Linear Regression and Calibration

The Sum of Squares Function

Ordinary Least Squares

Definition of the Sum of Squares Function

Start with a set of replicate values x_i and make a guess for the mean μ of the distribution: *a*.

We can now compute the deviations (residual) $\delta_i = x_i - a$.

We take the *squares* and add them up: This produces the *sum of squares*

$$SS = \sum_i \delta_i^2 = \sum_i (x_i - a)^2$$

If our guess is poor then SS will be large. A good guess will give a small value of SS. By minimizing the SS function we will find the **least squares estimate** (LSE)for the average a_{LSE}. We can easily find the LSE value for *a* by setting the derivative d(SS)/d*a* =0

We find:

$$\frac{dSS}{da} = \frac{d\sum_i (x_i - a)^2}{da} = -2\sum_i (x_i - a) = 0$$

Definition of the mean

We can divide both sides by 2 to give:

$$-\sum_i x_i + \sum_i a = 0$$

$$\sum_i x_i = na$$

$$a_{LSE} = \bar{x} = \frac{\sum_i x_i}{n}$$

In other words the sample average (or mean) indeed minimizes the sum of squares. The median by contrast does not have this nice property.

Ordinary Least Squares

Linear data are no longer pure replicates, because we vary the value of x.
For linear data we guess the slope *b* and intercept *a*, calculate deviations and SS. To minimize SS we must now take **two** derivatives (dSS/d*a* and dSS/d*b*) and put them zero simultaneously. Matrix notation is a great help when dealing with this kind of problem. We can write the above model as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Or:

$$\boldsymbol{y} = \boldsymbol{a} + \boldsymbol{x} \cdot \boldsymbol{b} + \boldsymbol{\varepsilon}$$

Ordinary Least Squares

The **X** matrix records for what values of x we choose to take a measurement. We generally assume that there is no error in these *set points* or *independent* variables. **Y** contains the *dependent* variable, the *measured* values. The matrix ε contains the random errors that we assume to be a normal distribution. The matrix β contains the parameters we wish to estimate, the slope ***b* and intercept *a* of our line.**
Finding the LSE for β can be done quite elegantly in matrix notation.

$$Y = X \cdot \beta$$

Ordinary Least Squares

Notice that the only unknowns left are in β. The **X** and **Y** matrices are known because they are either ***set* or *measured.* Solving for** β now requires some simple matrix algebra:

$$X^T Y = X^T X \cdot \beta$$

$$(X^T X)^{-1} X^T Y = \beta_{LSE}$$

The **regression** formula minimizes the sum of squares for a great many different models: point, line, circle, parabola or polynomial. It is one of the most powerful equations in statistics. Let's first look at a simple straight line.

To construct the X matrix we take the derivative with respect to x of both of the variables in the equation for a line.

$$y = \frac{\partial}{\partial a} a + \frac{\partial}{\partial b} x \cdot b$$

Simple linear least squares

Suppose there are *n* data points $\{(x_i, y_i), i = 1, ..., n\}$. The function that describes *x* and *y* is:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

The goal is to find the equation of the straight line $y = \alpha + \beta x$

which would provide a "best" fit for the data points. In the linear least squares approach, α (the *y*-intercept) and β (the slope) solve the following minimization problem:

$$Find \min Q(\alpha, \beta)$$

which can be done using the least squares criterion by minimizing

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Slope, intercept as least squares parameters

By expanding the quadratic expression in *α* and *β*, and taking derivatives with respect to *α* and *β* in order to minimize the objective function *Q* we find:

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

After some algebra

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{Cov[x,y]}{Var[x,y]}$$

$$\hat{\beta} = r_{xy}\frac{s_y}{s_x}$$

$$\hat{\alpha} = y - \hat{\beta}x$$

The difference between the model and the data

The calculated values using the regression model are called f

$$f = \hat{\alpha} + \hat{\beta}x$$

The mean of the observed data is defined as:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The criterion of a good fit will be to compare the difference between the actual data and the mean to the calculated model and the mean.

Sums of squares formulas

The variability of the data set can be measured using three sum of squares (SS) formulas:

1. The total sum of squares (proportional to the variance of the data):

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

2. The regression sum of squares:

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

3. The sum of squares of residuals:

$$SS_{res} = \sum_{i=1}^{n} (f_i - y_i)^2 = \sum_{i=1}^{n} {e_i}^2$$

Correlation coefficient

The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

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

$$SS_{res} = \sum_{i=1}^{n} (f_i - y_i)^2$$