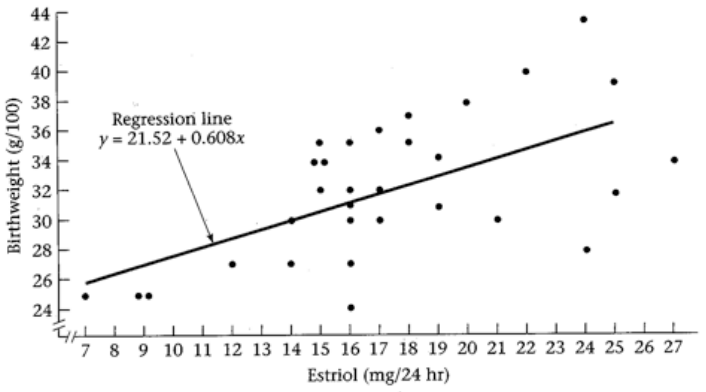
Statistics

March 6, 2014 Debdeep Pati

Regression methods

Greene and Touchstone conducted a study to relate birthweight to the estriol level of pregnant women. How can this relationship be quantified?

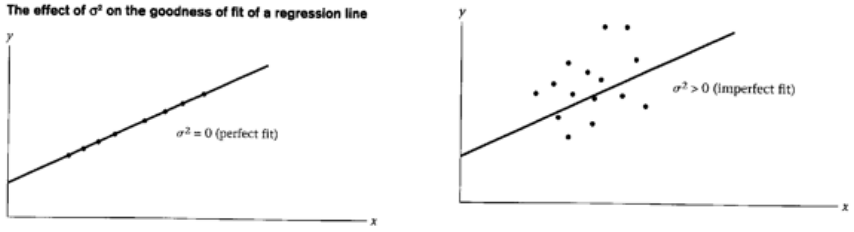
Data from the Greene-Touchstone study relating birthweight and estriol level in pregnant women near term



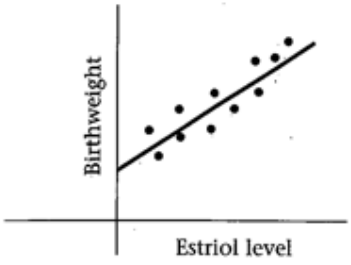
Source: Reprinted with permission of the *American Journal of Obstetrics and Gynecology, 85*(1), 1–9, 1963.

1. $y = \alpha + \beta x + \epsilon$.
2. ϵ is normally distributed with mean 0 and variance σ^2. y is called the dependent variable. x is called the independent variable.

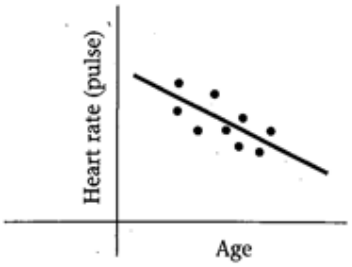
The effect of σ^2 on the goodness of fit of a regression line



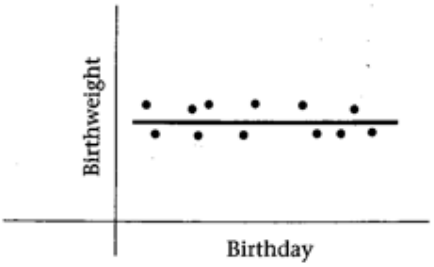
The interpretation of the regression line for different values of β



(a) $\beta > 0$



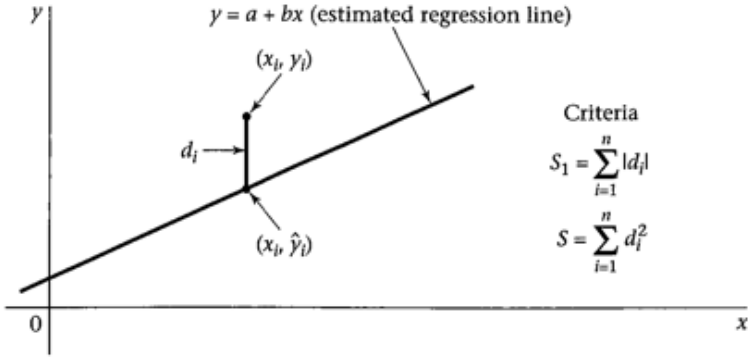
(b) $\beta < 0$



(c) $\beta = 0$

Criteria for fitting a regression line

Possible criteria for judging the fit of a regression line



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

$$a = \bar{y} - b\bar{x}$$

Prediction using Fitted Regression Line

The predicted, or average, value of y for a given value of x, as estimated from the fitted regression line, is denoted by $\hat{y} = a + bx$

R-squared

For any sample point (x_i, y_i), the residual, or residual component, of that point about the regression line is defined by $\hat{y}_i = a + bx_i$. For any sample point (x_i, y_i), the regression component of that point about the regression line is defined by $\hat{y}_i - \bar{y}$. Decompose the total sum of squares (TSS) into regression (Reg(SS) and residual components (Res(SS)).

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

Coefficient of determination, $R^2 = Reg(SS)/TSS$. The proportion of variation in the dependent variable y that can be explained by the independent variable x.

1 Case Study

A group of children who lived near a lead smelter in El Paso, Texas, were identified and their blood levels of lead were measured. An exposed group of 46 children were identified who had blood-lead levels $\geq 40\mu g/ml$. A control group of 78 children were also identified who had blood-lead levels $< 40\mu g/ml$. Two outcome variables were studied. The number of finger-wrist taps in the dominant hand and the Wechsler full-scale IQ score.

Treatment of Outliers

Figure 8.11 Stem-and-leaf and box plots of finger–wrist tapping score by group, El Paso Lead Study

(a) Control

Stem	Leaf	#	Boxplot
8	4	1	0
7	69	2	|
7	224	3	|
6	55558	5	|
6	01122344	8	+------+
5	566677778999	12	| |
5	000000011222334	15	*--+--*
4	566666888999	12	+------+
4	02	2	|
3	8	1	|
3			
2	6	1	0
2	3	1	0
1			
1	3	1	0

----+----+----+----+

Multiply Stem.Leaf by 10**+1

(b) Exposed

Stem	Leaf	#	Boxplot
8	3	1	0
7			
7	0	1	|
6			|
6	2	1	|
5	567789	6	+------+
5	0122244	7	| |
4	56889	5	*--+--*
4	0012244	7	+------+
3	5788	4	|
3	4	1	|
2			
2			
1			
1	34	2	0

----+----+----+----+

Multiply Stem.Leaf by 10**+1

21

Detecting outliers

The Extreme studentized deviate (or ESD statistic) $= \max_{i=1,\ldots,n} |x_i - \bar{x}|/S$. Suppose we have a sample $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$, but feel that there may be some outliers present. To test the hypothesis H_0 : no outliers present versus H_1 : that a single outlier is present, with a type I error of α.

1. We compute the Extreme Studentized Deviate test statistic (ESD). The sample value x_i, such that $ESD = |x_i - \bar{x}|/S$ is referred to as $x^{(n)}$.

2. We refer to Table 10 in the Appendix of BR to obtain the critical value $= ESD_{n,1-\alpha}$.

3. If $ESD > ESD_{n,1-\alpha}$, then we reject H_0 and declare that $x^{(n)}$ is an outlier. If $ESD < ESD_{n,1-\alpha}$, then we declare no outliers are present

Evaluate whether outliers are present for the finger-wrist tapping scores in the control group. The sample mean for control group is 54.4, sample standard deviation is 12.1, and n = 64. $ESD = |13-54.4|/12.1 = 3.44$ with 13 being the most extreme value. ($|84-54.4| < |13-54.4|$). From Table 10, $ESD_{70,.95} = 3.26$. $3.44 > ESD_{70,.95} = 3.26 > ESD_{64,.95}$. The finger-wrist tapping score of 13 is an outlier.

Multiple Regression

1. Hypertension: How the relationship between the blood-pressure levels of newborns and blood-pressure levels of infants relates to subsequent adult blood pressure.

2. The blood pressure of newborn is affected by several extraneous factors that make this relationship difficult to study. In particular, newborn blood pressures are affected by birthweight the day of life on which blood pressure is measured.

3. In this study, the infants were weighed as the time of the blood-pressure measurements. This birthweight is different from actual birthweight. Birthweights at 5 days are different from those at 2 days. We want to adjust the observed blood pressure for these two factors before we look at other factors.

Model

$$y_i = \alpha + \sum_{i=1}^{k} \beta_j x_{ij} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$$

F test for simple linear regression

To test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$ vs. H_1 : at least one $\beta_j \neq 0$. In general consider the problem where two models, 1 and 2, where model 1 is 'nested' within model 2. Model 1 is the Restricted model, and Model 2 is the Unrestricted one. That is, model 1 has p_1 parameters, and model 2 has p_2 parameters, where $p_2 > p_1$. The model with more parameters will always be able to fit the data at least as well as the model with fewer

parameters. Thus typically model 2 will give a better (i.e. lower error) fit to the data than model 1. But one often wants to determine whether model 2 gives a significantly better fit to the data. One approach to this problem is to use an F test.

If there are n data points to estimate parameters of both models from, then one can calculate the F statistic, given by

$$F = \frac{(RSS_1 - RSS_2)/(p_2 - p_1)}{RSS_2/(n - p_2)}$$

where RSS_i is the residual sum of squares of model i. Under the null hypothesis that model 2 does not provide a significantly better fit than model 1, F will have an F distribution, with $(p_2 - p_1, n - p_2)$ degrees of freedom. The null hypothesis is rejected if the F calculated from the data is greater than the critical value of the F-distribution for some desired false-rejection probability (e.g. 0.05). The F-test is a Wald test.

T Test for an individual independent variable

$H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$.
Compute $t = \hat{\beta}_j / SE(\hat{\beta}_j)$ which follows a t distribution with $(n - k - 1)$ degrees of freedom under H_0. If $t > t_{n-k-1,1-\alpha/2}$ or $t < -t_{n-k-1,1-\alpha/2}$, then reject H_0, otherwise accept H_0. The exact p-value is given by

$$\begin{cases} 2P(t_{n-k-1} > t) \,\text{if}\, t > 0 \\ 2P(t_{n-k-1} < t) \,\text{if}\, t < 0 \end{cases}$$