Introduction to Regression Analysis

Goals

- To review/introduce the calculation and interpretation of the least-squares regression coefficients in simple regression.
- To review/introduce the calculation and interpretation of the regression standard error and the simple correlation coefficients.
- To introduce the standard statistical inference and assumptions for simple linear regression.
- To describe properties of the least-squares coefficients as estimators of the parameters of the regression model.
- To introduce flexible and general procedures for statistical inference based on least-squares estimators.

Regression analysis examines the relationship between a quantitative response (or dependent) variable (denoted by Y) and one or more quantitative explanatory (or independent or predictor) variables, $X_1, ..., X_k$. Regression analysis traces the conditional distribution of Y, or some aspect of this distribution, such as its mean, as a function of the X's.

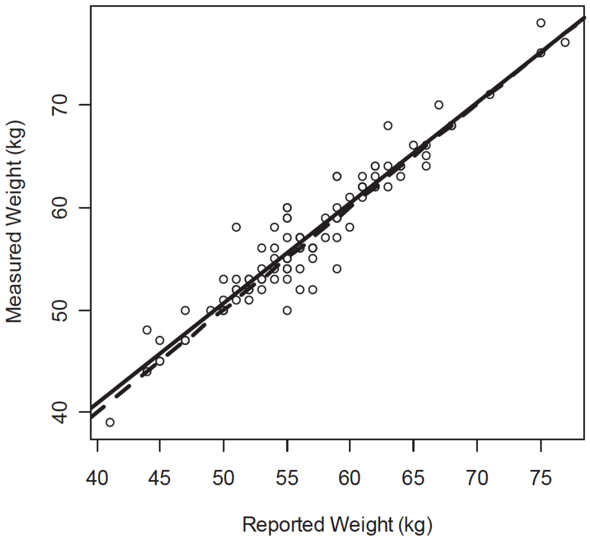
Linear Least Squares Regression

Linear least squares lies at the very heart of applied statistics:

- Often, data are adequately summarized by linear least-squares regression.
- The effective application of linear regression is expanded by data transformations and diagnostics.
- The general linear model — an extension of least-squares linear regression — is able to accommodate a very broad class of specifications.
- Linear least-squares provides a computational basis for a variety of generalizations (such as generalized linear models).

Imagine we have a continuous X, say reported weight in kilograms for a sample of 101 women who engage in regular exercise, and Y is their measured weight in kilograms. We want to use reported weight to predict actual weight.

We see that a straight line adequately summarizes the relationship and that we need only estimate two quantities, the intercept and slope.



The relationship between measured and reported weight appears to be linear, so it is reasonable to fit a line to the plot. Denoting measured weight by Y and reported weight by X, a line relating the two variables has the equation $Y = \beta_0 + \beta_1 X + \epsilon$

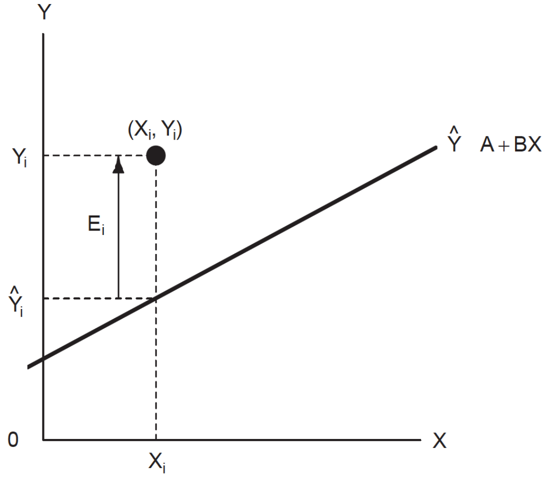
No line can pass perfectly through all of the data points. A residual, ϵ, reflects this fact.

The regression equation for the ith of the $n = 101$ observations is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

The residual

$$\epsilon_i = Y_i - \widehat{Y}_i = Y_i - (\beta_0 + \beta_1 X_i)$$

is the signed vertical distance between the point and the line.



Least-Squares Fit

A line that fits the data well makes the residuals small.

Simply requiring that the sum of residuals, $\sum_{i=1}^{n} \epsilon_i$, be small is futile, since large negative residuals can offset large positive ones.

Indeed, any line through the point $(\overline{X}, \overline{Y})$ has $\sum_{i=1}^{n} \epsilon_i = 0$.

Two possibilities immediately present themselves:

- Find β_0 and β_1 to minimize the absolute residuals, $\sum |\epsilon_i|$, which leads to least-absolute-values (LAV) regression.
- Find β_0 and β_1 to minimize the squared residuals, $\sum \epsilon_i^2$, which leads to least-squares (LS) regression.

Least-Squares Regression

We seek the values of β_0 and β_1 that minimize:

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

Solving the normal equations produces the least-squares coefficients:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

$$\widehat{\beta}_1 = \frac{cov_{XY}}{s_X^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

The formula for β_0 implies that the least-squares line passes through the point-of-means of the two variables. The least-squares residuals therefore sum to zero.

The formula for β_1 implies that $\sum X_i \epsilon_i = 0$; similarly, $\sum \widehat{Y}_i \epsilon_i = 0$. These properties imply that the residuals are uncorrelated with both the X's and the Y's.

For the data on measured weight (Y) and reported weight (X):

$$n = 101$$

$$\overline{Y} = \frac{5780}{101} = 57.23$$

$$\overline{X} = \frac{5731}{101} = 56.74$$

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = 4435$$

$$\sum (X_i - \overline{X})^2 = 4539$$

$$\beta_1 = \frac{4435}{4539} = 0.977$$

$$\beta_0 = 57.23 - 0.9771 \times 56.74 = 1.789$$

The least-squares regression equation is

$$\widehat{\text{Measured Weight}} = 1.79 + 0.977 \times \text{Reported Weight}$$

Interpretation of the least-squares coefficients:

$\beta_1 = 0.977$: A 1 kg increase in reported weight is associated on average with just under a 1 kg increase in measured weight.

Since the data are not longitudinal, the phrase "a unit increase" here does not imply a literal change over time, but rather a static comparison between two individuals who differ by 1 kg in their reported weights.

Ordinarily, we may interpret the intercept as the fitted value associated with $X = 0$, but it is impossible for an individual to have a reported weight equal to zero.

The intercept is often of little direct interest, because the fitted value for $X = 0$ is rarely important.

Here, however, if individuals' reports are unbiased predictions of their actual weights, then we should have $\widehat{Y} = X$ and $\beta_0 = 0$. The intercept, $\beta_0 = 1.79$ is close to zero, and the slope, $\beta_1 = 0.977$ is close to one.

Correlation

It is of interest to determine how closely the line fits the scatter of points.

The standard deviation of the residuals, S_ϵ, called the standard error of the regression, provides one index of fit.

Because of estimation considerations, the variance of the residuals is defined using $n-2$ degrees of freedom:

$$S_\epsilon^2 = \frac{\sum \epsilon_i^2}{n-2}$$

The standard error is therefore

$$S_\epsilon = \sqrt{\frac{\sum \epsilon_i^2}{n-2}}$$

Because it is measured in the units of the response variable, the standard error represents a type of 'average' residual.

For the regression of measured on reported weight, the sum of squared residuals is $\sum \epsilon_i^2 = 418.9$, and the standard error is

$$S_\epsilon = \sqrt{\frac{418.9}{101-2}} = 2.05 \text{ kg}$$

The correlation coefficient provides a relative measure of fit: To what degree do our predictions of Y improve when we base that prediction on the linear relationship between Y and X?

A relative index of fit requires a baseline — how well can Y be predicted if X is disregarded?

To disregard the explanatory variable is implicitly to fit the equation

$$Y_i = \beta_0' + \epsilon_i'$$

We can find the best-fitting constant β_0' by least-squares, minimizing

$$\sum \epsilon_i'^2 = \sum (Y_i - \beta_0')^2$$

The value of β_0' that minimizes this sum of squares is the response-variable mean, $\overline{Y}$.

The residuals, $\epsilon_i = Y_i - \widehat{Y}_i$ from the linear regression of Y on X will generally be smaller than the residuals, $\epsilon'_i = Y_i - \overline{Y}$, and it is necessarily the case that

$$\sum(Y_i - \widehat{Y}_i)^2 \leq \sum(Y_i - \overline{Y})^2$$

This inequality holds because the 'null model,' $Y_i = \beta'_0 + \epsilon'_i$ is a special case of the more general linear-regression model, $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$, setting $\beta_1 = 0$.

We call

$$\sum \epsilon'^2_i = \sum(Y_i - \overline{Y})^2$$

the total sum of squares (TSS) for Y, and

$$\sum \epsilon_i^2 = \sum(Y_i - \widehat{Y}_i)^2$$

is called the residual sum of squares (RSS).

The difference between the two, termed the regression sum of squares, $RegSS \equiv TSS - RSS$ gives the reduction in squared error due to the linear regression.

The ratio of $RegSS$ to TSS, the proportional reduction in squared error, defines the square of the correlation coefficient:

$$r^2 \equiv \frac{RegSS}{TSS}$$

To find the correlation coefficient, r, we take the positive square root of r^2 when the simple-regression slope β_1 is positive, and the negative square root when β_1 is negative.

If there is a perfect positive linear relationship between Y and X, then $r = 1$.

A perfect negative linear relationship corresponds to $r = -1$.

If there is no linear relationship between Y and X, then $RSS = TSS$, $RegSS = 0$, and $r = 0$.

Between these extremes, r gives the direction of the linear relationship between the two variables, and r^2 may be interpreted as the proportion of the total variation of Y that is 'captured' by its linear regression on X.

Decomposition of Variance

The decomposition of total variation into 'explained' and 'unexplained' components, paralleling the decomposition of each observation into a fitted value and a residual, is typical of linear models.

The decomposition is called the *analysis of variance* for the regression:

$$TSS = RegSS + RSS$$

The regression sum of squares can also be directly calculated as

$$RegSS = \sum(\widehat{Y}_i - \overline{Y})^2$$

It is also possible to obtain the sample correlation directly from the sample variances and covariances.

Define the sample covariance between X and Y,

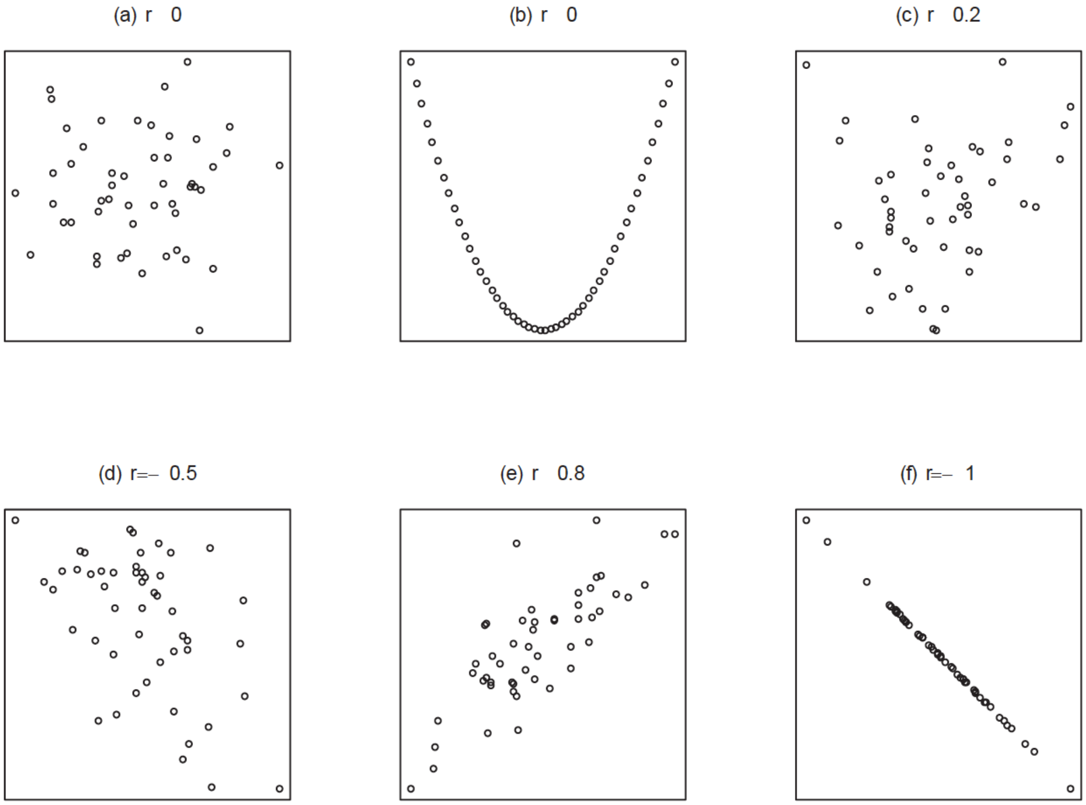


Figure 1: Figure 3 depicts several different levels of correlation.

$$S_{XY} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

then

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} = \frac{S_{XY}}{S_X S_Y}$$

where S_X and S_Y are, respectively, the sample standard deviations of X and Y.

Comparisons between r and β_1:

The correlation coefficient r is symmetric in X and Y, but the least-squares slope β_1 is not.

The slope coefficient β_1 is measured in the units of the response variable per unit of the explanatory variable. For example, if dollars of income are regressed on years of education, then the units of β_1 are dollars/year. The correlation coefficient r, however, is unitless.

A change in scale of Y or X produces a compensating change in β_1, but does not affect r. If, for example, income is measured in thousands of dollars rather than in dollars, the units of the slope become $1000s/year, and the value of the slope decreases by a factor of 1000, but r remains the same.

For the regression of measured on reported weight,

$$TSS = 4753.8$$

$$RSS = 418.87$$

$$RegSS = 4334.9$$

$$r^2 = \frac{4334.9}{4753.8} = .91188$$

Since β_1 is positive, $r = +\sqrt{.91188} = .9549$.

The linear regression of measured on reported weight captures 91 percent of the variation in measured weight.

Equivalently,

$$S_{XY} = \frac{4435.9}{101-1} = 44.359$$

$$S_X^2 = \frac{4539.3}{101-1} = 45.393$$

$$S_Y^2 = \frac{4753.8}{101-1} = 47.538$$

$$r = \frac{44.359}{\sqrt{45.393 \times 47.538}} = .9549$$

Gauss-Markov Theorem

Standard statistical inference in simple regression is based on a statistical model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The coefficients β_0 and β_1 are the population regression parameters to be estimated.

The error ϵ_i represents the aggregated, omitted causes of Y:

- Other explanatory variables that could have been but were not included.
- Measurement error in Y.

The key assumptions of the simple regression model concern the behavior of the errors — or, equivalently, of the distribution of Y conditional on X:

Linearity: $E(\epsilon_i) \equiv E(\epsilon|x_i) = 0$

Equivalently, the expected value of Y is a linear function of X:

$$\mu_i \equiv E(Y_i) \equiv E(Y|x_i) = E(\beta_0 + \beta_1 x_{i1} + \epsilon_i) = \beta_0 + \beta_1 x_i + E(\epsilon_i) = \beta_0 + \beta_1 x_i$$

Constant Variance: $V(\epsilon|x_i) = \sigma_\epsilon^2$

Equivalently, the variance of Y around the regression line is constant.

Normality: $\epsilon_i \sim N(0, \sigma_\epsilon^2)$

Equivalently, the conditional distribution of Y given x is normal: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$.

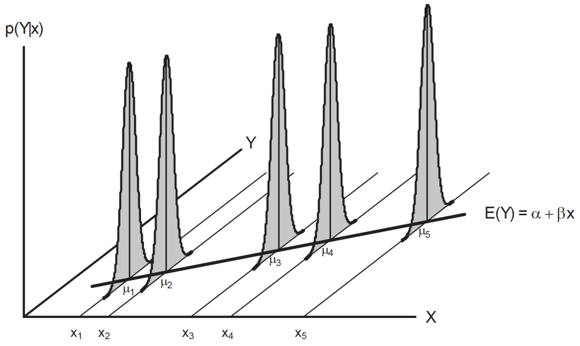


Figure 2: The assumptions of linearity, normality, and constant variance in the simple regression model.

Independence: The observations are sampled independently: Any pair of errors ϵ_i and ϵ_j (or, equivalently, of conditional response-variable values, Y_i and Y_j) are independent for $i \neq j$. The assumption of independence needs to be justified by the procedures of data collection.

Fixed X or X is independent of the error: Depending upon the design of a study, the values of the explanatory variable may be fixed in advance of data collection or they may be sampled along with the response variable.

Fixed X corresponds almost exclusively to experimental research.

When, as is more common, X is sampled along with Y, we assume that the explanatory variable and the error are independent in the population from which the sample is drawn: That is, the error has the same distribution $N(0, \sigma_\epsilon^2)$ for every value of X in the population.

Properties of the Least-Squares Estimator

Under the strong assumptions of the simple regression model, the sample least-squares coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have several desirable properties as estimators of the population regression coefficients β_0 and β_1:

The least-squares intercept and slope are linear estimators, in the sense that they are linear functions of the observations Y_i. This result is not important in itself, but it makes the distributions of the least-squares coefficients simple.

Under the assumption of linearity, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are unbiased estimators of β_0 and β_1:

$$E(\widehat{\beta}_0) = \beta_0 E(\widehat{\beta}_1) = \beta_1$$

An estimate is unbiased if over many repeated samples drawn from the population, the average value of the estimates in the samples equals the population value of the parameter being estimated.

Under the assumptions of linearity, constant variance, and independence, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ have sampling variances:

$$V(\widehat{\beta}_0) = \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \overline{x})^2} V(\widehat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum (x_i - \overline{x})^2}$$

$V(\widehat{\beta}_1)$ can be rewritten as:

$$V(\widehat{\beta}_1) = \frac{\sigma_\epsilon^2}{(n-1) S_X^2}$$

The Gauss-Markov theorem: Of all linear unbiased estimators, the least-squares estimators are most efficient. Under normality, the least-squares estimators are most efficient among *all* unbiased estimators, not just among linear estimators. This is a much more compelling result.

Under the full set of assumptions, the least-squares coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the maximum-likelihood estimators of β_0 and β_1.

Under the assumption of normality, the least-squares coefficients are themselves normally distributed:

$$\widehat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \overline{x})^2}\right) \widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \overline{x})^2}\right)$$

Even if the errors are not normally distributed, the distributions of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are approximately normal, with the approximation improving as the sample size grows (the central limit theorem).

Statistical Inference for Simple Linear Regression

```
library(car)
```

Confidence Intervals and Hypothesis Tests

The distributions of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ cannot be directly employed for statistical inference since σ_ϵ^2 is never known in practice.

The variance of the residuals provides an unbiased estimator of σ_ϵ^2

$$\widehat{\sigma}_\epsilon^2 = \frac{\sum \widehat{\epsilon}_i^2}{(n-2)}$$

and a basis for estimating the variances of $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\widehat{V(\beta_0)} = \frac{\widehat{\sigma}_\epsilon^2 \sum x_i^2}{n \sum (x_i - \overline{x})^2} \widehat{V(\beta_1)} = \frac{\widehat{\sigma}_\epsilon^2}{\sum (x_i - \overline{x})^2}$$

The added uncertainty induced by estimating the error variance is reflected in the use of the t-distribution, in place of the normal distribution, for confidence intervals and hypothesis tests.

To construct a $100(1-\alpha)\%$ confidence interval for the slope, we take

$$\beta = \widehat{\beta} \pm t_{\alpha/2} \widehat{SE(\beta)}$$

where $t_{\alpha/2}$ is the critical value of t with $n-2$ degrees of freedom and a probability of $\alpha/2$ to the right, and $\widehat{SE(\beta)}$ is the square root of $\widehat{V(\beta)}$. (This is just like a confidence interval for a population mean.)

Similarly, to test the hypothesis $H_0 : \beta = 0$, calculate the test statistic

$$t = \frac{\widehat{\beta}}{\widehat{SE(\beta)}}$$

which is distributed as t with $n-2$ degrees of freedom under H_0.

For Davis's regression of measured on reported weight, for example:

$$\widehat{\sigma}_\epsilon = \sqrt{\frac{418.87}{(101-2)}} = 2.0569 \widehat{SE(\beta_0)} = \frac{2.0569 \times \sqrt{329,731}}{\sqrt{101 \times 4539.3}} = 1.7444 \widehat{SE(\beta_1)} = \frac{2.0569}{\sqrt{4539.3}} = 0.030529$$

```
cDavis <- Davis
cDavis[12, c(2, 3)] <- Davis[12, c(3, 2)]  # correct the recording error
mod <- lm(weight ~ repwt, subset = sex=="F", data=cDavis)
summary(mod)
```

```
##
## Call:
## lm(formula = weight ~ repwt, data = cDavis, subset = sex == "F")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5248 -0.7526 -0.3654  0.6118  6.3841
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.77750    1.74441   1.019    0.311
## repwt        0.97722    0.03053  32.009   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.057 on 99 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.9119, Adjusted R-squared:  0.911
## F-statistic:  1025 on 1 and 99 DF,  p-value: < 2.2e-16
```

Since $t_{.025}$ for $101-2=99$ degrees of freedom is 1.984.

```r
qt(.975, 99)
```

```
## [1] 1.984217
```

95% CIs for β_0 and β_1 are:

$$\beta_0 = 1.778ś1.984 \times 1.744 = 1.778ś3.460\beta_1 = 0.9772ś1.984 \times 0.03053 = 0.9772ś0.06057$$

You can obtain the confidence intervals using `confint()`

```r
confint(mod)
```

```
##                  2.5 %   97.5 %
## (Intercept) -1.6837802 5.238787
## repwt        0.9166458 1.037803
```