Linear regression, continued

CIs and Hypothesis tests for $\boldsymbol{\sigma^2}$

Result: $\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim$ chi-square$(n-2)$

$100(1-\alpha)\%$ CI for σ^2:

$$\left(\frac{(n-2)\widehat{\sigma}^2}{\chi^2_{(\alpha/2)}}, \frac{(n-2)\widehat{\sigma}^2}{\chi^2_{1-(\alpha/2)}} \right)$$

where the chi-square percentiles correspond to $n-2$ df.

Hypothesis test:

$$\mathrm{H_0}\text{: } \sigma^2 = c$$

$$\mathrm{H_a}\text{: } \sigma^2 \left\{ \begin{array}{c} > \\ < \\ \neq \end{array} \right\} c$$

Test statistic: $\chi^2 = \frac{(n-2)\widehat{\sigma}^2}{c}$

$$\text{Reject } \mathrm{H_0} \text{ if } \chi^2 \left\{ \begin{array}{c} \geq \chi^2_\alpha \\ \leq \chi^2_{1-\alpha} \\ \geq \chi^2_{\alpha/2} \text{ or } \leq \chi^2_{1-(\alpha/2)} \end{array} \right\} \quad (n-2 \text{ df})$$

Inferences about $\boldsymbol{\mu = \beta_0 + \beta_1 x}$

Result: $\widehat{\mu} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ has a normal distribution

$$\mathrm{E}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right) = \beta_0 + \beta_1 x$$

$$\mathrm{V}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right) = \sigma^2 \left[\tfrac{1}{n} + \tfrac{(x-\overline{x})^2}{S_{xx}}\right]$$

estimate with $\widehat{\sigma}^2$

$100(1-\alpha)\%$ CI for $\beta_0 + \beta_1 x$:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{\alpha/2}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{xx}}\right]}$$

where the Student's t percentile corresponds to $n-2$ df

Prediction

A new observation of Y at x, denoted Y_{n+1}, is to be made. We can construct a **prediction interval** in such a way that $100(1-\alpha)\%$ of such intervals in the long run would contain Y_{n+1}. (Long run: hypothetical repetitions of the whole process of obtaining observations of Y at x_1, x_2, ..., x_n, fitting the model, and issuing the prediction interval)

Now, $Y_{n+1} \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$, but we do not know the values of the parameters.

Point prediction of Y_{n+1}: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

Error of prediction:

$$\widehat{Y} - Y_{n+1} \quad \text{(difference of two normals)}$$

$$\mathrm{E}\left(\widehat{Y} - Y_{n+1}\right) = 0 \quad \text{(unbiased prediction)}$$

$$\mathrm{V}\left(\widehat{Y} - Y_{n+1}\right) = \mathrm{V}\left(\widehat{Y}\right) + \mathrm{V}(Y_{n+1})$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}} \right] + \sigma^2$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}} + 1 \right]$$

(estimate with $\widehat{\sigma}^2$)

$100(1-\alpha)\%$ prediction interval for Y_{n+1}:

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \ \pm \ t_{\alpha/2} \sqrt{\widehat{\sigma}^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{xx}} + 1 \right]}$$

Model evaluation

Model evaluation centers around the **residuals** or errors:

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

As in AOV, the residuals should resemble a sample from a normal distribution with a constant variance. In particular, the variability of the residuals should not depend on y or x, nor should there be any systematic patterns in the signs of the residuals. Common diagnostic plots:

- normal probability plot of residuals
- scatter plot of residuals (vertical axis) vs. predicted y values ($\widehat{y}_i$'s)

Outliers are observations that are not well-described by the model, that is, they seem to occur outside of the usual range of variability predicted by the model. Fitted statistical models should be screened for outliers, and the outliers identified should be investigated. Outliers might be recording errors, or might have resulted from identifiable circumstances different from those common to the remaining data; outliers might also simply reflect inadequacy of the model.

One outlier measure is the **externally studentized residual**. This is obtained by dropping the observation from the data and using the remaining $n-1$ observations to refit the model, recalculate the residual, and calculate the estimated standard deviation (the estimate is not the sample standard deviation of the residuals, but rather involves the so-called "hat matrix" which will be displayed later in the course). An externally studentized residual has a Student's t distribution with $n-3$ df. A simple rule of thumb is to suspect an observation is an outlier if the absolute value of its esr is greater than 2.

Influential observations exert inordinate effects on the parameter estimates and model predictions. Omitting an influential observation from the regression analysis changes the results substantially, and therein arise the definitions of the various measures of influence. Observations tagged as influential should be investigated.

Note: an observation can be influential but not an outlier, an outlier but not influential, both, or neither.

Common measures of influence:

- DFFITS is the standardized change in the predicted value for y_i when that observation is dropped from the data set

- DFBETAS are the standardized changes in the parameter estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ when the observation is dropped

In both cases, the simple rule of thumb is to suspect an overly influential observation when the measure exceeds 2 in absolute value