SIMPLE LINEAR REGRESSION

Simple linear regression is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of *n* points in such a way that makes the sum of squared *residuals* of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.

Suppose there are *n* data points $\{y_i, x_i\}$,$i = 1, 2, \ldots, n$, which are *i*-th realizations of the random variables *Y* and *X* respectively. The goal is to find the equation of the straight line

$$y_i = a + bx_i + \varepsilon_i, \quad i = 1,2,\cdots,n$$

which would provide a "best" fit for the data points. In the above model, the intercept *a* and the slope *b* are unknown constants and ε_i is a random error component. The errors are assumed to have mean 0 (*zero*) and unknown variance σ^2. Additionally, we usually assume that the errors are uncorrelated. That is, the value of one error does not depend on the value of any other error. So, we assume that, given $X = (x_1, x_2, \cdots, x_n)$ be the set independent variables

$$E\left(\varepsilon_i \middle| X\right) = E\left(\varepsilon_i\right) = 0, \;\; Var\left(\varepsilon_i \middle| X\right) = Var\left(\varepsilon_i\right) = \sigma^2 \text{ and } Cov\left(\varepsilon_i \varepsilon_j\right) = 0.$$

Thus the conditional variance of error term is constant and does not depend on explanatory variable *X*. (Homoscedasticity Property)

It is convenient to view the regressor *x* as a mathematical variable controlled by the data analyst and measured with negligible error, whereas the response *y* is a random variable due to the presence of error component. So, there is probability distribution for *y* at each possible value for *x*. The mean of this distribution is

$$E\left(y|x\right)=a+bx$$

and the variance is

$$Var\left(y|x\right)=Var\left(a+bx+\varepsilon\right)=\sigma^2 .$$

Thus, the mean of *y* is a linear function of *x* although the variance of *y* does not depend on the value of *x*. Furthermore, because the errors are uncorrelated, the responses are also uncorrelated.

LEAST-SQUARES ESTIMATION OF THE PARAMETERS

Here the "best" will be understood as in the **least-squares** approach: such a line that minimizes the sum of squared residuals of the linear regression model. In other words, *a* and *b* solves the following minimization problem:

$$\min_{a,\, b} Q\left(a,b\right), \text{ where } Q\left(a,b\right)=\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\sum_{i=1}^{n}\left(y_i-a-bx_i\right)^2$$

The estimates of *a* and *b* are obtained by minimize the objective function *Q*. Differentiating *Q* partially with respect to *a* and *b* and equating them to zero, we get

$$\frac{\partial Q}{\partial \hat{a}}=\sum_{i=1}^{n}(-2)\left[y_i-\left(\hat{a}+\hat{b}x_i\right)\right]=0, \textbf{ and}$$

$$\frac{\partial Q}{\partial \hat{b}}=\sum_{i=1}^{n}(-2)x_i\left[y_i-\left(\hat{a}+\hat{b}x_i\right)\right]=0.$$

On simplification, they give us

$$\sum_{i=1}^{n} y_i = \hat{a}n + \hat{b}.\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{a}.\sum_{i=1}^{n} x_i + \hat{b}.\sum_{i=1}^{n} x_i^2$$

These equations are known as **least square normal equations.**

Solving this system of equations by method of elimination, we can find the least square estimates of a and b.

From first normal equation, we have

$$\hat{a} = \frac{\sum_{i=1}^{n} y_i - \hat{b}\sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{b}\bar{x}$$

Now, putting the expression for $\hat{a}$ in the second normal equation and simplifying, we have

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

$$\Rightarrow \hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

To verify whether the solution is really a minimum, the matrix of second order derivatives of Q, the Hessian matrix, must be positive definite. It is easy to show that

$$\boldsymbol{H}(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) = \begin{bmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix} = 2\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

Now, $|\boldsymbol{H}| = 2n\sum_{i=1}^{n}(x_i - \bar{x})^2$

and above is clearly positive definite, since $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$.

Example 1

To illustrate, let us consider the following data on the number of hours which ten persons studied for a French test and their scores on the test:

Hours Studied, x	Test Score, y	x^2	$x \times y$	y^2	
4	31	16	124	961	
9	58	81	522	3364	
10	65	100	650	4225	
14	73	196	1022	5329	
4	37	16	148	1369	
7	44	49	308	1936	
12	60	144	720	3600	
22	91	484	2002	8281	
1	21	1	21	441	
17	84	289	1428	7056	
100	**564**	**1376**	**6945**	**36562**	**←TOTALS**

Therefore, S_{xx} = 1376 – 1/10*(100)^2 = 376 and

S_{xy} = 6945 – 1/10*(100)×(564) = 1305.

Thus, $\hat{b}$ = 1305/376 = 3.471 and $\hat{a}$ = 56.4 – 3.471*10 = 21.69.

So, least square regression equation is $\hat{y} = 21.69 + 3.471x$.

PROPERTIES OF REGRESSION COEFFICIENTS

Linearity

Regression coefficient estimates are random variables, since they are just linear combinations of y_i and the y_i are random variables, since

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} y_i\left(x_i - \bar{x}\right)}{S_{xx}} = \sum_{i=1}^{n} w_i y_i, \quad where \quad w_i = \frac{x_i - \bar{x}}{S_{xx}}.$$

Unbiasedness

Let us now investigate the bias and variance properties of these estimates.

We have, from above

$$\hat{b} = \sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n} w_i\left(a + bx_i + \varepsilon_i\right) = a\sum w_i + b\sum w_i x_i + \sum w_i \varepsilon_i$$

Now, clearly,

$$\sum w_i = 0 \quad \text{and} \quad \sum w_i x_i = 1$$

So, the above expression reduces to $\hat{b} = b + \sum w_i \varepsilon_i$.

Therefore,

$$E(\hat{b}) = E\left(b + \sum w_i \varepsilon_i\right)$$
$$= E(b) + \sum w_i E(\varepsilon_i) = b, \quad [\because E(\varepsilon_i) = 0]$$

Thus, $\hat{b}$ is an unbiased estimate of the true slope b.

Similarly,

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{1}{n}\sum\left(a + bx_i + \varepsilon_i\right) - \hat{b}\bar{x}$$
$$= a + b\bar{x} + \bar{\varepsilon} - \hat{b}\bar{x}$$
$$= a - \left(\hat{b} - b\right)\bar{x} + \bar{\varepsilon}$$

Since $E\left(\hat{b}\right) = b$ and $E(\bar{\varepsilon}) = 0,$ we get $E(\hat{a}) = a$. Thus, $\hat{a}$ is also an unbiased estimate of the true intercept a.

Now, let us try to obtain the variances of the estimates. We have,
$\hat{b} = b + \sum_{i=1}^{n} w_i\varepsilon_i \Longrightarrow \hat{b} - E\left(\hat{b}\right) = \sum_{i=1}^{n} w_i\varepsilon_i.$

$$\operatorname{var}\left(\hat{b}\right) = E\left[\left\{\hat{b} - E\left(\hat{b}\right)\right\}^2\right] = E\left[\left(\sum_{i=1}^{n} w_i\varepsilon_i\right)^2\right]$$

$$= E\left[\sum_{i=1}^{n} w_i^2\varepsilon_i^2 + \text{cross product terms involving } \varepsilon_i \& \ \varepsilon_j, \ \ i \neq j\right]$$

$$= \sum_{i=1}^{n} w_i^2 E\left(\varepsilon_i^2\right) = \sigma^2\sum_{i=1}^{n} w_i^2 = \sigma^2 \times \frac{S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

Again,

$$\operatorname{var}(\hat{a}) = E[\{\hat{a} - E(\hat{a})\}^2] = E\left[\left\{-\left(\hat{b} - b\right)\bar{x} + \bar{\varepsilon}\right\}^2\right]$$
$$= \bar{x}^2 E\left[\left(\hat{b} - b\right)^2\right] + E[\bar{\varepsilon}^2] - 2\bar{x}E\left[\left(\hat{b} - b\right)\bar{\varepsilon}\right]$$
$$= \bar{x}^2\operatorname{var}\left(\hat{b}\right) + E[\bar{\varepsilon}^2] - 2\bar{x}E\left[\left(\hat{b} - b\right)\bar{\varepsilon}\right]$$

Now, $E\left(\bar{\varepsilon}^2\right) = \frac{\sigma^2}{n}$ and

$$E\left[\left(\hat{b}-b\right)\bar{\varepsilon}\right] = E\left[\left(\sum w_i\varepsilon_i\right)\left(\frac{1}{n}\sum \varepsilon_i\right)\right]$$
$$= E\left[\frac{1}{n}\left(\sum w_i\varepsilon_i^2 + \text{cross} - \text{product terms involving } \varepsilon_i \& \ \varepsilon_j,\ \ i \neq j\right)\right]$$
$$= \frac{1}{n}\sum w_i E\left(\varepsilon_i^2\right) = \frac{1}{n}\sigma^2\sum w_i = 0$$

Therefore,

$$\operatorname{var}(\hat{a}) = \bar{x}^2\frac{\sigma^2}{S_{xx}} + \frac{\sigma^2}{n} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right].$$

Putting together what we know so far, we can describe $\hat{b}$ as a *linear, unbiased estimator* of b, with a variance given by σ^2/S_{xx}. Similarly, $\hat{a}$ can be described as a linear, unbiased estimator of a, with a variance given by $\sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$.

To show that OLS estimates are best (i.e. has least variation), we will show that if there exist another linear unbiased estimator other than $\hat{b}$, then its variation must be greater than that of $\hat{b}$.

Let $\hat{b}^* = \sum k_i y_i$ be any other linear estimator of b. Suppose that, $k_i = w_i + c_i$, where c_i is a non-zero constant and w_i is as defined earlier.

$$\hat{b}^* = \sum k_i y_i = \sum\left(w_i + c_i\right)(a + bx_i + \varepsilon_i)$$
$$= a\sum w_i + a\sum c_i + b\sum w_i x_i + b\sum c_i x_i + \sum\left(w_i + c_i\right)\varepsilon_i$$
$$= a\sum c_i + b + b\sum c_i x_i + \sum\left(w_i + c_i\right)\varepsilon_i$$

Taking mathematical expectation of $\hat{b}^*$and noting that $E(\varepsilon_i) = 0$, we find that in order the above estimate to be unbiased it is necessary that $\sum c_i = \sum c_i x_i = 0$. So, in order for $\hat{b}^* = \sum k_i y_i$ to be in the class of *linear unbiased estimators*, it has to be

$$\hat{b}^* = b + \sum (w_i + c_i)\varepsilon_i$$

Now,

$$var(\hat{b}^*) = var\left[b + \sum (w_i + c_i)\varepsilon_i\right] = \sum (w_i + c_i)^2\, var(\varepsilon_i)$$

$$= \sigma^2 \sum (w_i + c_i)^2$$

$$= \sigma^2 \sum w_i^2 + \sigma^2 \sum c_i^2 \qquad \left[\sum c_i w_i = \sum c_i \frac{x_i - \bar{x}}{S_{xx}} = 0\right]$$

$$= var(\hat{b}) + \sigma^2 \sum c_i^2$$

$$\geq var(\hat{b})$$

Above establishes that, for the family of linear and unbiased estimators $\hat{b}^*$, each of the alternative estimators has variance that is greater than or equal to that of the least squares estimator $\hat{b}$. It may be noted from above expression that if $c_i = 0$ for all i, then only $\mathrm{var}(\hat{b}^*) = \mathrm{var}(\hat{b})$, and in that case $\hat{b}^* = \hat{b}$. Thus, *there is no other linear and unbiased estimator* of b that is better than $\hat{b}$. Hence the OLS estimate $\hat{b}$ is BLUE.

ESTIMATION OF $\boldsymbol{\sigma^2}$

The estimate of σ^2 can be obtained from the residuals $e_i = y_i - \hat{y}_i$. So, sum of squares of residuals or error sum of squares will be

$$SS_E = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

This can be simplified to

$$\begin{aligned} \text{SS}_\text{E} &= \sum_i \left[y_i - \hat{a} - \hat{b}x_i\right]^2 \\ &= \sum_i \left[y_i - \bar{y} - \hat{b}(x_i - \bar{x})\right]^2 \\ &= \sum_i \left[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})\right]^2 \\ &= S_{yy} + \hat{b}^2 S_{xx} - 2\hat{b}S_{xy} \\ &= S_{yy} + \hat{b}S_{xy} - 2\hat{b}S_{xy} \qquad \left[\because\ \hat{b} = \frac{S_{xy}}{S_{xx}} \Rightarrow \hat{b}S_{xx} = S_{xy}\right] \\ &= S_{yy} - \hat{b}S_{xy} \end{aligned}$$

It can be proved that $\boldsymbol{E(SS_E) = (n-2)\sigma^2}$ and hence $\frac{SS_E}{n-2}$, i.e. MS_E gives an unbiased estimate of σ^2.

In simple linear regression the estimated standard error of the slope is $SE(\hat{b}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ and the estimated standard error of the intercept is $SE(\hat{a}) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$, where $\hat{\sigma}^2$ = MS$_\text{E}$.

TESTING THE SLOPE OF REGRESSION EQUATION

Let the hypothesis be $H_0: b = b_0$ vs $H_1: b \neq b_0$

Since y_i are independent normal random variables and $\hat{b}$ is a linear combination of them, so $\hat{b}$ is $N(b, \sigma^2/S_{xx})$. So, we can test validity of above hypotheses using the statistic

$$t_0 = \frac{\hat{b}-b_0}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{b}-b_0}{\sqrt{MS_E/S_{xx}}},$$

which has a *t distribution* with *n* -2 degrees of freedom under the null hypothesis. Thus we would reject null hypothesis if

$$|t_0| > t_{\alpha/2,n-2}.$$

A similar procedure can be used to test the hypothesis about the intercept. To test

$$H_o: a = a_0$$
$$H_1: a \neq a_0$$

we would use the statistic

$$t_0 = \frac{\hat{a} - a_0}{\sqrt{MS_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$$

and reject the null hypothesis if $|t_0| > t_{\alpha/2,n-2}$.

Testing the following hypothesis can be used to ensure the significance of regression

$$H_o: b = 0$$
$$H_1: b \neq 0$$

Failure to reject above null hypothesis is equivalent to concluding that there is no linear relationship between *x* and *y*.

Example 2

Let us test the significance of regression using the following data and model parameters.

$\hat{b} = 3.471$, n =10, S_{xx} = 376, S_{xy} = 1305 and S_{yy} = 4752.4.

The hypotheses to be tested are $H_0\colon b = 0 \; vs \; H_1\colon b \neq 0$ and we test at 0.01 level of significance.

Mean square error is given by

$$MS_E = \frac{S_{yy} - \hat{b}S_{xy}}{n-2} = \frac{4752.4 - 3.471 \times 1305}{8} = 27.84.$$

So, the test statistic becomes

$$t_0 = \frac{\hat{b}}{\sqrt{\frac{MS_E}{S_{xx}}}} = \frac{3.471}{\sqrt{\frac{27.84}{376}}} = 12.76.$$

Now, since above computed value of t_0 = 12.76 is much greater than t(0,005,8) = 3.36, we reject the null hypothesis and conclude that regression is significant.

ANOVA APPROACH FOR TESTING SIGNIFICANCE OF REGRESSION

ANOVA procedure partitions the total variability in the response variable into two components as described below.

$$\sum(\boldsymbol{y_i} - \bar{\boldsymbol{y}})^2 = \sum[(\hat{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}) + (\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i)]^2$$

$$= \sum(\hat{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^2 + \sum(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i)^2 + \mathbf{2}\sum(\hat{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i)$$

Now, from the normal equations, we get

$\sum(y_i - \hat{y}_i) = \sum e_i = 0$ (from the first normal equation), and

$\sum x_i(y_i - \hat{y}_i) = \sum x_i e_i = 0$ (from the second normal equation).

Also, it is easy to show that $\sum(\hat{y}_i - \bar{y}) = \hat{b}\sum(x_i - \bar{x})$.

So, the above basic identity reduces to

$$\sum(\boldsymbol{y_i} - \bar{\boldsymbol{y}})^2 = \sum(\hat{\boldsymbol{y}}_i - \bar{\boldsymbol{y}})^2 + \sum(\boldsymbol{y_i} - \hat{\boldsymbol{y}}_i)^2$$

The two components in the right hand side of above equation, respectively, measures the amount of variability in y_i accounted for by the regression line [*called* regression sum of squares, *denoted by* SS$_R$] and the residual variation left unexplained by the regression line [*called* error sum of squares, *denoted by* SS$_E$]. Thus, above equation can equivalently be written as

$$S_{yy} = SS_R + SS_E \qquad (1),$$

where $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total corrected sum of squares of y. But we have already shown that

$$SS_E = S_{yy} - \hat{b}S_{xy}, \qquad \text{or equivalently}$$

$$S_{yy} = \hat{b}S_{xy} + SS_E \qquad (2)$$

therefore comparing (1) and (2), we get $SS_R = \hat{b}S_{xy}$.

The Total SS has n-1 degrees of freedom as one degree of freedom is lost as a result of the constraint $\sum(y_i - \bar{y}) = 0$. Sum square regression SS_R has 1 degree of freedom as SS_R is completely determined by one parameter, namely, b. Finally, SS_E has $n-2$ degrees of freedom because two constraints are imposed on the deviation $(y_i - \hat{y}_i)$ that lead to the estimation of $\hat{a}$ and $\hat{b}$.

It can be shown that $E[SS_E/(n-2)] = \sigma^2$, $E(SS_R) = \sigma^2 + b^2 S_{xx}$ and that under H_0, SS_E/σ^2 and SS_R/σ^2 are independent χ^2 random variable with n -2 and 1d.f. respectively.

Cochran's Theorem: If all n observations y_i come from the same normal distribution with mean μ and variance σ^2 and total SS is decomposed into k sum of squares SS_i, each with degrees of freedom df_i, then SS_i/σ^2 $(i = 1, 2, \cdots, k \text{ and } k \leq n)$ are independent chi-square variables with df_i degrees of freedom if and only if $\sum df_i = \text{total } df, \text{i.e.} (n-1)$.

In SLR, $TSS = SS_R + SS_E$ and $df_R + df_E = 1 + (n-2) = n-1$. Also *under null hypothesis*, each y_i is distributed as Normal with mean $\hat{a} = \bar{y}$ and variance σ^2. So by virtue of Cochran's Theorem, both SS_R/σ^2 and SS_E/σ^2 are independent χ^2 variables with 1 and (n-2) degrees of freedom respectively .

The null hypothesis H_0: $b = 0$ is thus tested by the statistic $F_0 = \dfrac{SS_R/1.\sigma^2}{SS_E/(n-2).\sigma^2} = \dfrac{MS_R}{MS_E}$, which follows the $F_{1,\, n\text{-}2}$ distribution and we would reject$H_0, if\ f_0 > f_{\alpha,1,n-2}$. The test procedure is usually represented in an ANOVA table, as given below.

Analysis of Variance for Testing Significance of Regression

Source of Variation	***Degrees of Freedom***	***Sum of Squares***	***Mean Square***	***F_0***
Regression	**1**	$SS_R = \hat{b}S_{xy}$	$MS_R = SS_R/1$	MS_R/MS_E
Error	$n-2$	SS_E $= S_{yy} - \hat{b}S_{xy}$	$MS_E = \dfrac{SS_E}{n-2}$	
TOTAL	$n-1$	S_{yy}		

Note: $\hat{\sigma}^2 = MS_E$

It may be noted that for testing significance of regression ANOVA procedure is equivalent to *t*-test. We have, $T_0 = \frac{\hat{b}}{\sqrt{\sigma^2/S_{xx}}}$. Squaring both sides and using $\hat{\sigma}^2 = MS_E$, we get

$$T_0^2 = \frac{\hat{b}^2 S_{xx}}{MS_E} = \frac{\hat{b} S_{xy}}{MS_E} = \frac{SS_R}{MS_E} = \frac{MS_R}{MS_E} = F_0.$$

It is worthwhile to note that square of a *t*-random variable with ν degrees freedom is an *F*-random variable with (1,ν) degrees of freedom. Thus, the test using T_0 is equivalent to test based on F_0.

Exercise:

Show that $\boldsymbol{E(S_{yy}) = b^2 S_{xx} + (n-1)\sigma^2}$, where $\boldsymbol{S_{yy} = \sum (y_i - \bar{y})^2}$

CONFIDENCE INTERVALS

Confidence interval is a measure of overall quality of the regression line. If the error terms, ε_i in the regression model are $NID(0, \sigma^2)$, then it is already shown that (using $\hat{\sigma}^2 = MS_E$)

$$\frac{\hat{b} - b}{\sqrt{\frac{MS_E}{S_{xx}}}} \text{ and } \frac{\hat{a} - a}{\sqrt{MS_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$$

follow *t-distribution* with (*n*-2) degrees of freedom.

So, for a given level of significance (α), we have

$$P\left\{-t_{\alpha/2,n-2} \leq \frac{\hat{b} - b}{\sqrt{MS_E/S_{xx}}} \leq t_{\alpha/2,n-2}\right\} = 1 - \alpha$$

i.e. $P\left\{-\hat{b} - t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}} \leq -b \leq -\hat{b} + t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}\right\} = 1 - \alpha$

Above leads to a $100(1-\alpha)\%$ confidence interval on the slope *b* in simple linear regression as

$$\hat{b} - t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}} \le b \le \hat{b} + t_{\alpha/2,n-2}\sqrt{\frac{MS_E}{S_{xx}}}.$$

Similarly, a $100(1-\alpha)\%$ confidence interval on the intercept a in simple linear regression is

$$\hat{a} - t_{\alpha/2,n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \le a \le \hat{a} + t_{\alpha/2,n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

INTERVAL ESTIMATION OF THE MEAN RESPONSE

A major use of a regression model is to estimate the mean response $E(y)$ for a particular value of the regressor variable x. Let x_0 be a value of the regressor variable within the region in which the variable is explored and we wish to estimate the mean response, say $\mu_{y|x_0}$ at $x = x_0$. A point estimator of $\mu_{y|x_0}$can be obtained from the fitted model as

$$\hat{\mu}_{y|x_0} = \hat{a} + \hat{b}x_0$$

which is an unbiased estimator since

$$E\left(\hat{\mu}_{y|x_0}\right) = E\left(\hat{a} + \hat{b}x_0\right) = a + bx_0 = \mu_{y|x_0}\,.$$

We can write $\hat{\mu}_{y|x_0}$ as

$$\begin{aligned}\hat{\mu}_{y|x_0} &= \bar{y} - \hat{b}\bar{x} + \hat{b}x_0 \\ &= \bar{y} + \hat{b}(x_0 - \bar{x}) \\ &= \bar{y} + \tfrac{S_{xy}}{S_{xx}}(x_0 - \bar{x}) \\ &= \textstyle\sum_{i=1}^{n} y_i \left[\tfrac{1}{n} + \tfrac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right]\end{aligned}$$

Since y_i are independent normal random variables and $\hat{\mu}_{y|x_0}$ is linear combination of them, so $\hat{\mu}_{y|x_0}$ is normally distributed.

The variance of $\hat{\mu}_{y|x_0}$is

$$V\left(\hat{\mu}_{y|x_0}\right) = \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right]^2 \mathrm{Var}(y_i)$$

$$= \sigma^2 \textstyle\sum_{i=1}^{n}\left[\frac{1}{n^2} + \frac{(x_i-\bar{x})^2(x_0-\bar{x})^2}{S_{xx}^2} + 2\,\frac{(x_i-\bar{x})(x_0-\bar{x})}{nS_{xx}}\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}^2}\textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]$$

Hence, we have shown that,

$$\hat{\mu}_{y|x_0} \sim N\left(\mu_{y|x_0}, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

.

So, it follows that

$$\frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{\sqrt{\sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}} \sim N(0,1)$$

Therefore, taking MS_E as an estimator of σ^2, we get

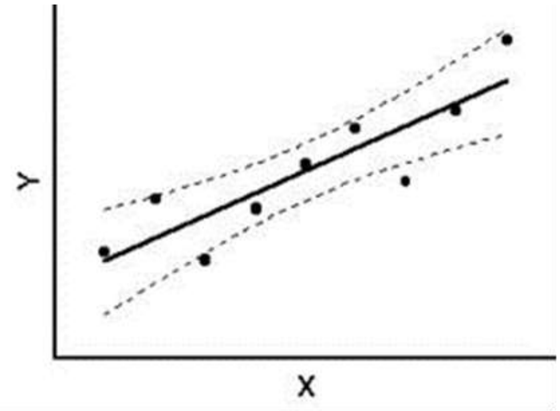
$$\frac{\hat{\mu}_{y|x_0} - \mu_{y|x_0}}{\sqrt{MS_E\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$$

Thus, $100(1 - \alpha)\%$ confidence interval about the mean response at $x = x_0$ (or about the regression line) is given by

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,\,n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \le \mu_{y|x_0}$$

$$\le \hat{\mu}_{y|x_0} + t_{\alpha/2,\,n-2}\sqrt{MS_E\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

It is to be noted that width of the confidence interval is a function of x_0 and the width is minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases.



Note: It may be noted that confidence interval gives the uncertainty in estimating the mean response due to the presence of uncertainty in the estimated model parameters.

PREDICTION OF NEW OBSERVATIONS

Let y_0 be the prediction of a new observation y corresponding to a specific level of the regressor variable x, say x_0. Clearly, the point estimate of the single future observation y_0 is

$$\hat{y}_0 = \hat{a} + \hat{b}x_0$$

To find the prediction interval that gives range of plausible values of the single future observation, we need to find the variation of error in prediction. In assessing the variation, we concentrate on the distribution of error in prediction, i.e. $y_0 - \hat{y}_0$.

Clearly, $E(\hat{y}_0) = E(\hat{a} + \hat{b}x_0) = a + bx_0 = y_0$. So $\hat{y}_0$ is an unbiased estimator of y_0. Moreover, $y_0 \sim N(a + bx_0, \sigma^2)$.

Again, we can write $y_0 - \hat{y}_0 = y_0 - (\hat{a} + \hat{b}x_0)$.

Therefore, $E(y_0 - \hat{y}_0) = E(y_0) - E(\hat{a} + \hat{b}x_0)$
$$= a + bx_0 - (a + bx_0) = 0.$$

It may be noted that y_0 is independent of the data values $y_1, y_2, \cdots, y_n$ that were used to estimate $\hat{a}$ and $\hat{b}$ and hence y_0 is independent of $\hat{a} + \hat{b}x_0$, i.e. $\hat{y}_0$.

Also, $var(y_0 - \hat{y}_0) = var(y_0) + var(\hat{y}_0)$

$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Hence we find that

$$y_0 - \hat{y}_0 \sim N\left(0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

So, it follows that

$$\frac{y_0 - \hat{y}_0}{\sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim N(0,1)$$

Therefore, taking MS_E as an estimator of σ^2, we get

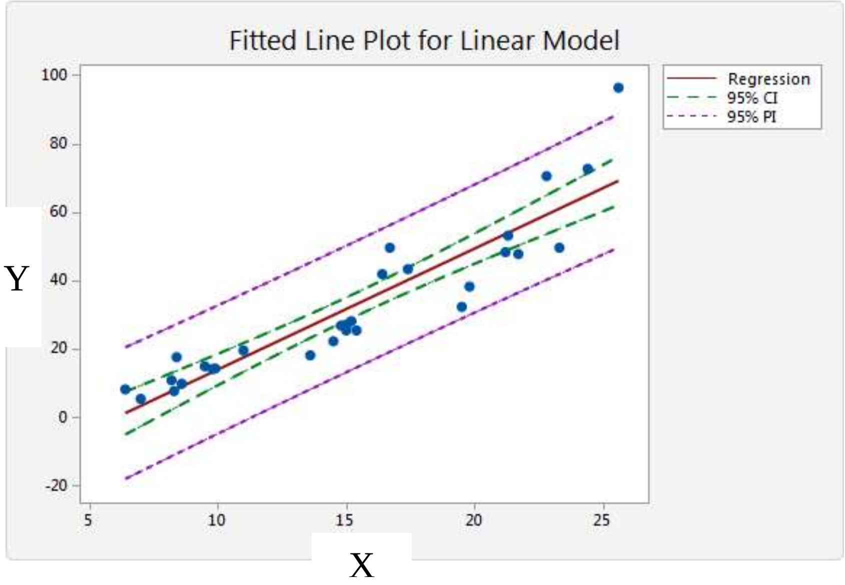
$$\frac{y_0 - \hat{y}_0}{\sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \sim t_{n-2}$$

$$P\left\{-t_{\alpha/2,n-2} \le \frac{y_0 - \hat{y}_0}{\sqrt{MS_E \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} \le t_{\alpha/2,n-2}\right\} = 1 - \alpha$$

Thus, $100(1 - \alpha)\%$ prediction interval on a future observation y_0 at $x = x_0$ is given by

$$\begin{aligned}\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_E\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]} &\le y_0 \\ &\le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_E\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}\end{aligned}$$

The prediction interval is of minimum width at $x_0 = \bar{x}$ and widens as $\left|x_0 - \bar{x}\right|$ increases.



ASSESSMENT OF REGRESSION MODEL

Throughout the discussion we have assumed that

1. Errors are
   a) normally distributed,
   b) distributed with mean ‘0’ and constant variance σ^2, and
   c) uncorrelated.

Above conditions are conveniently written as errors are $NID(0, \sigma^2)$.

2. Linear fit is the adequate fit.

We will now examine the adequacy of these assumptions.

Residual Analysis for Normality

Analysis of the residuals, i.e., $e_i = y_i - \hat{y}_i$, $i = 1, 2, \cdots, n$, is helpful in checking the assumption that errors are approximately normally distributed with constant variance.

As an approximate check of normality one can construct a histogram of the residuals or a normal probability of residuals.

The residuals, unlike the errors, do not all have the same variance: the variance depends on how further the corresponding *x*-value is from the average *x*-value. The fact that the variances of the residuals differ, even though the variances of the true errors are all equal to each other, it does not make sense to compare residuals at different data points without some sort of standardization.

Standardized Residuals

One may also standardize the residuals by computing

$$\boxed{d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}} = \frac{e_i}{\sqrt{MS_E}},} \quad i = 1, 2, \cdots, n.$$

If the errors are normally distributed, then approximately 95% of the standardized residuals should fall in the interval (-2, +2). Residuals that are far outside this interval may indicate the presence of outlier, i.e. an observation that is abnormal to the rest of the data. Sometimes outliers may provide important information about unusual circumstances of interest to experimenter and should be given due importance.

Studentized Residuals

Studentized Residuals are defined as $r_i = \frac{e_i}{SD(e_i)}$, $i = 1, 2, \cdots, n.$

Now,

$Var(e_i) = Var(y_i - \hat{y}_i) = Var(y_i) + Var(\hat{y}_i) - 2Cov(y_i, \hat{y}_i).$

Again, $Cov(y_i, \hat{y}_i) = Cov(\hat{y}_i + e_i, \hat{y}_i) = Var(\hat{y}_i) + Cov(\hat{y}_i, e_i).$

$$\begin{aligned} Cov(\hat{y}_i, e_i) &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}_i\right)(e_i - \bar{e}) \\ &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}(\hat{y}_i - \bar{y})e_i \qquad \text{since } \sum e_i = 0\,, i.e. \sum \hat{y}_i = \sum y_i \\ &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}\hat{y}_i e_i \qquad\qquad \text{since } \bar{y}\sum e_i = 0 \\ &= \tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}\left(\hat{a} + \hat{b}x_i\right)e_i \\ &= \tfrac{\hat{a}}{n-1}\textstyle\sum e_i + \tfrac{\hat{b}}{n-1}\sum x_i\, e_i \\ &= 0 \end{aligned}$$

Therefore,

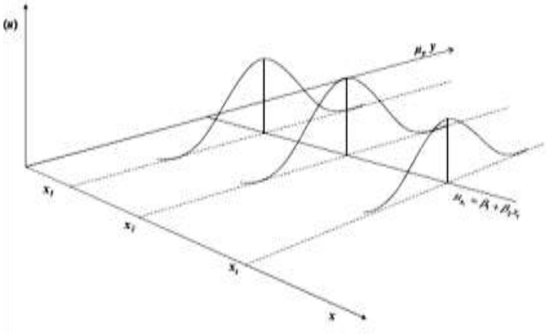
$$\begin{aligned} Var(e_i) = Var(y_i) - Var(\hat{y}_i) &= \sigma^2 - \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 - \left(\tfrac{1}{n} + \tfrac{(x_i-\bar{x})^2}{S_{xx}}\right)\right] \end{aligned}$$

So,

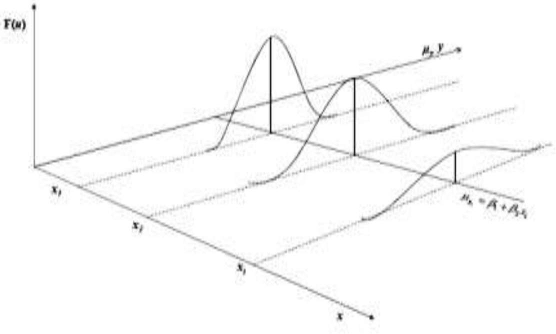
$$r_i = \frac{e_i}{\sqrt{\sigma^2\left[1 - \left(\frac{1}{n} + \frac{(x_i-\bar{x})^2}{S_{xx}}\right)\right]}}$$

Residual Analysis for Homoscedasticity

This assumption of constant variation is called the *homoscedasticity* assumption**.** The word comes from the Greek: *homo* (equal) and *scedasticity* (spread). This means that the variation of *y* around the regression line is the same across the *x* values; that is to say, it neither increases or decreases as *x* varies.

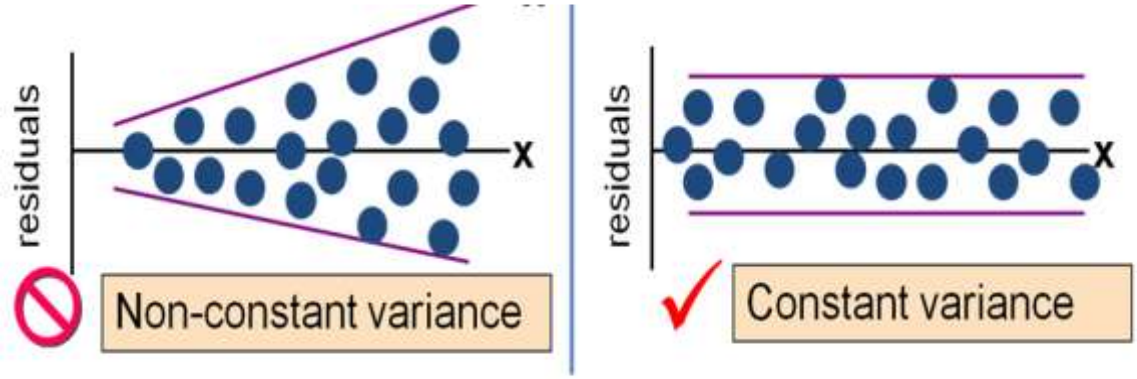


Homoscedastic



Heteroscedastic

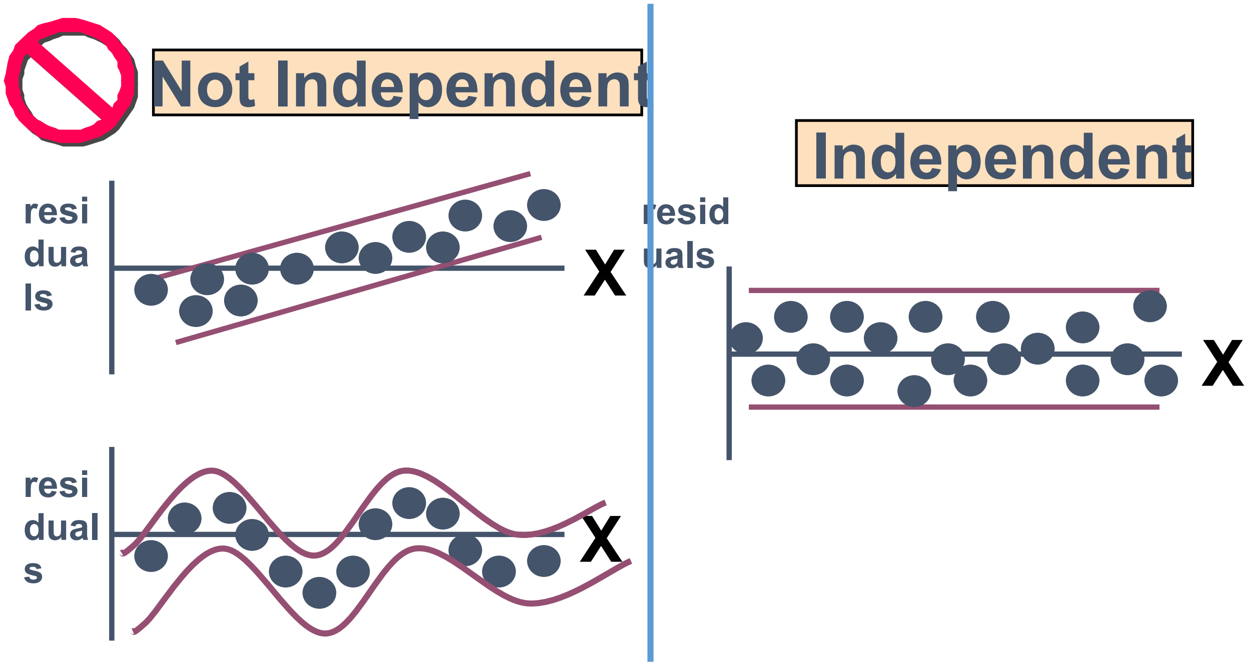
It is frequently helpful to plot the residuals (1) against the $\hat{y}_i$ and (2) against the x_i. If the plot is evenly and randomly distributed around the zero-residual-line, we will assume that there no abnormal pattern in the residuals. If the plot is funnel-shaped around the zero-residual-line, the variance of the observations is not remaining same over magnitude y_i or x_i.



Data transformation on the response y is often used to eliminate this problem. Widely used variance-stabilizing transformations include the use of $\sqrt{y}, ln(y), or\ \frac{1}{y}$ as the response. If the residual plot is found to be non-linear, the model requires higher order terms or possibility of including other independent variables should be explored.

Residual analysis for Independence

Here residuals are plotted against time sequence, i.e. in order of data collection.



Second plot in first column suggest auto correlation, as adjacent observations tend to have residuals of same sign.

Coefficient of Determination (R^2)

The quantity

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}}$$

is called the coefficient of determination, and is often used to judge the adequacy of the regression model. It should be noted that R^2 represents amount of variability in the data explained or accounted for by the regression model and since $0 \leq SS_R \leq S_{yy}$, $0 \leq R^2 \leq 1$.

Lack of Fit Test

Here we will test for the goodness of fit of the regression model. Specifically, we wish to test

H_0: The simple linear regression model is correct
H_1: The simple linear regression model is not correct.

The test involves partitioning the error SS into SS attributable to two components, namely, pure error and lack of fit of the model, that is,

$$SS_E = SS_{PE} + SS_{LOF}.$$

To compute $SS_{PE,}$ we must have repeated observations in the response for at least one level of x. Suppose we have *n* total observations such that

$$
\begin{array}{ll}
y_{11}, y_{12}, \cdots, y_{1n_1} & \text{repeated observations at } x_1 \\
\quad\cdots & \\
y_{j1}, y_{j2}, \cdots, y_{jn_j} & \text{repeated observations at } x_j \\
\quad\cdots & \\
\quad\cdots & \\
y_{m1}, y_{m2}, \cdots, y_{mn_m} & \text{repeated observations at } x_m
\end{array}
$$

The TSS for pure error would be obtained by summing over those levels of $x's$ that contain repeat observations.

$$SS_{PE} = \sum_{i} \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2$$

Note: In above the first sum is valid for those values of i for which x_i values have repeated y_{iu} values.

The degrees of freedom associated with the pure error SS is sum of degrees of freedoms of those x_i values having repeated y values.

The lack of fit sum square is simply $SS_{LOF} = SS_E - SS_{PE}$ with $(df_E - df_{PE})$ degrees of freedom. The test statistic for lack of fit would then be

$$F_0 = \frac{SS_{LOF}/df_{LOF}}{SS_{PE}/df_{PE}} = \frac{MS_{LOF}}{MS_{PE}}$$

and we would reject the hypothesis that model adequately fits the data if $f_0 > f_{\alpha, df_{\mathrm{LOF},} df_{\mathrm{PE}}}$.

Example 2

Consider the data on two variables, y and x shown below. Fit a simple linear regression model and test for lack of fit using α = 0.05

x	y
1.0	2.3, 1.8
2.0	2.8
3.3	1.8, 3.7
4.0	2.6, 2.6, 2.2
5.0	2.0
5.6	3.5, 2.8, 2.1
6.0	3.4, 3.2
6.5	3.4
6.9	5.0

The regression model is $\hat{y} = 1.697 + 0.259x$, the regression SS is SS_R = 3.4930, total SS is TSS = 10.83 and error SS is SS_E = 7.3372. The pure-error SS is computed as follows:

Level of x	$\sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2$	*Degrees of freedom*
1.0	0.1250	1
3.3	1.8050	1
4.0	0.0166	2
5.6	0.9800	2
6.0	0.0200	1
Totals	3.0366	7

So, lack of fit SS is $SS_{LOF} = SS_E - SS_{PE}$ = 7.3372 – 3.0366 = 4.3006.

ANOVA table for this data analysis is given below:

Source	DF	SS	MS	F_0	P-value
Regression	1	3.4930	3.4930	6.66	0.0218
Error	14	7.3372	0.5241		
(Lack of Fit)	*7*	*4.3006*	*0.6144*	*1.42*	*0.3276*
(Pure Error)	*7*	*3.0366*	*0.4338*		
Total	15	10.8302			

Since lack of fit is not significant, we cannot reject the null hypothesis that the tentative model adequately describes the data. Moreover, since regression is significant, we conclude that $b \neq 0.$

CORRELATION

We have so far assumed x to be a mathematical variable and y to be a random variable. But many applications of regression analysis involve situations where both x *and* y are random variables. In such situations, it is usually assumed that (x_i, y_i) are jointly distributed random variable obtained from bivariate normal distribution *f(x, y)* with μ_x and σ_x^2 as mean and variance of x and μ_y and σ_y^2 as mean and variance of y. For example, suppose we wish to develop a regression model relating the shear strength of spot welds to the weld diameters. Here welds diameter cannot be controlled and we would randomly select n spot welds and observe their diameters (x_i) and shear strength (y_i).

Correlation coefficient in such cases is defined as $= \frac{cov(x,y)}{\sigma_x \sigma_y}$. The estimate of ρ is the simple correlation coefficient and can be given by

$$r = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sqrt{\left[\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 . \sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2 \right]}} = \frac{S_{xy}}{\sqrt{S_{xx} . S_{yy}}}$$

We may also write

$$r^2 = \frac{S_{xy}^2}{S_{xx} . S_{yy}} = \frac{S_{xy}}{S_{xx}} . \frac{S_{xy}}{S_{yy}} = \frac{\hat{\beta} . S_{xy}}{S_{yy}} = \frac{SS_R}{S_{yy}} = R^2 .$$

Thus, the correlation coefficient is the square root of the coefficient of determination.

Example 1

Following table gives the purity of oxygen produced in a chemical distillation process and the % of hydrocarbons present at that time in the main condenser of the distillation unit.

Obs #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	**Total**	**Average**	**Sum_SQ**	**Sum_Prod**
HC Level %	0.99	1.02	1.15	1.29	1.46	1.36	0.87	1.23	1.55	1.40	1.19	1.15	0.98	1.01	1.11	1.20	1.26	1.32	1.43	0.95	23.92	1.1960	29.2892	2214.6566
Purity %	90.01	89.05	91.43	93.74	96.73	94.45	87.59	91.77	99.42	93.65	93.54	92.52	90.56	89.54	89.85	90.39	93.25	93.41	94.98	87.33	1843.21	92.1605	170044.5321	

a) Calculate the least square estimates of slope and intercept.

b) What % of total variability in Purity% is accounted for by the model?

c) Test the significance of the model thus obtained using ANOVA.

d) Obtain 95% confidence interval on i) slope and ii) intercept.

e) Construct a 95% confidence interval of mean purity level at HC level of 1.01.

f) Construct a 95% prediction interval at HC level % of 1.00.

Soln.

a) S_{xx}= 0.681, S_{xy} = 10.177, S_{yy} = 173.377

$\hat{b}$ = 14.944 $\hat{a}$ = 74.287

$$\hat{y} = 74.287 + 14.944\,x$$

$SS_R = \hat{b}S_{xy}$ = 152.085 $SS_E = S_{yy} - SS_R = 21.292$

b) $R^2 = \frac{SS_R}{S_{yy}}$ = 152.085/173.377 = 0.877192.

Thus, 87.7% of variability in purity % accounted for by the model.

c)

ANOVA table

Source of Variation	DF	SS	MS	f_0	Remark
Regression	1	152.085	152.085	128.559	Significant
Error	18	21.292	1.183		
Total	19	173.377			

d) 95% confidence interval on

i) slope: $\hat{b} - t_{0.025,18}\sqrt{\frac{MS_E}{S_{xx}}} \le b \le \hat{b} - t_{0.025,18}\sqrt{\frac{MS_E}{S_{xx}}}$

$$\Rightarrow 14.944 - 2.101\sqrt{1.183/0.681} \le b \le 14.944 + 2.101\sqrt{1.183/0.681}$$

⇨ [12.715, 17.713]

ii) Intercept: [70.936, 77.638]

e) Mean purity at x = 1.01 is 89.380 and the confidence interval is [89.095, 89.665].

f) Predicted value at x = 1.00 is 89.231 and the prediction interval is [86.827, 91.635]

Exercise 1

Show that an equivalent way to define the test for significance of regression in simple linear regression is to base the test on R^2as follows:

To test $H_0\colon b = 0$ versus $H_1\colon b \neq 0$, calculate

$$F_0 = \frac{R^2(n-2)}{1-R^2}$$

and reject the null hypothesis, if the computed value $f_0 > f_{\alpha,1,n-2}$.

Hence test the significance of regression at $\alpha = 0.05$ for a simple linear regression fit based on $n = 25$ observations with $R^2 = 0.90$.