Linear regression, continued

The quality of fit of a linear regression model can be measured by the **coefficient of determination**. Recall that

$$\text{SS(total)} = \text{SS(regression)} + \text{SS(residual)}$$

The coefficient of determination, universally called "r^2", is

$$r^2 = \frac{\text{SS(regression)}}{\text{SS(total)}} = 1 - \frac{\text{SS(residual)}}{\text{SS(total)}}$$

It is the proportion of total variability in the y_i's that is described or accounted for by the regression model. The value of r^2 is between 0 and 1; if $r^2 = 1$, all the data points lie on a line.

Interestingly, the likelihood ratio statistic for testing H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$ can be written in terms of the ratio of variances, or the t statistic, or r^2:

$$\frac{\widehat{L}_0}{\widehat{L}_\text{a}} = \left(\frac{\widehat{\sigma}_0^{\,2}}{\widehat{\sigma}_\text{a}^{\,2}}\right)^{-n/2} = \left[\frac{\text{SS(total)}/(n-1)}{\text{SS(residual)}/(n-2)}\right]^{-n/2}$$

$$= \left[1 + \frac{t^2}{(n-1)}\right]^{-n/2} = \left(1 - r^2\right)^{n/2}$$

Correlation

Correlation is a measure of linear association between two random variables. If X and Y are random variables, then the correlation between them is a constant defined by

$$\rho_{XY} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

The value of the correlation is bounded between -1 and $+1$. The expectation in the numerator is called the **covariance** of X and Y; it is real-valued & unbounded (negative or positive).

A correlation (or covariance) of zero does *not* imply that the random variables are independent. Exception: if X and Y have a bivariate normal distribution, then $\rho_{XY} = 0$ implies independence.

Model: X and Y have a **bivariate normal distribution** with means μ_X, μ_Y, variances σ_X^2, σ_Y^2, and correlation ρ_{XY}. Pdf (joint) is a bell-shaped, elongated dome.

ex:
- height and weight
- mother's height and daughter's height
- SAT/ACT score and college GPA

Data: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

Estimate of ρ_{XY} is the **sample correlation coefficient**:

$$\widehat{\rho}_{XY} = r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \widehat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}}$$

Yes: the estimate of ρ_{XY} is identical to the square root of r^2 in a regression.

Hypothesis test:

$$\mathrm{H}_0\text{: } \rho_{XY} = 0 \ \ (X \text{ and } Y \text{ are independent})$$

$$\mathrm{H}_\mathrm{a}\text{: } \rho_{XY} \left\{ \begin{array}{c} > \\ < \\ \neq \end{array} \right\} 0 \ \ (X \text{ and } Y \text{ dependent})$$

Test: identical to the Student's t-test for testing for $\beta_1 = 0$ in a regression using y_i's as dependent variable and x_i's as independent variable. Procedure: perform the regression and use the printed t-test for β_1.

Multiple regression

Situation: more than one independent variable; want to predict Y from x_1, x_2, ..., x_p.

ex:
- IRS predicts the amount of money to be recovered in an audit using (among other variables) amt. of deduction for charitable gifts, amt. of real estate losses, etc.

- House appraiser predicts sale price of a house based on sq. ft., # bedrooms, ave. sale price in neighborhood, etc.

Idea: mean of Y is taken to be a linear function of the predictor variables:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p$$

With just two predictor variables (not functionally dependent), this equation is a **plane**.

Model:

$$Y \sim \text{normal}\big(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p, \sigma^2\big)$$

Different types of predictor variables:

- ordinary quantitative variables

- indicator variables (AOV is a regression!)

  3 treatments; means μ_1, μ_2, μ_3

  $$x_1 = \begin{cases} 1 \text{ if observation is from trt 1} \\ 0 \text{ otherwise} \end{cases}$$

  $$x_2 = \begin{cases} 1 \text{ if observation is from trt 2} \\ 0 \text{ otherwise} \end{cases}$$

  $$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  $$\mu_1 = \beta_0 + \beta_1$$

  $$\mu_2 = \beta_0 + \beta_2$$

  $$\mu_3 = \beta_0$$

- nonlinear terms, e.g.

  $$\mathrm{E}(Y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

- interactions, e.g.

  $$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Estimates of unknown parameters are conveniently represented with **matrix notation**

(your task: learn to multiply matrices; learn what an inverse matrix is)