Regression Analysis

GIVEN: x, an ordinary variable which can be measured without appreciable error, and

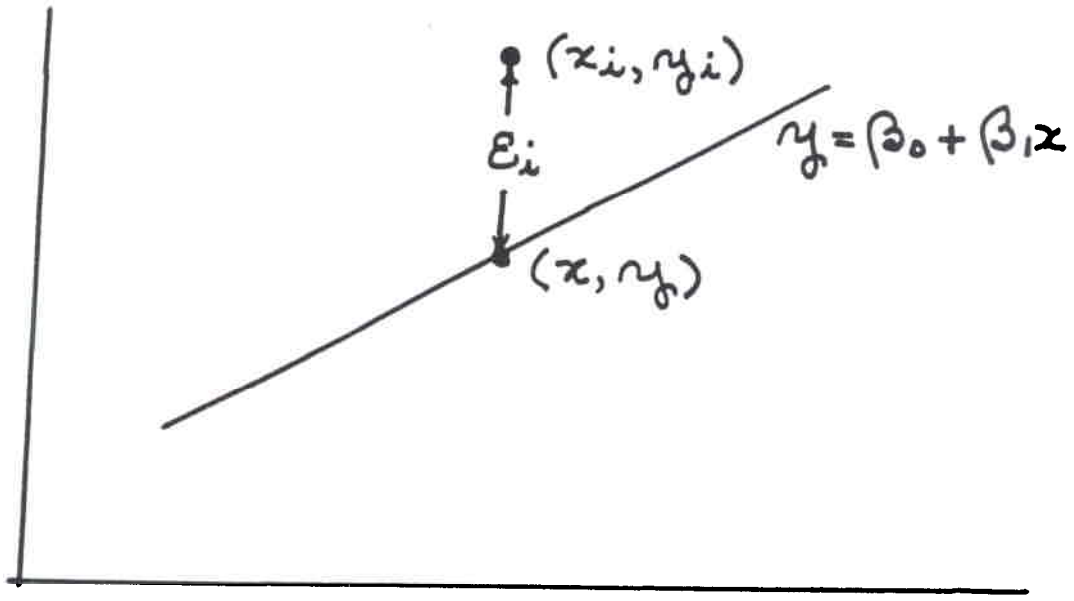
Y, a random variable,

where Y is hypothesized to have some dependence on x.

Method of Least Squares

- developed by Gauss
- "fits" hypothesized curve to the (x, Y) pairs of data
- minimizes the sum of the squared distances, measured in the Y direction, from the points to the curve.

Consider a simple case in which the curve is a straight line, $\eta = \beta_0 + \beta_1 x$



The distance from the point (x, η), on the line $\eta = \beta_0 + \beta_1 x$, to the point (x_i, η_i), where $x = x_i$, may be written as

$$\varepsilon_i = \eta_i - \eta = \eta_i - \beta_0 - \beta_1 x_i .$$

The error, ε_i, is called the <u>residual</u>

Squaring and summing

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i ,$$

we obtain

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

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

To minimize, we must have

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \beta_1} = 0$$

Thus

$$\partial \mathcal{L} / \partial \beta_0 = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\partial \mathcal{L} / \partial \beta_1 = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

So that

$$\sum y_i - n\beta_0 - \beta_1 \sum x_i = 0$$

$$\sum x_i y_i - \beta_0 \sum x_i - \beta_1 \sum x_i^2 = 0$$

These two equations are referred to as <u>normal equations</u>

Using the likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$,

Solving for $\hat{\beta}_0$ and $\hat{\beta}_1$ yields

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \frac{\hat{\beta}_1}{n}\sum x_i$$

and $$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

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

Use of this relationship for inferential purposes requires that an assumption be made about the distribution of the random variable, Y.

It is assumed that

$$Y_i,\ i = 1, \cdots, n_i$$

are distributed independently and Normally

with means $\beta_0 + \beta_1 x_i$

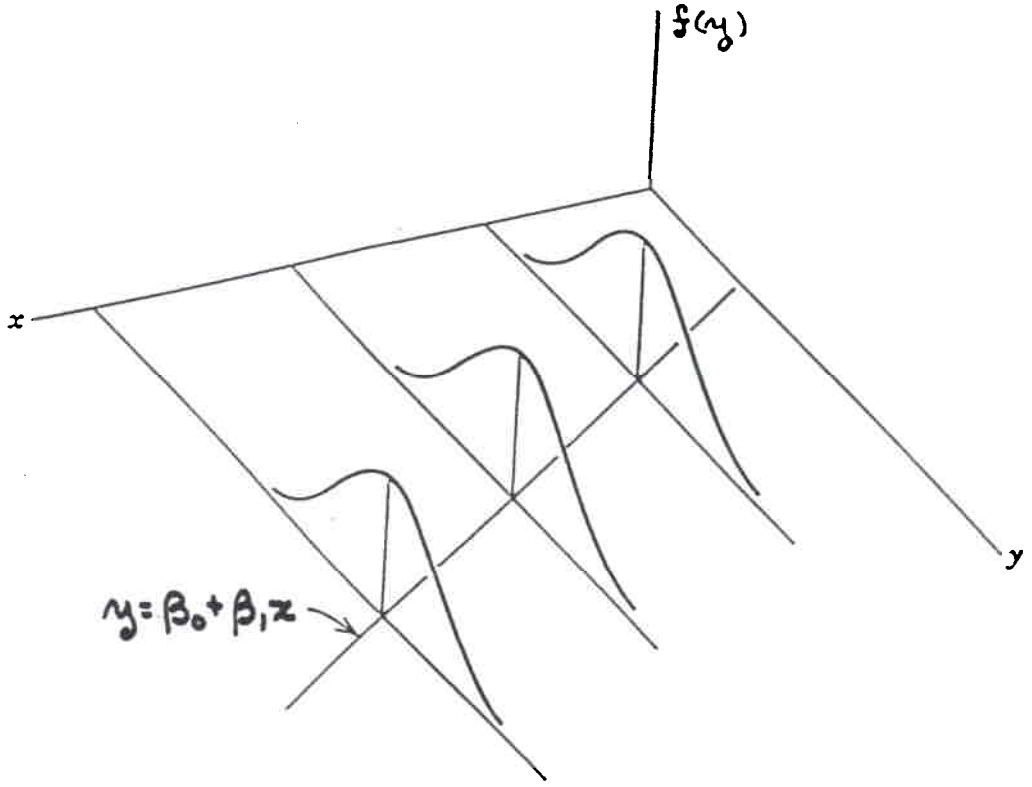
and common variance σ^2.

It is further assumed that the n performances of the experiment by which we obtain a sample

$$(x_1, y_1), \cdots, (x_n, y_n)$$

are independent.

A graphical interpretation of the distribution assumption is shown below



These assumptions lead to maximum likelihood estimates of β_0 and β_1 that are consistent with the normal equations.

Error Sum of Squares

When a line is fit to a set of data points (x_i, y_i), we can then estimate

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

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of β_0 and β_1, respectively

Then for every data pair, (x_i, y_i),

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

where $\varepsilon_i = y_i - \hat{y}_i$ is the residual

The estimate of the variance of these residuals is given by

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

where $\hat{\sigma}^2$ is referred to as the standard error of estimate

The value $SS_E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

is referred to as the error sum of squares.

Certain expressions are frequently encountered in linear regression analysis and also have useful interpretations.

- S_{xx} = Sum of Squared Deviations of x

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

$$= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left\{\sum_{i=1}^{n} x_i\right\}^2$$

which is related to the sample variance of x in accord with

$$s_x^2 = \hat{\sigma}_x^2 = \frac{1}{n-1} S_{xx}$$

- S_{yy} = Sum of Squared Deviations of y

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

$$= \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left\{\sum_{i=1}^{n} y_i\right\}^2$$

which is related to the sample variance of y in accord with

$$s_y^2 = \hat{\sigma}_y^2 = \frac{1}{n-1} S_{yy}$$

And, last but not least,

- S_{xy} = Sum of "Squared" Deviations of x and y

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

$$= \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left\{\sum_{i=1}^{n} x_i\right\}\left\{\sum_{i=1}^{n} y_i\right\}$$

which is related to the sample <u>covariance</u> of x and y in accord with

$$s_{xy} = \hat{\sigma}_{xy} = \frac{1}{n-1} S_{xy}$$

Parameter Estimation
Confidence Intervals
(2 sided)

for the Regression Coefficient (i.e., Slope), β_1:

Determine n, $\hat{\beta}_1$, S_{xx} and $\hat{\sigma}^2$
as previously defined

Select α = significance level
Then c = t-distribution value (critical value) corresponding to $\alpha/2$ with $(n-2)$ degrees of freedom

Determine

$$k = c * \underbrace{\left\{ \frac{\hat{\sigma}^2}{S_{xx}} \right\}^{1/2}}_{\text{standard error of the slope}}$$

Then

$$\text{CONF}_{(1-\alpha)} \left\{ (\hat{\beta}_1 - k) \leq \beta_1 \leq (\hat{\beta}_1 + k) \right\}$$

Parameter Estimation
Confidence Intervals
2 sided

For the Intercept, β_0:

Determine n, $\bar{x}$, $\hat{\beta}_0$, S_{xx} and $\hat{\sigma}^2$ as previously defined

Select α = significance level

Then c = critical value
= t-distribution value corresponding to $\alpha/2$ with $(n-2)$ degrees of freedom

Determine

$$k = c * \underbrace{\left\{ \hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] \right\}^{1/2}}_{\text{standard error for the intercept}}$$

Then

$$\text{CONF}_{(1-\alpha)} \left\{ (\hat{\beta}_0 - k) \leq \beta_0 \leq (\hat{\beta}_0 + k) \right\}$$

Parameter Estimation
Confidence Intervals

For the Regression Line (i.e. Mean),

$$\mu_{y} = \beta_0 + \beta_1 x$$

Determine n, $\bar{x}$, S_{xx} and $\hat{\sigma}^2$
as previously defined,
and $\hat{\mu}_{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$
→ for a particular x_0 ← ←

Select α = significance level
then c = t-distribution value
corresponding to $\alpha/2$ with
$(n-2)$ degrees of freedom

Determine

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

standard error of the mean on the regression line

→ For the particular x_0 ←

Then

$$\text{CONF}_{(1-\alpha)} \left\{ [\hat{\mu}_{y} - k(x_0)] \leq \mu_{y} \leq [\hat{\mu}_{y} + k(x_0)] \right\}$$

Parameter Estimation
Confidence Intervals

For the Prediction Interval
(on a future value y_0 at the value x_0)

Determine n, $\bar{x}$, S_{xx} and $\hat{\sigma}^2$
as previously defined
and $\hat{\mu}_{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$
→ for a particular x_0 ←

Select α = significance level
then c = t-distribution value
corresponding to $\alpha/2$ with
$(n-2)$ degrees of freedom

Determine

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

Standard error for the prediction interval

→ for the particular x_0 ←

Then

$$\text{CONF}_{(1-\alpha)} \left\{ [\hat{\mu}_{y} - k(x_0)] \leq \mu_{y} \leq [\hat{\mu}_{y} + k(x_0)] \right\}$$

Most Useful Hypothesis Test

(aka Test of Significance of the Regression)

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

Test statistic:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

Select α = significance level
then c = t-distribution score corresponding to $\alpha/2$ with $(n-2)$ degrees of freedom

If $|t_0| > |c|$, then Reject H_0

Note that failure to Reject H_0 is equivalent to concluding that there is no useful linear relationship between x and Y.

HAZARDS OF FITTING REGRESSION EQUATIONS TO "HAPPENSTANCE" DATA

INCONSISTENT DATA

A long record of data from any process is rarely consistent and comparable; e.g., standards are modified, suppliers are changed, and calibrations drift. Such events are infrequently recorded.

RANGE OF VARIABLES LIMITED BY CONTROL

If a process is under control, then of course a complete range of data will not exist, even though the relationship does, and thus there is likely to appear to be no relationship between certain variables.

SEMICONFOUNDING OF EVENTS

A process may be such that a particular change in one variable is usually (or sometimes invariably) accompanied by a corresponding change in a second variable, both of which are related to changes in a third variable. It is often impossible to identify the separate contributions of the first two variables to the third.

HAZARDS OF FITTING REGRESSION EQUATIONS TO "HAPPENSTANCE" DATA

NONSENSE RELATIONSHIPS

Some form of regression equation can be "fit" to virtually any set of data pairs. This does not provide evidence that the variables are necessarily causally related.

SERIALLY CORRELATED ERRORS

For sets of data that occur as a sequence in time, the errors are often not independent. When such dependence occurs, the ordinary method of least squares is inappropriate.

DYNAMIC RELATIONSHIPS

Occurs when variable values are affected by the recent past of other variables, in which cases "memory functions" must be developed.

CORRELATION ANALYSIS

- The correlation coefficient is simply an indicator of the extent to which there is a linear relationship between two variables

- The variables must be normally distributed

- The correlation coefficient cannot "prove" or "disprove" cause-and-effect relationships

- There may be underlying causal variables producing a strong correlation

Correlation Analysis

- Both X and Y are random variables
- The joint distribution of X and Y is bivariate Normal
- Assumes that the association between X and Y is linear
- Define $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ as the population correlation coefficient,
  where σ_x^2 = variance of X
  σ_y^2 = variance of Y
  σ_{xy} = covariance of X and Y
- Compute sample correlation coefficient by

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\left[\Sigma xy - \frac{1}{n}\Sigma x \Sigma y\right]}{\left\{\left[\Sigma x^2 - \frac{1}{n}(\Sigma x)^2\right]\left[\Sigma y^2 - \frac{1}{n}(\Sigma y)^2\right]\right\}^{1/2}}$$

$$-1 \leq r \leq +1$$

Correlation Coefficient
Test of Hypotheses

Test Statistic:

$$t_0 = r\sqrt{\frac{n-2}{1-r^2}}$$

with $(n-2)$ degrees of freedom
where n = # (x, y) pairs, and
r = sample correlation coefficient

$H_0: \rho = 0$ $H_1: \rho \neq 0$	$H_0: \rho = 0$ $H_1: \rho < 0$	$H_0: \rho = 0$ $H_1: \rho > 0$
two tailed test	one tailed test	one tailed test
Reject H_0 if $\lvert t_0 \rvert > c_1$	Reject H_0 if $t_0 < c_2$	Reject H_0 if $t_0 > c_3$

c_1 = critical value for $\alpha/2$ from t-distribution with $(n-2)$ dof

c_2 = critical value for $(1-\alpha)$ from t-distribution with $(n-2)$ dof

c_3 = critical value for α from t-distribution with $(n-2)$ dof

Linear Regression Analysis

Computational format:

From the original data set obtain

number of points (i.e., data pairs) → n, Σx, Σy, Σx^2, Σy^2, Σxy

From the preceding obtain

$\bar{x} = \Sigma x / n$

$\bar{y} = \Sigma y / n$

always + $S_{xx} = \Sigma x^2 - n\bar{x}^2$

always + $S_{yy} = \Sigma y^2 - n\bar{y}^2$

may be − $S_{xy} = \Sigma xy - n\bar{x}\bar{y}$

if slope is going to be negative

$\hat{\beta}_1 = S_{xy} / S_{xx}$ slope

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

$SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$

$\hat{\sigma}^2 = SS_E / (n-2)$ — always +

Then

Confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

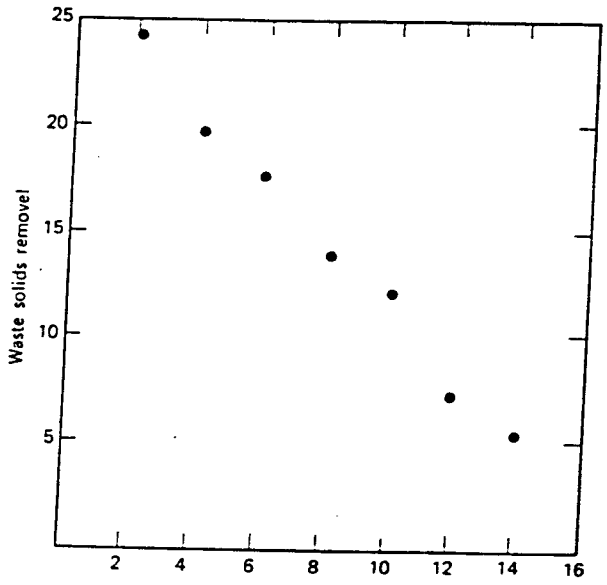
Confidence interval for β_0:

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

Confidence interval for $\mu_y = \beta_0 + \beta_1 x$ at $x = x_0$:

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Linear Regression Analysis



Feed flow in gallons per minute.

Results of a study to determine the percent of waste solids removed in a filtration system as a function of the flow rate of the effluent being fed into the system.

The percent of waste solid removed, y, was observed when each of the flow rates $2, 4, \cdots, 14$ gal/min was used.

From the original data:

$n = 7$

$\Sigma x = 56$

$\Sigma y = 100.8$

$\Sigma x^2 = 560$

$\Sigma y^2 = 1724.8$

$\Sigma xy = 632.6$

Linear Regression Analysis

Now we obtain

$$\bar{x} = \Sigma x / n = 56/7 = 8.0$$
$$\bar{y} = \Sigma y / n = 100.8/7 = 14.4$$

$$S_{xx} = \Sigma x^2 - n\bar{x}^2 = 56 - 7(8)^2 = 112$$
$$S_{yy} = \Sigma y^2 - n\bar{y}^2 = 1724.8 - 7(14.4)^2 = 273.28$$
$$S_{xy} = \Sigma xy - n\bar{x}\bar{y} = 632.6 - 7(8)(14.4) = -173.8$$

$$\hat{\beta}_1 = S_{xy}/S_{xx} = -173.8/112 \doteq -1.55$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 14.4 - (-1.55)(8) = 26.8$$

$$SS_E = S_{yy} - \hat{\beta}_1 S_{xy} = 273.28 - (-1.55)(-173.8) \doteq 3.89$$
$$\hat{\sigma}^2 = SS_E/(n-2) = 3.89/(7-2) \doteq 0.78$$

Suppose $\alpha = 0.05$. Since $n = 7$, $dof = (n-2) = (7-2) = 5$
For a two-sided confidence interval, then,
from Table IV (pg A-6), using $\alpha/2 = 0.025$, $t_{crit} = 2.571$

Thus
Confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{crit}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = -1.55 \pm 2.571\sqrt{\frac{0.78}{112}}$$

so CONF 95% $\{-1.76 \le \beta_1 \le -1.34\}$

Confidence interval for β_0:

$$\hat{\beta}_0 \pm t_{crit}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = 26.8 \pm 2.571\sqrt{0.78\left(\frac{1}{7} + \frac{(8)^2}{112}\right)}$$

so CONF 95% $\{24.9 \le \beta_0 \le 28.7\}$

Confidence interval for $\mu_4 = \beta_0 + \beta_1 x$ at $x = x_0$:

$$\hat{\mu}_4 = \hat{y}_0 \; ; \quad \hat{y}_0 \pm t_{crit} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

Our "best fit" regression line is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x = 26.8 - 1.55x$$

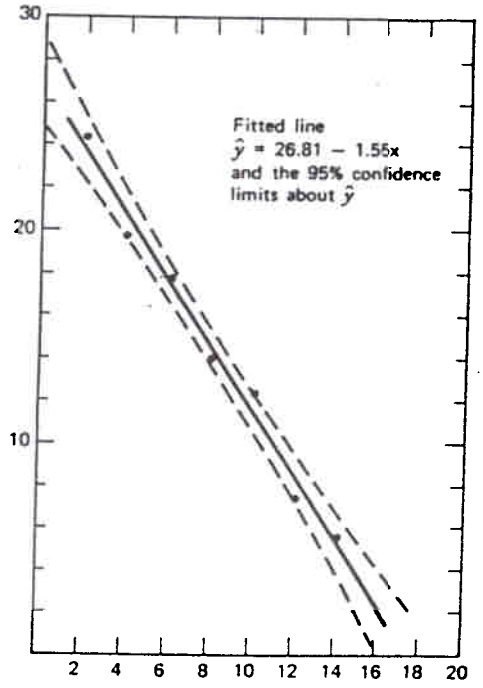
Suppose we set $x_0 = 4$;

then $\hat{y}_0 = 26.8 - 1.55(4) = 20.6$

Our confidence interval at $(4, 20.6)$ is thus

$$20.6 \pm 2.571 \sqrt{0.78 \left(\frac{1}{7} + \frac{(4-8)^2}{112} \right)}$$

so $CONF_{95\%} \{ 19.4 \leq \mu_4 \leq 21.8 \}$



Plot of fitted line and 95% confidence limits for predicted values. (Y_0)

Correlation

$$r = S_{xy} / \sqrt{S_{xx} S_{yy}}$$

For $H_0: \rho = 0$ critical value: $t_{crit} = t_{n-2, \alpha/2}$

$H_1: \rho \neq 0$ test statistic:

$$t_{test} = r \sqrt{\frac{n-2}{1-r^2}}$$

[note: r^2 called determination coefficient]

For our example:

$$r = S_{xy} / \sqrt{S_{xx} S_{yy}}$$

$S_{xx} = 112$

$S_{yy} = 273.28$

$S_{xy} = -173.8$

$$r = (-173.8) / \sqrt{(112 \times 273.28)} = -0.9934$$

$H_0: \rho = 0$ $n = 7 \Rightarrow$ dof for t-statistic $= (n-2) = 5$

$H_1: \rho \neq 0$ let $\alpha = 0.05$; 2-sided test $\Rightarrow \alpha/2 = 0.025$

$t_{crit} = t_{n-2, \alpha/2} = 2.571$ [Table III, pg A-6]

$$t_{test} = r \sqrt{\frac{n-2}{1-r^2}} = (-0.9934) \sqrt{5 / \{1 - (-0.9934)^2\}} = -19.37$$

Since $|t_{test}| > |t_{crit}|$, Reject H_0

Linear Regression and ANOVA:

Source	dof	Sums of Squares	Mean Squares	$\mathcal{F}$
Regression	1	$SS_R\ [=(S_{xy})^2/S_{xx}]$	$MS_R = SS_R/dof_R$	MS_R/MS_E
Error	$(n-2)$	$SS_E\ [=SS_T - SS_R]$	$MS_E = SS_E/dof_E$	
Total	$(n-1)$	$SS_T\ [=S_{yy}]$		

[note that $MS_E = \hat{\sigma}^2$]

From our example: $S_{xx} = 112$

$S_{yy} = 273.28$

$S_{xy} = -173.8$

$n = 7$

$SS_T = S_{yy} = 273.3$

$SS_R = (S_{xy})^2/S_{xx} = (-173.8)^2/112 = 269.7$

$SS_E = SS_T - SS_R = 273.3 - 269.7 = 3.6$

Source	dof	SS	MS	$\mathcal{F}_{1,5}$
Regression	1	269.7	269.7	374.6
Error	5	3.6	0.72	
Total	6	273.3		