

Session 7.5: Trainee Notes

Correlation and Linear Regression

Content

1. Correlation
2. Linear Regression

Learning Objective

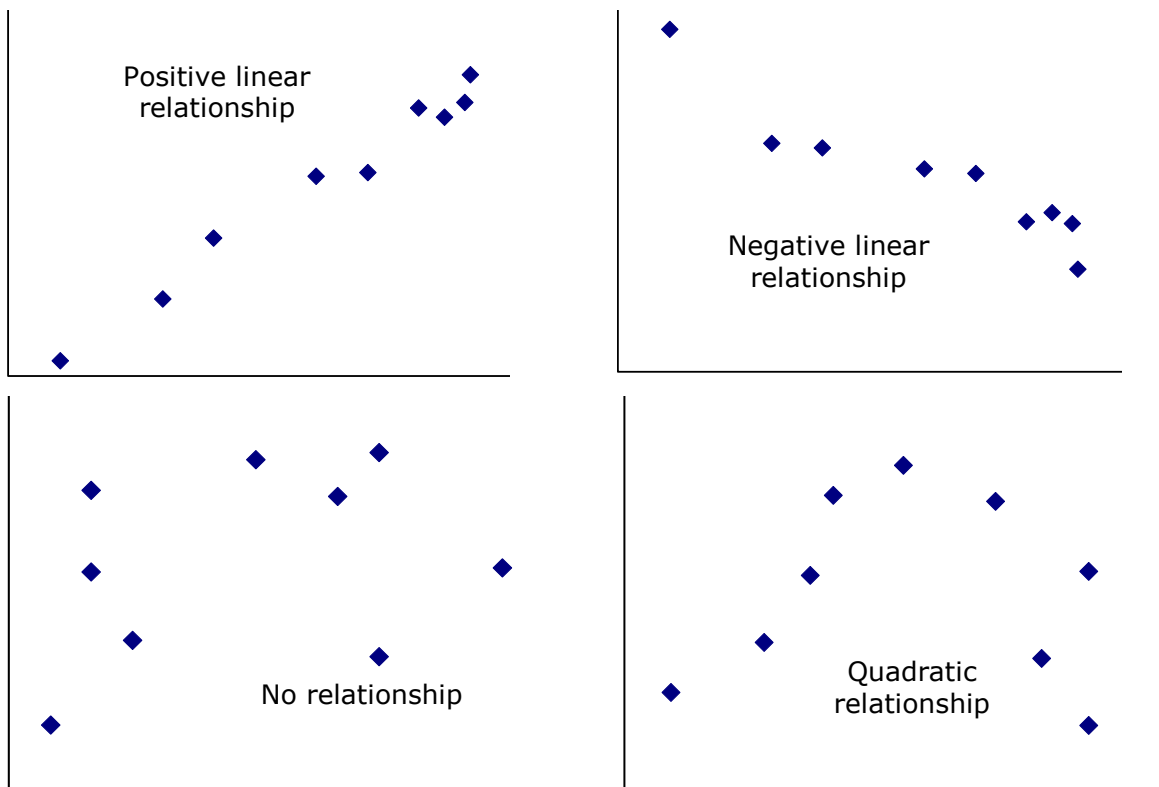
By the end of this session you will be able to:

- Calculate and understand correlation coefficients
- Predict linear regression parameters
- Use of two types of confidence interval on regression estimates

1. Correlation

Correlation is a statistical measurement of the strength of linear relationship between two variables. Possible correlations range from $+1$ to -1 . A zero correlation indicates that there is no relationship between the variables. A negative correlation indicates that as one variable goes up, the other goes down. A positive correlation means that both variables move in the same direction together. The size of the correlation provides an indication of the relationship between any two variables, the larger (in absolute terms) the value the stronger the nature of the relationship with values of 1 indicating a perfect linear relationship. The correlation can be used to provide a guide on how useful a linear regression equation might be in explaining one measure in terms of another.

Question: Is there a relationship between two continuous variables?



Examples:

Positive linear: e.g. weight and height

Negative linear: e.g. alcohol consumed and level of sobriety

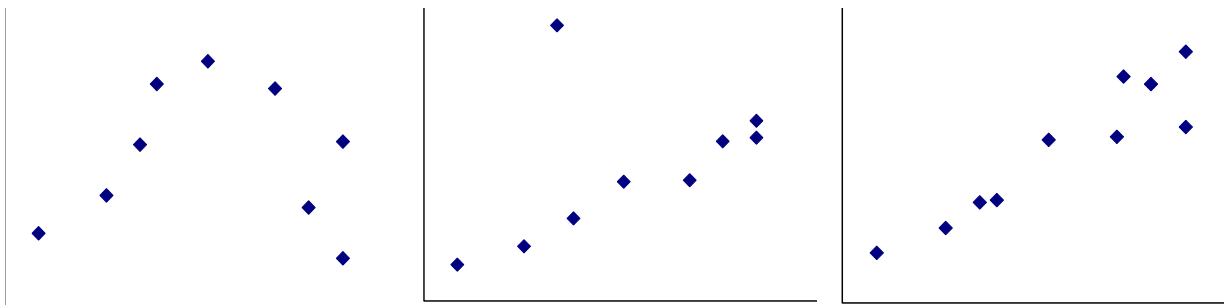
No relationship: e.g. weight and number of your house

Quadratic relationship: e.g. side length of a square and area

As illustrated above, drawing a scatter plot of your data points will indicate the direction and strength of any relationship and is a useful first step when comparing two measures.

In the scatter plot, look for:

- A non linear relationship, as in the first scatter plot given below - if not linear then transform the x values to produce a linear relationship
- Any outliers, as indicated in the second scatter plot below, - investigate why they appear as outliers, i.e. is it a data recording error?
- Equal variance, as indicated in the third scatter plot below, - approximate 'cigar shape' of points
- Check if the data points are approximately normally distributed using a p-p plot, (a p-p plot, probability-probability plot or percent-percent plot, is a plot for assessing how closely two data sets agree, and plots the two cumulative distribution functions against each other).



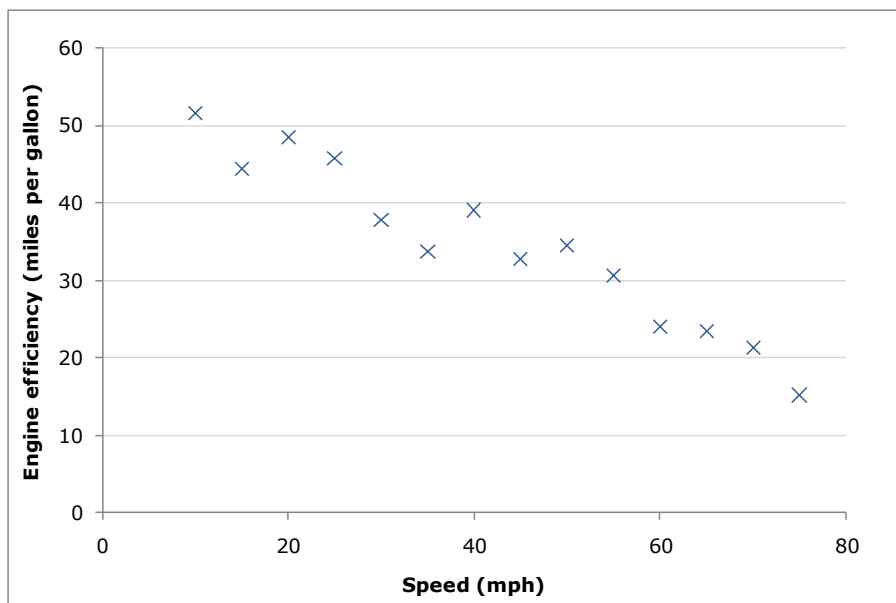
IF you are satisfied with all of the above then:

Calculate a correlation coefficient

Correlation coefficients tell you the direction and strength of the linear relationship between two variables.

- A coefficient takes values between -1 and +1.
- The +/- tells you whether your relationship is positive or negative.
- The value tells you how strong the relationship is.

Example of an Excel scatter plot (mpg v speed):



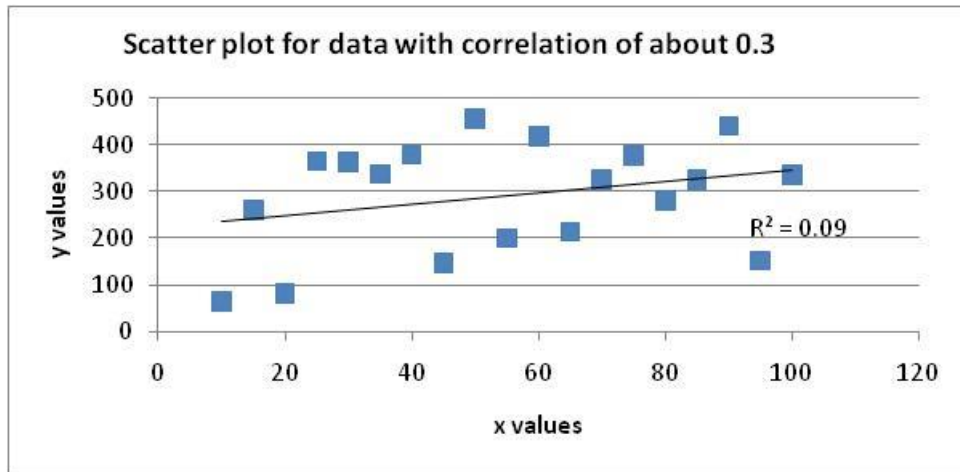
Excel Add-in 'Correlation' produces the following:

	<i>speed</i>	<i>mpg</i>
<i>speed</i>	1	
<i>mpg</i>	-0.96597	1

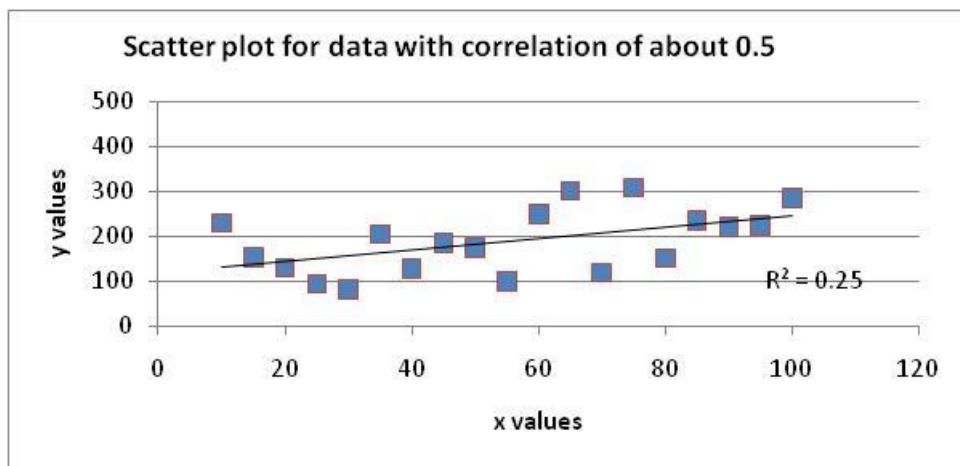
How to interpret your correlation coefficient

Interpretation of the size of a correlation coefficient suggests that:

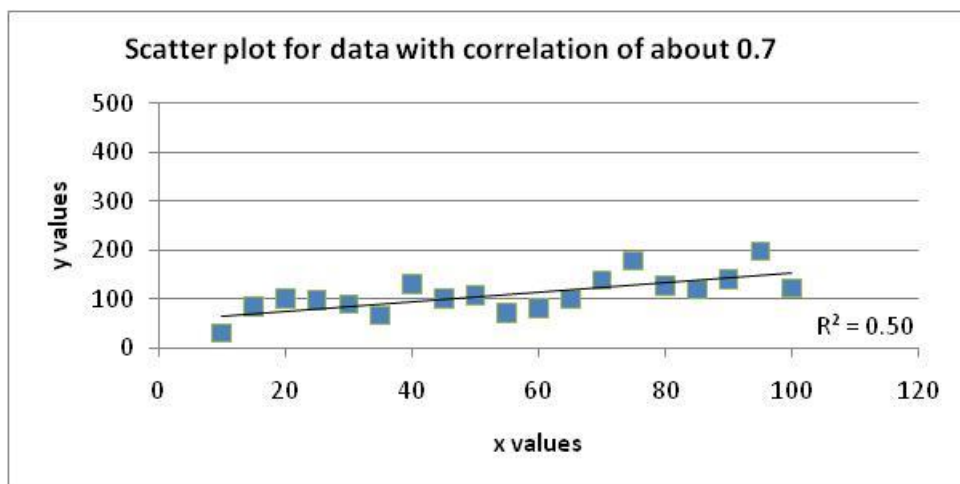
- A correlation between (0.10 and 0.29) or (-0.29 and -0.10) is small, i.e. quite a weak relationship, for example:



- A correlation between (0.30 and 0.59) or (-0.59 and -0.30) is medium, i.e. a reasonable relationship, for example:



- A correlation between (0.60 and 1.00) or (-1.00 and -0.60) is large, i.e. a fairly strong relationship, for example:



Calculate a coefficient of determination:

The *coefficient of determination*, (as calculated by r^2 or R^2), is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y . The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit.

For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by a regression equation). The other 15% of the total variation in y remains unexplained. The *coefficient of determination* is a measure of how well the regression line represents the data.

If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

- It shows how much variation the two variables share
- Calculated by squaring r and multiply by 100%
- For example, two variables with $r=0.5$ share $(0.5*0.5)*100\%=25\%$ of the variation

The earlier example of mpg and speed has a coefficient of determination of $(0.966)^2 = 93.3\%$, which is high but not surprising given the strength of the relationship as seen in the scatter plot. Only 6.7% of the 'noise' in the mpg data is unexplained by the speed being travelled.

Statistical significance of correlation:

- Treat this with care!
- Statistical significance of r (i.e. the probability that the null-hypothesis of a zero correlation), is strongly affected by the number of cases.
- Large samples may have a small correlation value but still be statistically significantly different from zero. It is useful to know that some of the variation between two measures is being explained but if it is very low then it may not be very helpful.
- Suggest that even low (but statistically significant) values are reported, but it is better to focus on the amount of shared variance (coefficient of determination) which will be small – and so not necessarily useful.

2. Linear regression

The theory

Once you have discovered some strength of linear relationship we can estimate the relationship.

In regression the two variables are called the dependent (or response / y) variable and the explanatory (or independent / regressor / x) variable.

- Explanatory variable:
 - values controlled or selected by the person experimenting
 - **not** subject to experimental variation
- Dependent variable:
 - observed to change in response to the explanatory variable
 - has some associated variability because of measurement error

We want to investigate whether the dependent variable depends on the explanatory variable by fitting a regression equation to the data to summarise the relationship.

Simple linear regression is the technique for fitting the 'best' model for a straight line relationship between the two variables. The nature of the relationship can be seen from the scatter plot, i.e. if it is linear, if not then it may be necessary to transform the 'x' values to produce a linear relationship. A linear relationship is summarised by the equation:

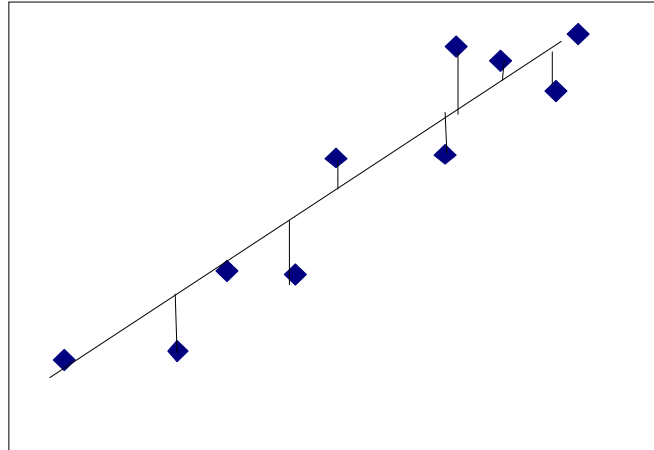
$$y = \beta_0 + \beta_1 x$$

where

- β_0 is the intercept; and
- β_1 is the slope of the line (rate of increase of y with x)

There is variation in the observed responses, the y_i values, $i=1\dots n$. The 'best line' is one that minimises the vertical distances between the observed data and the fitted line. The figure shows the vertical lines, which are referred to as the *errors*,

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



The procedure is called the *method of least squares*, where we fit a model of the following form:

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

such that we minimise the function $S = \sum \varepsilon_i^2$.

The 'best' fitted line is then denoted by:

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

where the 'hats' on β_0 and β_1 are to indicate that these are *estimates* of the intercept and slope, and are therefore subject to variability. The \hat{y} are referred to as *fitted values*, or *predicted values*. The differences between the observed and fitted values are called the *residuals*, i.e. $(y_i - \hat{y}_i)$ or ε_i .

Calculating and testing β_1

The *slope* of the regression equation estimated as:

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

The precision of this estimate is given by its *standard error*

$$\text{s.e.}(\hat{\beta}_1) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

where s is the standard error of the estimate = $\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$.

We can use this to test the hypothesis about the slope of the relationship, using a t-test with $(n-2)$ degrees of freedom.

$H_0: \beta_1 = 0$ (i.e. no significant slope)

$H_1: \beta_1 \neq 0$ (i.e. significant slope)

$$t = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)}$$

A 95% confidence interval can be calculated for the slope using the following expression:

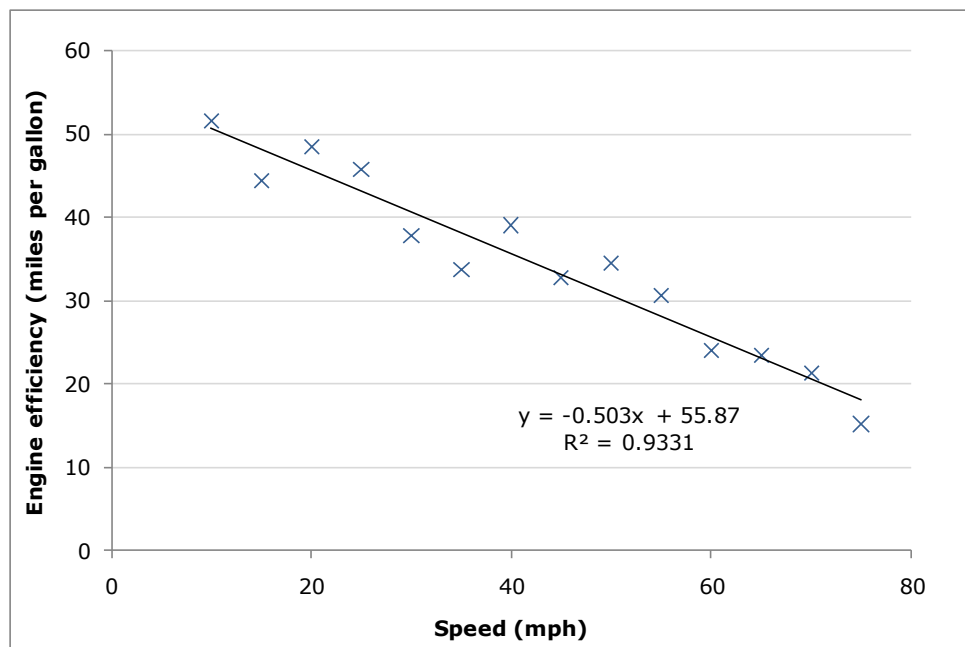
$$\hat{\beta}_1 \pm t_{n-2} * se(\hat{\beta}_1)$$

Computing a regression equation with Excel

Example

In Excel:

Option 1: Within Excel, insert a scatter plot and fitting the trend line (selecting the 'fit trend line' option within the scatter plot), request the equation and R² value. The example shows a strong negative relationship between speed and mpg, the linear relationship explains 93.3% of the variability, i.e. coefficient of determination.



Option 2: Using the Excel Add-in 'Regression', requesting 95% confidence interval on parameters, residuals and a residual plot:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.966
R Square	0.933
Adjusted R Square	0.927
Standard Error	2.939
Observations	14

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1441.33	1441.34	166.866	2.12E-08
Residual	12	103.65	8.6377		
Total	13	1544.98			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	55.900	1.8331	30.495	9.7E-13	51.906	59.894
speed	-0.5034	0.0390	-12.918	2.12E-08	-0.588	-0.418

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted mpg</i>	<i>Residuals</i>
1	50.86588	0.740158
2	48.34883	-3.92979
3	45.83179	2.677644
4	43.31474	2.482039
5	40.79769	-2.94948
6	38.28064	-4.56772
7	35.76359	3.386155
8	33.24654	-0.41052
9	30.7295	3.811667
10	28.21245	2.502038
11	25.6954	-1.73301
12	23.17835	0.282544
13	20.6613	0.675477
14	18.14425	-2.9672



Interpreting 'regression' output

Model summary shows:

- Correlation coefficient (R) – multiple correlation between the observed values and those fitted from the linear regression.
- Coefficient of determination (R square).
- Adjusted R square is an R squared value which takes into account the number of parameters in the regression (the coefficient of determination to use).
- Standard error allows significance tests to be made on R.

Coefficients table shows:

- Un-standardized coefficients and standard errors show the regression line
 - Estimated coefficient for constant is β_0 - the intercept
 - Estimated coefficient for explanatory measure is β_1 - the slope, i.e. speed
- t-tests to determine whether the coefficient values are significant
 - A significance level with $p < 0.05$ means that the explanatory variable is needed in the regression model
- Confidence interval for coefficients

Residual output shows the difference from the observed and predicted values using the regression equation. The plot of residuals by speed shows that the residuals lie equally about the '0' axis and that there is no obvious bias, i.e. similar size and numbers of residuals about the '0' axis across the whole range of speeds and suggests that they are Normally distributed.

Prediction using the regression line

If there is a significant linear regression, then it is possible to predict likely responses at certain values of the explanatory variable.

Prediction for the *average response* about the fitted line for the population at value x_0 is given by:

$$\text{Predicted value} \quad \hat{y}_o = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\text{With standard error} \quad s.e.(\hat{y}_o) = \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$$

The standard error takes into account the prediction error from fitting the

regression equation, i.e. $s.e.m = \sqrt{s^2 \left(\frac{1}{n} \right)}$, where s^2 is the residual error from

fitting the regression, plus a measure of the 'noise' error when not at the average x-value.

The prediction for an *individual point* at value x_0 is given by the same equation, but has an increased associated standard error:

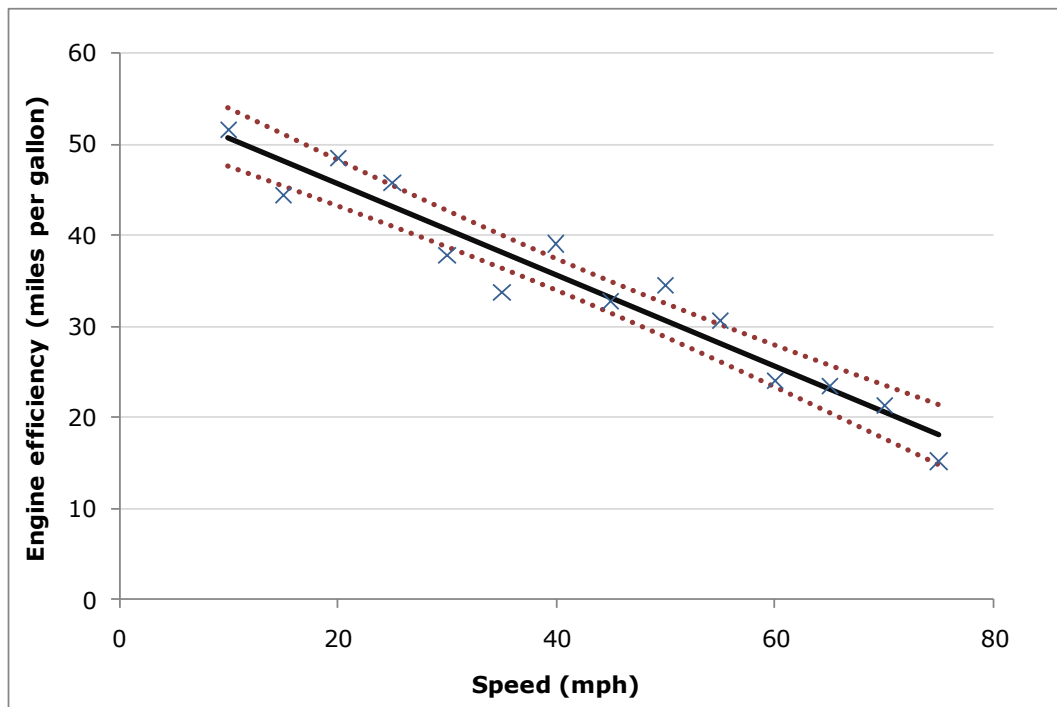
i.e. Predicted value $\hat{y}_o = \hat{\beta}_0 + \hat{\beta}_1 x_0$

With standard error $s.e.(\hat{y}_o) = \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$

The precision for an individual point is greater than that for a value predicted from the mean value since the 'residual noise' (s^2) is included within the estimated point standard error.

The precision is greatest closer to the mean value of the explanatory variable (x) and decreases towards the extremes as is illustrated in the following figures.

Predicted regression line with 95% confidence interval for 'average' responses



The standard error is higher for the individual point estimate because it needs to take into account both the error from the fitted model and the error associated with individual observations. This is illustrated by the two figures where the 95% confidence interval for the average response of values is the 'dotted' curve and the 95% confidence interval for predicted individual points is the 'dashed' curve.

Predicted regression line with 95% confidence interval for 'point' responses

