Testing Correlation

We use ρ to denote the **true** correlation between X and Y.

We can test the hypothesis that $\rho = 0$ (that is, that X and Y are uncorrelated using r. For testing

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \end{aligned}$$

we can use the test statistic

$$t = \frac{r}{\sqrt{(1 - r^2)/(n-2)}}$$

If H_0 is true, then approximately

$$t \sim \text{Student}(n-2)$$

Alternately, we could use

$$z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$$

and then, if H_0 is true, as (approximately)

$$Z \sim N \left(\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{n-3} \right)$$

when $\rho = 0$, so that (approximately)

$$\sqrt{n-3}\, Z \sim N(0,1)$$

A related quantity is the

Coefficient of Determination

or **R^2 Statistic**

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

Note that the *total variation* in y is recorded via

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

and the *random variation* is recorded via

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Therefore the **variation explained by the linear regression** is

$$SSR = SS_{yy} - SSE \qquad \text{as} \qquad SS_{yy} = SSR + SSE$$

Thus

$$r^2 = \frac{SSR}{SS_{yy}} = \frac{\text{Variation explained by Regression}}{\text{Total Variation}}$$

r^2 is a measure of model adequacy, that is, if $r^2 \approx 1$, then the linear model is a **good fit**.

Example: Blood Viscosity vs PCV.

We have

- $n = 32$
- $r = 0.879$
- $R^2 = r^2 = (0.879)^2 = 0.772$

Test of $\rho = 0$:

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = 10.087$$

We compare with a Student$(n-2) \equiv$ Student(30) distribution; the p-value is 3.73×10^{-11}, so there is strong evidence that $\rho \neq 0$.

2.1.6 Prediction

After the linear model is fitted, it can be used for **forecasting** or **prediction**. That is, given a new x value we can predict the corresponding y.

As before, we see that at any value of x_p, the prediction $\hat{y}_p$ is

$$\hat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$$

This is the best predictor of y at this x value.

We can also compute the standard error of this prediction; if the value of the random error variance σ^2 is known, then

$$\text{s.e.}(\hat{y}_p) = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

If σ is unknown, we estimate σ by $\widehat{\sigma} = s$ as defined previously

$$s^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2}$$

so that

$$\text{e.s.e.}(\hat{y}_p) = s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Note: This prediction is the expected value of y at $x = x_p$. That is, we have worked out

$$Var[\widehat{Y}_p] = Var[\widehat{\beta}_0 + \widehat{\beta}_1 x_p]$$

to compute the s.e. for $\widehat{Y}_p$.

But we can actually predict an **error corrupted** version of $\widehat{Y}_p$, $\widehat{Y}_p^\star$ say, where

$$\widehat{Y}_p^\star = \widehat{Y}_p + \epsilon_p$$

where ϵ_p is a new random error.

But

$$Var[\widehat{Y}_p^\star] = Var[\widehat{Y}_p + \epsilon_p] = Var[\widehat{Y}_p] + Var[\epsilon_p] = Var[\widehat{Y}_p] + \sigma^2$$

that is, there is an **extra** piece of variation due to ϵ_p.

Thus

$$\text{e.s.e.}(\hat{y}_p^\star) = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}} > \text{e.s.e.}(\hat{y}_p)$$

Prediction Intervals

A $100(1-\alpha)\%$ prediction interval for the **mean** value at $x = x_p$ is

$$\hat{y}_p \pm St_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

whereas for an individual new value (predicted with error) at $x = x_p$ is

$$\hat{y}_p \pm St_{\alpha/2}(n-2)s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Prediction Intervals