STAT6038 week 4 lecture 12

Rui Qiu

2017-03-17

Correlation (association) & causality

Ref. Ch. 5, Faraway test

If X does cause Y ($X \to Y$), then we should observe some association (not necessarily linear) between X and 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 and also rules out alternative causes (say Z).
  Some times you see spurious association or correlation between X and Y, but in fact, X and Y could be both caused by 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 and Y.
  note: relationship may not be linear

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

Coefficient of Determination (R^2)
a sample quality.

"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 generalize to multiple regression of Y on 2 or more X's.

$$R^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}} = 0.7388 \text{ or } 74\% = 1 - \frac{SS_{\text{Error}}}{SS_{\text{Total}}}$$

  where SS_{Error} is calculated by $s^2 = \hat{\sigma^2} = \sum e_i^2$
  $s_y^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$
  so, $SS_{\text{Regression}} = (n-1) \cdot var(y)$
  Note: $R^2 = (r)^2$ where r is the coefficient of correlation between Y and X

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

$$\overline{R}^2 = 1 - \frac{MS_{\text{Error}}}{MS_{\text{Total}}} = 1 - \frac{SS_{\text{Error}}/df_{\text{Error}}}{SS_{\text{Total}}/df_{\text{Total}}} = \cdots = R^2 - (1-R^2)\frac{df_{\text{Regression}}}{df_{\text{Total}}}$$

  where $(1-R^2)\frac{df_{\text{Regression}}}{df_{\text{Total}}}$ is called the adjusted factor.

- $F = \frac{MS_{\text{Regression}}}{MS_{\text{Error}}} \sim F_{k,n-p}$ overall F statistics
  These are all **summary** measures. But the F statistics has some advantages.
  - like $\overline{R}^2$ it does adjust for the df. (Exercise: show you derive $\overline{R}^2$ from F)
  - F is comparable to a known standard distribution (still have to choose X????). **HERERERERERER**

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 $X = 0$" (i.e. it is the intercept coefficient).
- An 95% confidence interval for β_1 is

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

  i.e. estimate plus/minus the critical value times standard error (for SLR)

- Similarly a 95% confidence interval for β_0 is

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