14 Hypothesis tests in regression models

Statisticians do it with a little deviance.

14.1 Distributions of the least squares estimators

Suppose as before that $Y_1, \dots, Y_n$ are independent and $Y_i = \alpha + \beta w_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, equivalently $Y_i \sim N(\alpha + \beta w_i, \sigma^2)$, 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 14.1

(i) $\hat\alpha = \bar Y$ *is distributed as* $N(\alpha, \sigma^2/n)$*;*

(ii) $\hat\beta$ *is distributed as* $N\big(\beta, (\mathbf{w}^\top\mathbf{w})^{-1}\sigma^2\big)$ *independently of* $\hat\alpha$*;*

(iii) the **residual sum of squares** R*, the minimised value of* S*, is distributed as* $\sigma^2\chi^2_{n-2}$ *independently of* $\hat\alpha$ *and* $\hat\beta$*, and is equal to*

$$R = \sum Y_j^2 - n\bar Y^2 - (\mathbf{w}^\top\mathbf{w})\hat\beta^2;$$

(iv) $\hat\sigma^2 = R/(n-2)$ *is an unbiased estimator of* σ^2*.*

Proof. Let

$$A = \begin{pmatrix} 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ (\mathbf{w}^\top\mathbf{w})^{-1/2}w_1 & \cdots & (\mathbf{w}^\top\mathbf{w})^{-1/2}w_n \\ \vdots & & \vdots \\ \cdot & \cdots & \cdot \end{pmatrix}$$

be an orthogonal matrix by appropriate choice of rows $3, \dots, n$. Then $Z_1, \dots, Z_n$ are independent with $\mathbf{Z} = A\mathbf{Y} \sim N\big(A(\alpha\mathbf{1} + \beta\mathbf{w}), \sigma^2\mathbf{I}\big)$, and

$$\begin{array}{ccccc} Z_1 & = & \sqrt{n}\hat\alpha & \sim & N\big(\sqrt{n}\alpha, \sigma^2\big) \\ Z_2 & = & (\mathbf{w}^\top\mathbf{w})^{1/2}\hat\beta & \sim & N\big((\mathbf{w}^\top\mathbf{w})^{1/2}\beta, \sigma^2\big) \\ Z_3 & = & \cdot & \sim & N\big(0, \sigma^2\big) \\ \vdots & & & & \vdots \\ Z_n & = & \cdot & \sim & N\big(0, \sigma^2\big) \end{array}$$

from which all the statements in (i) and (ii) follow.

(iii) and (iv) follow from

$$\sum_{i=1}^n Z_i^2 = n\bar Y^2 + (\mathbf{w}^\top\mathbf{w})\hat\beta^2 + \sum_{i=3}^n Z_i^2$$

and

$$\begin{aligned} \sum_{i=1}^n Z_i^2 &= \sum_{i=1}^n Y_i^2 \\ &= \|(Y - \hat\alpha\mathbf{1} - \hat\beta\mathbf{w}) + \hat\alpha\mathbf{1} + \hat\beta\mathbf{w}\|^2 \\ &= \|Y - \hat\alpha\mathbf{1} - \hat\beta\mathbf{w}\|^2 + n\hat\alpha^2 + \hat\beta^2\|\mathbf{w}\|^2 \\ &\qquad \text{(since all cross-product terms vanish)} \\ &= R + n\bar Y^2 + (\mathbf{w}^\top\mathbf{w})\hat\beta^2 \end{aligned}$$

So $R = \sum_3^n Z_i^2 \sim \sigma^2\chi^2_{n-2}$ and is independent of Z_1 and Z_2, i.e., of $\hat\alpha$ and $\hat\beta$. ■

14.2 Tests and confidence intervals

(a) A t-statistic may be constructed for testing the hypothesis that β takes a particular value β_0, since if β_0 is the true value of β:

$$T_0 = \frac{(\hat\beta - \beta_0)\sqrt{\mathbf{w}^\top\mathbf{w}}}{\sqrt{R/(n-2)}} = \frac{(\hat\beta - \beta_0)\sqrt{\mathbf{w}^\top\mathbf{w}}}{\hat\sigma} \sim t_{n-2}.$$

Therefore, to test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$, we compute t_0 and reject H_0 in a test of size α if $t_0 > t^{(n-2)}_{\alpha/2}$ or $t_0 < -t^{(n-2)}_{\alpha/2}$.

(b) A $(1-\alpha)100\%$ confidence interval may be found for β. Starting from the distributional result in (a) above, we find similarly as in Section 11.1,

$$P\left(\hat\beta - t^{(n-2)}_{\alpha/2}\hat\sigma/\sqrt{\mathbf{w}^\top\mathbf{w}} < \beta < \hat\beta + t^{(n-2)}_{\alpha/2}\hat\sigma/\sqrt{\mathbf{w}^\top\mathbf{w}}\right) = 1 - \alpha.$$

(c) We predict the value of Y that would be observed at a given w_0 by $\hat Y = \hat\alpha + \hat\beta w_0$. Then $Y - \hat Y \sim N\big(0, \sigma^2(1 + 1/n + w_0^2(\mathbf{w}^2\mathbf{w})^{-1})\big)$. Hence a $(1-\alpha)100\%$ **predictive confidence interval** for Y at w_0 is

$$\left[\hat Y - t^{(n-2)}_{\alpha/2}\hat\sigma\sqrt{1 + 1/n + w_0^2(\mathbf{w}^2\mathbf{w})^{-1}}\,,\ \hat Y + t^{(n-2)}_{\alpha/2}\hat\sigma\sqrt{1 + 1/n + w_0^2(\mathbf{w}^2\mathbf{w})^{-1}}\right]$$

14.3 The correlation coefficient

The **sample correlation coefficient** of x and y is defined as $r = S_{xy}/(S_{xx}S_{yy})^{\frac{1}{2}}$. Suppose $Y_i \sim N(a + \beta x_i, \sigma^2)$, independently for each i. The hypothesis that Y does not vary with x is $H_0 : \beta = 0$. The test statistic in (a) can be rewritten as follows:

$$t_0 = \frac{\hat\beta\sqrt{S_{xx}}}{\sqrt{R/(n-2)}} = \frac{(S_{xy}/S_{xx})\sqrt{S_{xx}}\sqrt{(n-2)}}{\sqrt{S_{yy} - S_{xy}^2/S_{xx}}} = \frac{\sqrt{n-2}\,r}{\sqrt{1-r^2}}\,,$$

and so H_0 should be rejected if r^2 is near 1.

Note that the variation in the data is $S_{yy} = \sum_j (y_j - \bar{y})^2$. The regression model 'explains' variation of $\sum_j (\hat{y}_j - \bar{y})^2$ where $\hat{y}_i = \hat{a} + \hat{\beta}x_i$. One can check that $\sum_j (\hat{y}_j - \bar{y})^2 = S_{xy}^2/S_{xx}$ and so the ratio of these is r^2. We say that 'the regression explains $100r^2$% of the variation in the data'.

14.4 Testing linearity

If we are fitting a linear regression, how can we know how good the fit of the estimated regression line is? In general we cannot be sure: a bad fit could quite well be caused by a large value of σ^2, which is unknown. We can test linearity if we are able to replicate the readings, so that, say, we take m readings at each value x_i, and get,

$$Y_{ij} = a + \beta x_i + \epsilon_{ij}, \quad j = 1, \dots, m,$$

for each $i = 1, \dots, n$. Then averaging over j for fixed i we have

$$\bar{Y}_i = a + \beta x_i + \eta_i = \alpha + \beta(x_i - \bar{x}) + \eta_i$$

where the η_i are IID $N(0, \sigma^2/m)$, independently of $\sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2$, which are IID $\sigma^2\chi_{m-1}^2$. Now, if we do a linear regression of $\bar{Y}_i$ on x_i, the residual sum of squares is

$$\sum_{i=1}^n \left(\bar{Y}_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\right)^2 \sim \frac{\sigma^2}{m}\chi_{n-2}^2,$$

if the means are indeed linearly related. Thus to test linearity we consider

$$F = \frac{m\sum_{i=1}^n \left(\bar{Y}_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\right)^2/(n-2)}{\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2/n(m-1)} \sim F_{n-2,n(m-1)},$$

if the model of linearity holds. We reject the hypothesis if $f > F_\alpha^{(n-2),n(m-1)}$.

14.5 Analysis of variance in regression models

Example 14.2 *This tables shows other data for male fruitflies.*

Groups of 25 males kept with	mean life (days)	s.e.	length (mm)	s.e.	sleep (%/day)	s.e.
no companions	63.56	16.4522	0.8360	0.084261	21.56	12.4569
1 uninterested female	64.80	15.6525	0.8256	0.069886	24.08	16.6881
1 interested female	56.76	14.9284	0.8376	0.070550	25.76	18.4465
8 uninterested females	63.36	14.5398	0.8056	0.081552	25.16	19.8257
8 interested females	38.72	12.1021	0.8000	0.078316	20.76	10.7443

'Length' is the length of the fruitfly's thorax. It turns out that longevity (y) is positively correlated to thorax size (x) (as plots of the data show).

Suppose we consider only the data for rows 2 and 3 and adopt a model that for $i = 2, 3$,

$$y_{ij} = a_i + \beta x_{ij}, \quad j = 1 \dots, 25.$$

Let $\bar{a} = \frac{1}{2}(a_2 + a_3)$. Our model 'explains' the observed variation in longevity within group i in terms of the sum of two effects: firstly, an effect due to thorax size, $\bar{a} + \beta x_{ij}$; secondly, an effect specific to group i, $a_i - \bar{a}$. We would like to test

$$H_0 : a_2 = a_3 \quad \text{against} \quad H_1 : a_2 \neq a_3\,.$$

To do this we need to fit the appropriate regression models under the two hypotheses by minimizing the residual sum of squares

$$S = \sum_{j=1}^{25} (y_{2j} - a_2 - \beta x_{2j})^2 + \sum_{j=1}^{25} (y_{3j} - a_3 - \beta x_{3j})^2\,.$$

Under H_1 we minimize freely over a_2, a_3, β and get $\hat{a}_2 = -46.04$, $\hat{a}_3 = -55.69$, $\hat{\beta} = 134.25$, with residual sum of squares $R_1 = 6962.90$.

Under H_0 we minimize subject to $a_2 = a_3$ and get $\hat{a}_2 = \hat{a}_3 = -45.82$, $\hat{\beta} = 128.18$, with residual sum of squares $R_0 = 8118.39$. We can write

$$R_0 = (R_0 - R_1) + R_1.$$

The degrees of freedom of H_0 and H_1 are 2 and 3 respectively. It can be shown that $R_1 \sim \sigma^2\chi_{50-3}^2$, whether or not H_0 is true. Also R_1 and $R_0 - R_1$ are independent. If H_0 is true, then $R_0 - R_1 \sim \sigma^2\chi_{3-2}^2$. If H_0 is not true then $R_0 - R_1$ is inflated.

As we have done previously for ANOVA in Section 12.4, we compute an F statistic

$$f = \frac{(R_0 - R_1)/(3-2)}{R_1/(50-3)} = 7.80,$$

which upon comparison to $F_{0.05}^{(1,47)} = 4.21$ leads us to reject H_0; there is indeed a significant difference between the longevities of the males in the two groups. This is the opposite to what we found with a t-test for equality of means in Example 11.3. The explanation is that the mean thorax size happens to be greater within the group of the males exposed to interested females. This is usually associated with greater longevity. When we take into account the fact that this group did not show the greater longevity that would be appropriate to its greater mean thorax size then we do find a difference in longevities between males in this group and those in the group that were kept with a nonreceptive female.

Thus we see that the analysis in Example 11.3 was deficient. There is a lesson in this example, which might be compared to that in Simpson's paradox.