Correlation and Simple Linear Regression[1]

November 30, 2017

[1] HMS, 2017, v1.6

Chapter References

- Diez: Chapter 7
- Navidi, Chapter 7

I don't expect you to learn the proofs what will follow.

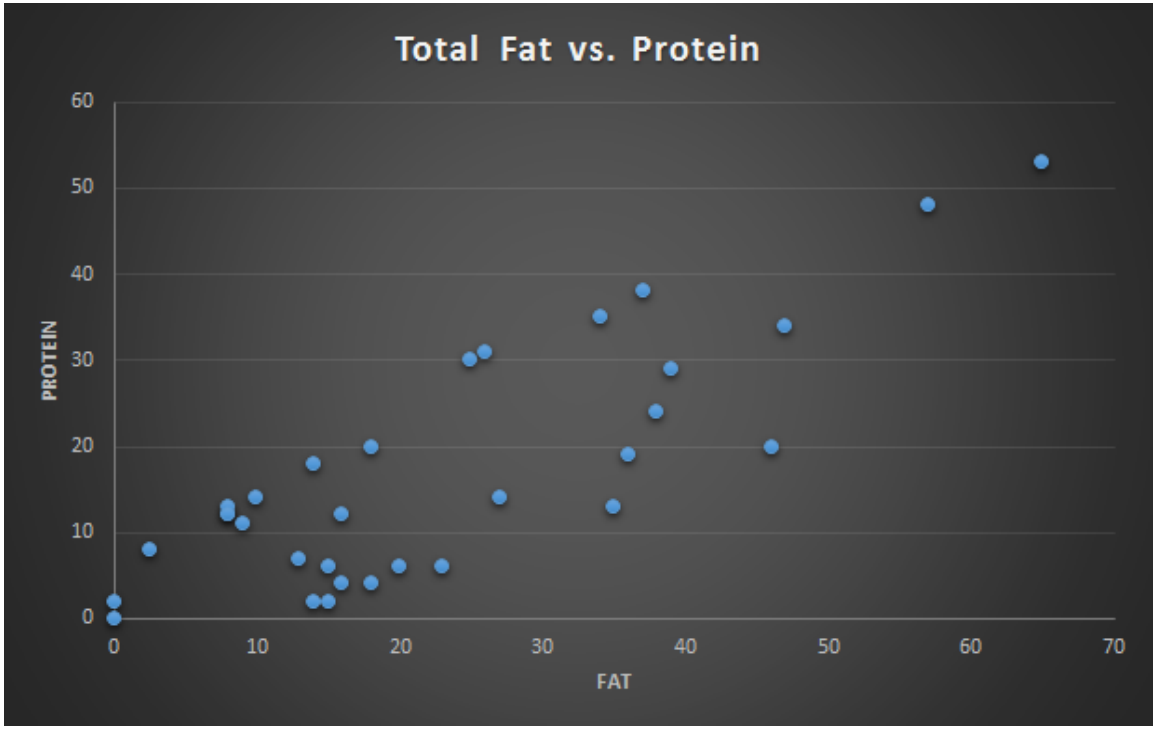
Linear Regression

Total fat versus protein for 30 Items on the Burger King Menu

Item	Total Fat	Protein
Whopper	39	29
Whopper w/Cheese	47	34
Double Whopper	57	48
Double Whopper w/Cheese	65	53
Hamburger	14	18
Cheeseburger	18	20
Double Hamburger	26	31
Double Cheeseburger w/ Bacon	37	38
Veggie Burger	10	14
BK Big Fish	38	24
BK Broiler Chicken	25	30
Chicken Tenders Sandwich	27	14
Chicken Tenders (4pc)	9	11
Fries (med)	18	4
Onion rings (med)	16	4
Jalapeno poppers (4pc)	13	7
Mozzarella Sticks (4pc)	16	12
Apple Pie	14	2
Croissan wich w/Sausage, Egg and Cheese	36	19
Biscuit	15	6
Biscuit w/Sausage, Egg and Cheese	46	20
French Toast Stix (5)	20	6
Cini-minis (4)	23	6
Hash Brown Rounds (small)	15	2
Vanilla Shake (med)	8	12
Chocolate Shake (Med w/ Syrup)	8	13
Strawberry Shake (Med)	8	12
Coke (med)	0	0
Tropicana OJ	0	2
1% Milk	2.5	8

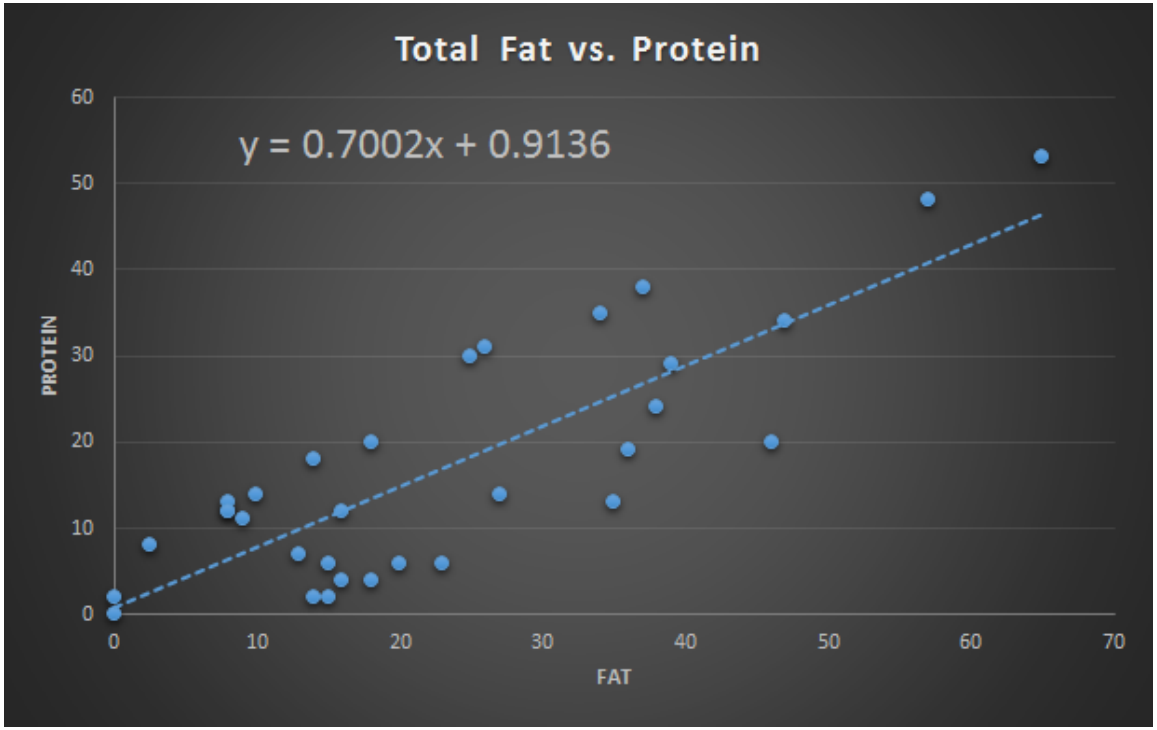
Linear Regression

Plot of total fat versus protein



Linear Regression

Plot of total fat versus protein with trend line from Excel



Fat content when protein is 35 grams is:

$$y = 1.2145 \times 35 - 3.298 = 39.21 \text{ grams}$$

Linear Regression

Some questions:

1. How did it compute the trend line?
2. Is a straight line the best curve to describe the data?
3. Given there is noise in the data how does this noise propagate to the answer of 39.21 grams?
4. What about other aspects such as the slope and intercept, what confidence do we have in those?
5. Is there a way to measure the quality of a linear fit or any other fit?

$$y = 1.2145 \times 35 - 3.298 = 39.21 \text{ grams}$$

How did it compute the trend line?

A straight line fit can be described using the relationship:

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

x is called the **independent variable**
y is called the **dependent variable**
β_0 is the y **intercept**.
β_1 is the **slope**

How did it compute the trend line?

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

β_0 is the y **intercept**.

β_1 is the **slope**

Linear Regression

When there is a single independent variable, we refer to the model as **simple regression**.

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2) \text{ and independent}$$

Two key questions:

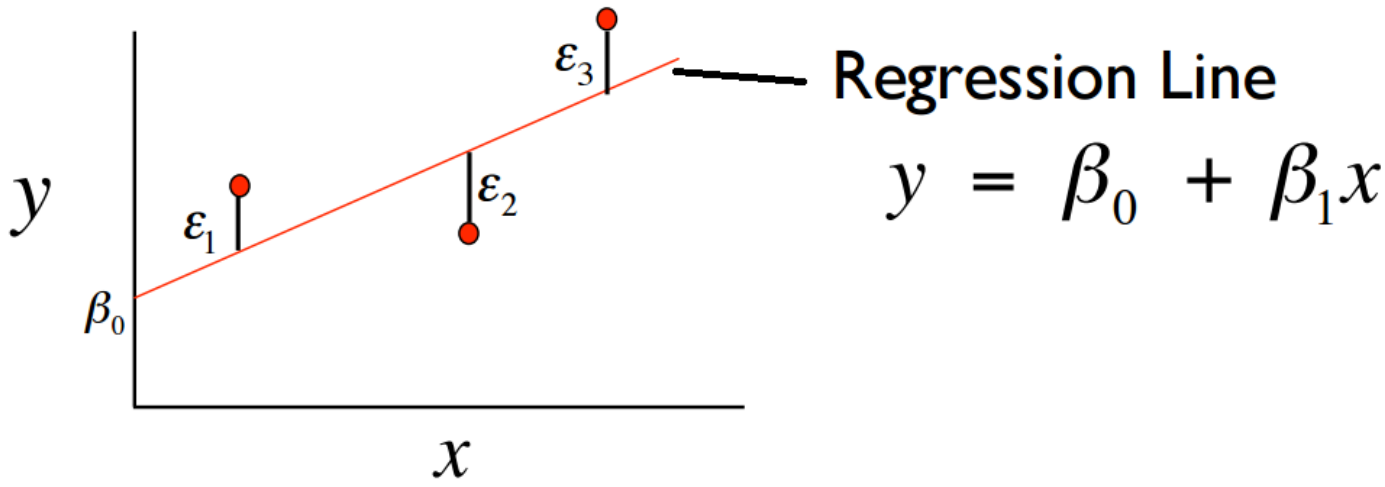
1. Is there actually a relationship between x and y and how strong is it?
2. If there is, can we formalize the relationship to make predictions?

Notation

y_i An experimental data point

$\hat{y}_i$ A fitted y value

How did it compute the trend line?



A data point on the graph will have the form:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where ε_i are called the **residuals** and is the result of random errors when collecting y_i. We assume these errors to be of the form (i.e identical mean and standard deviation):

$$\varepsilon : N(0, \sigma^2)$$

and independent, i.e one error does not influence other errors.

Interlude: Multiple Regression Models

The following is a linear regression model:

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

so is:

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

where we now have **multiple** independent variables, x_1, x_2, etc. Such equations are called **multiple linear regression models.**

These models describe hyperplanes.

Interlude: Multiple Regression Models

Multiple linear regression models can also include **interaction** terms:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \ldots + \varepsilon$$

Interlude: Multiple Regression Models

Multiple linear regression models can also include **polynomial** terms:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \beta_3 x_3^3 \ldots + \varepsilon$$

These may also include interaction terms.

We'll come back to multiple regression later.

Simple Regression

Consider:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

The difference between y_i and the point $\beta_0 + \beta_1 x_i$ is ε. Then:

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

Squaring both sides and summing over all points yields:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)^2)$$

We define the best fit line as:

The values for β_o and β_1 that minimizes $\sum \varepsilon_i^2$.

Simple Regression

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 = \chi^2$$

It can be shown that when L is minimized, the line described by β_0 and β_1 is the best fit.

Note: This is a chi-square statistic!

Simple Regression

The minimum of the least squares is the point where:

$$\frac{\partial L}{\partial \beta_0} = 0 \qquad \frac{\partial L}{\partial \beta_1} = 0$$

Differentiating yields:

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) x_i) = 0$$

Where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the best intercept and slope.

Update on Notation

x_i	Value of the i^{th} independent variable
y_i	An experimental data point
$\hat{y}_i$	A fitted y value
$\hat{\beta}_0$	Fitted intercept
$\hat{\beta}_1$	Fitted slope

Simple Regression

We can simplify the equations to form the **least squares normal equations**:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.

Simple Regression

First let's make the following substitutions:

$$S_X = \sum_{i=1}^{n} x_i \qquad S_Y = \sum_{i=1}^{n} y_i$$

$$S_{X^2} = \sum_{i=1}^{n} x_i^2 \qquad S_{XY} = \sum_{i=1}^{n} x_i y_i$$

This allows us to write the normal equations as:

$$\hat{\beta}_0 n + \hat{\beta}_1 S_X = S_Y$$

$$\hat{\beta}_0 S_X + \hat{\beta}_1 S_{X^2} = S_{XY}$$

Simple Regression

With the previous substitutions it is easier to solve for $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\Delta = nS_{X^2} - (S_X)^2$$

$$\hat{\beta}_0 = \frac{S_{X^2}S_Y - S_X S_{XY}}{\Delta}$$

$$\hat{\beta}_1 = \frac{nS_{XY} - S_X S_Y}{\Delta}$$

Estimating $\hat{\beta}_0$ from $\hat{\beta}_1$

$$y_1 = \hat{\beta}_1 x_1 + \hat{\beta}_0$$

$$\sum^{n} y_i = \sum \hat{\beta}_1 x_i + \sum \hat{\beta}_0$$

$$\sum \frac{y_i}{n} = \sum \frac{\hat{\beta}_1 x_i}{n} + \sum \frac{\hat{\beta}_0}{n}$$

$$\bar{y} = \hat{\beta}_1 \sum^{n} \frac{x_i}{n} + \hat{\beta}_0 \sum^{n} \frac{1}{n}$$

$$\bar{y} = \hat{\beta}_1 \bar{x} + \hat{\beta}_0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple Regression

Let's try one:

x_i		y_i
1		1
2		1
3		2
4		2
5		4
$\sum x_i$	$(\sum x_i)^2$	$\sum y_i$
15	225	10

Simple Regression

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum^{n} x_i^2 - \left(\sum^{n} x_i\right)^2}$$

$$n = 5$$

$$\sum x_i = 15 \quad \sum y_i = 10 \quad \sum x_i^2 = 55 \quad \sum x_i y_i = 37 \quad \left(\sum x_i\right)^2 = 225$$

$$\hat{\beta}_1 = \frac{5 \times 37 - 15 \times 10}{5 \times 55 - 225}$$

$$\hat{\beta}_1 = \frac{35}{50} = 0.7 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{10}{5} - 0.7\frac{15}{5} = -0.1$$

$$y = -0.1 + 0.7x$$

Interlude

Expand $\sum^n (x_i - \bar{x})^2$ Note: $\bar{x} = \sum^n x_i/n$ therefore $\sum^n x_i = n\bar{x}$

$$\begin{aligned}
\sum^n (x_i - \bar{x})^2 &= \sum^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum^n x_i^2 - 2\bar{x}\sum^n x_i + \sum^n \bar{x}^2 = \sum^n x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 \\
&= \sum^n x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum^n x_i^2 - n\bar{x}^2 \\
&= \sum^n x_i^2 - n\left(\sum^n \frac{x_i}{n}\right)^2 \\
&= \sum^n x_i^2 - \frac{\left(\sum^n x_i\right)^2}{n}
\end{aligned}$$

Interlude

By a similar expansion one can show that:

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

Interlude

SS stands for 'sum of squares', The following are sometimes called the short-cut expressions:

$$SS_X = \sum^n (x_i - \bar{x})^2 = \sum^n x_i^2 - \frac{\left(\sum^n x_i\right)^2}{n}$$

$$SS_Y = \sum^n (y_i - \bar{y})^2 = \sum^n y_i^2 - \frac{\left(\sum^n y_i\right)^2}{n}$$

$$SS_{XY} = \sum^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

We'll be using some of these later on.

Summary of Notation

S stands for Sum, eg $S_X = \sum x_i$

SS_X stands for sum of squares, eg $SS_X = \sum (x_i - \bar{x})^2$

Short-cut formula:

$$SS_X = \sum^n (x_i - \bar{x})^2 = \sum^n x_i^2 - \frac{\left(\sum^n x_i\right)^2}{n}$$

$$SS_{XY} = \sum^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$$

Simple Regression

$$\Delta = nS_{X^2} - (S_X)^2$$

$$\hat{\beta}_0 = \frac{S_{X^2}S_Y - S_X S_{XY}}{\Delta} \qquad \hat{\beta}_1 = \frac{nS_{XY} - S_X S_Y}{\Delta}$$

Multiply both sides by n:

$$n\ SS_X = n\sum(x_i - \bar{x})^2 = n\sum x_i^2 - \left(\sum^n x_i\right)^2$$

$$n\ SS_Y = n\sum(y_i - \bar{y})^2 = n\sum y_i^2 - \left(\sum^n y_i\right)^2$$

$$n\ SS_{XY} = n\sum^n(x_i - \bar{x})(y_i - \bar{y}) = n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)$$

Simple Regression

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X}$$

To derive $\hat{\beta}_0$ its easier to rearrange this:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In Summary

The final result is:

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum^n (x_i - \bar{x})(y_i - \bar{y})}{\sum^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Simple Regression

An alternative approach using Matrices:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = \sum_{i=1}^{n} x_i y_i$$

Define:

$$\mathbf{b} = \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \mathbf{X}^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Such that:

Simple Regression

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} (x_i)^2 = \sum_{i=1}^{n} x_i y_i$$

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$$

Simple Regression

Multiply both sides by the inverse of $\mathbf{X}^T\mathbf{X}$

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{Y}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\text{But } (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{I}$$

Therefore:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

These are called the matrix normal equations.

Using Python

There are a number of different packages that can be used to do linear regression, eg `scipy.stats`, `scikit-learn`

```
import numpy as np
import scipy.stats
import pylab

x = [1,2,3,4,5]; y = [1,1,2,2,4]
l = scipy.stats.linregress (x, y)
print l.slope
print l.intercept

xi = np.arange(0,7)
line = l.slope*xi + l.intercept
pylab.xlim(-1, 7); pylab.ylim(-1, 7)
pylab.plot (xi, line, 'b-', x, y, 'o')
```

Diagnostics: Looking at the Residuals

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i$$

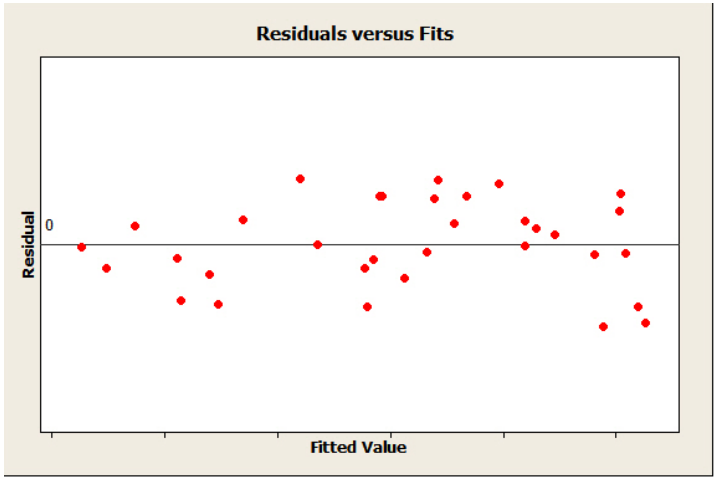
Recall that ε_i are called the **residuals.**

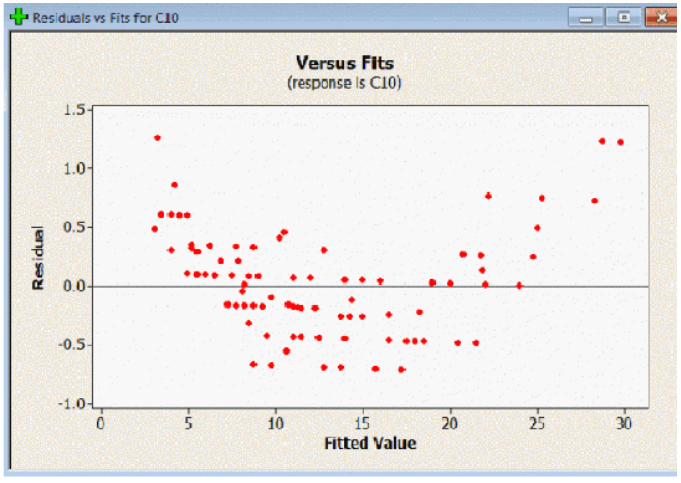
$$\varepsilon_i = \hat{y}_i - y_i$$

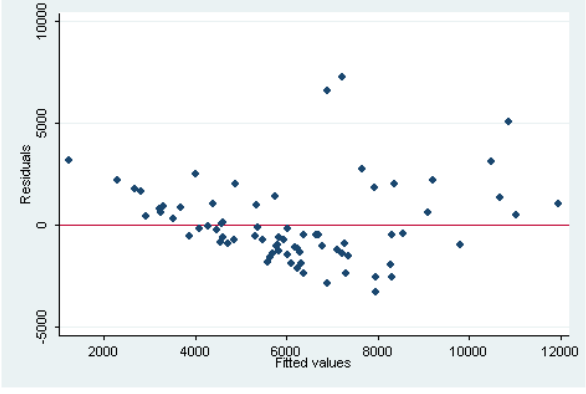
x_i	y_i	Fitted y_i	Residuals, ε_i
1	1	0.6	0.4
2	1	1.3	-0.3
3	2	2	0
4	2	2.7	-0.7
5	4	3.4	0.6

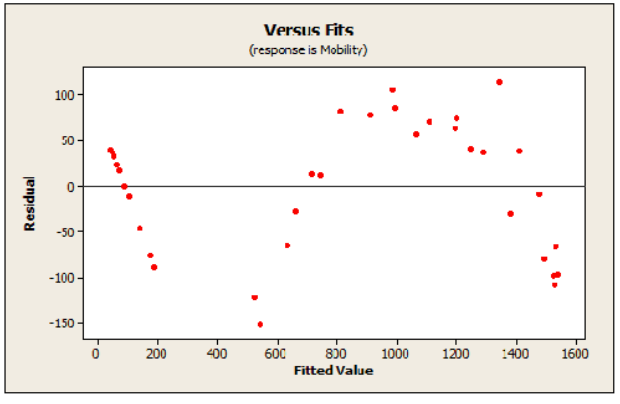
Looking at the Residuals

Plot the residual versus the fitted value:



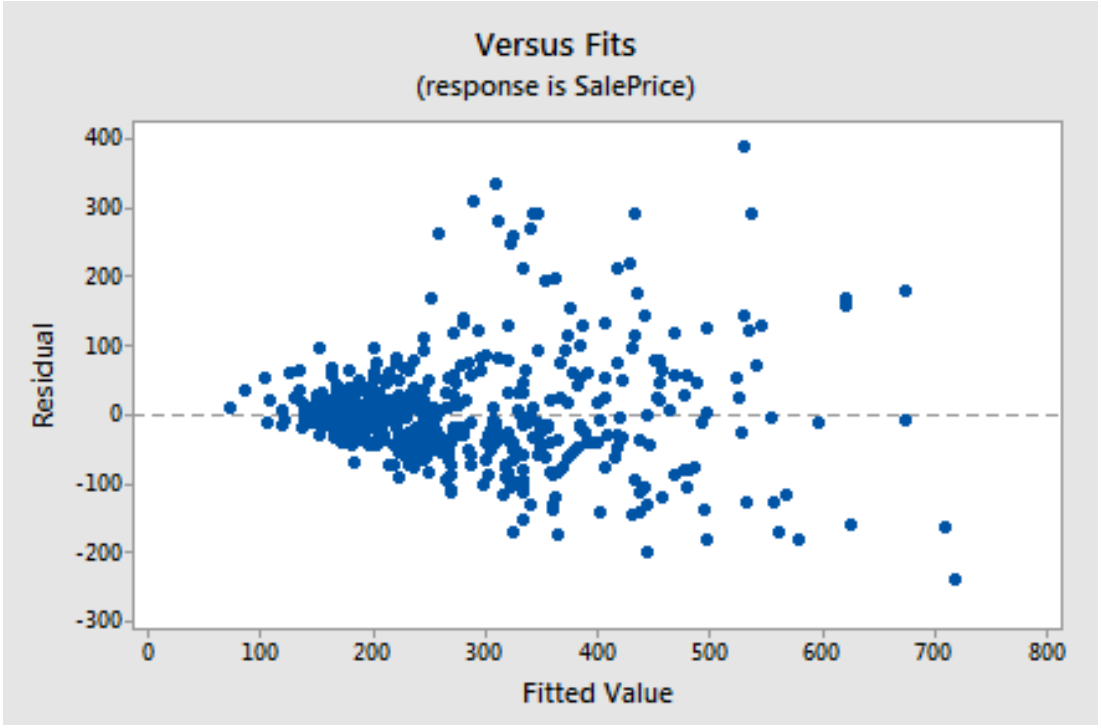






Unequal Variances

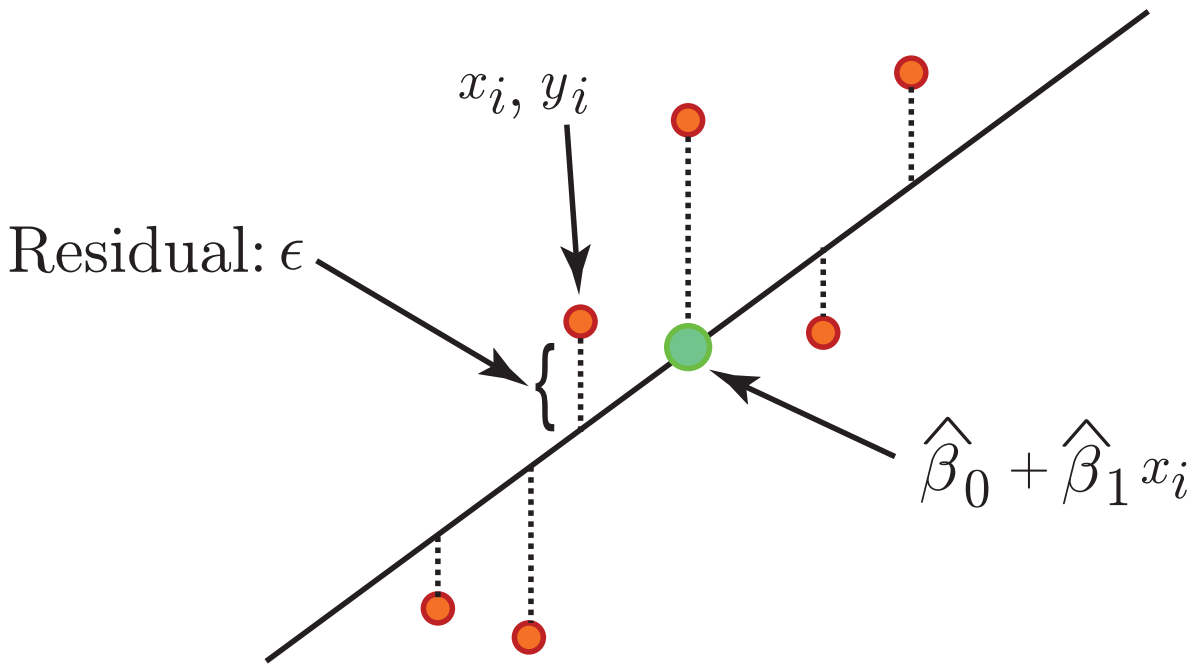
Plot the residual versus the fitted value:



Looking at the Residuals

Residual sum of squares, also called the Error Sum of Squares or SSE:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)]^2$$



Looking at the Residuals

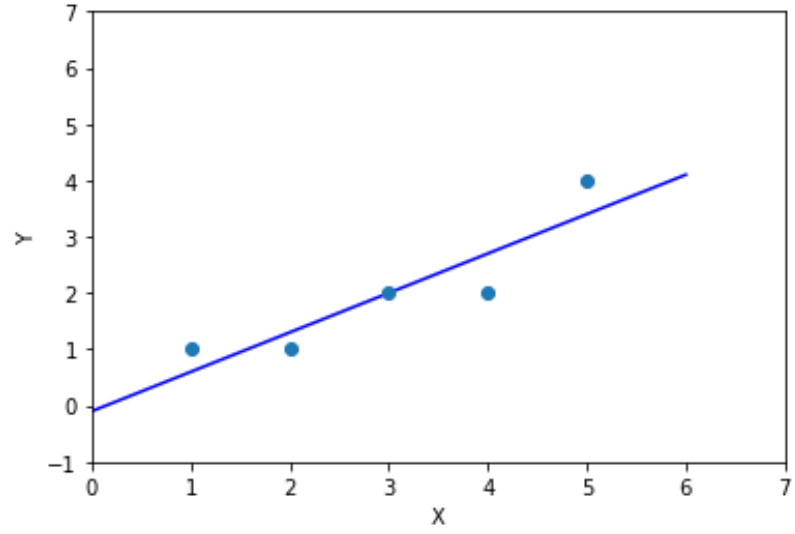
Residual Mean Square:

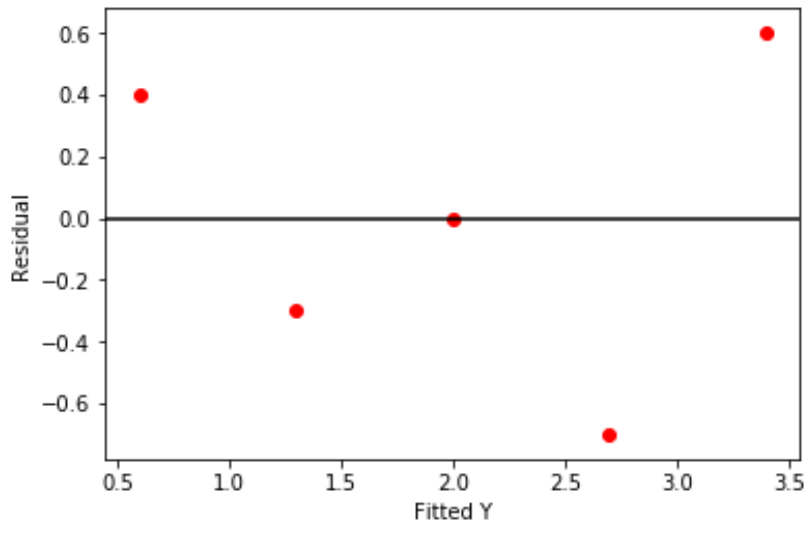
$$\text{MSE} = \frac{\sum (y_i - \hat{y}_i)^2}{N - 2}$$

This is the variance of the residuals. The smaller this is the tighter the fit. Unfortunately it has units and is not a suitable measure of how well the data are correlated. Moreover, it is possible for the SSE to be small and at the same time there be no correlation between x and y.

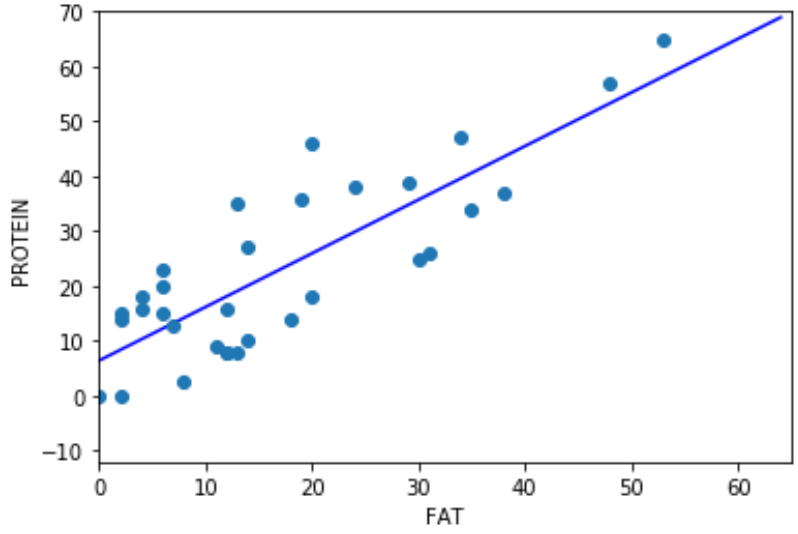
We will use it however when we come to investigate the confidence we have in the estimated slope and intercept.

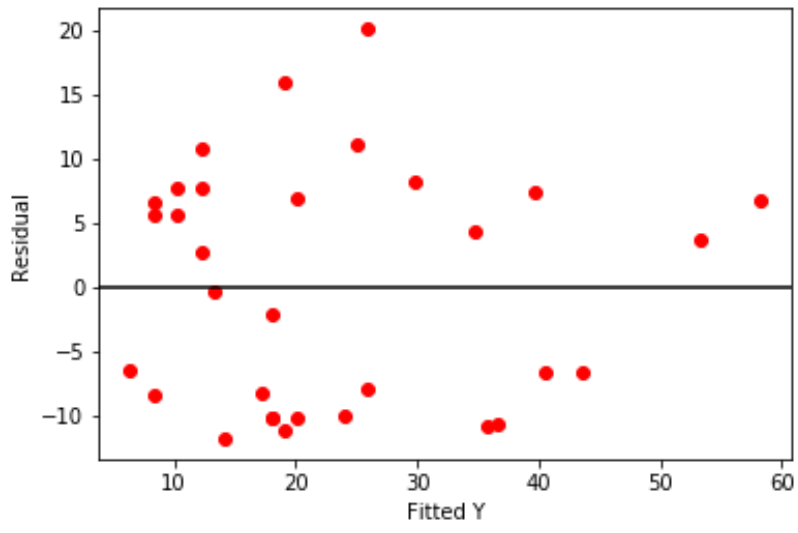
Using Python





Using Python





Always look at the residuals

Always look at the residuals when you fit a line.

Plot the residuals versus the fitted value.

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\Delta = n\ S_{X^2} - (S_X)^2$$

$$\hat{\beta}_0 = \frac{S_{X^2} S_Y - S_X S_{XY}}{\Delta}$$

$$\hat{\beta}_1 = \frac{n\ S_{XY} - S_X S_Y}{\Delta}$$

We need to propagate errors in y_i into $\hat{\beta}_o$ and $\hat{\beta}_1$.

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

Errors in y_i contribute to uncertainty in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\sigma^2 = \sum \sigma_i^2 \left(\frac{\partial f}{\partial y_i} \right)^2$$

$$\hat{\beta}_0 = \hat{\beta}_0(y_1, y_2, \ldots)$$

$$\hat{\beta}_1 = \hat{\beta}_1(y_1, y_2, \ldots)$$

$$\sigma^2_{\hat{\beta}_o} = \sigma_1^2 \left(\frac{\partial \hat{\beta}_0}{\partial y_1} \right)^2 + \sigma_2^2 \left(\frac{\partial \hat{\beta}_0}{\partial y_2} \right)^2 + \ldots$$

$$\sigma^2_{\hat{\beta}_1} = \sigma_1^2 \left(\frac{\partial \hat{\beta}_1}{\partial y_1} \right)^2 + \sigma_2^2 \left(\frac{\partial \hat{\beta}_1}{\partial y_2} \right)^2 + \ldots$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\Delta = n\ S_{X^2} - (S_X)^2$$

$$\hat{\beta}_0 = \frac{S_{X^2} S_Y - S_X S_{XY}}{\Delta}$$

$$\hat{\beta}_1 = \frac{n\ S_{XY} - S_X S_Y}{\Delta}$$

First thing to do is compute the derivatives:

$$\frac{\partial \hat{\beta}_0}{\partial y_i} = \frac{S_{X^2} - S_X x_i}{\Delta}$$

$$\frac{\partial \hat{\beta}_1}{\partial y_i} = \frac{n x_i - S_X}{\Delta}$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\frac{\partial \hat{\beta}_0}{\partial y_i} = \frac{S_{X^2} - S_X x_i}{\Delta} \qquad \frac{\partial \hat{\beta}_1}{\partial y_i} = \frac{n x_i - S_X}{\Delta}$$

We now insert these into

$$\sigma^2_{\hat{\beta}_o} = \sigma_1^2 \left(\frac{\partial \beta_0}{\partial y_1} \right)^2 + \sigma_2^2 \left(\frac{\partial \beta_0}{\partial y_2} \right)^2 + \dots$$

We can simplify by assuming all variances are equal, so that $\sigma_i^2 = \sigma^2$, and after some tedious algebra:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{S_{X^2}}{\Delta} \qquad \sigma^2_{\beta_1} = \sigma^2 \frac{n}{\Delta}$$

What is σ^2? $\sigma^2 = \text{MSE}$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

Let's fully write out $\hat{\beta}_1$:

$$\sigma^2_{\hat{\beta}_1} = \sigma^2 \frac{n}{n\ S_{X^2} - (S_X)^2} = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

$$\sigma^2_{\hat{\beta}_1} = \frac{n\sigma^2}{n\sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

Divide top and bottom by n:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - \frac{\sum (x_i)^2}{n}} = \frac{\sigma^2}{SS_X}$$

This is the you'll find in the text book for $\hat{\beta}_1$ though expressed as a standard deviation (take square root on both sides).

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

Let's fully write out $\hat{\beta}_0$:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{S_{X^2}}{n\ S_{X^2} - (S_X)^2} = \sigma^2 \frac{\sum x_i^2}{n \sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

This is not what you find in most text books. Let change the numerator to:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum x_i^2 + n\bar{x}^2 - 2n\bar{x}^2 + n\bar{x}^2}{n \sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

Expand one of the $\bar{x}$ in the $\bar{x}^2$ term:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum x_i^2 + n\bar{x}^2 - 2n\bar{x}\frac{\sum x_i}{n} + n\bar{x}^2}{n \sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

Cancel the common n:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i + n\bar{x}^2}{n \sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum x_i^2 + n\bar{x}^2 - 2\bar{x}\sum x_i + n\bar{x}^2}{n\sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

Collect the first three terms into a common summation:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum (x_i^2 + \bar{x}^2 - 2\bar{x}x_i) + n\bar{x}^2}{n\sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

The term in the common summation is: $\sum (x_i - \bar{x})^2$, therefore

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum (x_i - \bar{x})^2 + n\bar{x}^2}{n\sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2}$$

Divide top and bottom by n:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{\sum (x_i - \bar{x})^2/n + \bar{x}^2}{\sum_{i=1}^n x_i^2 - \left(\sum x_i\right)^2/n}$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \frac{1/n + \bar{x}^2/(\sum(x_i - \bar{x})^2)}{1}$$

Finally:

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right]$$

This is the form you usually find in the text books. Is there an easier way to get there?

$$\bar{y} = \hat{\beta}_1 \bar{x} + \hat{\beta}_0$$

Then

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Remember that for a linear equation $y = ax + b$ and assuming x and b are independent random variables ($\bar{y}$ and $\hat{\beta}_1$ are random variables, $\bar{x}$ is assumed to be a constant):

$$Var(y) = a^2 Var(x) + Var(b)$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$Var(\bar{\beta}_0) = Var(\bar{y}) + \bar{x}^2 Var(\bar{\beta}_1)$$

Note: $\sum Var(y_i) = n\sigma^2$ and the variance of y_i is σ^2, therefore the variance in $\bar{y}$ is:

$$Var(\bar{y}) = Var\left(\frac{1}{n}\sum y_i\right) = \frac{1}{n^2}\sum Var(y_i) = \frac{\sigma^2}{n}$$

We know what the variance in β_1 is, therefore we can write:

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum (x_i - \bar{x})^2}$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

To summarize (assuming $\sigma_i^2 = \sigma^2$):

$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum (x_i - \bar{x})^2}$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum^n (x_i - \bar{x})^2}$$

What can we use as σ_2 assuming we don't have an estimate from our measurements? In the absence of any estimate for the variance of your measured data, use the MSE estimate:

$$\text{MSE} = \sigma^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

MSE for our example:

$$\text{MSE} = \sigma^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

y_i	$\hat{y}$	$y_i - \hat{y}$	$(y_i - \hat{y})^2$
1	0.6	0.4	0.16
1	1.3	-0.3	0.09
2	2	0	0
3	2.7	-0.7	0.49
4	3.4	-0.6	0.36
Sum (SSE)			1.1

$$\text{MSE} = 1.1/(5-2) = 0.366666$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$

x_i	1	2	3	4	5
y_i	1	1	2	2	4

Note $\sigma^2 = 0.3666$ from previous slide.

Using:

$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{x}^2}{\sum (x_i - \bar{x})^2}$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum^n (x_i - \bar{x})^2}$$

$$\sigma^2_{\hat{\beta}_0} = 0.4033$$

$$\sigma^2_{\hat{\beta}_1} = 0.0367$$

`http://www.livephysics.com/tools/mathematical-tools/calculate-linear-regression-graph-scatter-plot-line-fit/`

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$: So what?

First thing we can do is construct confidence limits:

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\ \sigma_{\hat{\beta}_0}$$

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\ \sigma_{\hat{\beta}_1}$$

Uncertainties in $\hat{\beta}_0$ and $\hat{\beta}_1$: So what?

The second thing we can do is get estimates on predicted values, ie if I have a x_i value what is the corresponding fitted y_i and what is the degree of uncertainty in y_i?

Since $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, all we need do is propagate the uncertainty in the slope and intercept to get the uncertainty in y as we did before:

$$\sigma^2_{\hat{y}} = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

where σ^2 is the variance of y: $[1/(n-2)] \sum (y_i - \bar{y})^2$ – called MSE (see previous slides)

When we have $\sigma^2_{\hat{y}}$ we can also compute its confidence limit:

$$\hat{y} \pm t_{n-2,\alpha/2} \; \sigma_{\hat{y}}$$

Uncertainty in $\hat{y}$

Example

$$\sigma^2_{\hat{y}} = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$

where σ^2 is the variance of y: $[1/(n-2)] \sum (y_i - \bar{y})^2$ (MSE)
Note $\sigma^2 = 0.366$ from previous slide. Let us predict y at an x value of 3.4:

$$y = 0.7 \times 3.4 - 0.1 = 2.28$$

$$\sigma^2_{\hat{y}} = 0.366 \left(1 + \frac{1}{5} + \frac{0.16}{10} \right) = 0.446 \qquad \sigma_{\hat{y}} = \sqrt{0.446} = 0.667$$

$$y = 2.28 \pm 0.667$$

Confidence limits : $\hat{y} \pm t_{\alpha/2, n-2} \sigma_{\hat{y}}$

$$2.28 \pm 3.181 \times 0.667 = 2.28 \pm 2.12$$

Hypothesis testing on $\hat{\beta}_1$

If we have some idea of the distribution of $\hat{\beta}_1$ we can do a hypothesis test.

The easiest thing to do is use the following null hypothesis:

$$H_o : \hat{\beta}_1 = 0 \quad H_1 : \hat{\beta}_1 \neq 0$$

that is there is no relationship. The test indicates if the fitted regression model is of value in explaining variations in the observations or if you are trying to impose a regression model when no true relationship exists between x and y.

We can now do a t-test:

$$t = \frac{\hat{\beta}_1 - 0}{\sigma_{\hat{\beta}_1}}$$

Using:

$$\sigma_{\hat{\beta}_1} = \sqrt{\frac{\sigma^2}{SS_{X^2}}}$$

Hypothesis testing on $\hat{\beta}_1$

Given the previous example with 5 data points, test H_o that the slope is zero at the 5% level. This is a two tail test.

$$H_o : \hat{\beta}_1 = 0$$

$$y = 0.7x - 0.1 \qquad \hat{\beta}_1 = 0.1916$$

$$t = \frac{0.7}{0.1916} = 3.65$$

Look up t-table using $n - 2$ degrees of freedom: df=3.

The corresponding p-value at this t is: 0.0355

At the 5% significance level, $p < 0.05$, therefore there is little evidence to support the hypothesis that the slope is zero. We therefore propose that there is a real relationship between the two variables.

Hypothesis testing on $\hat{\beta}_1$

Output from Python (See next slide for code):

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.817
Model:                            OLS   Adj. R-squared:                  0.756
Method:                 Least Squares   F-statistic:                     13.36
Date:                Sat, 25 Nov 2017   Prob (F-statistic):             0.0354
Time:                        12:01:23   Log-Likelihood:                -3.3094
No. Observations:                   5   AIC:                             10.62
Df Residuals:                       3   BIC:                             9.838
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1000      0.635     -0.157      0.885      -2.121       1.921
x1             0.7000      0.191      3.656      0.035       0.091       1.309
==============================================================================
Omnibus:                        nan   Durbin-Watson:                   2.509
Prob(Omnibus):                  nan   Jarque-Bera (JB):                0.396
Skew:                        -0.174   Prob(JB):                        0.821
Kurtosis:                     1.667   Cond. No.                         8.37
==============================================================================
```

Code for Results on Previous Slide

```
import numpy as np
import scipy.stats
import statsmodels.api as sm

x = [1,2,3,4,5]; y = [1,1,2,2,4]
# Uncomment the following line to get the results on the next page
#x = [1,2,3,4,5]; y= [3,3.5,2.5,3.,2.8]
l = scipy.stats.linregress (x, y)

x = sm.add_constant(x) # So that we have an intercept

model = sm.OLS(y,x)
results1 = model.fit()
print results1.summary()
```

Hypothesis testing on $\hat{\beta}_1$

x = [1,2,3,4,5]; y= [3,3.5,2.5,3.,2.8]
Output from Python:

OLS Regression Results

Dep. Variable:	y	R-squared:	0.152
Model:	OLS	Adj. R-squared:	-0.130
Method:	Least Squares	F-statistic:	0.5388
Date:	Sun, 26 Nov 2017	Prob (F-statistic):	0.516
Time:	16:55:18	Log-Likelihood:	-1.0804
No. Observations:	5	AIC:	6.161
Df Residuals:	3	BIC:	5.380
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.2300	0.407	7.943	0.004	1.936	4.524
x1	-0.0900	0.123	-0.734	0.516	-0.480	0.300

Omnibus:	nan	Durbin-Watson:	3.407
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.162
Skew:	-0.050	Prob(JB):	0.922
Kurtosis:	2.125	Cond. No.	8.37

Simple Regression

For a more general form which we might come back to later:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma_i} \right)^2 = \chi^2$$

This is called the **Weighted Least Squares Estimation**. Each difference is weighted by the standard deviation, σ_i, at that point.

Summary of Equations

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum^n (x_i - \bar{x})(y_i - \bar{y})}{\sum^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\text{MSE} = \frac{\sum (y_i - \hat{y}_i)^2}{N - 2}$$

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \left(\frac{1}{n} \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum^n (x_i - \bar{x})^2}$$

$$\sigma^2_{\hat{y}} = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)$$