Correlation (association) & causality

Ref: Ch 5, Faraway text

If X does cause Y ($X \to Y$), then we should observe some association (not necessarily linear) between X & Y, but the converse is not necessarily true

"correlation does not imply causation"

Theories of causality differ between disciplines, but all share some common features:

- underlying theory: the "science" suggest some mechanism by which X might cause Y & also rules out alternative causes (say Z)

- temporal order: X must precede Y (so that it is $X \to Y$, not $Y \to X$)
- association: $X \to Y$ will usually result in some correlation (linear association) between X & Y
  - note: relationship may not be linear

If we discover "associations" in observational data & suspect that it is because $X \to Y$ then in the next iteration of the research process we might some try more structured approach (eg. a designed experiment)

Coefficient of Determination (R^2)

"Proportion of the variation in Y that can be explained by the model involving the X(s)"

- In the R output this is Multiple R-squared; called "multiple" as it does generalise to multiple regression of Y on 2 or more Xs

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 0.7388 \text{ or } 74\%$$

$$= 1 - \frac{SS_{Error}}{SS_{Total}} \quad \leftarrow s^2 = \hat{\sigma}^2 = \sum e_i^2$$

$\leftarrow s_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2$

so, $SS_T = (n-1) * var(y)$

NB: $R^2 = (r)^2$ ← coefficient of correlation between Y & X

- Adjusted R^2 (adjusted for the degrees of freedom)

$$\bar{R}^2 = 1 - \frac{MS_{Error}}{MS_{Total}} = 1 - \frac{SS_{Error}/df_{Error}}{SS_{Total}/df_{Total}}$$

$$= \ldots = R^2 - \underbrace{(1-R^2)\frac{df_{Regression}}{df_{Error}}}_{\text{adjustment factor}}$$

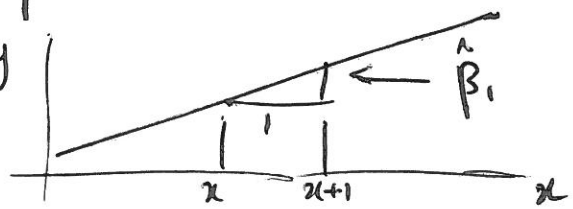
- $F = \frac{MS_{Regression}}{MS_{Error}} \sim F_{k, n-p}$ (known dist!) — Overall F statistic

These are all summary measures. But the F statistic has some advantages

- like $\bar{R}^2$ it does adjust for the df (Ex: show you can derive $\bar{R}^2$ from F)
- F is comparable to a known standard distribution (still have to choose α)

Interpreting the regression coefficients

- Interpretation of $\hat{\beta}_1$ (or any slope coefficient) is "the expected increase in Y as X increases by 1"



- Interpretation of $\hat{\beta}_0$ is "the expected value of Y when $X = 0$" (ie it is the intercept coefficient)

A 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\text{error df}}(0.975) \cdot se(\hat{\beta}_1)$$

$\hat{\beta}_1$: estimate; $t_{\text{error df}}(0.975)$: critical value; $se(\hat{\beta}_1)$: standard error $= \frac{s}{\sqrt{S_{xx}}}$ (for SLR)

Similarly a 95% confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{\text{error df}}(0.975) \cdot se(\hat{\beta}_0)$$

$$se(\hat{\beta}_0) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$