Simple linear regression: residuals

- Residuals are a measure of how much the observed (response) value differ fitted value
- Residuals are used to assess model assumptions (an important part of model diagnostic)
- E.g. Regressing the time of day on temperature of data with a line will give a poor fit (this is likely a non-linear relationship)

Residuals

Whenever a statistical model is assumed, it is important to determine whether the assumptions are realistic. In the case of the regression model, a key idea is that model checking should be based on *residuals*.

Suppose $Y_1, Y_2, \ldots, Y_n$ satisfy the regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent with

$$\epsilon_i \sim N(0, \sigma^2).$$

To test the assumptions of the model we use the residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$.

Residuals

The **residuals** are defined as

$$\boxed{\hat{e}_i = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\hat{y}_i},} \qquad i = 1,2,\dots,n.$$

$\varepsilon_i \sim \text{iid } N(0, \sigma^2)$

The distribution of the residuals $\hat{e}_i$:

① should (approximately) not depend on x,

② should be (approximately) normal

If either of these two conditions are not satisfied, then we say that the regression model may not be appropriate for our data.

Properties of residuals

① $$\sum_{i=1}^{n} \hat{e}_i = 0,$$

② $$\sum_{i=1}^{n} \hat{e}_i x_i = 0,$$

③ $$E[\hat{e}_i] = 0, \text{and}$$

④ $$var(\hat{e}_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right).$$

Proof of properties of residuals

③ $$E[\hat{e}_i] = E[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]$$
$$= E[Y_i] - E[\hat{\beta}_0] - E[\hat{\beta}_1] x_i$$
$$= (\beta_0 + \beta_1 x_i) - \beta_0 - \beta_1 x_i$$
$$= 0$$

④ $$var(\hat{e}_i) = var(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))$$
$$= var(Y_i) + var(\hat{\beta}_0 + \hat{\beta}_1 x_i) - 2\,cov(Y_i, \hat{\beta}_0 + \hat{\beta}_1 x_i)$$
$$= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right] - 2\sigma^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$$
$$= \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right]$$

Standardized residuals

Standardizing $\hat{e}_i$, we have

$$\boxed{\tilde{e}_i = \frac{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{\sigma\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}}$$

which satisfies

$$E[\tilde{e}_i] = 0 \quad \text{and} \quad var(\tilde{e}_i) = 1.$$

Studentized residuals

In practice, σ^2 is often unknown and estimated by S_e^2.

$$e_i^* = \frac{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)}{s_e\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}$$

Technically, this is known as internally studentized residuals. It does not have a t-distribution, as S_e^2 and $\hat{e}_i$ are not independent.

To mitigate this, we can use the externally studentized residuals instead, which has the same form as e_i^* above but the regression model was fitted to the data with the i^{th} observation removed. Then this will have a t-distribution.

Leverage

$$\mathrm{var}(\hat{e}_i) = \sigma^2 \left[1 - \underbrace{\boxed{\frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}_{\text{leverage } (h_{ii})} \right]$$

$$= \sigma^2 (1 - h_{ii})$$

Leverage is a measure of how far x_i is from $\bar{x}$.

It is useful for identifying influential observations (observations that have a strong influence on the the estimated parameters of the regression model).