Regression Analysis

Regression analysis is a statistical technique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical sciences, economics, management, life and biological sciences, and the social sciences. In fact, regression analysis may be the most widely used statistical technique.

The Strategy Of Regression Analysis

A comprehensive regression analysis involves several different phases, each of which is important to the successful fitting of regression models. We describe these phases by the acronym ISEAS: Data **investigation**, model **specification**, parameter **estimation**, model **assessment**, variable **selection**.

Investigation

The first step in a regression analysis is to investigate (I) the data base. A **listing** of the data base, especially if it is entered into a computer file, should be scanned in order to detect any obvious errors in the data entrye.g., negative values for variates that can only be positive, or decimal points entered in the wrong position. Erroneous or incorrectly coded data values can have a serious effect on the estimation of the model parameters; consequently, the detection of outliers at this stage of the analysis can save much time and effort later on and can lessen the chance of erroneous inferences due to outlier effects. Along with the scanning of the data base, **summary statistics** for each of the variables should be computed. Incorrectly coded or erroneous data values often can be identified by examining the maximum and minimum data values for each variable. This is especially important when data bases are so large that scanning the individual data values is difficult.

Plotting response and predictor variables is another important step in the first phase of a regression analysis. A sequence plot of the response variable against its run number or against a time index can often identify nonrandom patterns in the experiment. For example, abrupt shifts could indicate unforeseen or unplanned changes in the operating conditions of an experiment. Cyclic trends could indicate a time dependence among the responses.

The response and predictor variables should also be plotted against one another. Such scatterplots may indicate that one or more of the variates need to be reexpressed prior to the fitting of a regression model. For example, plotting the response variable against each of the predictor variables may indicate the need to reexpress the response variable if several of the plots show the same nonlinear trend. If only one or two of the plots show nonlinear trends, the predictors involved should be reexpressed.

Box plots and stem-and-leaf plots are valuable aids for identifying possible outliers in a single variable. Scatterplots are useful for identifying combinations of data values that might not be apparent from an examination of each variable separately. In a scatterplot an observation might appear in a corner without having an extremely large or small value on either of the variates.

Specification

The second stage of a regression analysis is to initially specify (S) the form of the regression model. Not only must functional forms for each of the variables be specified, but one must consider whether interaction terms, polynomial terms, or nonlinear functions of the predictor variables should be included in the model.
The specification of a regression model relies heavily on one's knowledge of the experiment or process being studied. Not only does such knowledge aid one in selecting the variables to be included in the experiment, and thus in the model, but it is also important when questions concerning the functional form of the model are discussed. Plots of the variables, as suggested above, can aid in the initial specification of the model, especially when an experiment is being conducted or a process is being studied for which little is known about the effects of variables on the response.
Ordinarily, if no indications to the contrary are shown by plots or by one's knowledge of the problem being studied, all variables are entered into the model in a linear fashion with no transformations.

Estimation

The next stage in a regression model is to estimate (E) the model parameters, ordinarily using appropriate computer software. In addition to just the estimates of the model parameters, regression programs usually provide statistics that assist one in assessing the fit. Thus the estimation stage leads directly to the assessment stage of the analysis.

Assessment

When assessing (A) the adequacy of the model fit one must consider several questions.

- Are the model assumptions reasonable, including the functional form of the response and predictor variables ?
- If so, does the model adequately fit the data ?
- Can the errors be considered to be normally distributed with zero means and constant standard deviations?
- Are any observations unduly influencing the fit ?

Selection

Following the assessment of the specified model, perhaps following a respecification of the model or some other remedial action, one proceeds to a selection (S) of the individual predictors. This is done because often one is unsure, prior to the experiment, which of the predictor variables affect the response.

Linear Regression model for data

A simple linear regression presents us with, n pairs in the plane (x_1, y_1),(x_2, y_2),...,(x_n, y_n):

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

where (x_i, y_i) is the i^{th} data point and

1. x_i is a realization of the "independent" or predictor random variable X.
2. y_i is a realization of the "dependent" or prediction random variable Y.
3. β_0 is the y-intercept parameter, generally unknown.
4. β_1 is the slope or *rate of change parameter*, generally unknown.
5. ϵ_i $i = 1, 2, \ldots, n$ are unobservable "noise" or "error" random variables with mean zero and constant variance σ^2
6. β_0 + β_1 x is called the true unknown "regression function" of y on X. ie, E[Y]=β_0 + β_1x.

X	Y	X^2	Y^2	XY
x_1	y_1	$x^2 1$	y_1^2	$x_1 y_1$
x_2	y_2	$x^2 2$	y_2^2	$x_2 y_2$
x_3	y_3	$x^2 3$	y_3^2	$x_3 y_3$
x_4	y_4	$x^2 4$	y_4^2	$x_4 y_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$x^2 n$	y_n^2	$x_n y_n$
$\sum_{i=1}^n x_i$	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n x_i^2$	$\sum_{i=1}^n y_i^2$	$\sum_{i=1}^n x_i y_i$

An Example

Example 2.1. A rocket motor is manufactured by bonding an igniter propellant and a sustainer propellant together inside a metal housing. The shear strength of the bond between the two types of propellant is an important quality characteristic. It is suspected that shear strength is related to the age in weeks of the batch of sustainer propellant. Twenty observations on shear strength and the age of the corresponding batch of propellant have been collected and are shown in Table 2.1.

	S-Strength(psi)	Age	S-Strength^2	Age^2	S-Strength $\times$ Age
	2158.70	15.50	4659985.690	240.250	33459.850
	1678.15	23.75	2816187.423	564.062	39856.062
	2316.00	8.00	5363856.000	64.000	18528.000
	2061.30	17.00	4248957.690	289.000	35042.100
	2207.50	5.50	4873056.250	30.250	12141.250
	1708.30	19.00	2918288.890	361.000	32457.700
	1784.70	24.00	3185154.090	576.000	42832.800
	2575.00	2.50	6630625.000	6.250	6437.500
	2357.90	7.50	5559692.410	56.250	17684.250
	2256.70	11.00	5092694.890	121.000	24823.700
	2165.20	13.00	4688091.040	169.000	28147.600
	2399.55	3.75	5757840.203	14.062	8998.312
	1779.80	25.00	3167688.040	625.000	44495.000
	2336.75	9.75	5460400.562	95.062	22783.312
	1765.30	22.00	3116284.090	484.000	38836.600
	2053.50	18.00	4216862.250	324.000	36963.000
	2414.40	6.00	5829327.360	36.000	14486.400
	2200.50	12.50	4842200.250	156.250	27506.250
	2654.20	2.00	7044777.640	4.000	5308.400
	1753.70	21.50	3075463.690	462.250	37704.550
Totals	42627.150	267.250	92547433.458	4677.688	528492.637
	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n x_i$	$\sum_{i=1}^n y_i^2$	$\sum_{i=1}^n x_i^2$	$\sum_{i=1}^n x_i y_i$

$$\widehat{\beta_1} = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{20(528,492.637) - (267.250)(42,627.15)}{20(4,677.688) - (267.25)^2} = -37.15359$$

and

$$\widehat{\beta_0} = \frac{\sum_{i=1}^n y_i - \widehat{\beta_1}\sum_{i=1}^n x_i}{n} = \overline{y} - \widehat{\beta_1}\overline{x} = 2,131.35 - (-37.15359)(13.3625) = 2627.822$$

Error random variable: ϵ_i's

There are general assumptions that are made on the error ϵ_i's and subsequently on the responses y_i's:

1. The ϵ_i's are normally distributed, this also implies that the y_i's are normally distributed.
2. $E(\epsilon_i) = 0$, the mean of the errors is zero, which implies $E[Y]=\beta_0 + \beta_1 x$.
3. $Var(\epsilon_i) = \sigma^2$, which also implies that $Var(y_i) = \sigma^2$, and
4. $Cov(\epsilon_i, \epsilon_j) = 0$, $i \neq j$, which also implies that $Cov(y_i, y_j) = 0$, $i \neq j$

Principle of Least Squares

One can fit any number of models using least squares

1. linear models

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

2. polynomials

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p$$

3. Other functions that are linear in the parameters to be estimated

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^{\frac{3}{2}} + \frac{\beta_3}{ln \mid x_{1i} \mid} \\ y_i &= \beta_0 + \beta_1 e^{-2x_{1i}} + \beta_2 sin\left(x_{1i} x_{2i}\right) + \beta_3 x_{1i} ln\left(x_{2i}^2\right) tan\left(x_{3i}\right) \end{aligned}$$

4. nonlinear models that have been “linearized”.

Suppose we let

$$\begin{aligned}
\epsilon_i &= y_i - (\beta_0 + \beta_1 x_i) \quad \text{then} \\
\epsilon_i^2 &= [y_i - (\beta_0 + \beta_1 x_i)]^2
\end{aligned}$$

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

We wish to find estimates of β_0 and β_1, call them $\widehat{\beta_0}$ and $\widehat{\beta_1}$, that would minimize Q.

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\
and & \\
\frac{\partial Q}{\partial \beta_1} &= -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)
\end{aligned}$$

After taking derivatives and setting equal to zero, and passing through the summation operator, we can solve the following equations:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= n\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{n} x_i \\
\text{and} \\
\sum_{i=1}^{n} x_i y_i &= \widehat{\beta_0}\sum_{i=1}^{n} x_i + \widehat{\beta_1}\sum_{i=1}^{n} x_i^2
\end{aligned}$$

Solving first for $\widehat{\beta_1}$ and them for $\widehat{\beta_0}$ we have the following estimates:

$$\begin{aligned}
\widehat{\beta_1} &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
\text{and} \\
\widehat{\beta_0} &= \frac{\sum_{i=1}^{n} y_i - \widehat{\beta_1}\sum_{i=1}^{n} x_i}{n} \\
&= \overline{y} - \widehat{\beta_1}\overline{x}
\end{aligned}$$

We can simplify the expression for β_1 to:

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

where

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} \end{aligned}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

and also

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2$$

The fitted regression line

The fitted regression line is given by after finding estimates for β_0 and β_1

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

Define a residual as:

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

that is the actual (original) value of the response minus the fitted (predicted) value. After getting $\widehat{\beta_0}$ and $\widehat{\beta_1}$ we can substitute these values back into Q to find this minimum.

$$\hat{Q} = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_i \right) \right]^2 = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^{\,2}$$

This sum which can be re-written as

$$S_{yy} - \widehat{\beta_1} S_{xy} = SS_E$$

is called the residual or error sum of squares (SS_E). An unbiased estimate of σ^2 is given by

$$\begin{aligned} \widehat{\sigma^2} &= \frac{SS_E}{n-2} = MS_E \\ &= S^2 \end{aligned}$$

which is referred to as the mean square error (or literally the mean of the squared errors)

Is the regression significant ?

One of the first question we need to answer is "Is the regression significant". In the linear case we are basically asking do we have a linear relationship. Is there statistical evidence to conclude that the slope of the true regression line is different from zero ?

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
$\beta_1 = 0$	Error assumptions (n > 2)		$\beta_1 > 0$	$T \geq t_{n-2,\alpha}$
		$\mathrm{T} = \frac{\widehat{\beta}_1 - 0}{S_{\widehat{\beta}_1}}$	$\beta_1 \neq 0$	$\mid T \mid \geq t_{n-2,\alpha/2}$
		$= \frac{\widehat{\beta}_1 - 0}{s\sqrt{\frac{1}{S_{xx}}}}$		
			$\beta_1 < 0$	$T \leq t_{n-2,\alpha}$

If the null hypothesis is rejected then the regression is generally considered "significant"

Table 1: Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square	F-test
Regression	1	$SS_R = \hat{\beta}_1 S_{xy}$	$MS_R = \frac{\hat{\beta}_1 S_{xy}}{1}$	$\frac{MS_R}{MS_E}$ —— $F_{1,n-2}$
Error	$n-2$	$SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$	$MS_E = \frac{SS_E}{n-2}$	
Total	S_{yy}	$n-1$		

Of lesser significance is the test for the intercept term $\beta_0 = 0$.

Null Hypothesis	Assumptions	Test Statistic	Alternative Hypothesis	Rejection Region
$\beta_0 = 0$	Error assumptions (n > 2)		$\beta_0 > 0$	$T \geq t_{n-2,\alpha}$
		T $= \frac{\widehat{\beta}_0 - 0}{S_{\widehat{\beta}_0}}$	$\beta_0 \neq 0$	$\mid T \mid \geq t_{n-2,\alpha/2}$
		$= \frac{\widehat{\beta}_1 - 0}{s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}}}$		
			$\beta_0 < 0$	$T \leq t_{n-2,\alpha}$

It should be noted that the linear model can be re-written so that the intercept term is zero.

Distributional properties of $\widehat{\beta_0}$ and $\widehat{\beta_1}$

The method of least squares along with the Gauss-Markov theorem that the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are the "best linear unbiased estimators" for β_0 and β_1. But what is the distribution of these estimates? Without going into the derivation of the variances of these estimates and following the assumptions made on the error and response random variables, we have that

$$\widehat{\beta_0} \quad \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n S_{xx}}\right)$$

Similarly, it can be shown for slope parameter estimate, $\widehat{\beta_1}$,

$$\widehat{\beta_1} \quad \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Example:

Inferences on the parameters $\widehat{\beta_0}$ and $\widehat{\beta_1}$

$$\frac{\widehat{\beta_0} - \beta_0}{S_{\widehat{\beta_0}}} \sim t_{n-2}$$

and

$$\frac{\widehat{\beta_1} - \beta_1}{S_{\widehat{\beta_1}}} \sim t_{n-2}$$

Confidence Interval for β_0.

Using the above a (1-α) (100 %) confidence interval for β_0 is given by

$$\begin{aligned}
\widehat{\beta_0} &\pm t_{n-2,\alpha/2} S_{\widehat{\beta_0}} \\
&\pm t_{n-2,\alpha/2} s \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n S_{xx}}}
\end{aligned}$$

Example:

Confidence Interval for β_1.

Using the above a (1-α) (100 %) confidence interval for β_1 is given by

$$\begin{aligned} \widehat{\beta_1} & \pm t_{n-2,\alpha/2} S_{\widehat{\beta_1}} \\ & \pm t_{n-2,\alpha/2} s \sqrt{\frac{1}{S_{xx}}} \end{aligned}$$

Example:

Inferences about the estimated regression

Suppose we wish to use or regression function to find the mean response, E(y), for a single measurement at a point x_0. Then we can use the fitted regression function as an estimate of the mean response,

$$\widehat{E(y)} = \widehat{y}_{x_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

Example:

to get this value. Note that it is important that this new point x_0 be within the range of the current x values.
If we wish to get a confidence interval for the mean reponse, E(y), at the "new" value x_0 we would have:

$$\widehat{y}_{x_0} \pm t_{\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

Example:

Predicting a response at a point $\mathbf{x}_p$

Suppose we wish to use our regression to "predict" a response for some future point, say x_p, of the current values then the best predictor is obviously,

$$\widehat{y}_{x_p} = \widehat{\beta_0} + \widehat{\beta_1} x_p$$

If we wish to get a confidence interval for the future value we would have:

$$\widehat{y}_{x_p} \pm t_{\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{xx}}}$$

Example:

Predicting a mean response of m future measurements at a point $\mathbf{x}_p$

Suppose we wish to use our regression to "predict" m responses at some future point, say x_p. Then best estimate of the the mean of the responses $\overline{y}$ of the current values then the best predictor is obviously,

$$\widehat{y}_{x_p} \quad = \quad \widehat{\beta_0} + \widehat{\beta_1} x_p$$

If we wish to get a confidence interval for the future value we would have:

$$\widehat{y}_{x_p} \pm t_{\alpha/2,n-2} s \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{S_{xx}}}$$

Example:

Coefficient of Determination

$$\begin{aligned} R^2 &= \frac{SS_R}{S_{yy}} \\ &= 1 - \frac{SS_E}{S_{yy}} \end{aligned}$$

proportion of variability in the responses that can be explained by or accounted for by the predictor X. Example:

Correlation

We can define the Pearson product moment correlation coefficient (r) as an estimate of the true population coefficient ρ.

$$\begin{aligned} r &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \\ &= \widehat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}} \end{aligned}$$

The correlation is a measure of the strength and direction of the relationship between x and y. Note that it does not imply that x causes y or influences y. It just measures the strength of the relationship.
For our example we have a correlation of $\approx$ -0.95, which says that there is a moderately strong relationship between x and y.