2.5 : Goodness of Fit

We have shown how to obtain (ordinary) least squares estimators of the slope and intercept of the best-fit straight line. We must still ask how good is our linear model in the first place; i.e. we can obtain the best-fit straight line but this may still be a very poor fit to the data.

How can we test if our model is a good one? Answering this question is tantamount to determining whether the residuals of the data are, indeed, drawn from the assumed distribution – i.e. a pdf with mean zero and variance σ^2 (or σ_i^2 in the case of weighted least squares). The **true** residuals are, in fact, unknown, since they are given by

$$\epsilon_i \quad = \quad y_i \, - \, a \, - bx_i$$

and the true values of the parameters a and b are, of course, unknown. We can *estimate* the residuals, however, in the obvious way simply by replacing a and b in the above formula by their least squares estimators, i.e.

$$\hat{\epsilon}_i \quad = \quad y_i \, - \, \hat{a}_{\mathrm{LS}} \, - \hat{b}_{\mathrm{LS}} x_i$$

and ask whether these estimated residuals are consistent with our model assumptions. This provides us with a means of assessing whether the linear model is a good one in the first place.

Our assumptions are known as our **hypothesis**, and we **test** this hypothesis when we test how well our data fit our model. We call such a hypothesis test a **goodness of fit test**. (We will consider more general hypothesis tests in the next section).

2.5.1 : The χ^2 statistic

We can test how well the data fits the linear model using the χ^2 statistic. For the simple case of one independent variable this is defined as

$$\chi^2 \quad = \quad \sum_{i=1}^{n} \left[\frac{y_i \, - \, \hat{a}_{\mathrm{LS}} \, - \, \hat{b}_{\mathrm{LS}} x_i}{\sigma_i} \right]^2$$

where σ_i^2 is the variance of the i^{th} residual and is assumed known *a priori*. In other words, χ^2 is the sum of the squared residuals, weighted by their variance.

Note that we **must** know the σ_i^2 *a priori*; if we don't then we can say nothing about the goodness of fit of the data to the model (see Figure 14 below).

If the residuals are distributed as $N(0, \sigma_i)$ then the statistic given above has the χ^2 pdf, given by

$$p(\chi^2) \quad = \quad p_0 \left(\chi^2\right)^{\nu/2} e^{-\chi^2/2} \quad \chi^2 \geq 0$$

Here ν is known as the number of **degrees of freedom** of the pdf. The mean value of the pdf is ν and the variance is 2ν.

For a sample size of n, the χ^2 statistic has $\nu = n - 2$ degrees of freedom: the number of degrees of freedom is smaller than n because the statistic is formed not from the true (and unknown) residuals, but from their *estimates* – i.e. we do not know the true values of a and b and must replace them by their least squares estimators when forming the χ^2 statistic.

2.5.2 : Using χ^2 to measure goodness of fit

We use tables of the cumulative distribution function of the χ^2 RV in order to determine whether the hypothesis that the data are well described by the linear model is justified[2]. If the value of the χ^2 statistic is found to be excessively large, or excessively small, compared to its expected value, then we reject our hypothesis and seek a better model.

How does this work in practice? Tables tell us the value of the χ^2 RV for which a certain percentage of the pdf lies to the **left** of that value (we call this a **percentile** of the CDF). For example:-

$$\chi^2_{0.995} \quad = \quad \text{value of } t \text{ for which Prob}(\chi^2 < t) = 0.995 \quad = \quad 32.8 \text{ for } \nu = 15$$

$$\chi^2_{0.90} \quad = \quad \text{value of } t \text{ for which Prob}(\chi^2 < t) = 0.90 \quad = \quad 9.24 \text{ for } \nu = 5$$

Thus we require to carry out the following steps to determine the goodness of fit for our linear model, $Y = a + bX$.

1. Using the real data and the formulae of Section 2.4, determine the least squares estimators of a and b.

2. Using these estimators of a and b, and the (assumed known) variance, σ_i^2, of each residual, calculated the observed value of the χ^2 statistic, i.e.

$$\chi^2_{\text{obs}} \quad = \quad \sum_{i=1}^{n} \left[\frac{y_i \; - \; \hat{a}_{\text{LS}} \; - \; \hat{b}_{\text{LS}} x_i}{\sigma_i} \right]^2$$

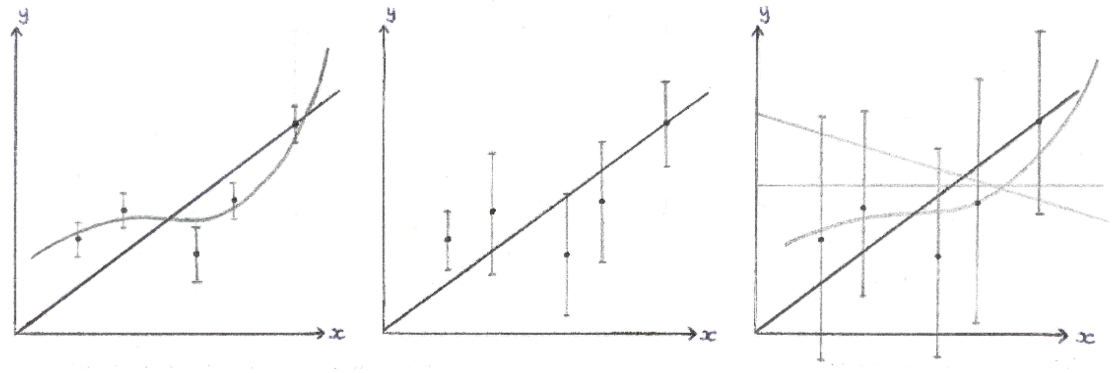
[2] Nowadays it is also common to use statistical packages on computer to determine the percentiles, rather than consulting tables

3. For the appropriate number of degrees of freedom (in this case $\nu = n - 2$, since we have two unknown parameters that we must replace with their estimators), compare χ^2_{obs} with various percentiles of the χ^2 CDF, in order to determine how likely it is that one would obtain as large a value of χ^2_{obs} (or indeed larger) if the linear model were correct.

4. Make a **decision** to **accept** or **reject** the hypothesis of a linear model, based on how likely χ^2_{obs} is found to be.

Figure 14 shows how changing the size of the σ_i, and hence changing the value of χ^2_{obs}, changes our interpretation of the goodness of fit of the *same* observed data to a linear model. In the left hand panel, the errors are sufficiently small (and hence the value of χ^2_{obs} sufficiently large) to indicate that a linear model is a **poor** model for the data – i.e. we need to consider a **curve**. In the right hand panel, conversely, the errors are so large (and hence the value of χ^2_{obs} so small) that the best-fit straight line is **consistent** with the data, but so too are many other straight line fits, and indeed other model curves. We need more or better data to tell if the linear model is the most appropriate. In the central panel, the errors are of a size consistent with our hypothesis, as borne out by a value of χ^2_{obs} which is close to the expected value for that number of degrees of freedom, and so we conclude that the linear model is a good one.

Figure 14: Best linear fits, with different σ_i and different χ^2, for the same data.



2.6 : Fitting General Models

We can apply the χ^2 goodness of fit test more generally than just to fit straight line relations. Suppose we have a physical model for the functional relationship between some variable, Y and another variable, X, i.e.

$$y_i^{\text{model}} \equiv y^{\text{model}}(x_i; \theta_1, ..., \theta_k)$$

where the θ_j are unknown parameters of the model. Suppose we now observe $\{y_i^{\text{obs}}; i = 1, ..., n\}$, where we suppose that

$$y_i^{\text{obs}} = y_i^{\text{model}} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_i)$, for all i, and the ϵ_i are mutually independent.

Suppose we obtain least squares estimators, $\hat{\theta}_1, ..., \hat{\theta}_k$, of the parameters of the model[3]. Then, if our model is correct, it follows that

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i^{\text{obs}} - y_i^{\text{model}}}{\sigma_i} \right]^2$$

has a χ^2 distribution with $n - k$ degrees of freedom.

If the residuals are *not* normally distributed then we can still construct a statistic with an approximately χ^2 distribution by first binning the data. The average value of y^{obs} in each bin, when compared to y^{model} for that bin, will have a residual which is approximately normal due to the Central Limit Theorem. The 'closeness' to normality depends on the original pdf of the residuals and the number of points in each of the bins. Usually if this number exceeds about 15 then the approximation to normality is quite adequate.

[3] the details of how we do this in practice need not concern us in this course. For a general model it is often not possible to write down the analytic expression for the least squares estimators, in the same way as for the linear model, but there exist computer packages for determining the least squares estimators numerically in the general *non-linear* case.