Linear Models: Linear Regression and Correlation

So far, we tested whether two subpopulations' means are equal.

For example, we can so far answer:

is fever reduction effect of a certain drug the same for children vs. adults? The answer would be:

$H_0 : \mu 1 = \mu 2$

But what should we do when we have more than 2 subpopulations?

Linear Models: Linear Regression and Correlation

What happens if we now wish to consider many age groups: infants (0-1), toddlers (1-3), pre-school (3-6), early school (6-9), … old age (65 and older)?

$H_0 : \mu 1 = \mu 2 = \mu 3 = \mu 4 = \ldots = \mu 15$

What would some of the alternatives be?

Linear Models: Linear Regression and Correlation

One alternative:

$H_a : \mu 1 < \mu 2 < \mu 3 < \mu 4 < \ldots < \mu 15$

This is doable, but it would be a cumbersome test.

Or, another:

$H_a : \mu 1 = \mu 2 - \text{b} = \mu 3 - 2\text{b} = \ldots = \mu 15 – 14\text{b}$;

Then, H_0: b = 0

In other words, we postulate a specific relationship between the subpopulation means, which simplifies the null hypothesis.

The Simple Linear Regression Model

The simplest deterministic mathematical relationship between two variables *x* and *y* is a linear relationship
$y = \beta_0 + \beta_1 x$.

The set of pairs (*x*, *y*) for which $y = \beta_0 + \beta_1 x$ determines a straight line with slope β_1 and *y*-intercept β_0.

The objective of this section is to develop an equivalent linear probabilistic model.

If the two variables are probabilistically related, then for a fixed value of *x*, there is uncertainty in the value of the second variable.

We say that the two variables are related linearly "on average"

The Simple Linear Regression Model

More generally, the variable whose value is fixed by the experimenter will be denoted by *x* and will be called the **independent, predictor,** or **explanatory variable.**

For fixed *x*, the second variable will be random; we denote this random variable and its observed value by *Y* and *y*, respectively, and refer to it as the **dependent** or **response variable.**

Usually observations will be made for a number of settings of the independent variable.

The Simple Linear Regression Model

Let $x_1, x_2, \ldots, x_n$ denote values of the independent variable for which observations are made, and let Y_i and y_i, respectively, denote the random variable and observed value associated with x_i.

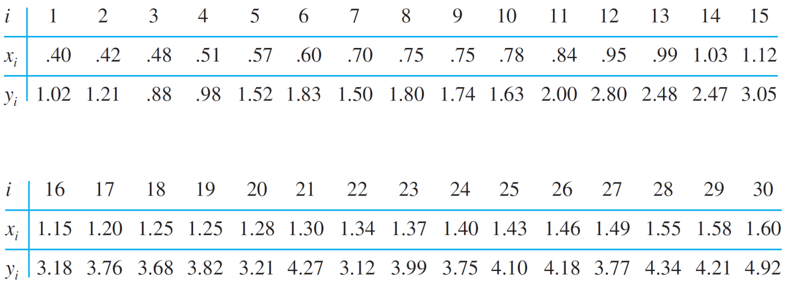
The available bivariate data then consists of the *n* pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

A picture of this data called a **scatter plot** gives preliminary impressions about the nature of any relationship. In such a plot, each (x_i, y_i) is represented as a point plotted on a two dimensional coordinate system.

Example 1

cont'd

The order in which observations were obtained was not given, so for convenience data are listed in increasing order of *x* values:



i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_i	.40	.42	.48	.51	.57	.60	.70	.75	.75	.78	.84	.95	.99	1.03	1.12
y_i	1.02	1.21	.88	.98	1.52	1.83	1.50	1.80	1.74	1.63	2.00	2.80	2.48	2.47	3.05

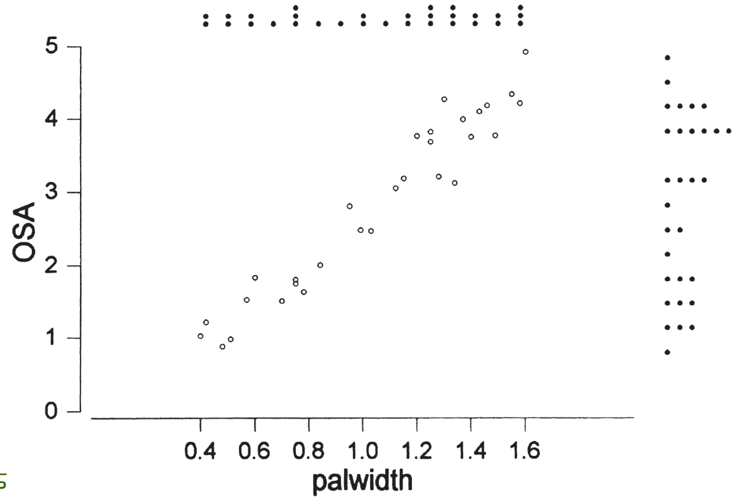
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
x_i	1.15	1.20	1.25	1.25	1.28	1.30	1.34	1.37	1.40	1.43	1.46	1.49	1.55	1.58	1.60
y_i	3.18	3.76	3.68	3.82	3.21	4.27	3.12	3.99	3.75	4.10	4.18	3.77	4.34	4.21	4.92

Thus $(x_1, y_1) = (.40, 1.02)$, $(x_5, y_5) = (.57, 1.52)$, and so on.

Example 1

cont'd

The scatter plot (OSA=y, palwidth = x):



Example 1

cont'd

- There is a strong tendency for y to increase as x increases. That is, larger values of OSA tend to be associated with larger values of width—a positive relationship between the variables.

- It appears that the value of y could be predicted from x by finding a line that is reasonably close to the points in the plot

In other words, there is indication of a substantial (though not perfect) linear relationship between the two variables.

A Linear Probabilistic Model

For the deterministic model $y = \beta_0 + \beta_1 x$, the actual observed value of y is a linear function of x.

The appropriate generalization of this to a probabilistic model assumes that

the average of Y is a linear function of x

-- ie, that for fixed x the actual value of variable$Y$ differs from its expected value by a random amount.

Simple Linear Regression Model

There are parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the **model equation**

$$Y = \beta_0 + \beta_1 x + \in$$

The quantity $\in$ in the model equation is the ERROR -- a random variable, assumed to be symmetrically distributed with

$$E(\in) = 0 \text{ and } V(\in) = \sigma^2.$$

A Linear Probabilistic Model

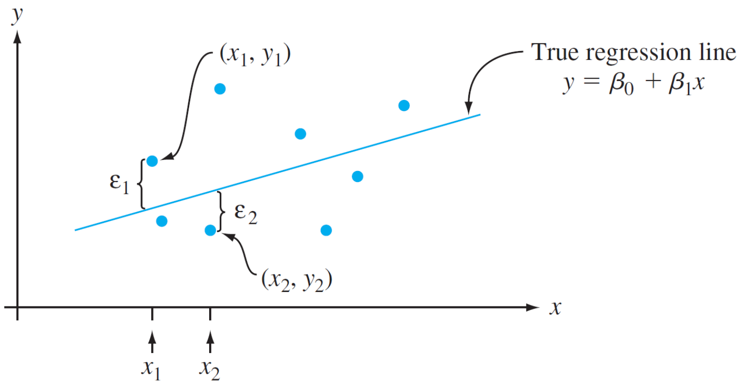
The variable $\in$ is sometimes referred to as the **random deviation** or **random error term** in the model.

Without $\in$, any observed pair (x, y) would correspond to a point falling exactly on the line $\beta_0 + \beta_1 x$, called the **true** (or **population**) **regression line.**

The inclusion of the random error term allows (x, y) to fall either above the true regression line (when $\in > 0$) or below the line (when $\in < 0$).

A Linear Probabilistic Model

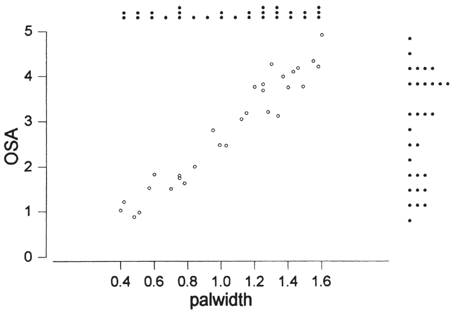
The points (x_1, y_1), …, (x_n, y_n) resulting from n independent observations will then be scattered about the true regression line:



A Linear Probabilistic Model

On occasion, the appropriateness of the simple linear regression model may be suggested by theoretical considerations (e.g., there is an exact linear relationship between the two variables, with ϵ representing measurement error).

Much more frequently, the reasonableness of the model is indicated by data -- a scatter plot exhibiting a substantial linear pattern.



A Linear Probabilistic Model

If we think of an entire population of (x, y) pairs, then $\mu_{Y|x*}$ is the mean of all y values for which $x = x*$, and $\sigma^2_{Y|x*}$ is a measure of how much these values of y spread out about the mean value.

If, for example, x = age of a child and y = vocabulary size, then $\mu_{Y|5}$ is the average vocabulary size for all 5-year-old children in the population, and $\sigma^2_{Y|5}$ describes the amount of variability in vocabulary size for this part of the population.

A Linear Probabilistic Model

Once x is fixed, the only randomness on the right-hand side of the linear model equation is in the random error ϵ.

Recall that its mean value and variance are 0 and σ^2, respectively, whatever the value of x. This implies that

$$\mu_{Y|x*} = E(\beta_0 + \beta_1 x* + \epsilon)$$

$$= \beta_0 + \beta_1 x* + E(\epsilon)$$

$$= \beta_0 + \beta_1 x*$$

A Linear Probabilistic Model

$$\sigma^2_{Y|x*} = V(\beta_0 + \beta_1 x* + \epsilon)$$

$$= V(\beta_0 + \beta_1 x*) + V(\epsilon)$$

$$= 0 + \sigma^2$$

$$= \sigma^2$$

Replacing $x*$ in $\mu_{Y|x*}$ by x gives the relation $\mu_{Y|x} = \beta_0 + \beta_1 x$, which says that the *mean value (expected value, average)* of Y, rather than Y itself, is a linear function of x.

A Linear Probabilistic Model

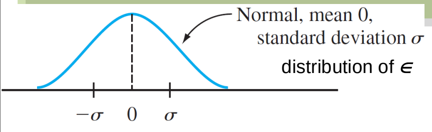
The true regression line $y = \beta_0 + \beta_1 x$ is thus the *line of mean values*; its height for any particular x value is the expected value of Y for that value of x.

The slope β_1 of the true regression line is interpreted as the ***expected (average)* change in *Y*** associated with a 1-unit increase in the value of x.

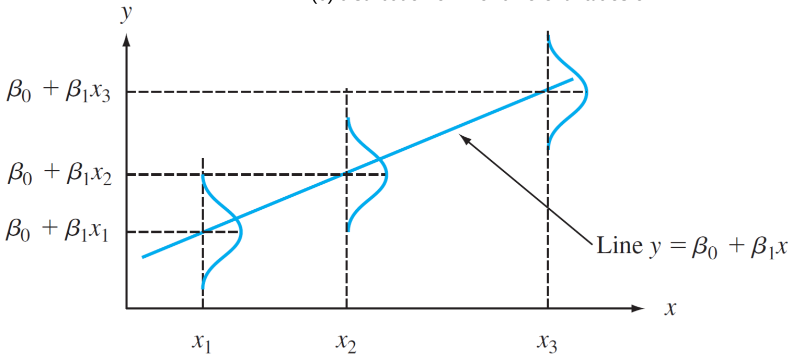
(This is equivalent to saying "the change in average (expected) Y associated with a 1-unit increase in the value of x.")

The variance (amount of variability) of the distribution of Y values is the same at each different value of x (homogeneity of variance).

When errors are normally distributed…



(b) distribution of Y for different values of x



The variance parameter σ^2 determines the extent to which each normal curve spreads out about the regression line

A Linear Probabilistic Model

When σ^2 is small, an observed point (x, y) will almost always fall quite close to the true regression line, whereas observations may deviate considerably from their expected values (corresponding to points far from the line) when σ^2 is large.

Thus, this variance can be used to tell us how good the linear fit is (we'll explore this notion of model fit later.)

Example 2

Suppose the relationship between applied stress *x* and time-to-failure *y* is described by the simple linear regression model with true regression line $y = 65 - 1.2x$ and $\sigma = 8$.

Then for any fixed value x^* of stress, time-to-failure has a normal distribution with mean value $65 - 1.2x^*$ and standard deviation 8.

Roughly speaking, in the population consisting of all (*x*, *y*) points, the magnitude of a typical deviation from the true regression line is about 8.

Example 2

cont'd

For *x* = 20, *Y* has mean value

$$\mu_{Y|20} = 65 - 1.2(20) = 41$$

So then Y for x=20 is **N(41, 64).**

From here,

$$P(Y > 50 \text{ when } x = 20) = P\left(Z > \frac{50 - 41}{8}\right)$$

$$= 1 - \Phi(1.13)$$

$$= .1292$$

Example 2

cont'd

The probability that time-to-failure exceeds 50 when applied stress is 25 is, because $\mu_{Y|25} = 65 - 1.2(25) = 35$,

$$P(Y > 50 \text{ when } x = 25) = P\left(Z > \frac{50 - 35}{8}\right)$$

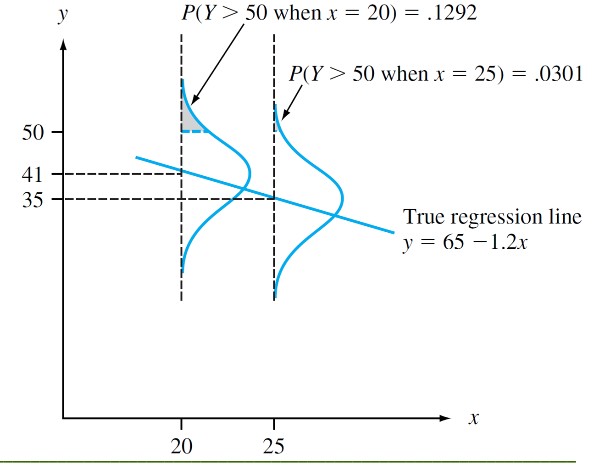
$$= 1 - \Phi(1.88)$$

$$= .0301$$

Example 2

cont'd

These probabilities are illustrated in:



Example 2

Suppose that Y_1 denotes an observation on time-to-failure made with x = 25 and Y_2 denotes an independent observation made with x = 24.

Then

$Y_1 - Y_2$ is normally distributed

with mean value

$E(Y_1 - Y_2) = \beta_0 + \beta_1(25) - \beta_0 + \beta_1(24) = \beta_1 = -1.2$,

variance

$V(Y_1 - Y_2) = \sigma^2 + \sigma^2 = 128$,

$\sqrt{128} = 11.314$

standard deviation

Example 2

cont'd

The probability that Y_1 exceeds Y_2 is

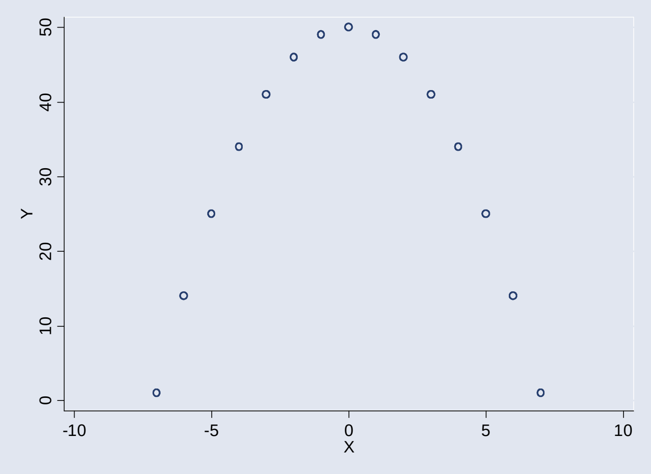
$$P(Y_1 - Y_2 > 0) = P\left(Z > \frac{0 - (-1.2)}{11.314}\right)$$

$$= P(Z > .11)$$

$$= .4562$$

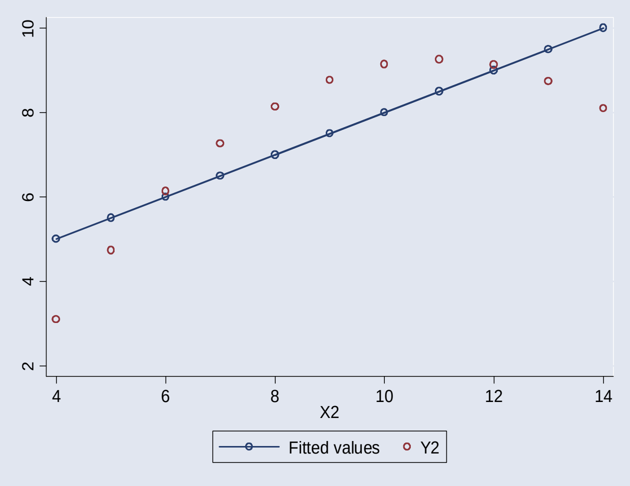
That is, even though we expected Y to decrease when x increases by 1 unit, it is not unlikely that the observed Y at x + 1 will be larger than the observed Y at x.

A couple of caveats with linear relationships…



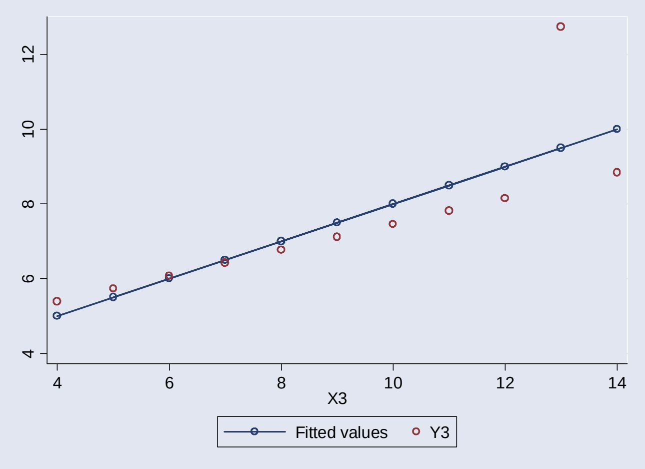
“no linear relationship” does not mean “no relationship”! Here, there is no linear, but there is a perfect quadratic relation between x and y

A couple of caveats…



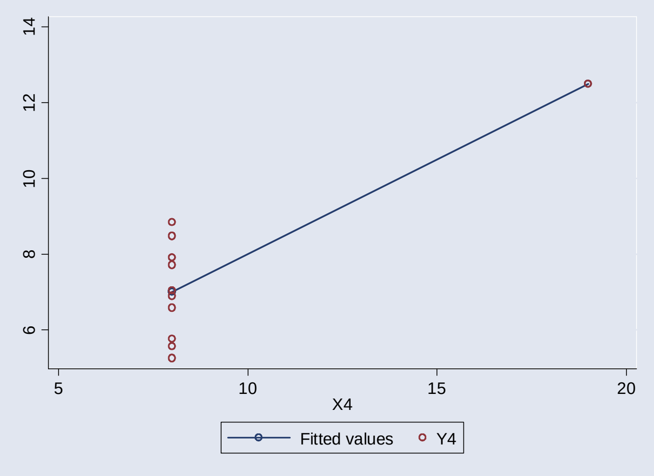
Relation not linear – but an ok approximation in this range of x

A couple of caveats…



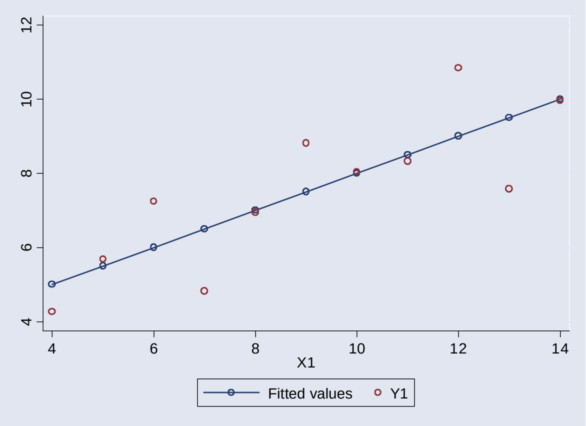
Too dependent on that one point out there

A couple of caveats…



Totally dependent on that one point out there

A couple of caveats…



Nice!

Estimating model parameters

Estimating Model Parameters

The values of β_0, β_1, and σ^2 will almost never be known to an investigator.

Instead, sample data consists of n observed pairs

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

from which the model parameters and the true regression line itself can be estimated.

The data (pairs) are assumed to have been obtained independently of one another.

Estimating Model Parameters

That is, y_i is the observed value of Y_i, where

$Y_i = \beta_0 + \beta_1 x_i + \in_I$

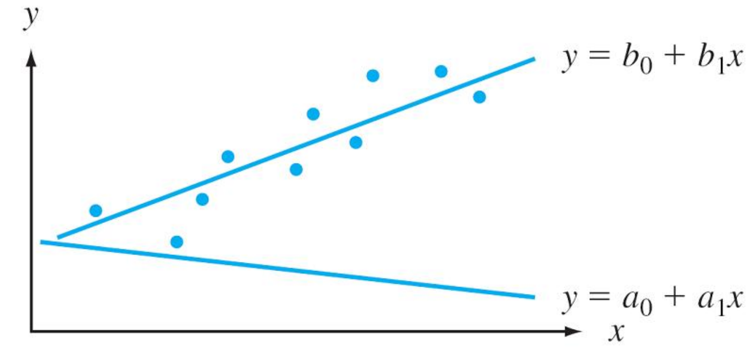
and the n deviations $\in_1, \in_2, \ldots, \in_n$ are independent rv's.

Independence of $Y_1, Y_2, \ldots, Y_n$ follows from independence of the $\in_i$'s.

According to the model, the observed points will be distributed around the true regression line in a random manner.

Estimating Model Parameters

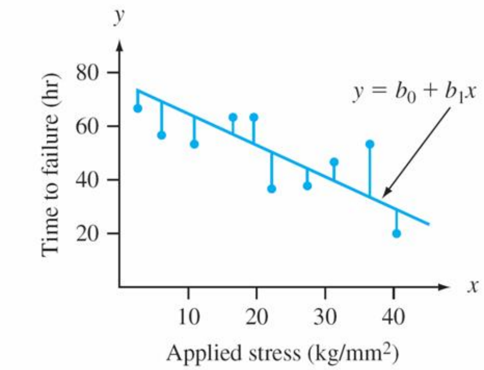
Figure shows a typical plot of observed pairs along with two candidates for the estimated regression line.



Estimating Model Parameters

The "best fit" line is motivated by the principle of **least squares**, which can be traced back to the German mathematician Gauss (1777–1855):

a line provides the ***best fit*** *to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.*



Estimating Model Parameters

The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$and $\hat{\beta}_1$, are called the **least squares estimates –** they are those values that minimize $f(b_0, b_1)$.

That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that $f(\hat{\beta}_0, \hat{\beta}_1) \le f(b_0, b_1)$ for any b_0 and b_1.

Estimating Model Parameters

The **estimated regression line** or **least squares line** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Estimating Model Parameters

Cancellation of the –2 factor and rearrangement gives the following system of equations, called the **normal equations:**

$$nb_0 + (\Sigma x_i)b_1 = \Sigma y_i$$

$$(\Sigma x_i)b_0 + (\Sigma x_i^2)b_1 = \Sigma x_i y_i$$

These equations are linear in the two unknowns b_0 and b_1.

Provided that not all x_i's are identical, the least squares estimates are the unique solution to this system.

Estimating Model Parameters

The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Shortcut formulas for the numerator and denominator of $\hat{\beta}_1$ are

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

Estimating Model Parameters

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\Sigma y_i - \hat{\beta}_1 \Sigma x_i}{n} = \overline{y} - \hat{\beta}_1 \overline{x}$$

The computational formulas for S_{xy} and S_{xx} require only the summary statistics Σx_i, Σy_i, Σx_i^2 and $\Sigma x_i y_i$ (Σy_i^2 will be needed shortly).

In computing $\hat{\beta}_0$, use extra digits in $\hat{\beta}_1$ because, if $\overline{x}$ is large in magnitude, rounding will affect the final answer.
In practice, the use of a statistical software package is preferable to hand calculation and hand-drawn plots.

Example

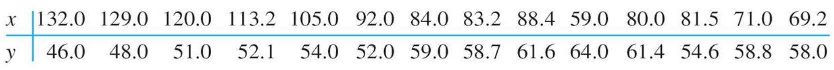
The cetane number is a critical property in specifying the ignition quality of a fuel used in a diesel engine.

Determination of this number for a biodiesel fuel is expensive and time-consuming.

The article "Relating the Cetane Number of Biodiesel Fuels to Their Fatty Acid Composition: A Critical Study" (*J. of Automobile Engr*., 2009: 565–583) included the following data on *x* = iodine value (g) and *y* = cetane number for a sample of 14 biofuels.

Example cont'd

The iodine value is the amount of iodine necessary to saturate a sample of 100 g of oil. The article's authors fit the simple linear regression model to this data, so let's follow their lead.



x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Example cont'd

Calculating the column sums gives

$\Sigma x_i = 1307.5$, $\Sigma y_i = 779.2$,

$\Sigma x_i^2 = 128{,}913.93$, $\Sigma x_i y_i = 71{,}347.30$,

$\Sigma y_i^2 = 43{,}745.22$, from which

$S_{xx} = 128{,}913.93 - (1307.5)^2/14 = 6802.7693$

$S_{xy} = 71{,}347.30 - (1307.5)(779.2)/14 = -1424.41429$

Example

cont'd

The estimated slope of the true regression line (i.e., the slope of the least squares line) is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -.20938742$$

We estimate that the expected change in true average cetane number associated with a 1g increase in iodine value is –.209—i.e., a decrease of .209.

Example

cont'd

Since $\bar{x}$ = 93.392857 and $\bar{y}$ = 55.657143, the estimated intercept of the true regression line (i.e., the intercept of the least squares line) is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 55.657143 - (-.20938742)(93.392857)$$

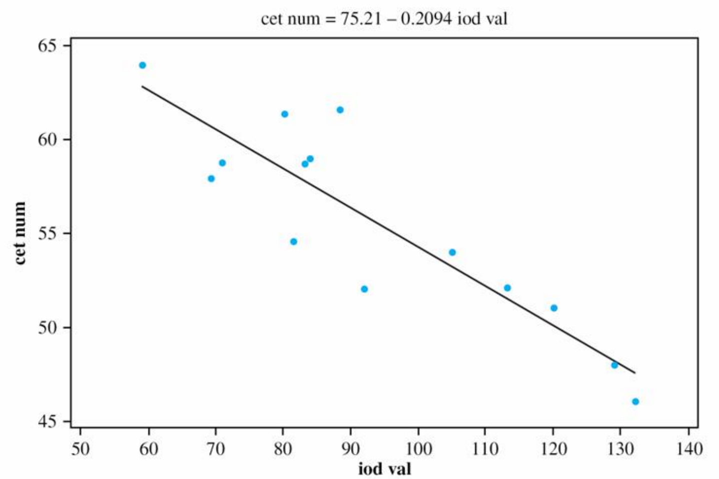
$$= 75.212432$$

The equation of the estimated regression line (least squares line) is $y = 75.212 - .2094x$, exactly that reported in the cited article.

Example

cont'd

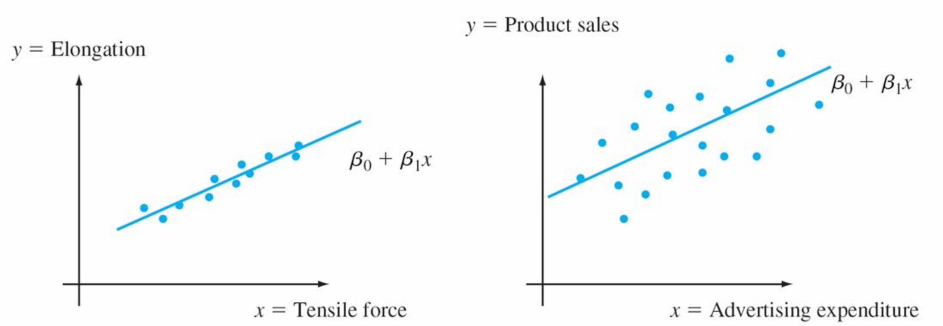
Scatter plot with the least squares line superimposed.



Estimating σ^2 and σ

Estimating σ^2 and σ

The parameter σ^2 determines the amount of spread about the true regression line. Two separate examples:



Estimating σ^2 and σ

An estimate of σ^2 will be used in confidence interval (CI) formulas and hypothesis-testing procedures presented in the next two sections.

Because the equation of the true line is unknown, the estimate is based on the extent to which the sample observations deviate from the estimated line.

Many large deviations (residuals) suggest a large value of σ^2, whereas deviations all of which are small in magnitude suggest that σ^2 is small.

Estimating σ^2 and σ

Definition
The **fitted** (or **predicted**) **values** $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by successively substituting $x_1,\ldots, x_n$ into the equation of the estimated regression line:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_2, \ldots, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$$

The **residuals** are the differences $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \ldots, y_n - \hat{y}_n$ between the observed and fitted y values.

Note – the **<u>residuals are estimates of the true error</u>**, since they are deviations of the y's from the estimated regression line, while the errors are deviations from the .

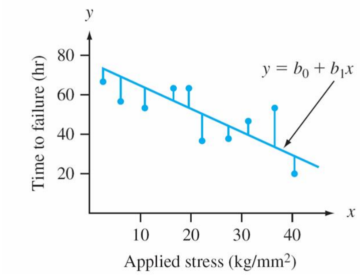
Estimating σ^2 and σ

In words, the predicted value $\hat{y}_i$ is the value of *y* that we would predict or expect when using the estimated regression line with $x = x_i$; $\hat{y}_i$ is the height of the estimated regression line above the value x_i for which the *i*th observation was made. It's just the mean for that population where $x = x_i$.

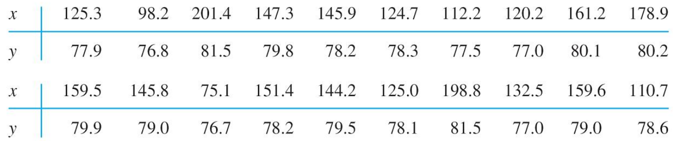
The residual $y_i - \hat{y}_i$ is the vertical deviation between the point (x_i, y_i) and the least squares line—a positive number if the point lies above the line and a negative number if it lies below the line.

Estimating σ^2 and σ

When the estimated regression line is obtained via the principle of least squares, the sum of the residuals should in theory be zero, if the error distribution is symmetric (and it is, due to our assumption).



Example



x	125.3	98.2	201.4	147.3	145.9	124.7	112.2	120.2	161.2	178.9
y	77.9	76.8	81.5	79.8	78.2	78.3	77.5	77.0	80.1	80.2
x	159.5	145.8	75.1	151.4	144.2	125.0	198.8	132.5	159.6	110.7
y	79.9	79.0	76.7	78.2	79.5	78.1	81.5	77.0	79.0	78.6

Relevant summary quantities (*summary statistics*) are

$\Sigma x_i = 2817.9$, $\Sigma y_i = 1574.8$, $\Sigma x^2_i = 415{,}949.85$,

$\Sigma x_i\, y_i = 222{,}657.88$, and $\Sigma y^2_i = 124{,}039.58$,

from which $x = 140.895$, $y = 78.74$, $S_{xx} = 18{,}921.8295$, $S_{xy} = 776.434$.

Example

cont'd

Thus

$$\hat{\beta}_1 = \frac{776.434}{18{,}921.8295} = .04103377 \approx .041$$

$$\hat{\beta}_0 = 78.74 - (.04103377)(140.895) = 72.958547 \approx 72.96$$

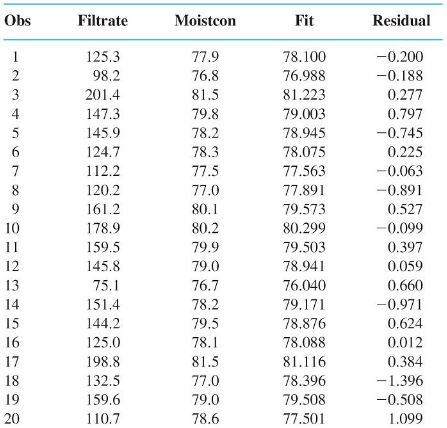
from which the equation of least squares line is
$y = 72.96 + .041x$.

For numerical accuracy, the fitted values are calculated from $\hat{y}_i = 72.958547 + .04103377x_i$

Example

cont'd

All predicted values (fits) and residuals appear in the accompanying table.



Obs	Filtrate	Moistcon	Fit	Residual
1	125.3	77.9	78.100	−0.200
2	98.2	76.8	76.988	−0.188
3	201.4	81.5	81.223	0.277
4	147.3	79.8	79.003	0.797
5	145.9	78.2	78.945	−0.745
6	124.7	78.3	78.075	0.225
7	112.2	77.5	77.563	−0.063
8	120.2	77.0	77.891	−0.891
9	161.2	80.1	79.573	0.527
10	178.9	80.2	80.299	−0.099
11	159.5	79.9	79.503	0.397
12	145.8	79.0	78.941	0.059
13	75.1	76.7	76.040	0.660
14	151.4	78.2	79.171	−0.971
15	144.2	79.5	78.876	0.624
16	125.0	78.1	78.088	0.012
17	198.8	81.5	81.116	0.384
18	132.5	77.0	78.396	−1.396
19	159.6	79.0	79.508	−0.508
20	110.7	78.6	77.501	1.099

Estimating σ^2 and σ

The **error sum of squares** (equivalently, residual sum of squares), denoted by SSE, is

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

Estimating σ^2 and σ

The divisor n – 2 in s^2 is the number of degrees of freedom (df) associated with SSE and the estimate s^2.

This is because to obtain s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of 2 df (just as μ had to be estimated in one sample problems, resulting in an estimated variance based on n – 1df).

Replacing each y_i in the formula for s^2 by the rv Y_i gives the estimator S^2.

It can be shown that S^2 is an unbiased estimator for σ^2

Example, cont.

The residuals for the filtration rate–moisture content data were calculated previously.

The corresponding error sum of squares is

$$\text{SSE} = (–.200)^2 + (–.188)^2 + \cdots + (1.099)^2 = 7.968$$

The estimate of σ^2 is then $\hat{\sigma}^2 = s^2 = 7.968/(20 – 2) = .4427$, and the estimated standard deviation is

$$\hat{\sigma} = s = \sqrt{.4427} = .665$$

Roughly speaking, .665 is the magnitude of a typical deviation from the estimated regression line—some points are closer to the line than this and others are further away.

Estimating σ^2 and σ

Computation of SSE from the defining formula involves much tedious arithmetic, because both the predicted values and residuals must first be calculated.

Use of the following computational formula does not require these quantities.

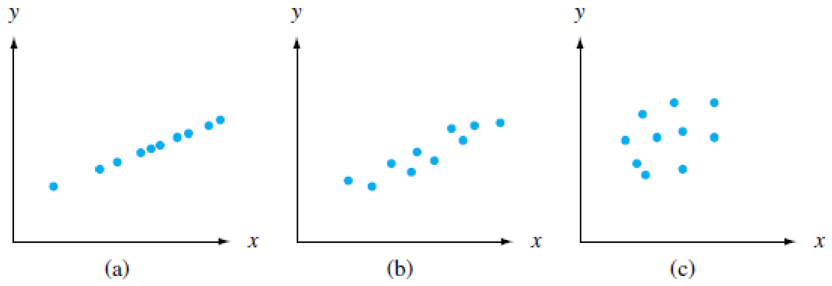
$$\text{SSE} = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

This expression results from substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into $\sum(y_i - \hat{y}_i)^2$, squaring the summand, carrying through the sum to the resulting three terms, and simplifying.

The Coefficient of Determination

The Coefficient of Determination

Different variability in observed *y* values:



Using the model to explain y variation:
(a) data for which all variation is explained;
(b) data for which most variation is explained;
(c) data for which little variation is explained

The Coefficient of Determination

The points in the first plot all fall exactly on a straight line. In this case, all (100%) of the sample variation in *y* can be attributed to the fact that *x* and *y* are linearly related in combination with variation in *x*.

The points in the second plot do not fall exactly on a line, but compared to overall *y* variability, the deviations from the least squares line are small.

It is reasonable to conclude in this case that much of the observed *y* variation can be attributed to the approximate linear relationship between the variables postulated by the simple linear regression model.

When the scatter plot looks like that in the third plot, there is substantial variation about the least squares line relative to overall *y* variation, so the simple linear regression model fails to explain variation in *y* by relating *y* to *x*.

The Coefficient of Determination

The error sum of squares SSE can be interpreted as a measure of how much variation in *y* is left unexplained by the model—that is, how much cannot be attributed to a linear relationship.

In the first plot SSE = 0, and there is no unexplained variation, whereas unexplained variation is small for second, and large for the third plot.

A quantitative measure of the total amount of variation in observed *y* values is given by the **total sum of squares**

$$\text{SST} = S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

The Coefficient of Determination

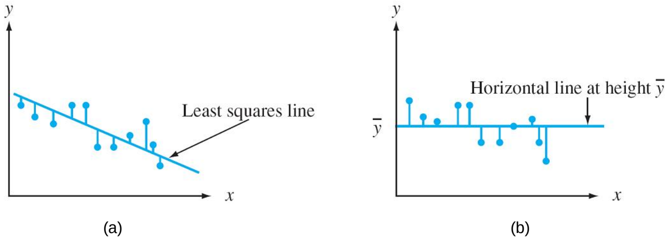
Total sum of squares is the sum of squared deviations about the sample mean of the observed *y* values – when no predictors are taken into account.

Thus the same number $\bar{y}$ is subtracted from each y_i in SST, whereas SSE involves subtracting each different predicted value $\hat{y}_i$ from the corresponding observed y_i.

This in some sense is as bad as SSE can get if there is no regression model (ie, slope is 0).

The Coefficient of Determination

Just as SSE is the sum of squared deviations about the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$, SST is the sum of squared deviations about the horizontal line at height (since then vertical deviations are $y_i - \bar{y}$)



Sums of squares illustrated: (a) SSE = sum of squared deviations about the least squares line; (b) SSE = sum of squared deviations about the horizontal line

The Coefficient of Determination

Furthermore, because the sum of squared deviations about the least squares line is smaller than the sum of squared deviations about *any* other line, SSE < SST unless the horizontal line itself is the least squares line.

The ratio SSE/SST is the proportion of total variation that cannot be explained by the simple linear regression model, and 1 – SSE/SST (a number between 0 and 1) is the proportion of observed *y* variation explained by the model.

The Coefficient of Determination

Definition
The **coefficient of determination**, denoted by r^2, is given by

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

It is interpreted as the proportion of observed *y* variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between *y* and *x*).

The higher the value of r^2, the more successful is the simple linear regression model in explaining *y* variation.

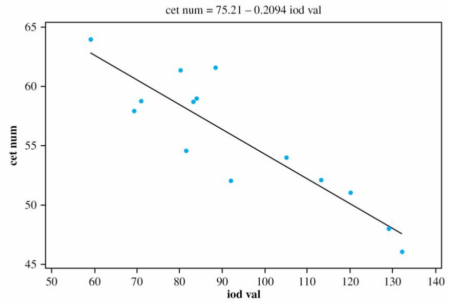
The Coefficient of Determination

When regression analysis is done by a statistical computer package, either r^2 or $100r^2$ (the percentage of variation explained by the model) is a prominent part of the output.

If r^2 is small, an analyst will usually want to search for an alternative model (either a nonlinear model or a multiple regression model that involves more than a single independent variable) that can more effectively explain *y* variation.

Example

The scatter plot of the iodine value–cetane number data portends a reasonably high r^2 value.



Example

cont'd

The coefficient of determination is then

r^2 = 1 – SSE/SST = 1 – (78.920)/(377.174) = .791

That is, 79.1% of the observed variation in cetane number is attributable to (can be explained by) the simple linear regression relationship between cetane number and iodine value (r^2 values are even higher than this in many scientific contexts, but social scientists would typically be ecstatic at a value anywhere near this large!).

The Coefficient of Determination

The coefficient of determination can be written in a slightly different way by introducing a third sum of squares—**regression sum of squares,** SSR—given by

$$\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 = \text{SST} - \text{SSE}.$$

Regression sum of squares is interpreted as the amount of total variation that *is* explained by the model.

Then we have

r^2 = 1 – SSE/SST = (SST – SSE)/SST = SSR/SST

the ratio of explained variation to total variation.

Inference About the Slope Parameter β_1

Inferences About the Slope Parameter β_1

In virtually all of our inferential work thus far, the notion of sampling variability has been pervasive.

In particular, properties of sampling distributions of various statistics have been the basis for developing confidence interval formulas and hypothesis-testing methods.

The key idea here is that the value of any quantity calculated from sample data—the value of any statistic—will vary from one sample to another.

Example 10

The following data is representative of that reported in the article “An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data” (*J. of Engr. for Power*, July 1973: 165–170), x = burner-area liberation rate (MBtu/hr-ft^2) and y = NO$_x$ emission rate (ppm).

There are 14 observations, made at the x values 100, 125, 125, 150, 150, 200, 200, 250, 250, 300, 300, 350, 400, and 400, respectively.

Simulation experiment cont’d

Suppose that the slope and intercept of the true regression line are $\beta_1 = 1.70$ and $\beta_0 = –50$, with $\sigma = 35$.

Let’s fix x values 100, 125, 125, 150, 150, 200, 200, 250, 250, 300, 300, 350, 400, and 400.

We then generate a sample of random deviations $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{14}$ from a normal distribution with mean 0 and standard deviation 35

and then add $\tilde{\epsilon}_i$ to $\beta_0 + \beta_1 x_i$ to obtain 14 corresponding y values.

LS calculations were then carried out to obtain the estimated slope, intercept, and standard deviation for this sample of 14 pairs (x_i, y_i).

Simulation experiment

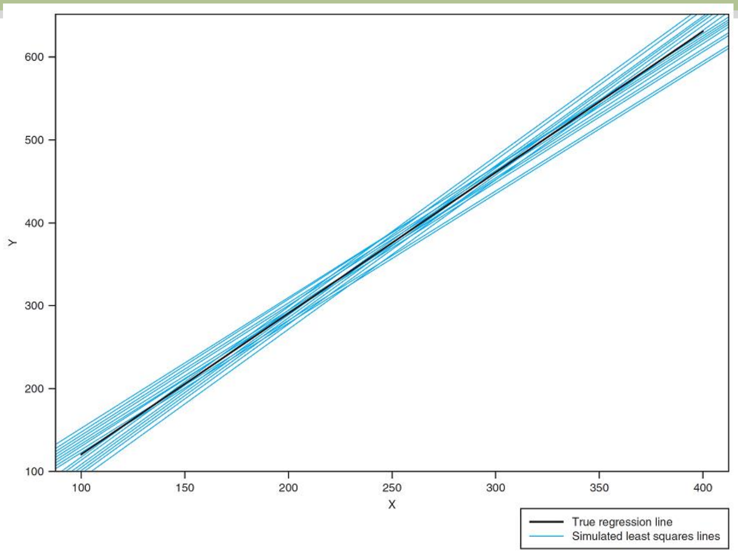
cont'd

This process was repeated a total of 20 times, resulting in the values:

	$\hat{\beta}_1$	$\hat{\beta}_0$	s		$\hat{\beta}_1$	$\hat{\beta}_0$	s
1.	1.7559	−60.62	43.23	11.	1.7843	−67.36	41.80
2.	1.6400	−49.40	30.69	12.	1.5822	−28.64	32.46
3.	1.4699	−4.80	36.26	13.	1.8194	−83.99	40.80
4.	1.6944	−41.95	22.89	14.	1.6469	−32.03	28.11
5.	1.4497	5.80	36.84	15.	1.7712	−52.66	33.04
6.	1.7309	−70.01	39.56	16.	1.7004	−58.06	43.44
7.	1.8890	−95.01	42.37	17.	1.6103	−27.89	25.60
8.	1.6471	−40.30	43.71	18.	1.6396	−24.89	40.78
9.	1.7216	−42.68	23.68	19.	1.7857	−77.31	32.38
10.	1.7058	−63.31	31.58	20.	1.6342	−17.00	30.93

There is clearly variation in values of the estimated slope and estimated intercept, as well as the estimated standard deviation.

Simulation experiment: graphs of the true regression line and 20 least squares lines



Inferences About the Slope Parameter β_1

The estimators are:

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \overline{x})(Y_i - \overline{Y})}{\Sigma(x_i - \overline{x})^2} \quad => \quad \hat{\beta}_1 = \frac{\Sigma(x_i - \overline{x})Y_i}{S_{xx}} = \Sigma c_i Y_i$$

That is, $\hat{\beta}_1$is a linear function of the independent rv's $Y_1, Y_2, \ldots, Y_n$, each of which is normally distributed.

$$\hat{\beta}_0 = \frac{\Sigma Y_i - \hat{\beta}_1 \Sigma x_i}{n}$$

$$\hat{\sigma}^2 = S^2 = \frac{\Sigma Y_i^2 - \hat{\beta}_0 \Sigma Y_i - \hat{\beta}_1 \Sigma x_i Y_i}{n - 2}$$

Inferences About the Slope Parameter β_1

Invoking properties of a linear function of random variables as discussed earlier, leads to the following results.

1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 (the distribution of $\hat{\beta}_1$ is always centered at the value of β_1).

2. The variance and standard deviation of β_1 are

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

where $S_{xx} = \Sigma(x_i - x)^2$

Inferences About the Slope Parameter β_1

Replacing σ by its estimate *s* gives an estimate for $\sigma_{\hat{\beta}_1}$ (the estimated standard deviation, i.e., estimated standard error, of $\hat{\beta}_1$):

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$$

(This estimate can also be denoted by $\hat{\sigma}_{\hat{\beta}_1}$.)

3. The estimator $\hat{\beta}_1$ has a normal distribution (because it is a linear function of independent normal rv's).

Inferences About the Slope Parameter β_1

The variance of $\hat{\beta}_1$ equals the variance σ^2 of the random error (or, equivalently, of any Y_i), divided by $\Sigma(x_i - \overline{x})^2$ This denominator is a measure of how spread out the x_i's are about $\overline{x}$.

We conclude that making observations at x_i values that are quite spread out results in a more precise estimator of the slope parameter (smaller variance of $\hat{\beta}_1$), whereas values of x_i all close to one another imply a highly variable estimator.

Of course, if the x_i's are spread out too far, a linear model may not be appropriate throughout the range of observation.

Inferences About the Slope Parameter β_1

Theorem

The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

has a *t* distribution with *n* – 2 df.

A Confidence Interval for β_1

A Confidence Interval for β_1

As in the derivation of previous CIs, we begin with a probability statement:

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate β_1 and substitution of estimates in place of the estimators gives the CI formula.

A 100(1 – α)% **CI for the slope β_1** of the true regression line is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

Example

Variations in clay brick masonry weight have implications not only for structural and acoustical design but also for design of heating, ventilating, and air conditioning systems.

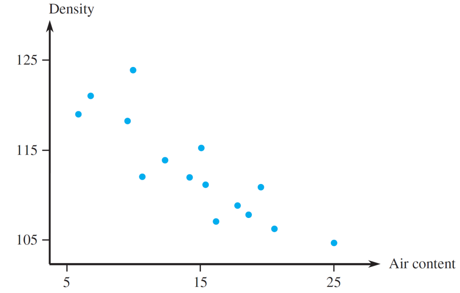
The article “Clay Brick Masonry Weight Variation” (*J. of Architectural Engr.*, 1996: 135–137) gave a scatter plot of *y* = mortar dry density (lb/ft³) versus *x* = mortar air content (%) for a sample of mortar specimens, from which the following representative data was read:

x	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0	15.3	16.2	17.8	18.7	19.7	20.6	25.0
y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1	111.3	107.2	108.9	107.8	111.0	106.2	105.0

Example

cont’d

The scatter plot of this data certainly suggests the appropriateness of the simple linear regression model; there appears to be a substantial negative linear relationship between air content and density, one in which density tends to decrease as air content increases.



Example

cont’d

The values of the summary statistics required for calculation of the least squares estimates are

$\Sigma x_i = 218.1$ $\Sigma y_i = 1693.6$ $\Sigma x_i y_i = 24{,}252.54$ $\Sigma x_i^2 = 3577.01$
$\Sigma y_i^2 = 191{,}672.90$; n = 15

from which $S_{xy} = –372.404$, $S_{xx} = 405.836$, $\hat{\beta}_1 = –.917622$, $\hat{\beta}_0 = 126.248889$,

SST = 454.163, SSE = 112.4432, and
$r^2 = 1 – 112.4432/454.1693 = .752$.

Example

cont'd

Roughly 75% of the observed variation in density can be attributed to the simple linear regression model relationship between density and air content. Error df is 15 – 2 = 13, giving s^2 = SSE/(n-2) = 112.44/13 = 8.65 and s = 2.941.

The estimated standard deviation of $\hat{\beta}_1$ is

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{2.941}{\sqrt{405.836}} = .1460$$

A confidence level of 95% requires $t_{.025,13}$ = 2.160. The CI is

–.918 ± (2.160)(.1460) = –.918 ± .315 = (–1.23, –.60)

Hypothesis-Testing Procedures

Hypothesis-Testing Procedures

The most commonly encountered pair of hypotheses about β_1 is H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. When this null hypothesis is true, $\mu_{Y\cdot x} = \beta_0$ independent of x. Then knowledge of x gives no information about the value of the dependent variable.

Null hypothesis: H_0: $\beta_1 = \beta_{10}$

Test statistic value: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Hypothesis-Testing Procedures

Alternative Hypothesis	**Alternative Hypothesis**
H_a: $\beta_1 > \beta_{10}$	$t \geq t_{\alpha,n-2}$
H_a: $\beta_1 < \beta_{10}$	$t \leq -t_{\alpha,n-2}$
H_a: $\beta_1 \neq \beta_{10}$	either $t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$

A P-value based on n – 2 can be calculated just as was done previously for t tests.

the test statistic value is the ***t* ratio**

$t = \hat{\beta}_1/s_{\hat{\beta}_1}$

Regression analysis in R

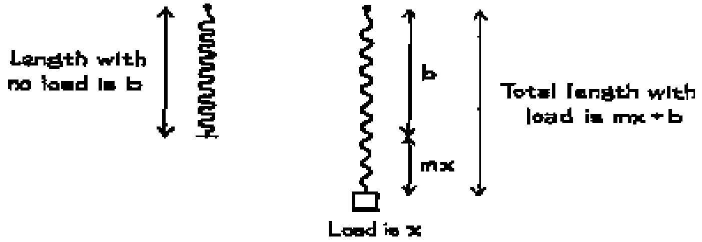
Regression analysis in R

```
Fitting Linear Models in R

Description

lm is used to fit linear models (regression)


Command:
lm(formula, data, subset, ...)

Example:
Model1 = lm(outcome ~ predictor)
```

Example: Regression analysis in R

Robert Hooke (England, 1653-1703) was able to assess the relationship between the length of a spring and the load placed on it. He just hung weights of different sizes on the end of a spring, and watched what happened. When he increased the load, the spring got longer. When he reduced the load, the spring got shorter. And the relationship was more or less linear.

Let *b* be the length of the spring with no load. When a weight of *x* kilograms is tied to the end of the spring, the spring stretches to a new length.



Example: Regression analysis in R

According to Hooke's law, the amount of stretch is proportional to the weight *x*. So the new length of the spring is

$$y = mx + b$$

In this equation, *m* and *b* are constants which depend on the spring.

Their values are unknown and have to be estimated using experimental data.

Hooke's Law in R

Data below come from an experiment in which weights of various sizes were loaded on the end of a length of piano wire.

The first column shows the weight of the load. The second column shows the measured length. With 20 pounds of load, this "spring" only stretched about 0.2 inch (10 kg is approximately 22 lb, 0.5 cm is approximately 0.2 in).

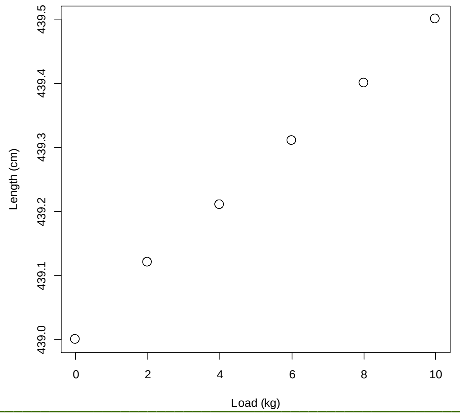
Piano wire is not very stretchy!

439 cm is about 14.4 feet
(the room needed to have a
fairly high ceiling)

Weight	Length
0 kg	439.00 cm
2	439.12
4	439.21
6	439.31
8	439.40
10	439.50

Regression analysis in R

```
x = c(0, 2, 4, 6, 8, 10)
y = c( 439.00, 439.12, 439.21, 439.31, 439.40, 439.50)
plot( x, y, xlab = 'Load (kg)', ylab = 'Length (cm)' )
```



Regression analysis in R

```
Hooke.lm = lm( y ~ x)

summary(Hooke.lm)

Residuals:
        1         2         3         4         5         6 
-0.010952  0.010762  0.002476  0.004190 -0.004095 -0.002381 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept) 4.390e+02  6.076e-03 72254.92  < 2e-16 ***
x           4.914e-02  1.003e-03    48.98 1.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.008395 on 4 degrees of freedom
Multiple R-squared: 0.9983,     Adjusted R-squared: 0.9979 
F-statistic:  2399 on 1 and 4 DF,  p-value: 1.04e-06
```

Regression analysis in R

```
coefficients(Hooke.lm)
# (Intercept)            x 
# 439.01095238   0.04914286 

Hooke.coefs = coefficients(Hooke.lm)

Hooke.coefs[1]
439.011

Hooke.coefs[2]
0.04914286

# expected stretch for x = 5kg
Hooke.coefs[1] + Hooke.coefs[2] * 5 = 439.2567
```

If we knew that line for certain, then the estimated standard deviation of the actual stretch around that expected value would be:

Residual standard error: 0.008395

Regression analysis in R

```
coefficients(Hooke.lm)
# (Intercept)           x
# 439.01095238   0.0491428
```

So, in the Hooke’s law, *m* = 0.05 cm per kg, and *b* = 439.01 cm.

The method of least squares estimates the length of the spring under no load to be 439.01 cm.

And each kilogram of load **makes** this particular spring stretch by an amount estimated as 0.05 cm on average.

Note that in general correlation is not causation, but in this controlled experimental setting, the causation is clear and simple. This is an exception more than a rule in statistical modeling.

Regression analysis in R

So, in the Hooke’s law, *m* = 0.05 cm per kg, and *b* = 439.01 cm.

The method of least squares estimates the length of the spring under no load to be 439.01 cm.
This is a bit longer than the measured length at no load (439.00 cm).

A statistician would trust the least squares estimate over the measurement. Why? Because the least squares estimate takes advantage of all six measurements, not just one. Some of the measurement error is likely to cancel out. Of course, the six measurements are tied together by a solid theory: Hooke's law. Without the theory, the least squares estimate wouldn't be worth much.

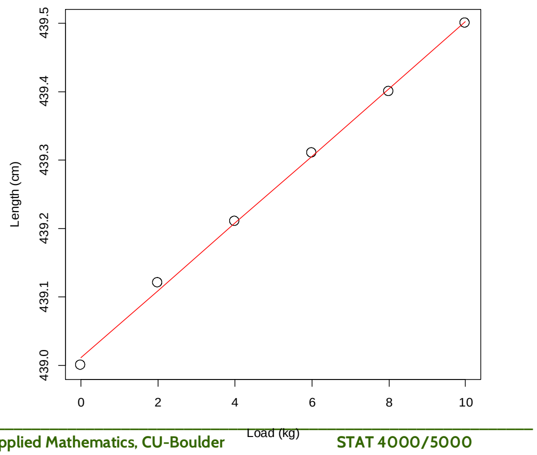
example taken from analytictech.com

Regression analysis in R

```
y.hat <- Hooke.coefs[1] + Hooke.coefs[2]* x
439.0110 439.1092 439.2075 439.3058 439.4041 439.5024

> predict(Hooke.lm)
439.0110 439.1092 439.2075 439.3058 439.4041 439.5024

e.hat <- y - y.hat

mean( e.hat )
# -2.842171e-14

cbind( y, y.hat, e.hat )
#           y    y.hat        e.hat
# [1,] 439.00 439.0110 -0.010952381
# [2,] 439.12 439.1092  0.010761905
# [3,] 439.21 439.2075  0.002476190
# [4,] 439.31 439.3058  0.004190476
# [5,] 439.40 439.4041 -0.004095238
# [6,] 439.50 439.5024 -0.002380952
```

Regression analysis in R

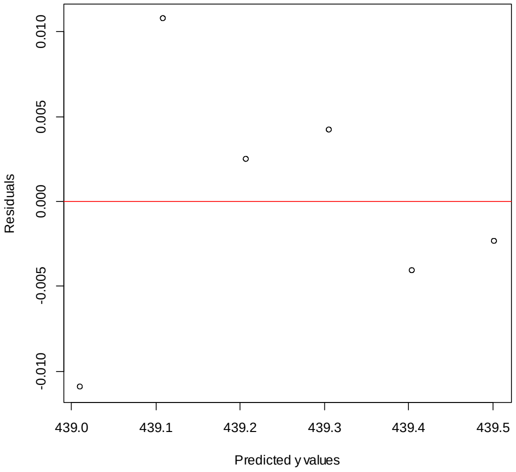
```
plot( x, y, xlab = 'Load (kg)', ylab = 'Length (cm)' )
lines( x, y.hat, lty = 1, col = 'red' )
```



Regression analysis in R

```
plot(y.hat, e.hat, xlab = 'Predicted y values',
  ylab = 'Residuals' )
abline(h = 0, col = 'red')
```

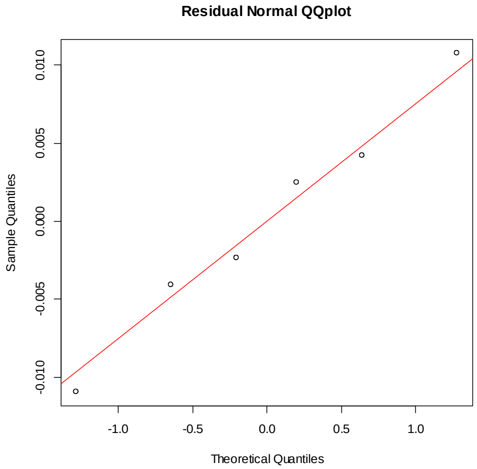
This plot doesn't look good, actually



Regression analysis in R

```
qqnorm( e.hat, main = 'Residual Normal QQplot' )
abline( mean( e.hat ), sd( e.hat ), col = 'red' )
```

This plot looks good



Inferences Concerning $\mu_{Y\cdot x^*}$ and the Prediction of Future Y Values

Inference Concerning Mean of Future Y

```
Hooke.coefs = coefficients(Hooke.lm)
# (Intercept)            x
# 439.01095238   0.04914286

# expected stretch for x = 5kg
Hooke.coefs[1] + Hooke.coefs[2] * 5 = 439.2567
```

If we knew that line for certain, then the estimated standard deviation of the actual stretch around that expected value would be:

```
Residual standard error: 0.008395
```

But we don’t know m and b for certain – we’ve just estimated them, and we know that their estimators are Normal random variables.

Inference Concerning Mean of Future Y

Let *x** denote a specified value of the independent variable *x*.

Once the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have been calculated, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ can be regarded either as a point estimate of $\mu_{Y \cdot x^*}$ (the expected or true average value of *Y* when *x* = *x**) or as a prediction of the *Y* value that will result from a single observation made when *x* = *x**.

The point estimate or prediction by itself gives no information concerning how precisely $\mu_{Y \cdot x^*}$ has been estimated or *Y* has been predicted.

Inference Concerning Mean of Future Y

This can be remedied by developing a CI for $\mu_{Y \cdot x^*}$ and a prediction interval (PI) for a single *Y* value.

Before we obtain sample data, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are subject to sampling variability—that is, they are both statistics whose values will vary from sample to sample.

Suppose, for example, that $\beta_0 = 439$ and $\beta_1 = 0.05$.

Then a first sample of (*x*, *y*) pairs might give $\hat{\beta}_0 = 439.35$, $\hat{\beta}_1 = 0.048$; a second sample might result in $\hat{\beta}_0 = 438.52$, $\hat{\beta}_1 = 0.051$; and so on.

Inference Concerning Mean of Future Y

It follows that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ itself varies in value from sample to sample, so it is a statistic.

If the intercept and slope of the population line are the aforementioned values 439 and 0.05, respectively, and *x** =5kgs, then this statistic is trying to estimate the value

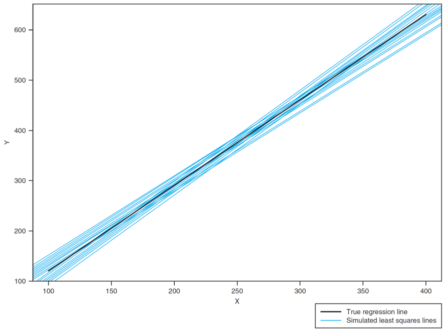
$$439 + 0.05(5) = 439.25$$

The estimate from a first sample might be 439.35 + 0.048(5) = 439.59, from a second sample might be 438.52 + 0.051(5) = 438.775 , and so on.

Inference Concerning Mean of Future Y

Consider the value *x** = 10.

Recall that because 10 is further than *x* = 5 from the "center of the data", the estimated "y.hat" values are more variable.



Inference Concerning Mean of Future Y

In the same way, inferences about the mean *Y* value $\hat{\beta}_0 + \hat{\beta}_1$ *x** are based on properties of the sampling distribution of the statistic $\hat{\beta}_0 + \hat{\beta}_1$ *x**.

Substitution of the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{\beta}_0 + \hat{\beta}_1$ *x** followed by some algebraic manipulation leads to the representation of $\hat{\beta}_0 + \hat{\beta}_1$ *x** as a linear function of the *Yi*'s:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] Y_i = \sum_{i=1}^{n} d_i Y_i$$

The coefficients d_1, d_2, …., d_n in this linear function involve the x_i's and *x**, all of which are fixed.

Inference Concerning Mean of Future Y

Application of the rules to this linear function gives the following properties.

Proposition

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$ where *x** is some fixed value of *x*. Then

1. The mean value of $\hat{Y}$ is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

Thus $\hat{\beta}_0 + \hat{\beta}_1$ *x** is an unbiased estimator for $\hat{\beta}_0 + \hat{\beta}_1$ *x** (i.e., for $\mu_{Y \cdot x^*}$).

Inference Concerning Mean of Future Y

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma^2_{\hat{Y}} = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum x_i^2 - (\sum x_i)^2/n} \right] = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

And the standard deviation $\sigma_{\hat{Y}}$ is the square root of this expression. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1$ *x**,

denoted by $s_{\hat{Y}}$ or $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$, results from replacing σ by its estimate *s*:

$$s_{\hat{Y}} = s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y}$ has a normal distribution.

Inference Concerning Mean of Future Y

The variance of $\hat{\beta}_0 + \hat{\beta}_1$ *x** is smallest when *x** = $\bar{x}$ and increases as *x** moves away from $\bar{x}$ in either direction.

Thus the estimator of $\mu_{Y \cdot x^*}$ is more precise when *x** is near the center of the x_i's than when it is far from the values at which observations have been made. This will imply that both the CI and PI are narrower for an *x** near *x* than for an *x** far from *x*.

Most statistical computer packages will provide both $\hat{\beta}_0 + \hat{\beta}_1$ *x** and $s_{\hat{\beta}_0 + \hat{\beta}_1 x^*}$ for any specified *x** upon request.

Inference Concerning Mean of Future Y

```
Hooke.coefs = coefficients(Hooke.lm)
# (Intercept)            x
# 439.01095238   0.04914286

> predict(Hooke.lm)
439.0110 439.1092 439.2075 439.3058 439.4041 439.5024

# expected stretch for x = 5kg
Hooke.coefs[1] + Hooke.coefs[2] * 5 = 439.2567


> predict(Hooke.lm, se.fit=TRUE)
$fit
       1        2        3        4        5        6
439.0110 439.1092 439.2075 439.3058 439.4041 439.5024

$se.fit
[1] 0.006075862 0.004561497 0.003571111 0.003571111
0.004561497 0.006075862
```

Inference Concerning Mean of Future Y

Just as inferential procedures for β_1 were based on the *t* variable obtained by standardizing β_1, a *t* variable obtained by standardizing $\hat\beta_0 + \hat\beta_1 x^*$ leads to a CI and test procedures here.

Theorem

The variable

$$T = \frac{\hat\beta_0 + \hat\beta_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat\beta_0 + \hat\beta_1 x^*}} = \frac{\hat Y - (\beta_0 + \beta_1 x^*)}{S_{\hat Y}} \qquad \textbf{(12.5)}$$

has a *t* distribution with *n* – 2 df.

Inference Concerning Mean of Future Y

A probability statement involving this standardized variable can now be manipulated to yield a confidence interval for $\mu_{Y\cdot x^*}$

A 100(1 – α)% **CI** for the expected value of *Y* when *x* = *x**, is

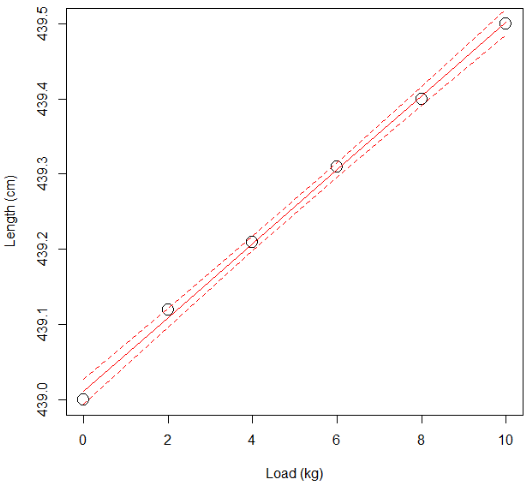
$$\hat\beta_0 + \hat\beta_1 x^* \pm t_{\alpha/2,n-2} \cdot s_{\hat\beta_0 + \hat\beta_1 x^*} = \hat y \pm t_{\alpha/2,\,n-2} \cdot s_{\hat Y}$$

This CI is centered at the point estimate for $\mu_{Y\cdot x^*}$ and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator on which the point estimate is based.

Inference Concerning Mean of Future Y

```
preds =  predict(Hooke.lm, se.fit=TRUE)
$fit
439.0110 439.1092 439.2075 439.3058 439.4041 439.5024

$se.fit
0.00608 0.00456 0.00357 0.00357 0.00456 0.00608

> plot( x, y, xlab = 'Load (kg)', ylab = 'Length (cm)',cex=2 )
> lines( x, y.hat, lty = 1, col = 'red' )
> preds =  predict(Hooke.lm, se.fit=TRUE)
> CI.ub = preds$fit + preds$se.fit*qt(.975,6-2)
> CI.lb = preds$fit - preds$se.fit*qt(.975,6-2)
> lines(x,CI.ub,lty=2,col="red")
> lines(x,CI.lb,lty=2,col="red")
```

Inference Concerning Mean of Future Y



Example: concrete

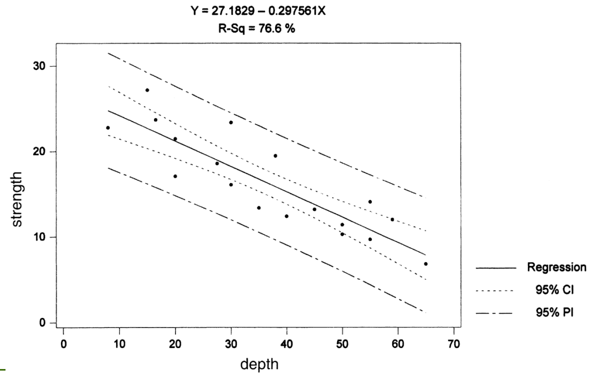
Corrosion of steel reinforcing bars is the most important durability problem for reinforced concrete structures.

Carbonation of concrete results from a chemical reaction that also lowers the pH value by enough to initiate corrosion of the rebar.

Representative data on *x* = carbonation depth (mm) and *y* = strength (MPa) for a sample of core specimens taken from a particular building follows

Example -- concrete cont'd

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4
x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8



Example -- concrete cont'd

Let's now calculate a 95% confidence interval, for the mean strength for all core specimens having a carbonation depth of 45. The interval is centered at

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(45) = 27.18 - .2976(45) = 13.79$$

The estimated standard deviation of the statistic $\hat{Y}$ is

$$s_{\hat{Y}} = 2.8640\sqrt{\frac{1}{18} + \frac{(45 - 36.6111)^2}{4840.7778}} = .7582$$

The 16 df *t* critical value for a 95% confidence level is 2.120, from which we determine the desired interval to be

$$13.79 \pm (2.120)(.7582) = 13.79 \pm 1.61 = (12.18, 15.40)$$

A Prediction Interval for a Future Value of *Y*

A Prediction Interval for a Future Value of *Y*

Rather than calculate an interval estimate for $\mu_{Y \cdot x^*}$ an investigator may wish to obtain a range or an interval of plausible values for the value of *Y* associated with some future observation when the independent variable has value *x**.

Consider, for example, relating vocabulary size *y* to age of a child *x*. The CI with *x** = 6 would provide a range that covers with 95% confidence the true average vocabulary size for all 6-year-old children.

Alternatively, we might wish an interval of plausible values for the vocabulary size of a particular 6-year-old child. How can you tell that a child is “off the chart” for example?

A Prediction Interval for a Future Value of *Y*

A CI refers to a parameter, or population characteristic, whose value is fixed but unknown to us.

In contrast, a future value of *Y* is not a parameter but instead a random variable; for this reason we refer to an interval of plausible values for a future *Y* as a **prediction interval** rather than a confidence interval.

A Prediction Interval for a Future Value of *Y*

The error of prediction is *Y* – ($\hat{\beta}_0$- $\hat{\beta}_1$x*), a difference between two random variables. Because the future value *Y* is independent of the observed Y_i’s,

$$V[Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)] = \text{variance of prediction error}$$

$$= V(Y) + V(\hat{\beta}_0 + \hat{\beta}_1 x^*)$$

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

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

A Prediction Interval for a Future Value of *Y*

Furthermore, because $E(Y) = \beta_0 + \beta_1 x^*$ and expectation of $\hat{\beta}_0 + \hat{\beta}_1 x^* = \beta_0 + \beta_1 x^*$, the expected value of the prediction error is $E(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*) = 0$.

It can then be shown that the standardized variable

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}}$$

has a *t* distribution with *n* – 2 df.

A Prediction Interval for a Future Value of *Y*

Manipulating to isolate *Y* between the two inequalities yields the following interval.

A 100(1 – α)% **PI for a future *Y* observation to be made when *x* = *x**** is

$$\begin{aligned} &\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} \qquad \textbf{(12.7)} \\ &= \hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{Y}}} \end{aligned}$$

A Prediction Interval for a Future Value of *Y*

The interpretation of the prediction level 100(1 – α)% is similar to that of previous confidence levels—if is used repeatedly, in the long run the resulting interval will actually contain **the observed *y* values** 100(1 – α)% of the time.

Notice that the 1 underneath the initial square root symbol makes the PI (12.7) wider than the CI (12.6), though the intervals are both centered at $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Also, as $n \to \infty$ the width of the CI approaches 0, whereas the width of the PI does not (because even with perfect knowledge of β_0 and β_1, there will still be randomness in prediction).

Example -- concrete

Let's return to the carbonation depth-strength data example and calculate a 95% PI for a strength value that would result from selecting a single core specimen whose depth is 45 mm. Relevant quantities from that example are

$\hat{y} = 13.79$ $s_{\hat{Y}} = .7582$ $s = 2.8640$

For a prediction level of 95% based on *n* – 2 = 16 df, the *t* critical value is 2.120, exactly what we previously used for a 95% confidence level.

Example -- concrete

cont'd

The prediction interval is then

$$13.79 \pm (2.120)\sqrt{(2.8640)^2 + (.7582)^2} = 13.79 \pm (2.120)(2.963)$$

$$= 13.79 \pm 6.28$$

$$= (7.51, 20.07)$$

Plausible values for a single observation on strength when depth is 45 mm are (at the 95% prediction level) between 7.51 MPa and 20.07 MPa.

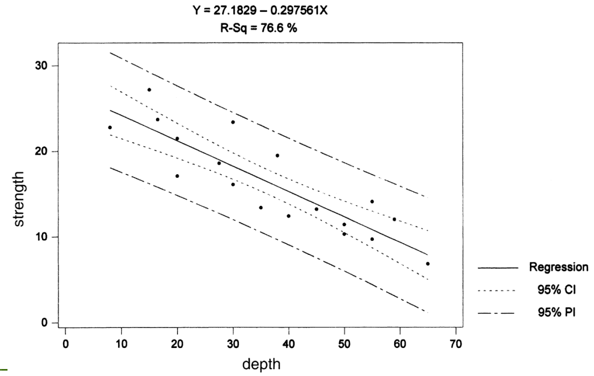
The 95% confidence interval for mean strength when depth is 45 was (12.18, 15.40). The prediction interval is much wider than this because of the extra $(2.8640)^2$ under the square root.

Example -- concrete

cont'd

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4

x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8



Diagnostic Plots

Diagnostic Plots

The basic plots that many statisticians recommend for an assessment of model validity and usefulness are the following:

1. e_i^* (or e_i) on the vertical axis versus x_i on the horizontal axis

2. e_i^* (or e_i) on the vertical axis versus $\hat{y}_i$ on the horizontal axis

3. $\hat{y}_i$ on the vertical axis versus y_i on the horizontal axis

4. A histogram of the standardized residuals

Diagnostic Plots

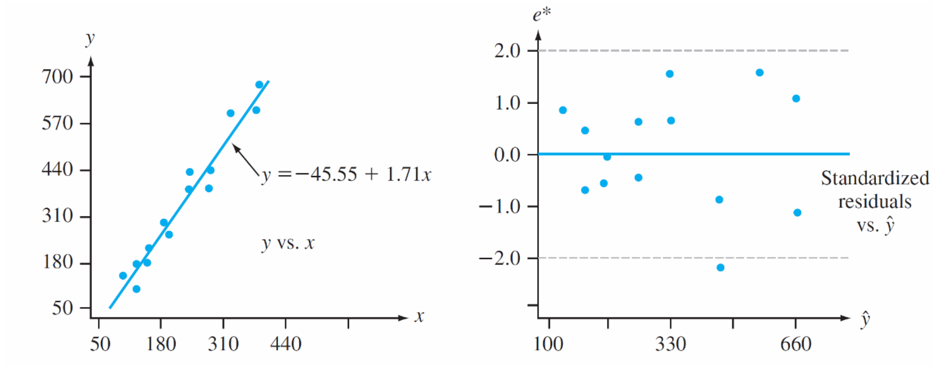
Plots 1 and 2 are called **residual plots** (against the independent variable and fitted values, respectively), whereas Plot 3 is fitted against observed values.

If Plot 3 yields points close to the 45-deg line [slope +1 through (0, 0)], then the estimated regression function gives accurate predictions of the values actually observed.

Provided that the model is correct, neither residual plot should exhibit distinct patterns.

The residuals should be randomly distributed about 0 according to a normal distribution, so all but a very few standardized residuals should lie between –2 and +2 (i.e., all but a few residuals within 2 standard deviations of their expected value 0).

Example



Difficulties and Remedies

Difficulties and Remedies

Although we hope that our analysis will yield plots like these, quite frequently the plots will suggest one or more of the following difficulties:

1. A nonlinear relationship between *x* and *y* is appropriate.

2. The variance of ϵ (and of *Y*) is not a constant σ^2 but depends on *x*.

3. The selected model fits the data well except for a very few discrepant or outlying data values, which may have greatly influenced the choice of the best-fit function.

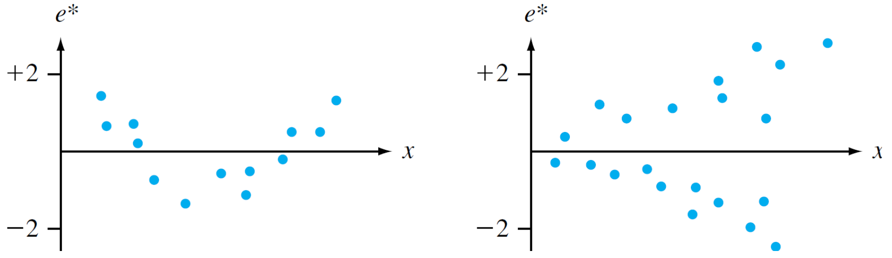
Difficulties and Remedies

4. The error term ϵ does not have a normal distribution.

5. When the subscript *i* indicates the time order of the observations, the ϵ_i's exhibit dependence over time.

6. One or more relevant independent variables have been omitted from the model.

Difficulties and Remedies

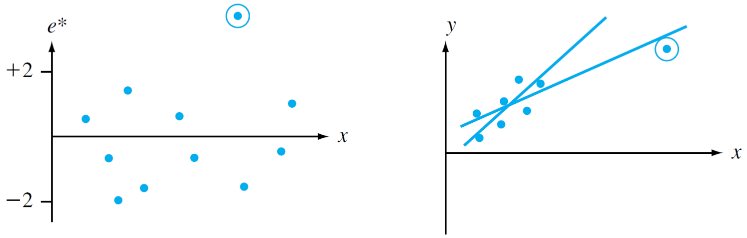


(a)

(b)

Plots that indicate abnormality in data:
(a) nonlinear relationship; (b) nonconstant variance;

Difficulties and Remedies cont'd

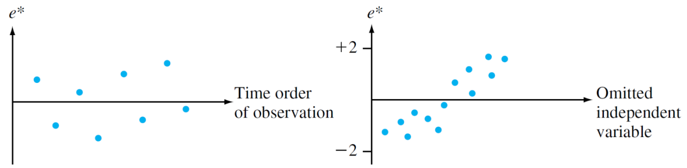


(c)

(d)

Plots that indicate abnormality in data:
(c) discrepant observation; (d) observation with large influence;

Difficulties and Remedies cont'd



(e)

(f)

Plots that indicate abnormality in data:
(e) dependence in errors; (f) variable omitted

Difficulties and Remedies

For a more comprehensive discussion, one or more of the references on regression analysis should be consulted. If the residual plot exhibits a curved pattern, then a nonlinear function of *x* may be used.

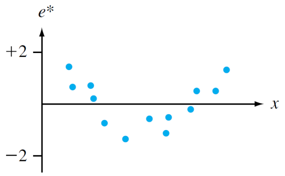
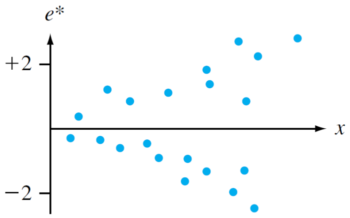


Figure 13.2(a)

Difficulties and Remedies

The residual plot below suggests that, although a straight-line relationship may be reasonable, the assumption that $V(Y_i) = \sigma^2$ for each *i* is of doubtful validity.



Using advanced methods like weighted LS (WLS), as can more advanced models, is recommended for inference.

Difficulties and Remedies

When plots or other evidence suggest that the data set contains outliers or points having large influence on the resulting fit, one possible approach is to omit these outlying points and recompute the estimated regression equation.

This would certainly be correct if it were found that the outliers resulted from errors in recording data values or experimental errors.

If no assignable cause can be found for the outliers, it is still desirable to report the estimated equation both with and without outliers omitted.

Difficulties and Remedies

Yet another approach is to retain possible outliers but to use an estimation principle that puts relatively less weight on outlying values than does the principle of least squares.

One such principle is MAD (minimize absolute deviations), which selects $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize $\Sigma\,|y_i - (b_0 + b_1x_i|$.

Unlike the estimates of least squares, there are no nice formulas for the MAD estimates; their values must be found by using an iterative computational procedure.

Difficulties and Remedies

Such procedures are also used when it is suspected that the ϵ_i's have a distribution that is not normal but instead have “heavy tails” (making it much more likely than for the normal distribution that discrepant values will enter the sample); robust regression procedures are those that produce reliable estimates for a wide variety of underlying error distributions.

Least squares estimators are not robust in the same way that the sample mean $\overline{X}$ is not a robust estimator for μ.

Difficulties and Remedies

When a plot suggests time dependence in the error terms, an appropriate model will include a time variable –

For that -- take the Time Series course taught by Math or the Applied Math department.

So far we have learned about LR…

-how to fit SLR – estimates the true linear relationship between the outcome and a predictor. Interpret slope as the Average Change in Y when X increases by 1 unit

-how to do inference in SLR (is the slope significant?)

-we learned to check for assumption violations (non-constant variance, non-linear patterns in residuals, autocorrelation?)

- we learned to use standardized residuals instead of ordinary residuals

-we learned about outliers (outliers in Y space – those points with large residuals)

-we also learned about the more sneaky kind: the outliers in the X space – they almost always have a small ordinary residual (that residual’s variance is tiny, since their X is far away from the mean of all X’s), so they won’t be visible on the ordinary residual plot (but they may be big on the standardized res plot). They are called leverage points (points with high leverage)

-When the removal of leverage points results in a very different line, then they are called influential points.

And today we will talk about MLR

-how to fit MLR – estimate the true linear relationship (regression surface) between the outcome and a predictor. Interpret slope on X1 as the Average Change in Y when X1 increases by 1 unit, holding all other X’s constant

-Will learn how to do inference for one slope, several slopes (is a subset of slopes significant – does this subset of variables ”matter”?), or all slopes.

- Same approach to checking for assumption violations (non-constant variance, non-linear patterns in residuals)

- Again, must use standardized residuals instead of ordinary residuals

- Must think about outliers (outliers in Y space – those points with large residuals)

-Must think about *any* outliers in the X space – leverage points -- they may have a small ordinary residual.

Multiple Regression Analysis

Definition

The **multiple regression model equation** is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

where $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.

In addition, for purposes of testing hypotheses and calculating CIs or PIs, it is assumed that ϵ is normally distributed.

This is not a regression line any longer, but a regression surface.

Multiple Regression Analysis

Let $x_1^*, x_2^*, \ldots, x_k^*$ be particular values of $x_1,\ldots,x_k$. Then

$$\mu_{Y \cdot x_1^*, \ldots, x_k^*} = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^*$$

Thus just as $\beta_0 + \beta_1 x$ describes the mean Y value as a function of x in simple linear regression, the **true** (or **population**) **regression function** $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ gives the expected value of Y as a function of $x_1,\ldots, x_k$.

The β_i*'s* are the **true** (or **population**) **regression coefficients.**

Multiple Regression Analysis

The regression coefficient β_1 is interpreted as the expected change in Y associated with a 1-unit increase in x_1 *while* $x_2,\ldots, x_k$ *are held fixed*.

Analogous interpretations hold for $\beta_2,\ldots, \beta_k$.

Thus, these coefficients are called partial or adjusted regression coefficients.

In contrast, the simple regression slope is called the marginal (or unadjusted) coefficient.

Models with Interaction and Quadratic Predictors

For example, f an investigator has obtained observations on y, x_1, and x_2, one possible model is $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

However, other models can be constructed by forming predictors that are mathematical functions of x_1 and/or x_2.

For example, with $x_3 = x_1^2$ and $x_4 = x_1 x_2$, the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

Models with Interaction and Quadratic Predictors

In general, it is not only permissible for some predictors to be mathematical functions of others but also often desirable in the sense that the resulting model may be much more successful in explaining variation in *y* than any model without such predictors. This is still “linear regression”, even though the relationship between outcome and predictors may not be.

For example, the model

$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \in$ is still MLR with $k = 2$, $x_1 = x$, and $x_2 = x^2$.

Example -- travel time

The article “Estimating Urban Travel Times: A Comparative Study” (*Trans. Res.*, 1980: 173–175) described a study relating the dependent variable *y* = travel time between locations in a certain city and the independent variable x_2 = distance between locations.

Two types of vehicles, passenger cars and trucks, were used in the study.

Let

$$x_1 = \begin{cases} 1 & \text{if the vehicle is a truck} \\ 0 & \text{if the vehicle is a passenger car} \end{cases}$$

Example – travel time cont’d

One possible multiple regression model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \in$$

The mean value of travel time depends on whether a vehicle is a car or a truck:

mean time = $\beta_0 + \beta_2 x_2$ when $x_1 = 0$ (cars)

mean time = $\beta_0 + \beta_1 + \beta_2 x_2$ when $x_1 = 1$ (trucks)

Example -- travel time cont’d

The coefficient β_1 is the difference in mean times between trucks and cars with distance held fixed; if $\beta_1 > 0$, on average it will take trucks the same amount of time longer to traverse any particular distance than it will for cars.

A second possibility is a model with an interaction predictor:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \in$$

Example – travel time

cont'd

Now the mean times for the two types of vehicles are

mean time = $\beta_0 + \beta_2 x_2$ when $x_1 = 0$

mean time = $\beta_0 + \beta_1 + (\beta_2 + \beta_3) x_2$ when $x_1 = 1$

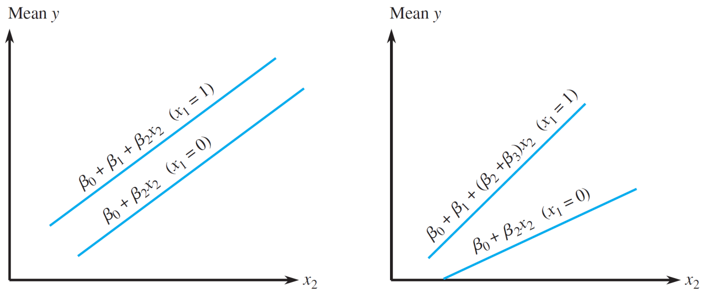
Does it make sense to have different intercepts for cars and truck? (Think about what intercept actually means).

So, what would be a third model you could consider?

Example – travel time

cont'd

For each model, the graph of the mean time versus distance is a straight line for either type of vehicle, as illustrated in Figure 13.14.



(a) no interaction

(b) interaction

Regression functions for models with one dummy variable (x_1) and one quantitative variable x_2

Figure 13.14

Example -- travel

cont'd

The two lines are parallel for the first (no-interaction) model, but in general they will have different slopes when the second model is correct.

For this latter model, the change in mean travel time associated with a 1-mile increase in distance depends on which type of vehicle is involved—the two variables "vehicle type" and "travel time" interact.

Indeed, data collected by the authors of the cited article suggested the presence of interaction.

Example -- travel

cont'd

But, does it make sense to have different intercepts for cars and truck? (Think about what intercept actually means).

So, what would be a third model you could consider?

$$Y = \beta_0 + \beta_2 x_2 + \beta_3 x_1 x_2 + \in$$

Now the mean times for the two types of vehicles are

mean time = $\beta_0 + \beta_2 x_2$ for cars

mean time = $\beta_0 + (\beta_2 + \beta_3) x_2$ for trucks