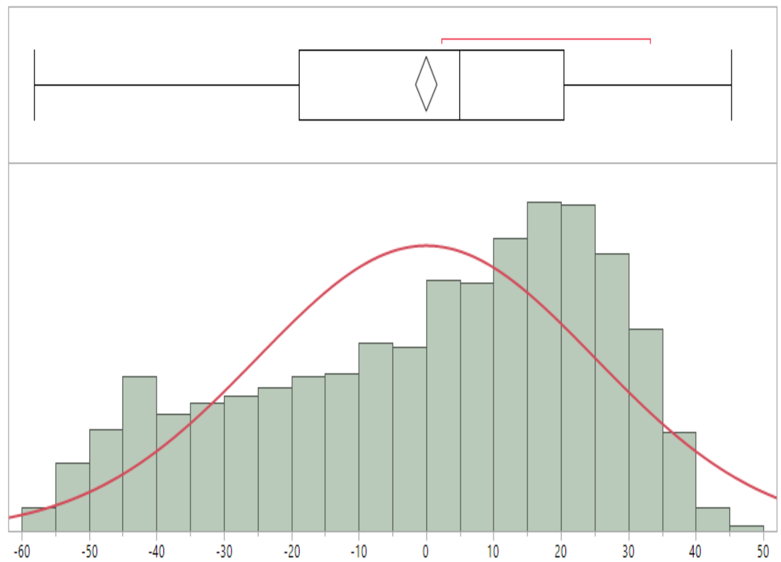
Normality of residuals

- 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)

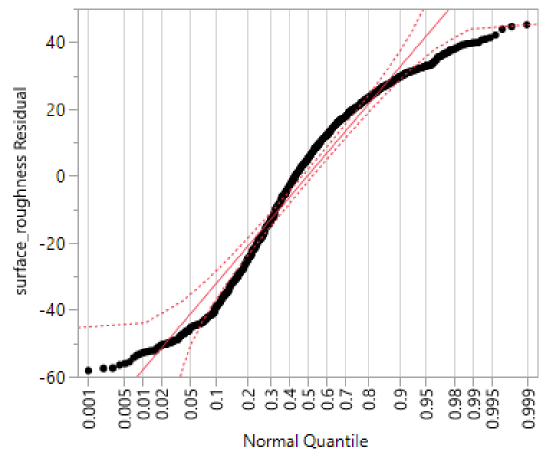


It seems the residuals are not normaly distributed in this example.The residuals have a left skewed distirbution.

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. experimental (x's)

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

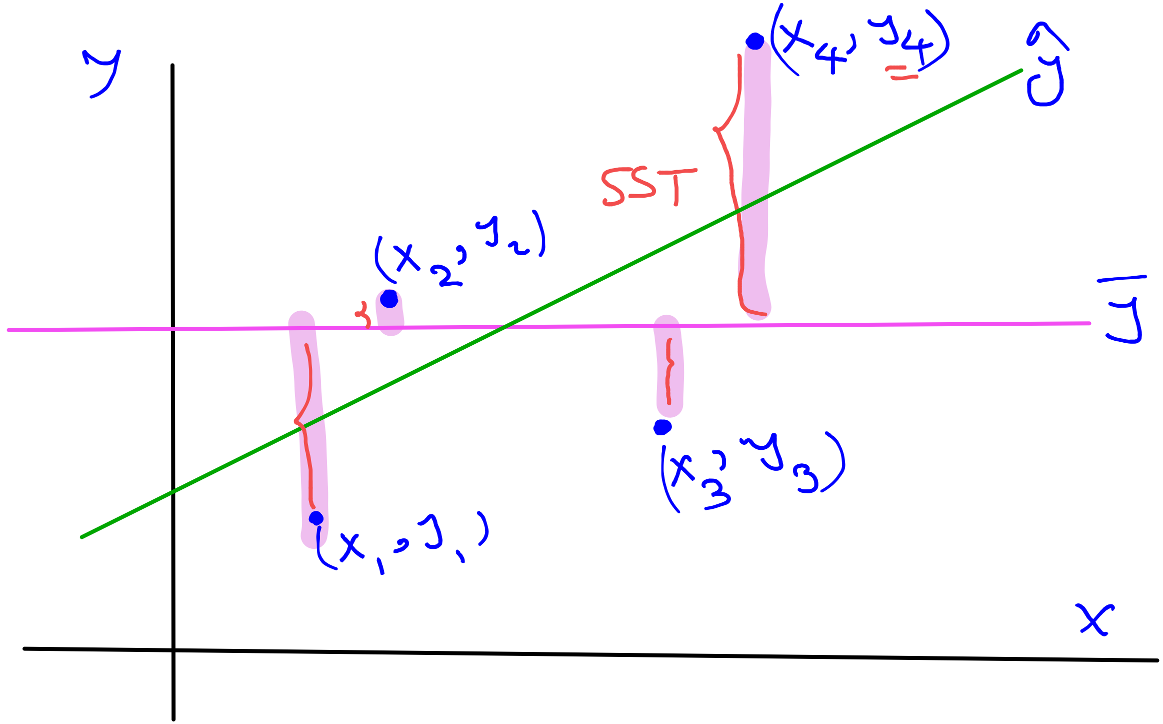
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

$$SSR = SSTO - SSE$$

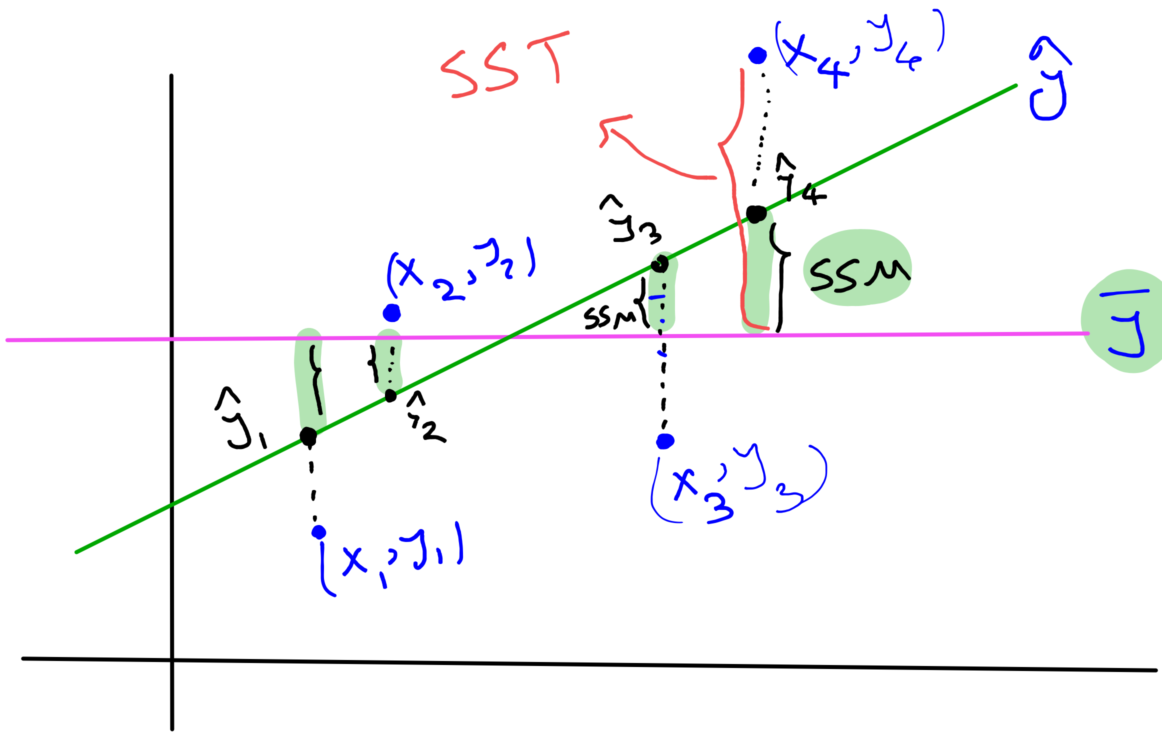
sums of squares breakdown:

$SSTO = \sum (y_i - \bar{y})^2$: measures variation of observed y_i values around their observed mean $\bar{Y}$.



SSR or SSM $= \sum(\hat{y}_i - \bar{y})^2$: Sum of Squares of Model (Regression) measures the relationship between x, y.

(SSR or SSM)

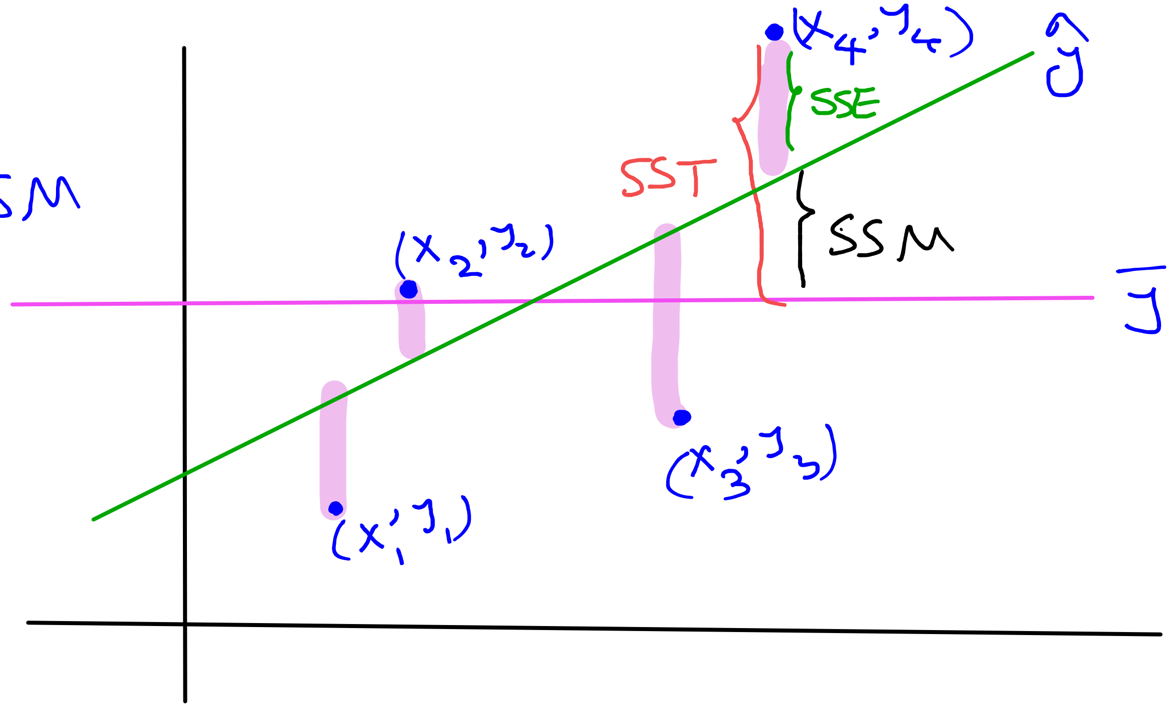


$\hat{y}_i = b_0 + b_1 x_i$

$SSE = \sum (y_i - \hat{y}_i)^2$: sum of squares of Error

measures factors other than the relationship between x, y. (what's left after fitting a linear model between x, y)

$SST = SSE + SSM$



Calculating Coeffecient of Determination (R^2

Method b. Using Sums of Squares

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

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

- **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:

$$R^2 = \boxed{(r)^2}$$

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

Calculating Coeffecient of Determination (R^2

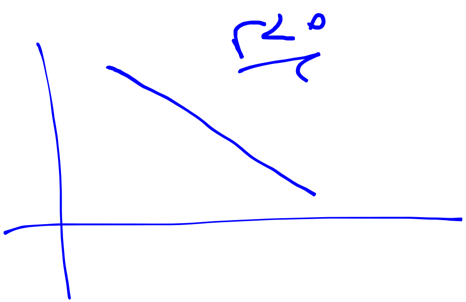
Example: Stress on Bars

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:

$$R^2 = (r)^2 = (-0.795)^2 = \boxed{0.633}$$



> 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.
course page

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).

simple linear regression

- 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

Describing Relationships

Idea

Fitting Lines

Best Estimate

Good Fit

Correlation

Residuals

Assessment

R^2

Fitting Curves

MLR

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

Describing Relationships

Idea

Fitting Lines

Best Estimate

Good Fit

Correlation

Residuals

Assessment

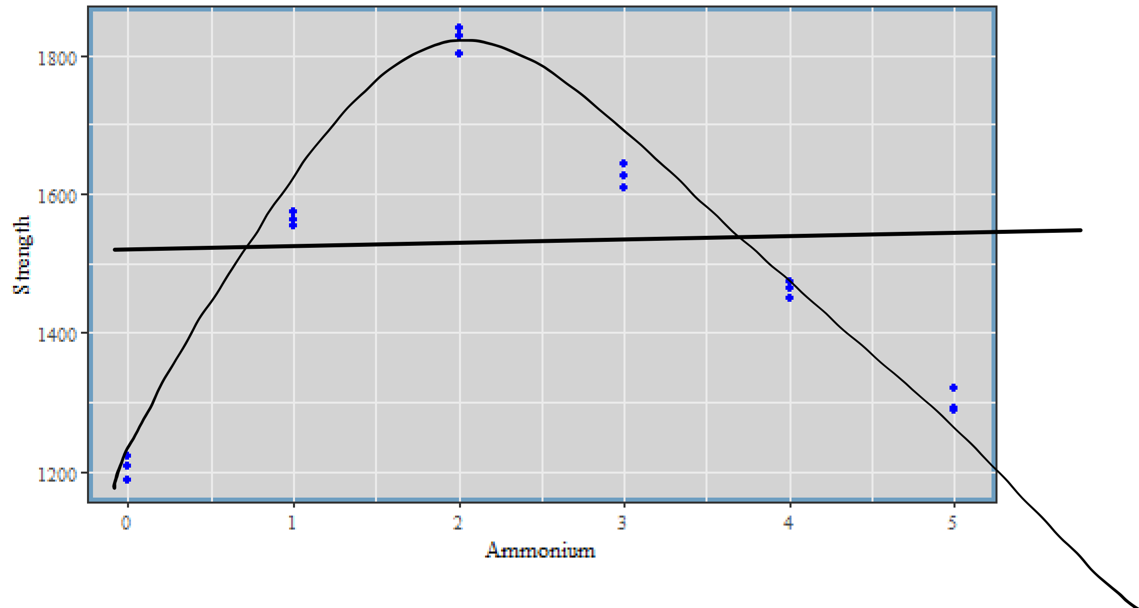
R^2

Fitting Curves

MLR

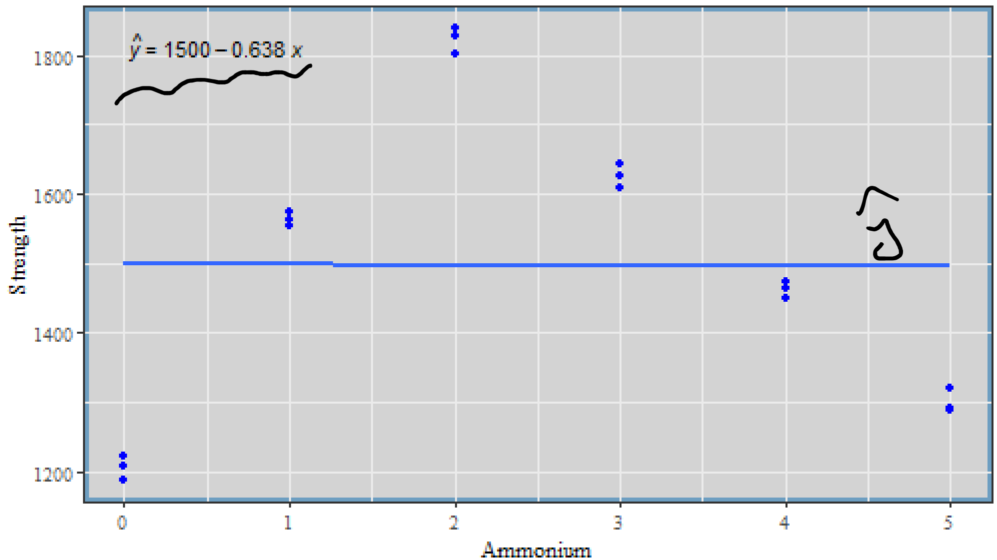
Example:

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



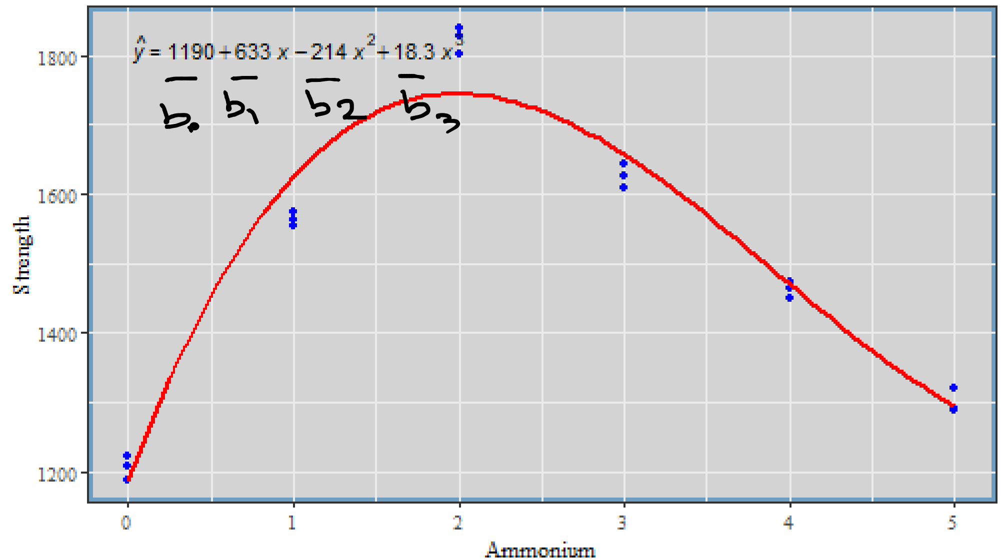
Example:

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



$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

One More Example   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.

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

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

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:

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

hardness

copper%.

temp.

Example: Hardness of Alloy

Enter the data in JMP

untitled

untitled

Columns (3/0)
- perce…copper
- temperature
- hardness

Rows

All rows	12
Selected	0
Excluded	0
Hidden	0
Labelled	0

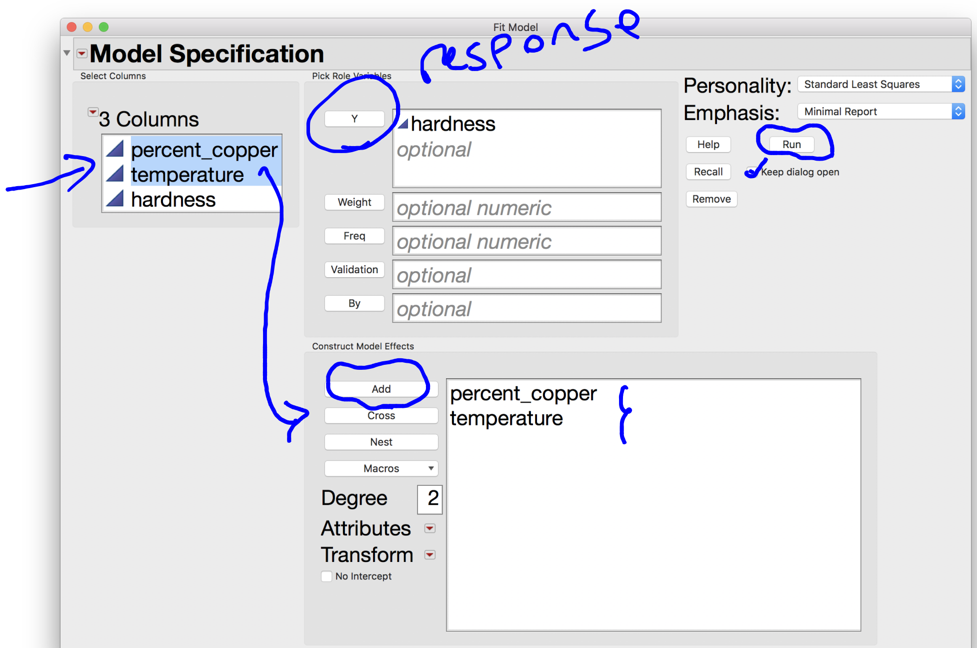
	x_1	x_2	y
	percent_copper	temperature	hardness
1	0.02	1000	78.9
2	0.02	1100	65.1
3	0.02	1200	55.2
4	0.02	1300	56.4
5	0.1	1000	80.9
6	0.1	1100	69.7
7	0.1	1200	57.4
8	0.1	1300	55.4
9	0.18	1000	85.3
10	0.18	1100	71.8
11	0.18	1200	60.7
12	0.18	1300	58.9

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:

untitled - Fit Least Squares

Response hardness

Effect Summary

Summary of Fit

RSquare	0.899073	R^2
RSquare Adj	0.876645	
Root Mean Square Error	3.790931	
Mean of Response	66.30833	
Observations (or Sum Wgts)	12	

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*

Parameter Estimates

	Term	Estimate	Std Error	t Ratio	Prob>|t|
b_0	Intercept	161.33646	11.43285	14.11	<.0001*
b_1	percent_copper	32.96875	16.75371	1.97	0.0806
b_2	temperature	-0.0855	0.009788	-8.74	<.0001*

Effect Tests

Effect Details

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.

SSR SSE SSTO

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:

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

Describing Relationships
Idea
Fitting Lines
Best Estimate
Good Fit
Correlation
Residuals
Assessment
R^2
Fitting Curves
MLR

Example: Hardness of Alloy

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

$b_0 =$ $b_1 =$ $b_2 =$

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:

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

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

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

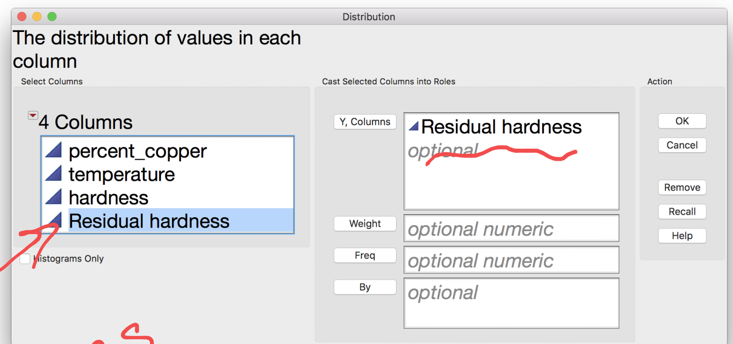
$$= 161.33646 + 3.296875 - 85.5$$

$$= 79.13333$$

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.

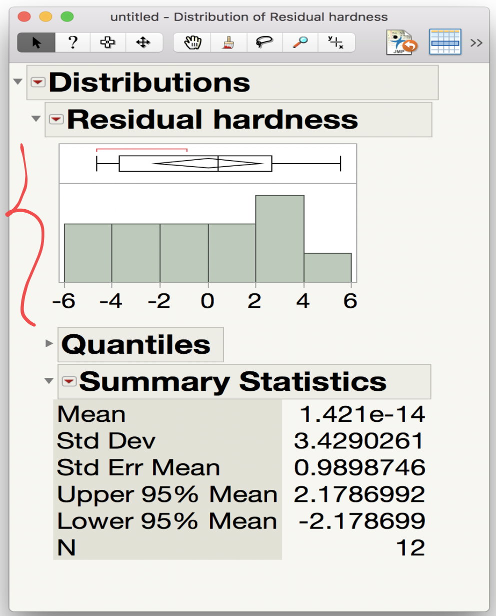
From Analyze > Distribution:



save residuals

Example: Hardness of Alloy

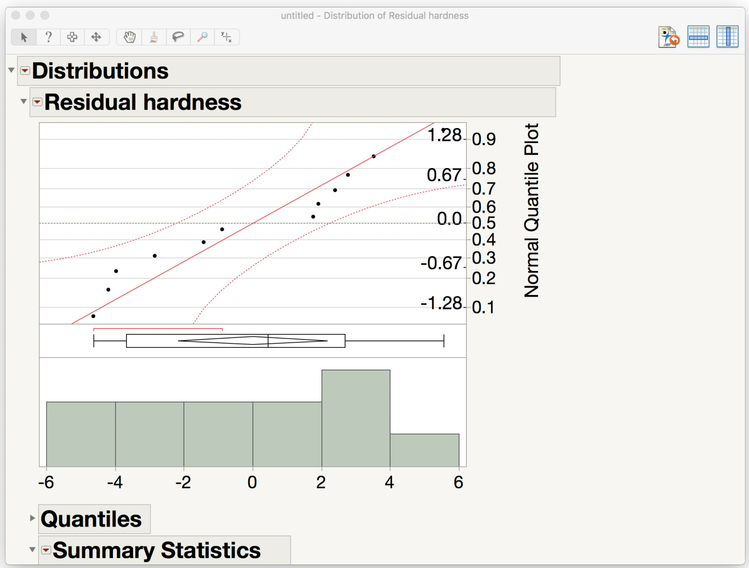
There aren't many residuals here (just 12) but we would like to make sure that the histogram has rough bell-shape (normal residuals are good). I would call this one inconclusive.



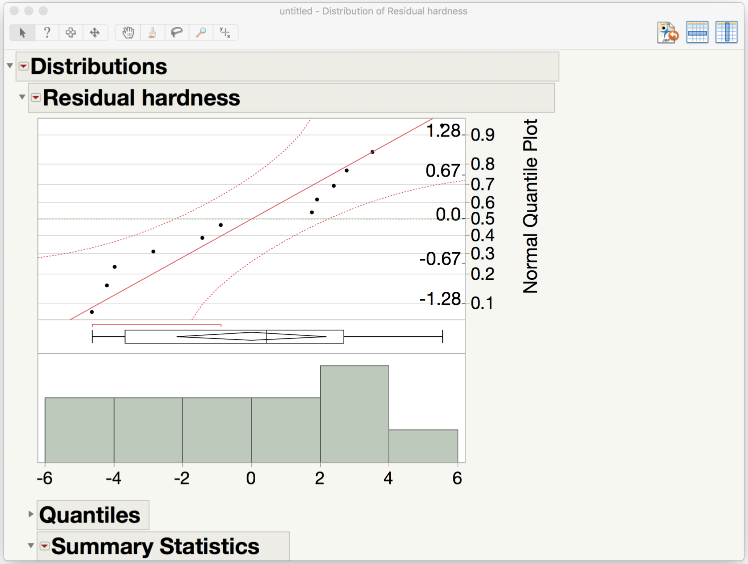
Example: Hardness of Alloy

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).