1. Applied Regression Analysis 3rd edition

Norman R. Draper, Harry Smith

Response variable = Model Function + Random error

Errors assumed to be independent

Linear first order model:

$$Y = \beta_0 + \beta_1 X_i + e$$

For a given X, a corresponding Y consists of the value $\beta_0 + \beta_1 X_i$ plus an amount ε, the increment by which an individual Y may fall off the regression line

Meaning of linear Model

- When we say a model is linear or nonlinear we are referring to linearity or nonlinearity in the parameters
- Value of highest power of a predictor variable in the model is called the order of the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

is a Second order (x) linear (β's) regression model.

Least Squares Estimation:

- β_0, β_1 and e are unknown & difficult to discover as it changes in each observation Y.
- We calculate estimate of β_0 and β_1 b_0, b_1:

$$\hat{Y} = b_0 + b_1 X_i$$

$\hat{Y}$ predicted value of Y for a given X.

Suppose we have n sets of observations $(X_1, Y_1) \ldots (X_n, Y_n)$ then we can write

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \text{For } i = 1, 2 \ldots n$$

So that the sum of squares of deviation from the true line is:

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- S is called the Sum of Squares function
- We shall choose as values of b_0 and b_1 that, when substituted for β_0 and β_1, produce the least possible value of S

- We determine b_0 and b_1 by differentiating first with respect to β_0 and then with respect to β_1 and setting results equal to zero:

$$\frac{dS}{d\beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)$$

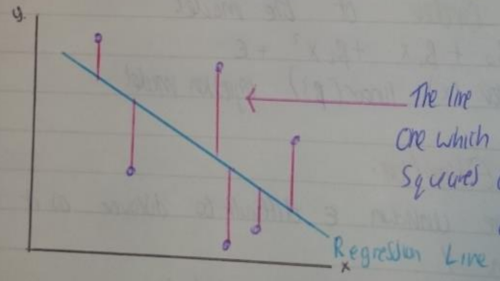
$$\frac{dS}{d\beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

So that the estimates b_0 and b_1 are solutions to 2 eq^n's:

$$\Sigma (Y_i - b_0 - b_1 x_i) = 0$$

$$\Sigma x_i (Y_i - b_0 - b_1 x_i) = 0$$

We sub (b_0, b_1) for (β_0, β_1) when we equate to zero



The line fitted by least squares is the one which makes the sum of squares of all vertical discrepancies as small as possible

The solutions to these eq^n's yield:

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2} \quad \text{where } \bar{x} = \frac{\Sigma x_i}{n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Also: $\Sigma (x_i - \bar{x})(Y_i - \bar{y}) = \Sigma x_i y_i - \bar{x} \Sigma Y_i - \bar{y} \Sigma x_i + n\bar{x}\bar{y}$

$$= \Sigma x_i y_i - n\bar{x}\bar{y}$$

$$= \Sigma x_i y_i - \left(\frac{\Sigma x_i \Sigma y_i}{n}\right)$$

Σx_i^2 is known as uncorrected sum of squares of x's

$(\Sigma x_i)^2/n$ is the correction for the mean of x's

The difference is called corrected sum of squares of x's $\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}$ S_{xx}

$\Sigma x_i y_i$ is called the uncorrected sum of products

$\frac{\Sigma x_i \Sigma y_i}{n}$ is called correction for the means $\Sigma(x_i-\bar{x})(y_i-\bar{y}) = S_{xy}$

The difference is called corrected sum of product of X and Y

Notation

$$
\begin{aligned}
S_{xy} &= \Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&= \Sigma(x_i-\bar{x})(y_i) \\
&= \Sigma x_i(y_i-\bar{y}) \\
&= \Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n} \\
&= \Sigma x_i y_i - n\bar{x}\bar{y}.
\end{aligned}
$$

$$
\begin{aligned}
S_{xx} &= \Sigma(x_i-\bar{x})^2 \\
&= \Sigma(x_i-\bar{x})x_i \\
&= \Sigma x_i^2 - (\Sigma x_i)^2/n \\
&= \Sigma x_i^2 - n\bar{x}^2
\end{aligned}
$$

$$
\begin{aligned}
S_{yy} &= \Sigma(y_i-\bar{y})^2 \\
&= \Sigma(y_i-\bar{y})y_i \\
&= \Sigma y_i^2 - (\Sigma y)^2/n \\
&= \Sigma y_i^2 - n\bar{y}^2
\end{aligned}
$$

The easily remembered formula for b_1 is then: $\frac{S_{xy}}{S_{xx}}$

$b_0 = \bar{y} - b_1\bar{x}$ Sub in to $\hat{y} = b_0 + b_1 x_i$

$\Rightarrow \hat{y} = \bar{y} + b_1(x-\bar{x})$

If we set $x=\bar{x}$, we see $\hat{y}=\bar{y}$. This means point $(\bar{x},\bar{y})$ lies on the fitted line.

In other words the least squared line contains the centre of gravity of the data

$(y_i - \hat{y}_i)$ equals the residual e_i

Since $\hat{y}_i = \bar{y} + b_1(x_i - \bar{x})$

$$y_i - \hat{y}_i = (y_i - \bar{y}) - b_1(x_i - \bar{x})$$

Which we can sum to give:

$$\Sigma(y_i - \hat{y}_i) = \Sigma(y_i - \bar{y}) - b_1 \Sigma(x_i - \bar{x}) = 0$$

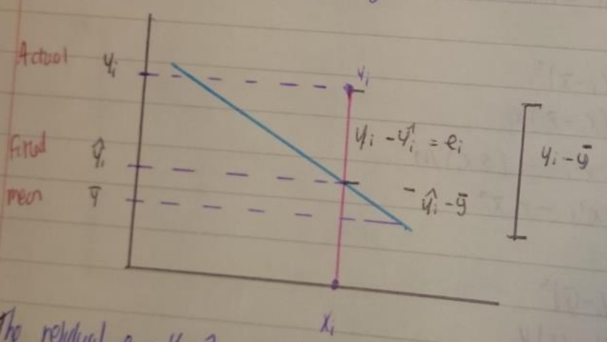
Residuals Sum to zero in theory

The Analysis Of Variance

We tackle question of how much of the [illegible] variation in the data has been explained by the regression line.

Consider. $y_i - \hat{y}_i = y_i - \bar{y} - (\hat{y}_i - \bar{y})$

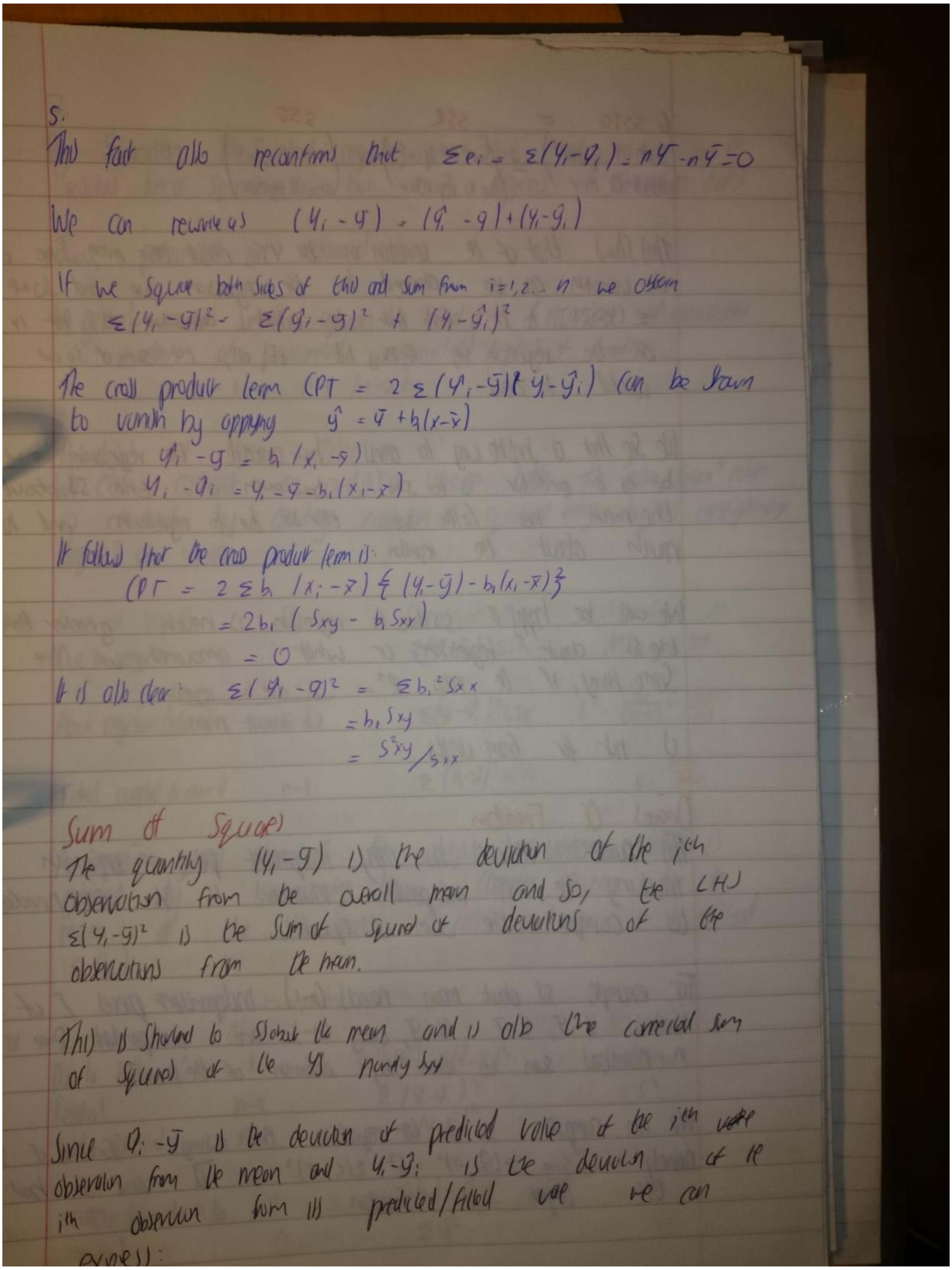
What this means geometrically is



The residual $e_i = y_i - \hat{y}_i$ is difference between 2 quantities (1) the deviation of the observed y_i from the overall mean $\bar{y}$ and (2) the deviation of the fitted $\hat{y}_i$ from the overall mean $\bar{y}$.

Note that average of the $\hat{y}_i$ is the same as average of y_i

$$\frac{\Sigma \hat{y}_i}{n} = \Sigma(b_0 + b_1 x_i)/n$$
$$= (nb_0 + b_1 n\bar{x})/n$$
$$= b_0 + b_1\bar{x}$$
$$= \bar{y}$$



This fact also reconfirms that $\Sigma e_i = \Sigma(y_i - \hat{y}_i) = n\bar{y} - n\bar{y} = 0$

We can rewrite as $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$

If we square both sides of this and sum from $i = 1, 2 \ldots n$ we obtain

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

The cross product term $CPT = 2\Sigma(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$ can be shown to vanish by applying $\hat{y} = \bar{y} + b_1(x - \bar{x})$

$$\hat{y}_i - \bar{y} = b_1(x_i - \bar{y})$$

$$y_i - \hat{y}_i = y_i - \bar{y} - b_1(x_i - \bar{x})$$

It follows that the cross product term is:

$$CPT = 2\Sigma b_1(x_i - \bar{x})\{(y_i - \bar{y}) - b_1(x_i - \bar{x})\}$$

$$= 2b_1(S_{xy} - b_1 S_{xx})$$

$$= 0$$

It is also clear: $\Sigma(\hat{y}_i - \bar{y})^2 = \Sigma b_1^2 S_{xx}$

$$= b_1 S_{xy}$$

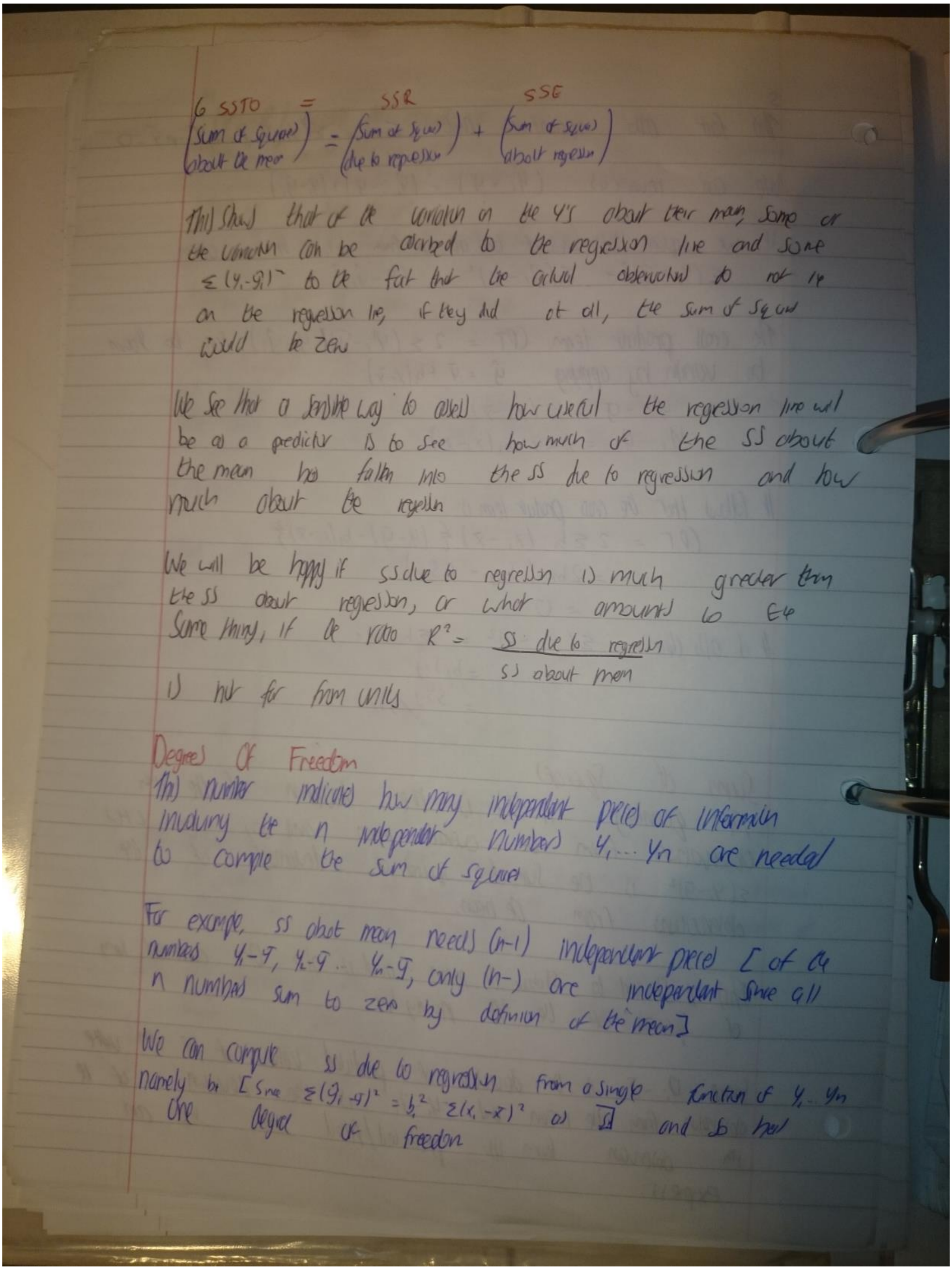
$$= S^2_{xy}/S_{xx}$$

Sum of Squares

The quantity $(y_i - \bar{y})$ is the deviation of the ith observation from the overall mean and so, the LHS $\Sigma(y_i - \bar{y})^2$ is the sum of squares of deviations of the observations from the mean.

This is shortened to SS about the mean and is also the corrected sum of squares of the Ys namely Syy

Since $\hat{y}_i - \bar{y}$ is the deviation of predicted value of the ith value observation from the mean and $y_i - \hat{y}_i$ is the deviation of the ith observation from its predicted/fitted value we can express:



6 SSTO = SSR + SSE

$$\begin{pmatrix}\text{Sum of Squares}\\ \text{about the mean}\end{pmatrix} = \begin{pmatrix}\text{Sum of Squares}\\ \text{due to regression}\end{pmatrix} + \begin{pmatrix}\text{Sum of Squares}\\ \text{about regression}\end{pmatrix}$$

This shows that of the variation in the y's about their mean, some of the variation can be ascribed to the regression line and some $\Sigma(y_i - \hat{y}_i)^2$ to the fact that the actual observations do not lie on the regression line, if they did at all, the sum of squares would be zero

We see that a sensible way to assess how useful the regression line will be as a predictor is to see how much of the SS about the mean has fallen into the SS due to regression and how much about the regression

We will be happy if SS due to regression is much greater than the SS about regression, or what amounts to the same thing, if the ratio $R^2 = \frac{\text{SS due to regression}}{\text{SS about mean}}$ is not far from unity

Degrees Of Freedom

This number indicates how many independant pieces of information involving the n independant numbers $y_1, \ldots y_n$ are needed to compile the sum of squares

For example, SS about mean needs (n-1) independant pieces [of the numbers $y_1-\bar{y}, y_2-\bar{y} \ldots y_n-\bar{y}$, only (n-) are independant since all n numbers sum to zero by definition of the mean]

We can compute SS due to regression from a single function of $y_1 \ldots y_n$ namely b_1 [Since $\Sigma(\hat{y}_i - \bar{y})^2 = b_1^2\,\Sigma(x_i - \bar{x})^2$ as $\boxed{?}$ and so has one degree of freedom

By subtraction, the SS about regression, which we shall in future call the residual sum of squares, has n-2 degrees of freedom (df).

Two parameters are estimated

In general, residual sum of squares is based on (numbers of observations - number of parameters estimated) degrees of freedom

$$n - 1 = 1 + (n-2)$$

Analysis of Variance Table

We can construct an analysis of variance table. The "mean square" column is calculated by dividing each sum of squares entry by its corresponding degrees of freedom

1.3

Source of Variation	Degrees of freedom	Sum of Squares SS	Mean Square MS
Due to regression	1	$\Sigma(\hat{Y}_i-\bar{Y})^2$ SSR	$MS_{Reg} = \frac{SSR}{1}$
About regression (residual)	n-2	$\Sigma(Y_i-\hat{Y}_i)^2$ SSE	$s^2 = \frac{SS}{(n-1)}$ $\frac{SSE}{n-2}$
Total corrected for mean $\bar{Y}$	n-1	$\Sigma(Y_i-\bar{Y})^2$ SSTO	$F = \frac{MSR}{MSE}$

A more general form of the analysis of variance table, is obtained by incorporating the correction factor for the mean of the Y's into the table where it is called $SS(b_0)$.

Incorporating $SS(b_0)$

1.4

Source	df	SS	$MS = SS/df$
Due to b_1/b_0	1	$SS(b_1\vert b_0) = \Sigma(\hat{Y}_i-\bar{Y})^2$	MS_{reg}
Residual	n-2	$\Sigma(Y_i-\hat{Y}_i)^2$	s^2
Total corrected.	n-1	$\Sigma(Y_i-\bar{Y})^2$	
correction factor due to b_0	1	$SS(b_0) = \frac{(\Sigma Y_i)^2}{n} = n\bar{Y}^2$	
Total	n	ΣY_i^2	

Alternative way to display 1.4 is to drop the line "total corrected".
Total line is sum of remaining 3 entries.

1.5

Source	df	SS	MS = SS/df
b_0	1	$n\bar{y}^2$	—
$b_1 \mid b_0$	1	S^2_{xy}/S_{xx}	MS_{reg}
Residual	$n-2$	By subtraction	s^2
TOTAL	n	Σy_i^2	

First SS entry due to b_0 is the amount of variation in $n\bar{y}^2$ explained by a horizontal straight line $\hat{y} = \bar{y}$.

If we model $Y = \beta_0 + \epsilon$ via least squares, fitted model is $\hat{y} = \bar{y}$.

If we subsequently fit the "with slope also" model $Y = \beta_0 + \beta_1 x_i + \epsilon$, the "due to $b_1|b_0$" SS entry $\frac{S^2_{xy}}{S_{xx}}$ is the extra variation picked up by the slope term over and above that picked up by the intercept alone

$$SS(b_1|b_0) = \Sigma(\hat{y}_i - \bar{y})^2 = b_1 \{ \Sigma(x_i - \bar{x})(y_i - \bar{y}) \} = b_1 S_{xy}$$

$$= \frac{(\Sigma(x_i - \bar{x})(y_i - \bar{y}))^2}{\Sigma(x_i - \bar{x})^2} = \frac{S^2_{xy}}{S_{xx}}$$

$$= \frac{(\Sigma x_i y_i - (\Sigma x_i \Sigma y_i)/n)^2}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{S^2_{xy}}{S_{xx}}$$

$$= \frac{(\Sigma(x_i - \bar{x})y_i)^2}{(\Sigma x_i - \bar{x})^2}$$

Note, total corrected SS, $\Sigma(y_i - \bar{y})^2$ can be written as:

$$S_{yy} = \Sigma y_i^2 - (\Sigma y_i)^2/n$$

$$\text{or } S_{xy} = \Sigma y_i^2 - n\bar{y}^2$$

The mean square about regression, s^2 will provide an estimate based on the degrees of freedom of the variance about the regression, which we will call $\sigma^2_{y \cdot x}$

If regression equation were estimated from an indefinitely large number of observations, the variance about the regression would represent a measure of the error with which any observed value of y could be predicted from a given value of X using the determined eq"

Skeleton Analysis of Variance Table

A skeleton analysis of variance table consists of the "Source" and "df" columns only.

R^2 Statistic

A useful statistic to check is the R^2 value of a regression fit:

$$R^2 = \frac{(\text{SS due to regress given } b_0)}{(\text{Total SS corrected for the mean } \bar{y})} \qquad \frac{SSR}{SSTO}$$

$$= \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} \qquad 1.3.15 \qquad i = 1,2 \ldots n$$

R^2 measures 'proportion of total variance about the mean $\bar{y}$ explained by the regression".

In fact R is the correlation between y and $\hat{y}$ and is usually called the multiple correlation coefficient

R^2 is then "the square of the multiple correlation coefficient".

For a straight line fit: $R^2 = SS(b_1|b_0)/S_{yy}$

$$= S_{xy}^2/(S_{xx}S_{yy})$$

Thus r^2 explains G% of the total variation in the data about the average $\bar{y}$.
This is quite a large proportion

R^2 can take values as high as $1 \rightarrow 100$ when all the x values are different

When repeat runs exist in the data, the value of R^2

is $V(a) = a_1^2 V(y_1) + a_2^2 V(y_2) + \dots + a_n^2 V(y_n)$

if the y_i are pairwise uncorrelated and the a_i are constant; furthermore, if $V(y_i) = \sigma$

$$V(a) = (a_1^2 + a_2^2 + \dots + a_n^2)\sigma^2$$
$$= (\Sigma a_i^2)(\sigma^2)$$

In the expression for b_1, $a_i = (x_i - \bar{x}) / \Sigma(x_i - \bar{x})^2$ since x_i can be regarded as constants. Hence after reduction:

$$V(b_1) = \frac{\sigma^2}{\Sigma(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

The Standard deviation of b_1, is the square root of the variance:

$$\text{s.d}(b_1) = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{S_{xx}}}$$

or if σ is unknown, and we use the estimate S in its place, assuming the model is correct, the estimated Standard deviation of b_1:

$$\text{est s.d}(b_1) = \frac{S}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{S}{\sqrt{S_{xx}}}$$

Alternative terminology for estimated Standard deviation is the Standard error

Confidence Interval for β_1

If we assume that the variations of the observations about the line are normal - that is the errors ε_i are all from the same normal distribution, $N(0, \sigma^2)$ - it can be shown that we can assign $100(1-\alpha)\%$ Confidence limits for β_1 ⇒

$$b_1 \pm \cancel{t-1} \; \frac{t(n-2, 1-\tfrac{1}{2}\alpha)\, S}{\sqrt{\Sigma(x_i - \bar{x})^2}}$$

where $t(n-2, 1-\tfrac{1}{2}\alpha)$ is the $100(1-\tfrac{1}{2}\alpha)$ percentage point of a t-dist, with $n-2$ df.

Test for $H_0\ \beta_1 = \beta_{10}$ Vers $H_1\ \beta_1 \neq \beta_{10}$

Test null hypothesis that β_1 is equal to β_{10}, where β_{10} is a Specified value that could be zero, against alternative hypothesis that β_1 is different to β_{10} by calculating

$$t = \frac{b_1 - \beta_{10}}{se(b_1)} = \frac{(b_1-\beta_{10})\left(\sqrt{\Sigma(x_i-\bar{x})^2}\right)}{s}$$

and comparing $|t|$ with $t_{n-2,\,1-\frac{1}{2}\alpha}$ from a t-table with n-2 df. The test will be a 2 Sided test conducted at 100 α % level in this form.

The value lies with $(\sim, \sim)$ and this Statement is made at 95% Confidence β_1

If $|t|$ exceeds tcritical H_0 is rejected

We could also examine CI to see if it includes the value or not.

Reject or do not reject.

- If $|t|$ value had been smaller than the critical value we could not reject the hypothesis.
- The most we can say is that on basis of certain obtained data we we cannot reject it.
- It may be that in another set of data we can find evidence that is contrary to our hypothesis and so reject it

Confidence Interval Represents a Set of Test.

Once we have CI for β_1 we do not actually have to compute the $|t|$ value for a particular 2 Sided-t-test at the Same α level

It is easier to examine interval for β_1 and see if it contains β_{10} If it does then hypothesis $\beta_1 = \beta_{10}$ cannot be rejected

If $|b_1 - \beta_{10}| > t(n-2, 1-\frac{1}{2}\alpha) \frac{s}{\sqrt{\Sigma(x_i-\bar{x})^2}}$

that is, β_{10} lies outside the limits of eq 1.4d

Standard Deviation of the Interval for β_0

A CI for β_0 and a test of whether or not β_0 is equal to some specified value can be constructed in a way similar to that just described for β_1

$$sd(b_0) = \sigma \sqrt{\frac{\Sigma x_i^2}{n\,\Sigma(x_i-\bar{x})^2}}$$

Replacement of σ by s provides the estimated $sd(b_0)$, that is $s.e(b_0)$

Thus the $(1-\alpha)\%$ confidence limits are given by:

$$b_0 \pm t(n-2, 1-\tfrac{1}{2}\alpha)\, s \sqrt{\frac{\Sigma x_i^2}{n\,\Sigma(x_i-\bar{x})^2}}$$

A test for $H_0: \beta_0 = \beta_{00}$ against $H_1: \beta_0 \neq \beta_{00}$ will be rejected at the α level if β_{00} falls outside the CI, or will not be rejected if it lies inside or may be conducted separately by finding quantity:

$$t = \frac{(b_0 - b_{00})}{s \cdot \sqrt{\frac{\Sigma x_i^2}{n\,\Sigma(x_i-\bar{x})^2}}}$$

and comparing it with $t(n-2, 1-\alpha\frac{1}{2})$.

Regression F-Test

F-Test for Significance of Regression

Since y_i are random variables any function of them is also a random variable; two particular functions are MSreg, the mean square due to regression, and s^2 the mean square due to residual variation, which are in the analysis of variance table),

These functions then have their own distribution, mean, variance and moments.

It can be shown that the mean values are:

$$E[MS_{reg}] = \sigma^2 + \beta_1^2 \Sigma(x_i - \bar{x})^2$$

$$E(s^2) = \sigma^2$$

Where if z is a random variable, $E[z]$ denotes its mean or expected value

Suppose that the errors ε_i are independent $N(0, \sigma^2)$ variables.

It can be shown that if $\beta_1 = 0$, the variable MSreg multiplied by its degrees of freedom (here one) and divided by σ^2 follows a χ^2 dist with the same (one) df

In addition $(n-2)s^2/\sigma^2$ follows χ^2 dist with $(n-2)$ df.

Since 2 variables are independent, a statistical theorem tells us that the ratio $F = \frac{MS_{reg}}{s^2}$

Follows an F distribution with 1 and n-2 df provided $\beta_1 = 0$

This fact can thus be used as a test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

We compare the ratio $F = MS_{reg}/s^2$ with the $100(1-\alpha)\%$ point of the tabulated $F(1, n-2)$ dist in order to determine whether β_1 can be considered nonzero on the basis of the data we have seen

$F = t^2$

We have 2 tests for the test of $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ a t-test and F-test.

In fact, the two test are equivalent and mathematically related here, due to the theoretical fact that $F(1, v) = \{t(v)\}^2$ that is, the square of a t-variable with v d.f is an F-variable with 1 and v df.

NOTE: This only happens when first df of F is 1

$$F = \frac{MS_{reg}}{s^2} = \frac{b_1 \Sigma(x_i - \bar{x})\Sigma(y_i - \bar{y})}{s^2}$$

$$= \frac{b_1^2 \Sigma(x_i - \bar{x})^2}{s^2} \quad \text{(by def of } b_1\text{)}$$

$$= \left[\frac{b_1\sqrt{\Sigma(x_i - \bar{x})^2}}{s}\right]^2$$

$$= t^2$$

Since variable $F(1, n-2)$ is the square of the $t(n-2)$ variable, and this carried over to the percentage points (upper α tail of F and 2-tailed t, total of α). Exact same result

When there are more regression coefs the overall F-test for reg, which is the extension of the one given here, does not correspond to the t-test of a coef. This is why we need to know both t and F-test.

However tests for an individual coef can be made in either t or F by a similar argument cannot attain 1 no matter how well the model fits.

This is because no model however good, can explain the variation in the data due to pure error

Confidence Intervals and Tests for β_0 and β_1

We now make the basic assumptions of model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, 2 \dots n$$

1 ϵ_i is a random variable with mean zero and variance σ^2 (unknown) that is: $E(\epsilon_i) = 0$ $\quad V(\epsilon_i) = \sigma^2$

2 ϵ_i and ϵ_j are uncorrelated, $i \neq j$, so that $Cov(\epsilon_i, \epsilon_j) = 0$ thus:

$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad V(Y_i) = \sigma^2$$

A further assumption, which is not immediately necessary and will be recalled when used is that:

3 ϵ_i is normally distributed random variable, with mean zero and variance σ^2 by assumption (1); that is:

$$\epsilon_i \sim N(0, \sigma^2)$$

Under this additional assumption, ϵ_i, ϵ_j are not only uncorrelated but necessarily independent.

Standard Deviation of Slope b_1; Confidence Interval for β_1

We know $b_1 = \Sigma(X_1 - \bar{X})(Y_1 - \bar{Y}) / \Sigma(X_1 - \bar{X})^2$

$$= \Sigma(X_1 - \bar{X}) Y_1 \ / \Sigma(X_1 - \bar{X})^2$$

[Since the other term removed from numerator is $\Sigma(X_1 - \bar{X})\bar{Y} = \bar{Y}\Sigma(X_i - \bar{X}) = 0$

$$b_1 = \{(X_1 - \bar{X})Y_1 + \dots + (X_n - \bar{X})Y_n\} / \Sigma(X_1 - \bar{X})^2$$

Now the variance of a function

$$a = a_1 Y_1 + a_2 Y_2 + \dots + a_n Y_n$$