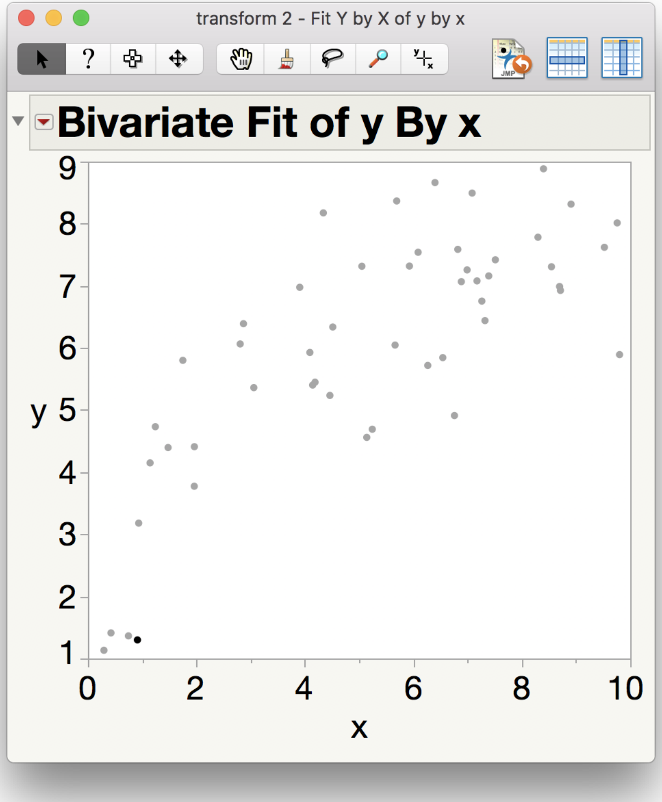


- If the points all fall on the line, then the residuals have the same spread as the normal distribution (i.e., the residuals follow a bell-shape, which is what we want).
- If they stay within the curves, then we can say the residuals follow a rough bell shape (which is good).
- If points fall outside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

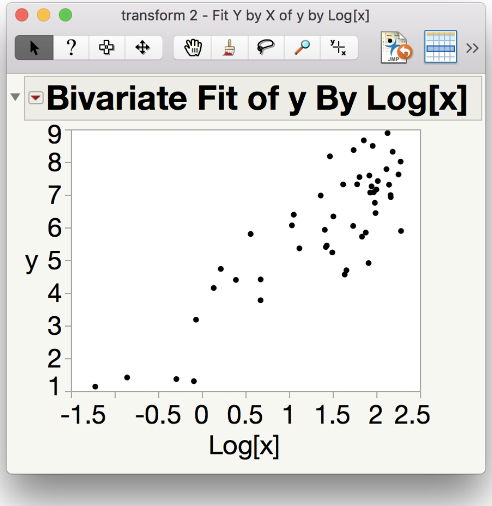
Now consider the residual plot you would get by trying to fit a quadratic. What would that look like?

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: $ln(x)$.


transform 2 - Fit Y by X of y by Log[x]
Bivariate Fit of y By Log[x]
y
9
8
7
6
5
4
3
2
1
-1.5
-0.5
0
0.5
1
1.5
2
2.5
Log[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

Fitting Curves

MLR

Ex: Hard Alloy

Dangers in Fits

Overfitting

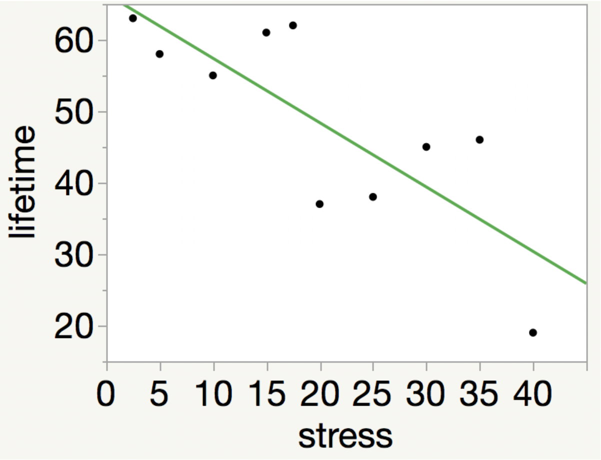
Dangers in Fitting Relationships

Example: Stress and Lifetime of Bars

Consider the bars example again

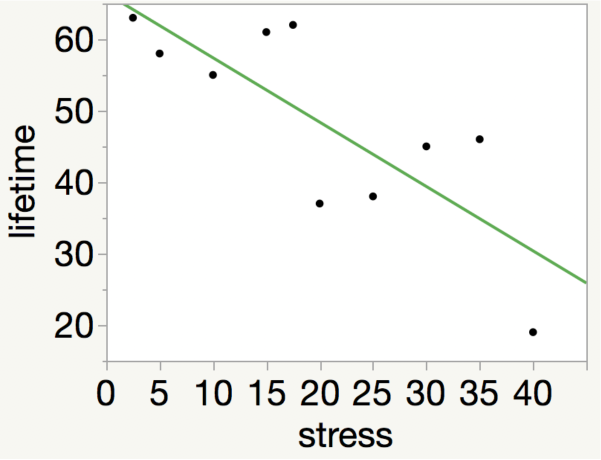
stress (kg/mm^2)	2.5	5.0	10.0	15.0	17.5	20.0	25.0	30.0	35.0	40.0
lifetime (hours)	63	58	55	61	62	37	38	45	46	19

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.

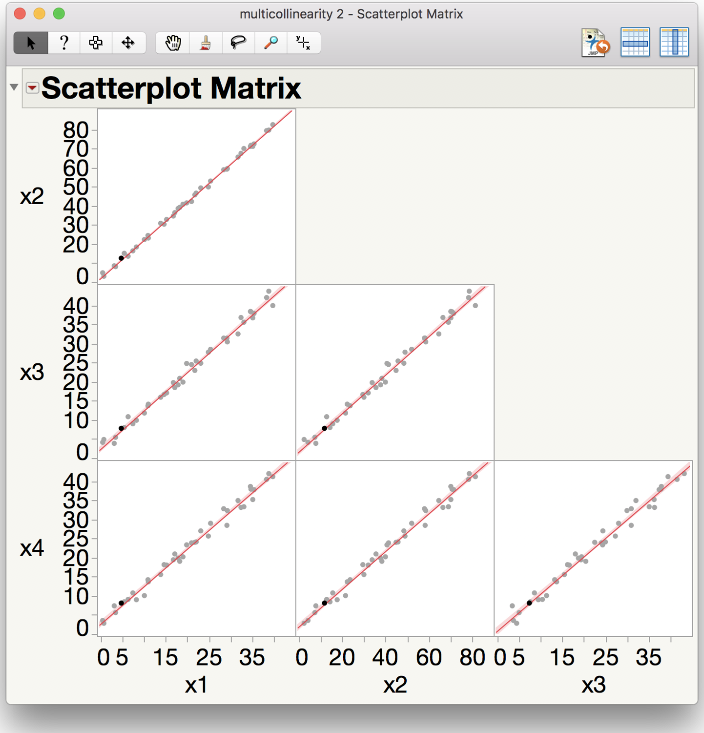
Everytime we add a new term to the polynomial, we give the fitted relationship the ability to make one more turn.

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

Dangers in Fitting Relationships

Multicollinearity

Multicollinearity occurs when you have strongly correlated experimental variables.



Dangers in Fitting Relationships

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).