Multicollinearity

Miaoyan Wang

Department of Statistics
UW Madison

Reading: Chapter 5 in J.F.. Chapter. 13.1-13.2 in R.C.

Multicollinearity

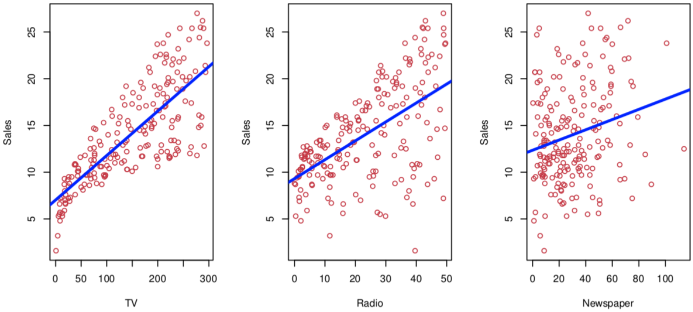
- Recall multiple linear regression:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma^2).$$

- We interpret β_j as the mean change in Y per unit change in X_j, holding all other predictors fixed.
- E.g., consider the relationship between sales and advertising budget on various media:

$$\text{Sale} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} + \beta_3 \text{newspaper} + \varepsilon, \ \varepsilon \sim N(0, \sigma^2).$$

Advertising data



- Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- Which media contribute most to sales?
- Is there synergy among the advertising media?

Results from advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Correlations:

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

How to we interpret $\hat{\beta}_3 < 0$, but $\mathrm{Cov}(\mathrm{newspage}, \mathrm{sales}) > 0$?

Multicollinearity

- The ideal scenario is when the predictors are uncorrelated
  - Each coefficient can be estimated and tested separately.
  - Interpretation such as "a unit change in X_j is associated with a β_j average change in Y, while holding all other predictors fixed".
- Correlation amongst predictors cause problem:
  - The variance of all coefficient estimates tends to increase, sometimes dramatically.
  - Interpretations become hazardous.

Two quotes by famous statisticians

- "Essentially, all models are wrong, but some are useful"
  George Box!

- "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"
  Fred Mosteller and John Tukey, paraphrasing George Box

Multicollinearity

- When the explanatory variables are correlated among themselves, **multicollinearity** among them is said to exist.
- Consider two extreme cases.
  - Case 1: Uncorrelated explanatory variables.
  - Case 2: Perfectly correlated explanatory variables.

Case 1: Uncorrelated Explanatory Variables

- Suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- Suppose X_1 and X_2 are orthogonal such that the sample correlation between X_1 and X_2 is 0.

$$\sum_{i=1}^{n} (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2) = 0$$

- We can show (why?)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_{i1} - \bar{X}_1)}{\sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2}, \quad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_{i2} - \bar{X}_2)}{\sum_{i=1}^{n} (X_{i2} - \bar{X}_2)^2}.$$

- That is, the LS estimate of β_1 is not affected by X_2 and the LS estimate of β_2 is not affected by X_1.
- Interpretation of regression coefficients is clear: β_1 (or β_2) is the expected change in Y for one unit increase in X_1 (or X_2) with X_2 (or X_1) held constant.

Case 2: Perfectly Correlated Explanatory Variables

- Again, suppose $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
- But $X_2 = 2X_1 + 1$.
- Suppose $\beta_0 = 3, \beta_1 = 2, \beta_2 = 5$.
- Then all the following models give the same fit for Y:
  - $Y = 3 + 2X_1 + 5X_2 + \varepsilon$.
  - $Y = 8 + 12X_1 + \varepsilon$.
  - $Y = 2 + 6X_2 + \varepsilon$.

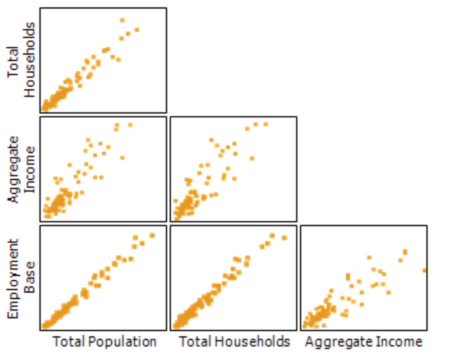
  For example, with 1 unit increase in X_1, there are 2 units increase in X_2 and $\beta_1 + 2\beta_2$ change in Y.

Consequences of Multicollinearity

- In practice, most cases are in between the two extreme cases.
- Effect of multicollinearity on the inference of regression coefficients.
  - Larger changes in the fitted $\hat{\beta}_k$ when another X is added or deleted.
  - More difficult to interpret $\hat{\beta}_k$ as the effect of X_k on Y, because the other X's cannot be held constant.
  - $\boldsymbol{X}^t\boldsymbol{X}$ ill-conditioned or rank-deficient
  - Estimates become sensitive to minor changes of data.
    (why?)

Diagnostics for Multicollinearity

- Large changes in $\hat{\beta}$'s when an explanatory variable (or an observation) is added or deleted.
- Significant joint effects for the affected variables, but wide confidence intervals for β's corresponding to important explanatory variables.
- The sign of $\hat{\beta}$ is counter-intuitive.
- Explanatory variables are highly correlated. e.g. scatter plot matrix R command: pairs(...)



Variance Inflation Factor (VIF)

- Variance inflation factor (VIF) for $\hat{\beta}_k$:

$$\mathrm{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p-1$$

  where R_k^2 is the coefficient of multiple determination when X_k is regressed on the $p-2$ other X explanatory variables.

- That is, R_k^2 is the coefficient of multiple determination R^2 of the model

$$X_k = \beta_0 + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} + \beta_{k+1} X_{k+1} + \cdots + \beta_{p-1} X_{p-1} + \varepsilon.$$

- If the mean VIF values of VIF_k $(k = 1, \ldots, p-1)$ is considerably greater than 1, there may be serious multicollinearity problems.
- If the largest VIF value among VIF_k $(k = 1, \ldots, p-1)$ is larger than 10, multicollinearity may have a large impact on the inference.