Linear Models

Recall $\underset{}{y_i} = \underset{\text{pop. intercept}}{\beta_0} + \underset{\text{pop. slope}}{\beta_1} \underset{\text{pred. variable}}{X_i} + \underset{\text{error}}{\varepsilon_i}$

$$y_1 = \beta_0 + \beta_1 X_1 + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 X_2 + \varepsilon_2$$
$$y_3 = \beta_0 + \beta_1 X_3 + \varepsilon_3$$
$$\ldots \quad \ldots \quad \ldots \quad \ldots$$

$\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots \equiv$ residuals

Deviation of y_i from expected $\hat{y}_i$

Recall $\varepsilon \sim N(0, \sigma^2)$

Then $Y \sim \beta_0 + \beta_1 X + \varepsilon$, which can be written as: $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$

Aside show that $N(10, \sigma) = 10 + N(0, \sigma)$

We now use OLS or ML to estimate β_0 and β_1

$$\beta_1 : \quad b_1 = \frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\beta_0 : \quad b_0 = \bar{y} - b_1 \bar{x}$$

$$\varepsilon_i : \quad e_i = y_i - \hat{y}_i$$

~~What if instead of continuous X_i, we had categorical values?~~

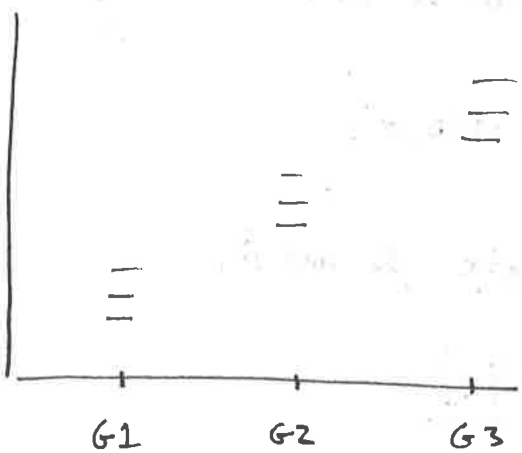
What if we have more than one predictor variable?

$$y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \ldots + \varepsilon_i$$

What if instead of continuous X_i, we had catagorical variables?

- Instead of β_0, use <u>overall mean</u> μ.
- Use dummy variables, one for each category/group.
- i treatment levels (1 to p)
- j replicates within each treatment (1 to n)

$$Y_{ij} = \mu + \beta_1 (\text{dummy}_1)_{ij} + \beta_2 (\text{dummy}_2)_{ij} + \cdots + \varepsilon_{ij}$$



9 observations

$$(\text{dummy}_1) = \begin{matrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{matrix}$$

$$(\text{dummy}_2) = \begin{matrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{matrix}$$

More typically this model expressed as an <u>effects model</u> in which treatment levels are represented by a single term that denotes the effect of each level on the overall mean

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

A problem with this is that if you have e.g. three groups, w/ three means, and the overall mean, you have 4 params and 3 groups of data

Correlation

Let's take a step back and ignore the potential cause-effect relationship implied by linear models (i.e., change in x, results in a change in y). Perhaps thier change is <u>correlated</u>, or we may say that <u>x</u> and <u>y</u> <u>covary</u>.

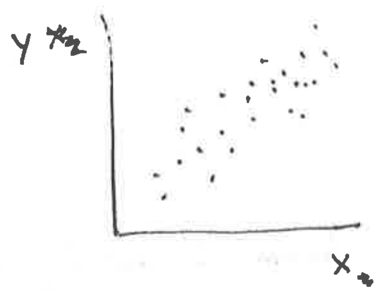
We measure the <u>strength</u> of this relationship as

<u>covariation</u>

or

<u>correlation</u>

How do we visualize co-variation?



How do we calculate co-variation?

Recall variance in x : $\dfrac{\sum (x_i - \bar{x})^2}{n-1}$

Similarly, variance in y : $\dfrac{\sum (y_i - \bar{y})^2}{n-1}$

Co-variance then is : $\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$ A.K.A. Sum of cross-products

$\boxed{\text{Go to example}}$

Co-variance (much like variance) is depent ~~of~~ on magnitude of two variables of interest.

[multiply sepal length by 10 and see the effect]

→ We can <u>standardize</u> co-variance by dividing by the <u>standard deviations</u> of both variables, yielding the <u>Pearson's correlation coefficent</u> (product-moment coeff)

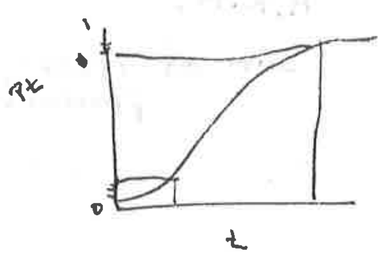
$$\rho_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

when using $\hat{\bar{x}}, \hat{\bar{y}}$; $\rho_{xy} = r_{xy}$

$\rho_{xy} \in [-1, 1]$

Also, when using <u>samples</u> from X and Y, then $\underline{\underline{r}}$ is the sample correlation - coeff statistic. It has a dist'n, just like all other sample statistics. We can write its <u>standard error</u> : $\sqrt{\frac{1 - r^2}{(n-2)}}$

And the we can test whether $\underline{\underline{r}}$ is significantly different from 0 or not, using a t test : $t = \frac{r}{s_r}$



Linear Regression

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Null Hypotheses:

$\beta_0 = 0$ (the population y-intercept equals zero)

$\beta_1 = 0$ (the populatio slope equals zero)

We are rarely interested in the first null hypothesis.

Both of these are tested using t-statistics.

Full model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Reduced model $Y_i = \beta_0 + \varepsilon_i$

Q: Which one will likely have greater error (variation)?

The difference in the variation explained by these two models is known as: Explained Variation

Does the full model explain significantly more variation?

For this, we use the F-test.

What's going on here visually?

Go to p. 172 in Logan.