Lecture 5: Linear regression with one predictor

Winter 2015

Ref: Hartmann Ch. 3

Suppose we have N measurements of one predictor variable x_j (e. g. car weight) and corresponding measurements of a predictand variable y_j (e. g. miles per gallon). We might ask how much of the variability in y_j can be explained by variability in x_j, using a linear fit of the form

$$\hat{y}_j = a_0 + a_1 x_j. \tag{5.0.1}$$

5.1 Mathematics of least-squares regression

Least-squares regression chooses a_0 and a_1 to minimize the mean square residual

$$Q(a_0, a_1) = N^{-1} \sum_{j=1}^{N} [y_j - \hat{y}_j]^2 = \sum_{j=1}^{N} [y_j - (a_0 + a_1 x_j)]^2.$$

Minimizing with respect to a_0 and a_1, we obtain

$$\begin{aligned} 0 &= \frac{\partial Q}{\partial a_0} = -N^{-1} \sum_{j=1}^{N} 2[y_j - (a_0 + a_1 x_j)] \\ 0 &= \frac{\partial Q}{\partial a_1} = -N^{-1} \sum_{j=1}^{N} 2x_j[y_j - (a_0 + a_1 x_j)] \end{aligned}$$

Defining

$$\overline{z} = N^{-1} \sum_{j=1}^{N} z_j,$$

we obtain a pair of simultaneous equations for the coefficients:

$$\begin{aligned} a_0 + a_1 \overline{x} &= \overline{y} \\ a_0 + a_1 \overline{x^2} &= \overline{xy} \end{aligned}$$

Further defining $x' = x - \overline{x}$ and $y' = y - \overline{y}$ (removing means), the solution is:

$$\begin{aligned} a_0 &= \overline{y} - a_1 \overline{x} \\ a_1 &= \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x'y'}}{\overline{x'^2}} \end{aligned} \tag{5.1.1}$$

a_1 is often called the **regression slope**. If we start with **demeaned** x and y data whose sample mean has already been removed ($\overline{x} = \overline{y} = 0$), then $a_0 = 0$.

Let $\sigma_x^2 = \overline{x'^2}$ be the sample variance of x and similarly for y. Let

$$r = \overline{x'y'}/\sigma_x \sigma_y \tag{5.1.2}$$

be the sample correlation coefficient between x and y. Then one-variable linear regression is all about r:

1. The regression slope
$$a_1 = r\sigma_y/\sigma_x \tag{5.1.3}$$
is proportional to r.

2. The regression explains a fraction r^2 of the variance of y:
$$\overline{\hat{y}^2} = a_1^2 \overline{x'^2} = (r^2 \sigma_y^2/\sigma_x^2)\sigma_x^2 = r^2 \sigma_y^2. \tag{5.1.4}$$

When the regression line is subtracted off of y, the **residuals** $\epsilon_j = y_j - a_1 x_j$ have a sample mean of zero. An unbiased estimate of their variance is

$$\sigma_\epsilon^2 = (1 - r^2)\sigma_y^2 \frac{N-1}{N-2} \tag{5.1.5}$$

The $N-1$ is the DOF for estimating variance, while $N-2$ is the residual DOF after estimating the two regression parameters.

5.2 Matlab regression example

The first part of the Matlab script **regression_example.m** on the class web page uses a dataset on compact cars from the 1970s and 1980s built into the Statistics toolbox to calculate the linear regression between car weight and car efficiency (in miles per gallon), both using the above mathematics and using toolbox functions.

5.3 Uncertainty in regression coefficients

We discussed the sampling uncertainty in the correlation coefficient in Lecture 3.6. As long as the observations (and hence the residuals ϵ_j) are linearly independent and identically distributed, one can use the CLT to show that for sufficiently large N ($> 10 - 30$ in practice, depending whether the PDF of the

observations is strongly non-normal), the regression coefficients derived from N samples will be normally distributed about their true values with variances:

$$\begin{aligned} \sigma^2(a_0) &= \sigma_\epsilon^2/N \\ \sigma^2(a_1) &= \sigma_\epsilon^2/(N\sigma_x^2) \end{aligned}$$

This allows us to derive approximate confidence intervals (e. g. $\pm 2\sigma$ for 95%) for the coefficients.

It is worth plotting the residuals ϵ_j vs. x_j to see if they look uncorrelated and of similar variance across the range of x. If successive residuals appear serially correlated, we should use an effective sample sizes N^* in place of N. For instance, If the residuals are given at uniformly spaced x_j and are AR1 with a substantial lag-1 autocorrelation r_1, use $N^* = N(1-r_1)/(1+r_1)$ (following Eqs. 3.5.1 and 3.6.4).

5.4 Regressing many variables on one predictor

The above approach trivially extends to regressing a whole array of data on one predictor, since each variable can just be separately regressed on the predictor.

5.5 Regression caveats

As shown in Figure 1, the same r can encompass a variety of relationships between two variables, not all of which embody the linear relationship with random scatter that was assumed in deriving linear regression. In particular, linear regression

1. does not accurately describe nonlinear relationships
2. can be affected by data clustering and especially outlier data points
3. does not demonstrate causality
4. can reflect codependence of X and Y on some third 'hidden' variable.
5. requires modification if X is uncertain as well as Y (in particular, if the uncertainty in X is a significant fraction of its standard deviation σ_x across the observations), otherwise the regression slope will be underestimated by our formulas.

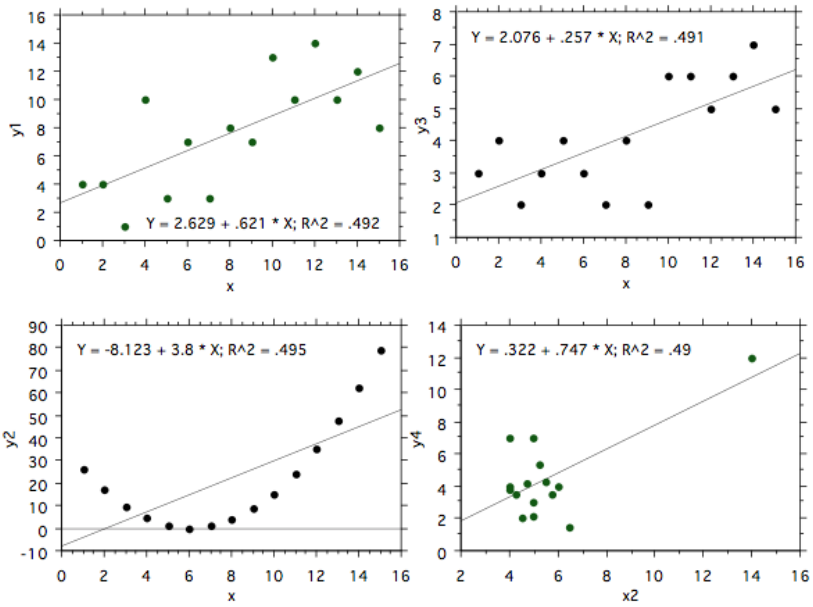

Figure 1: Four datasets with the same correlation coefficient of 0.7. From Hartmann, Ch. 3