Linear regression, continued

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

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

The coefficient of determination, universally called “\(r^2\)”, is

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

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 \(r^2\):

\[
\frac{\hat{L}_0}{\hat{L}_a} = \left( \frac{\hat{\sigma}_0^2}{\hat{\sigma}_a^2} \right)^{-n/2} = \left[ \frac{SS(\text{total})/(n - 1)}{SS(\text{residual})/(n - 2)} \right]^{-n/2}
\]

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

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

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 $\mu_X$, $\mu_Y$, variances $\sigma_X^2$, $\sigma_Y^2$, and correlation $\rho_{XY}$. 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:** $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
Estimate of $\rho_{XY}$ is the **sample correlation coefficient**:

\[
\hat{\rho}_{XY} = r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}
\]

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

Hypothesis test:

- $H_0$: $\rho_{XY} = 0$ ($X$ and $Y$ are independent)

- $H_a$: $\rho_{XY} \begin{cases} > 0 \\ < 0 \\ \neq 0 \end{cases}$ ($X$ and $Y$ 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 $\beta_1$. 
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:

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

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

**Model:**

$$Y \sim \text{normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \sigma^2)$$
Different types of predictor variables:

- ordinary quantitative variables
- indicator variables (AOV is a regression!)

3 treatments; means $\mu_1$, $\mu_2$, $\mu_3$

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

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

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

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

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

$$\mu_3 = \beta_0$$

- nonlinear terms, e.g.

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

- interactions, e.g.

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$
Estimates of unknown parameters are conveniently represented with matrix notation.

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