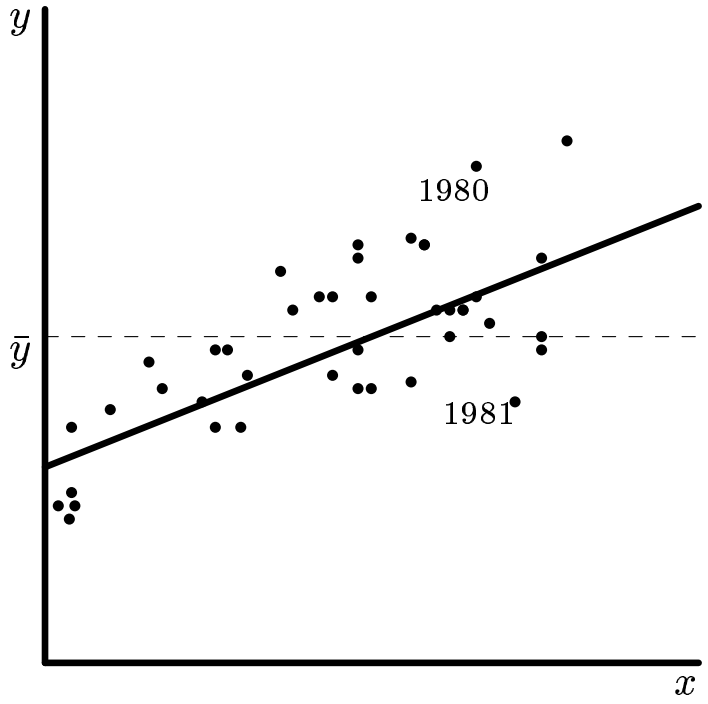
Least Squares - 2 variable case

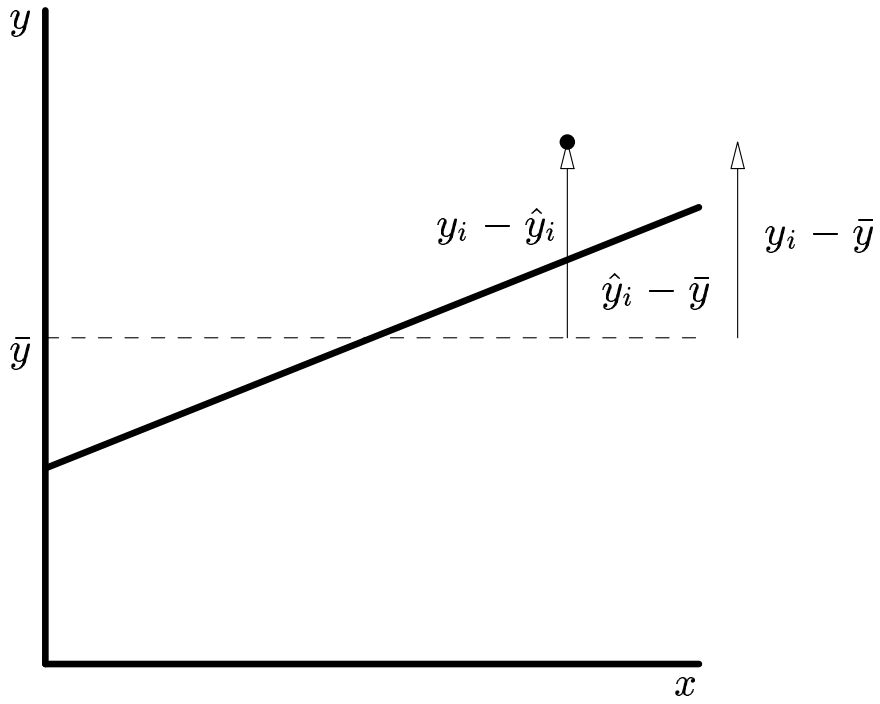
- A linear relationship is expected to exist between variables x and y,

$$y_i = \alpha + \beta x_i \,, \qquad i = 1 \ldots n$$

(e.g. consumption and income).



- We are looking to find estimates of the slope, $\hat{\beta}$, and the intercept, $\hat{\alpha}$, which minimise the the *sum of squared residuals* (RSS or $\sum(y_i - \hat{y})^2$). i.e. minimise the (sum of) unexplained deviations from the fitted line.
- The *explained sum of squares* (ESS) is the sum of the distances $(y_i - \bar{y})$.
- The *total sum of squares* (TSS) is the sum of the ESS and RSS.



Method of Least Squares

- The theoretical relationship between y and x is $y = \alpha + \beta x + u$.
- The *estimated* relationship is $y = \hat{\alpha} + \hat{\beta} x + \hat{u}$.
- Note that u are called *disturbances* and that $\hat{u}$ are known as *residuals*. That is, residuals result from the estimation procedure but disturbances are theoretical.
- To estimate $\hat{\alpha}$ and $\hat{\beta}$ we need to minimise the sum of squared residuals, ie:

$$\min \sum \hat{u}_i^2 = \sum (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

- This requires differentiating with respect to, in turn, $\hat{\alpha}$ and $\hat{\beta}$ and finding stationary points.

- The results of this minimisation give the *normal equations*:

$$\begin{aligned} \sum y_i &= \hat{\alpha} + \hat{\beta} \sum x_i \\ \sum x_i y_i &= \hat{\alpha} \sum x_i + \hat{\beta} \sum x_i^2 \end{aligned}$$

  which we can solve to find $\hat{\alpha}$ and $\hat{\beta}$

- Solving these equations gives the estimators:

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{aligned}$$

- It is *very* important to realise that $\hat{\alpha}$ and $\hat{\beta}$ are *random variables*, and as such have an expectation and variance.

- Appreciate the difference between an *estimator* and an *estimate*. The formulas above are general estimators which will give particular OLS estimates for a set of data.

The assumptions of classical OLS

- To do inference (assess adequacy) on the regression model, we need to make certain assumptions. The *classical* assumptions are:
  - u_i are normally distributed.
  - u_i have mean 0.
  - u_i have constant variance. (homoscedasticity)
  - $\text{Cov}(u_i, u_j) = 0$ for all i, j
  - the x_i (RHS variables) are not random
- Because, u_i are normally distributed, this will imply that the estimators $\hat{\beta}$ and $\hat{\alpha}$ are normally distributed.

Statistical Properties of OLS Estimators

- We would like our estimators to have as low a variance as possible – low variance implies higher accuracy.
- We would also like our estimators to be unbiased, that is, we want the expected value of the estimator to be equal to the true (theoretical) value.
- Note: In modern econometrics, it is of more concern that these requirements are met in large samples (as the sample size grows to infinity).
- Under certain assumptions which were discussed above, known as the classical assumptions, it can be shown that the OLS estimators defined above are unbiased and have the lowest possible variance of any possible linear unbiased estimators. They are *BLUE*.

- The assumptions as stated above are:

$$\begin{aligned} u_i &\sim N(0,\sigma^2) \quad \forall i \\ Cov(u_i,u_j) &= 0 \quad \forall i \neq j \end{aligned}$$

  We also require that the x_i are not random variables.

- To see the unbiasedness of $\hat{\beta}$:

$$E(\hat{\beta}) = E\left(\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right)$$

  Then substitute for y_i and $\bar{y}$ and use the fact that Eu_i=0 to get (on rearrangement):

$$\begin{aligned} E(\hat{\beta}) &= E\left(\frac{\sum(x_i-\bar{x})(\alpha-x_i\beta+u_i-(\alpha-\beta\bar{x}+\bar{u}_i)}{\sum(x_i-\bar{x})^2}\right) \\ &= \left(\frac{\beta\sum(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)=\beta \end{aligned}$$

- Similarly for $\hat{\alpha}$

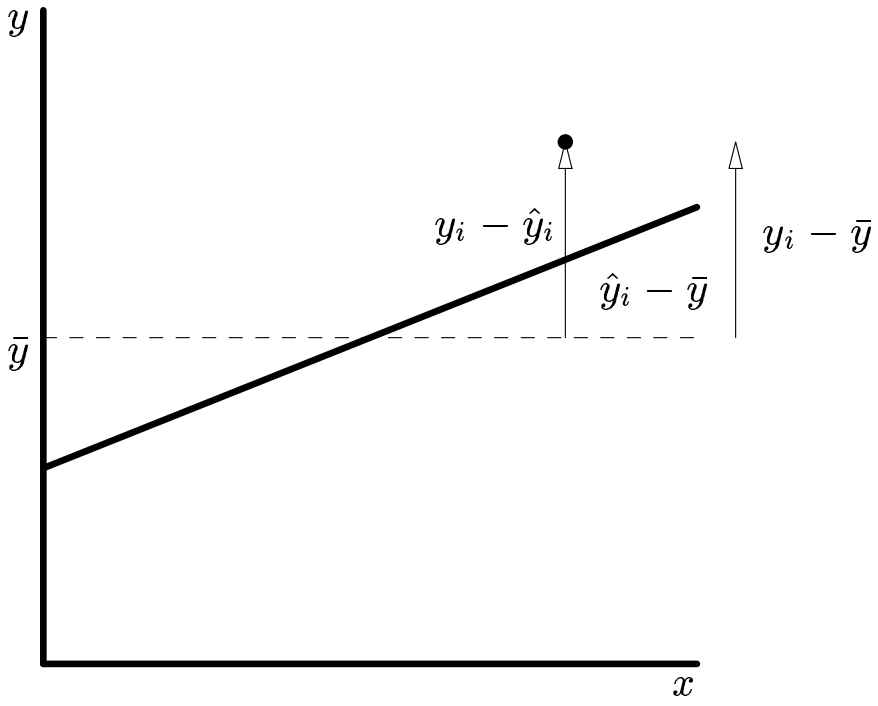
$$\begin{aligned} E(\hat{\alpha}) &= E(\bar{y} - \hat{\beta}\bar{x}) \\ &= E(\alpha + \beta\bar{x} + \bar{u} - \hat{\beta}\bar{x}) \\ &= E(\alpha + \bar{x}(\beta - \hat{\beta}) + \bar{u}) \\ &= \alpha \end{aligned}$$

Analysis of Variance

- One very basic measure of 'goodness of fit' is R^2. This is defined roughly as the proportion of data that is in the explained part of the regression rather than the residual (unexplained) part. Thus:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

- R^2 lies between 0 and 1 and indicates the proportion of the variation in y that has been explained in the regression.



- To perform any serious analysis about the adequacy of the model, we need to find the variances of $\hat{\alpha}$ and $\hat{\beta}$. We will also need to estimate the variance of the error term, $\hat{\sigma}^2$.
- The variance of $\hat{\beta}$:

$$\mathrm{V}(\hat{\beta}) = E\left(\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} - E\hat{\beta}\right)^2$$

- Substitute (as before) for y_i with the regression equation $y_i = \beta x_i + u_i$

$$\begin{aligned}
\mathrm{V}(\hat{\beta}) &= E\left(\frac{\sum(\beta(x_i - \bar{x})^2 + (x_i - \bar{x})(u_i - \bar{u}))}{\sum(x_i - \bar{x})^2} - \beta\right)^2 \\
&= E\left(\beta + \frac{\sum(x_i - \bar{x})(u_i - \bar{u})}{\sum(x_i - \bar{x})^2} - \beta\right)^2 \\
&= \frac{\sum(x_i - \bar{x})^2 \, E(u_i - \bar{u})^2}{(\sum(x_i - \bar{x})^2)^2} \\
&= \frac{\sigma^2}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

- As noted earlier, this variance is the lowest possible for a linear unbiased estimator.

- Unfortunately, σ^2 is an unknown parameter and must be estimated. Since it is the variance of the disturbances, it makes sense to use the sample variance of the residuals as an estimator. Thus,

  $$\hat{\sigma}^2 = \frac{\hat{u}^2}{n}$$

  may be used. However, it can be shown that this is a biased estimator of σ^2 in a small sample. This is because we have lost the independence in the residuals required by estimating two other parameters ($\hat{\alpha}$ and $\hat{\beta}$). It can be shown that an unbiased estimate of σ^2 is

  $$\hat{\sigma}^2 = \frac{\hat{u}^2}{n-2}$$

  because we have estimated 2 parameters. Showing that this is unbiased is not easy, but true.

- We say that we have used up two *degrees of freedom* in estimating α and β.