13 Linear regression and least squares

Numbers are like people; torture them enough and they'll tell you anything.

13.1 Regression models

One of the most widely used examples of estimation in statistics is provided by linear regression. For example, if Y_i is the number of unemployed in UK in the ith month after some date, we might make the hypothesis that

$$Y_i = a + \beta i + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_1, \dots, \epsilon_n$ are IID $N(0, \sigma^2)$ and a, β are some unknown constants. The business of estimating β is to do with detecting a trend in unemployment. A related problem is that of testing $H_0 : \beta = 0$, a test of whether there is any trend in unemployment.

The model above is a special case of the **simple linear regression model** in which, with the same assumptions on $\{\epsilon_i\}$, a, β, σ^2,

$$Y_i = a + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where the x_i are known constants. In the case above $x_i = i$.

A **multiple regression** model has more that one explanatory variable on the right hand side, e.g.,

$$Y_i = a + \beta_1 \log i + \beta_2 z_{i-5} + \epsilon_i, \quad i = 1, \dots, n,$$

where perhaps z_{i-5} is the number of unemployed people who were in training programmes five months earlier. The estimation of a, β_1 and β_2 and associated tests are similar to what we find for simple linear regression.

13.2 Least squares/MLE

Suppose $Y_1, \dots, Y_n$ are independent and $Y_i = \alpha + \beta w_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, or equivalently that $Y_i \sim N(\alpha + \beta w_i, \sigma^2)$, and where α, β and σ^2 are unknown parameters and the w_i are known constants such that $\sum_i w_i = 0$.

Theorem 13.1 *The MLEs of α and β are obtained by minimizing*

$$S = \sum_{i=1}^{n} \big(Y_i - \mathbb{E}(Y_i)\big)^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta w_i)^2$$

with respect to α and β. These are called the **least squares estimators** *and are given by:*

$$\boxed{\hat\alpha = \bar Y \qquad \textit{and} \qquad \hat\beta = S_{wY}/S_{ww}}$$

where $S_{ww} = \sum_i w_i^2$, and $S_{wY} = \sum_i w_i Y_i$.

Proof. Since $Y_i \sim N(\alpha + \beta w_i, \sigma^2)$ the likelihood of of $y_1, \dots, y_n$ is

$$f_Y(y \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta w_i)^2 \right) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-S/2\sigma^2}.$$

The maximum likelihood estimator minimizes S, and so at a minimum,

$$\left.\frac{\partial S}{\partial \alpha}\right|_{\substack{\alpha=\hat\alpha\\ \beta=\hat\beta}} = -2\sum_{i=1}^{n} (y_i - \hat\alpha - \hat\beta w_i) = 0\ , \quad \left.\frac{\partial S}{\partial \beta}\right|_{\substack{\alpha=\hat\alpha\\ \beta=\hat\beta}} = -2\sum_{i=1}^{n} w_i (y_i - \hat\alpha - \hat\beta w_i) = 0.$$

Hence

$$\sum_{i=1}^{n} Y_i - n\hat\alpha = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} w_i Y_i - \hat\beta \sum_{i=1}^{n} w_i^2 = 0,$$

from which the answers follow. ■

13.3 Practical usage

Given a linear regression model

$$Y_i = a + \beta x_i + \epsilon_i,$$

in which $\sum_i x_i \neq 0$ we make the transformation $w_i = x_i - \bar x$ and consider

$$Y_i = \alpha + \beta w_i + \epsilon_i,$$

where $\bar x = \sum_{i=1}^n x_i / n$ and $\alpha = a + \beta \bar x$. This gives the situation described in 13.1, and we can use results of 13.2 to estimate the regression and the results in 14.1 to perform tests. Making the necessary transformations we have

$$\boxed{\hat a = \bar Y - \hat\beta \bar x \qquad \text{and} \qquad \hat\beta = S_{xY}/S_{xx}}$$

where $S_{xx} = \mathbf{w}^\top \mathbf{w} = \sum_i (x_i - \bar x)^2$ and $S_{xY} = \sum_i (x_i - \bar x)(Y_i - \bar Y)$.

We speak of 'regressing y on x'. A package such as MINITAB will return the estimated regression line in the form

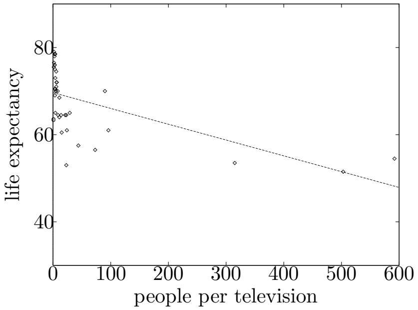
$$y = \hat a + \hat\beta x.$$

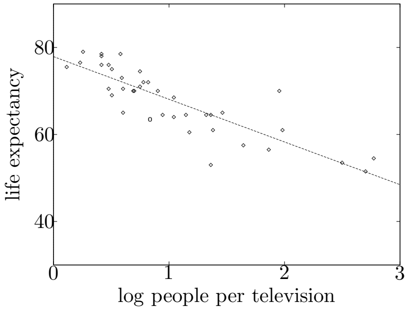
Note that the point $(\bar x, \bar Y)$ always lies on the regression line, i.e., $\bar Y = \hat a + \hat\beta \bar x$.

Example 13.2 *The following data for 40 nations has been extracted from a 1993 almanac.*

The correlations of life expectancy with people/television and people/doctor are -0.606 and -0.666 respectively. Scatter plots suggests that a better fit might be obtained by a regression of life expectancy on either the logarithm of people/television or logarithm of people/doctor. When this is done the correlations are respectively -0.855 and -0.832.

country	mean life expectancy, y	people per television, u	people per doctor, v
Argentina	70.5	4.0	370
Bangladesh	53.5	315.0	6166
Brazil	65.0	4.0	684
⋮			⋮
United Kingdom	76.0	3.0	611
United States	75.5	1.3	404
Venezuela	74.5	5.6	576
Vietnam	65.0	29.0	3096
Zaire	54.0	*	23193





Let $x_i = \log_{10} u_i$ and consider fitting a regression of y against x. There is data for 38 countries (as television data for Zaire and Tanzania is missing). We compute the following summary statistics

$$\bar{y} = 67.76, \quad \bar{x} = 1.0322, \quad S_{yy} = 2252.37, \quad S_{xx} = 17.120, \quad S_{xy} = -167.917.$$

These give

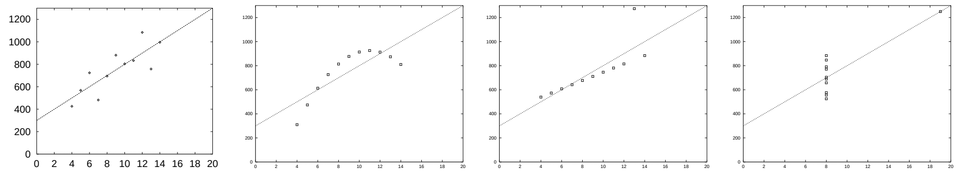
$$\hat{\beta} = S_{xy}/S_{xx} = -9.808, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 77.887, \quad r = S_{xy}/(S_{xx}S_{yy})^{\frac{1}{2}} = -0.855.$$

Although people/television appears to be a useful predictor of a country's life expectancy, we don't really expect that sending shiploads of televisions to countries with short life expectancies would cause their people to live longer. This points up the obvious, but sometimes forgotten fact, that there may be correlation between two variables without causation.

13.4 Data sets with the same summary statistics

Tha application of regression analysis requires care. The following data sets have nearly the same value of the sufficient statistics $\bar{x}, \bar{y}, S_{xx}$ and S_{xy}. The regression line is about $y = 300 + 50\,x$ in each case. However, a simple linear regression is only appropriate in the first case. In the second case a quadratic would be more

appropriate. The third case is affected by the presence of an **outlier** and the fourth case is really no more than a straight line fit through 2 points. The lesson is: plot the data!



10	804	8	695	13	758	9	881	11	833	14	996	6	724	4	426	12	1084	7	482	5	568
10	914	8	814	13	874	9	877	11	926	14	810	6	613	4	310	12	913	7	726	5	474
10	746	8	677	13	1274	9	711	11	781	14	884	6	608	4	539	12	815	7	642	5	573
8	658	8	576	8	771	8	884	8	847	8	704	8	525	19	1250	8	556	8	791	8	689

13.5 Other aspects of least squares

Least squares can be used to fit other models. For example, to fit a **regression through the origin** we would minimize

$$S = \sum_i (Y_i - \beta x_i)^2$$

and get $\hat{\beta} = \sum_i x_i Y_i / \sum_i x_i^2$. To fit a multiple regrssion model, such as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i\,,$$

in which y is predicted from p variables, $x_1, \ldots, x_p$, we would minimize

$$\sum_i (Y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2$$

with respect to $\beta_0, \ldots, \beta_p$. Stationarity conditions give $p+1$ simultaneous linear equations in $\hat{\beta}_0, \ldots, \hat{\beta}_p$, which can be solved for these estimators.

Least squares estimators have many nice properties, one of which is that they are **best linear unbiased estimators**. Consider simple linear regression, in which $Y_i = a + \beta x_i + \epsilon_i$, where the ϵ_i are independent with common mean 0 and variance σ^2 (but are now not necessarily normal RVs). Suppose we want to estimate β by a linear function of the observations, i.e., $\hat{\beta} = \sum_i c_i Y_i$. If this estimator is to be unbiased then we need

$$\mathbb{E}\hat{\beta} = \sum_i c_i(a + \beta x_i) = \beta, \quad \text{for all } a, \beta.$$

Hence we need $\sum_i c_i = 0$ and $\sum_i c_i x_i = 1$. Now the variance of the estimator is

$$\mathrm{var}(\hat{\beta}) = \mathrm{var}\left(\sum_i c_i Y_i\right) = \sigma^2 \sum_i c_i^2.$$

So we have the constrained optimization problem:

$$\text{minimize } \sum_i c_i^2 \text{ subject to } \sum_i c_i = 0 \text{ and } \sum_i c_i x_i = 1.$$

Using Lagrangian methods it is easy to find the solution: $c_i = (x_i - \bar{x})/S_{xx}$. This gives the usual LSE $\hat{\beta} = S_{xy}/S_{xx}$. A similar analysis shows that the best linear unbiased estimator of a is also the usual LSE of a, i.e., $\hat{a} = \bar{Y} - \hat{\beta}\bar{x}$.