Regression Models

A straight-line relationship of a response variable, y, to an explanatory variable x can be written as

$$y \;=\; \beta_0 \;+\; \beta_1 x \;+\; \epsilon$$

ϵ is the "residual", or "error" — the amount by which a particular data point departs from the straight line.

We may have many explanatory variables, $x_1, \ldots, x_k$. We can then try to explain the response by a "multiple regression" model:

$$y \;=\; \beta_0 \;+\; \beta_1 x_1 \;+\; \cdots \;+\; \beta_k x_k \;+\; \epsilon$$

Example: How does the yield of a wheat crop relate to the amount of fertilizer, the amount of rain, the average temperature, and which of two varieties were planted? The variety is coded as a numerical variable (eg, as 0 or 1).

Statistical Inference for Regression

The population regression equation describes the true relationship in the population. We won't ever know the true regression coefficients, β_0, β_1, ..., β_k, exactly.

We will just have estimates, b_0, b_1, ..., b_k, from our sample. If we want to know how good these are, we can find confidence intervals for them.

We may also want to perform a hypothesis test, such as:

$$H_0: \quad \beta_3 = 0$$
$$H_a: \quad \beta_3 \neq 0$$

For example: Does temperature affect yield (and if so, which way)?

Least Squares Estimates

We will estimate the regression coefficients (β_j) by *least squares*. The estimates (b_j) are chosen to minimize the total squared error

$$E \;=\; \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_{i,1} + \cdots + b_k x_{i,k})]^2$$

Here, $x_{i,j}$ is the value of variable x_j for unit i.

We do this by solving a set of linear equations that equate the the derivatives of E to zero. For instance:

$$\frac{\partial E}{\partial b_1} = \sum_{i=1}^{n} -2x_{i,1}[y_i - (b_0 + b_1 x_{i,1} + \cdots + b_k x_{i,k})] = 0$$

There are $k+1$ equations with $k+1$ unknowns. So the solution for b_0, b_1, ..., b_k is typically unique. (When won't it be?)

It turns out that the b_j are *linear* functions of the observed responses (y_i).

Sampling Distribution of Coefficients

To find confidence intervals and do hypothesis tests, we must find the sampling distribution of the estimated regression coefficients (the b_j).

Since we are modeling only how y relates to the x_j, only the y values are considered to be random.

We will assume that the distribution of the *residuals* in this relationship is $N(0, \sigma_\epsilon^2)$. (Note, we *don't* need to assume that the y_i and $x_{i,j}$ values are normally distributed.)

We also assume the residuals for different cases are *independent*

It then follows that the distribution of each b_j is also normal. The mean is β_j. The standard deviation (also called the standard error) is proportional to the std.dev. of the residuals, σ.

T Tests for Regression

We will seldom know the standard deviation of the residuals. Instead, we will have to estimate it from the actual residuals found with the estimated b_j. We use the estimate

$$s \;=\; \sqrt{\frac{\sum_i e_i^2}{n-k-1}}$$

where $e_i = y_i - (b_0 + b_1 x_{i1} + \cdots + b_k x_{i,k})$.

Why divide by $n-k-1$ rather than n? One reason: it makes s^2 an unbiased estimate of σ^2.

We can now form a t statistic

$$t \;=\; b_j / SE_{b_j}$$

where SE_{b_j} is the standard error for b_j, which will be s times a function of the $x_{i,j}$.

If $\beta_j = 0$, this statistic has a t distribution with $n-k-1$ df. We can use it to test $H_0 : \beta_j = 0$.

A Simulated Example

```
 ROW        y    f        r        t    v

   1  16.9306    0  20.7990  38.6096    0
   2  19.3094    0  29.2036  35.6736    0
   3  22.6540    0  26.0196  32.1261    1
   4  21.7794    0  26.8248  35.2651    1
   5  19.3538    1  27.2788  33.8707    0
   6  23.1051    1  28.7106  30.4276    0
   7  18.1631    1  20.3252  39.3550    1
   8  19.2454    1  20.0066  34.7420    1
   9  21.6882    2  27.5688  30.7554    0
  10  18.4430    2  24.9888  36.5555    0
  11  20.4656    2  26.6816  38.4984    1
  12  19.6138    2  21.2772  31.5124    1

MTB > regress 'y' 4 'f' 'r' 't' 'v'

The regression equation is
y = 22.2 - 0.138 f + 0.353 r - 0.337 t + 1.85 v

Predictor       Coef       Stdev    t-ratio        p
Constant      22.180       4.767       4.65    0.000
f            -0.1384      0.3136      -0.44    0.672
r            0.35297     0.09238       3.82    0.007
t           -0.33699     0.09258      -3.64    0.008
v             1.8546      0.5592       3.32    0.013

s = 0.8674      R-sq = 86.7%     R-sq(adj) = 79.2%
```

The real relationship was:

$$y \;=\; 25 + 0.3f + 0.3r - 0.4t + 2v + \epsilon$$

with $\sigma_\epsilon = 0.7$.