Outline

1. Multicollinearity
2. Extra Sums of Squares

Multicollinearity

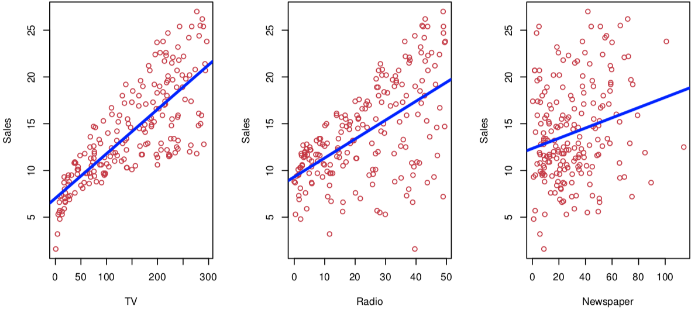
- Recall multiple linear regression:

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

- 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 $X_1, X_2, \ldots, X_p$ 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 $\text{Cov}(\text{newspage}, \text{sales}) > 0$?

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 coefficients 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:
  - $Y = 3 + 2X_1 + 5X_2 + \varepsilon$.
  - $Y = 8 + 12X_1 + \varepsilon$.
  - $Y = 2 + 6X_2 + \varepsilon$.
- Since different models give equally good fit, the interpretation of regression coefficients is difficult.

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.

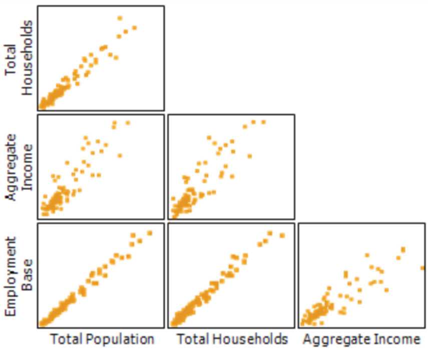
Consequences of Multicollinearity

- 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

  ______________________.
- 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 $\hat{\beta}_k$:

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

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.

Outline

Extra Sums of Squares

- Basic ideas: An **extra sum of squares** measures the marginal reduction (or increase) in the SSE (or SSR) when one or several explanatory variables are added to the regression model, given other explanatory variables are already in the model.
- Extra sums of squares are useful for

  ____.
- Recall the general linear test approach.

General Linear Test Approach

- Consider the **full model** (or, **unrestricted model**)

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \text{iid } N(0, \sigma^2)$$

  and obtain SSE(F).

- Consider the **reduced model** (or, **restricted model**) under the $H_0 : \beta_1 = 0$

$$Y = \beta_0 + \varepsilon, \quad \varepsilon \sim \text{iid } N(0, \sigma^2)$$

  and obtain SSE(R).

Example 1

- Response variable Y and 2 explanatory variables X_1, X_2.
- The full model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

  and denote its sum of squared error by $\text{SSE}(X_1, X_2)$.
- To test $H_0 : \beta_2 = 0$, what is the reduced model?

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon,$$

  and denote its sum of squared error by $\text{SSE}(X_1)$
- Compute $\text{SSE}(X_1)$ and $\text{SSE}(X_1, X_2)$.

$$\text{SSE}(X_1) \geq \text{SSE}(X_1, X_2)$$

Example 1

- Define

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_1) \end{aligned}$$

- Interpretation: $SSR(X_2|X_1)$ measures the decrease in the SSE when X_2 is added to the regression model, given X_1 is already in the model.

Partial F Test: Example 1

- The test statistic for $H_0 : \beta_2 = 0$ is

$$
\begin{aligned}
F^* &= \frac{\frac{\mathrm{SSE}(X_1) - \mathrm{SSE}(X_1, X_2)}{(n-2)-(n-3)}}{\frac{\mathrm{SSE}(X_1, X_2)}{n-3}} \\
&= \frac{\frac{\mathrm{SSR}(X_2 | X_1)}{1}}{\frac{\mathrm{SSE}(X_1, X_2)}{(n-3)}}
\end{aligned}
$$

- Under the H_0,

$$F^* \sim F_{1,n-3}.$$

- The decision rule is to reject H_0 if $f^* > f_{1,n-3,\alpha}$.
- Relation to a T-test for β_2 in the full model? as $(t^*)^2 = f^*$.

Example 2

- Response variable Y and 3 explanatory variables X_1, X_2, X_3.
- The full model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  and denote the sum of squared error by $\text{SSE}(X_1, X_2, X_3)$.
- To test $H_0 : \beta_1 = \beta_3 = 0$, what is the reduced model?

$$Y = \beta_0 + \beta_2 X_2 + \varepsilon.$$

  Denote the sum of squared error by $\text{SSE}(X_2)$.
- Compute $\text{SSE}(X_1, X_2, X_3)$ with $\text{SSE}(X_2)$.

Example 2

- The extra sum of squares $\mathrm{SSR}(X_1, X_3|X_2)$ is defined as

$$\mathrm{SSR}(X_1, X_3|X_2) = \mathrm{SSE}(X_2) - \mathrm{SSE}(X_1, X_2, X_3)$$

- Interpretation: $\mathrm{SSR}(X_1, X_3|X_2)$ measures the decrease in the SSE when X_1 and X_3 are added to the regression model, given X_2 is already in the model.
- Equivalently,

$$\mathrm{SSR}(X_1, X_3|X_2) = \mathrm{SSR}(X_1, X_2, X_3) - \mathrm{SSR}(X_2)$$

- Interpretation: Equivalently, $\mathrm{SSR}(X_1, X_3|X_2)$ measures the increase in the SSR when X_1 and X_3 are added to the regression model, given X_2 is already in the model.

Partial F Test: Example 2

- The test statistic for $H_0 : \beta_1 = \beta_3 = 0$ is

$$F^* = \frac{\frac{SSE(X_2) - SSE(X_1, X_2, X_3)}{(n-2)-(n-4)}}{\frac{SSE(X_1, X_2, X_3)}{n-4}} = \frac{\frac{SSR(X_1, X_3 | X_2)}{2}}{\frac{SSE(X_1, X_2, X_3)}{(n-4)}}$$

- Under the H_0, $F^* \sim F_{2,n-4}$.
- The decision rule is to reject H_0 if $f^* > f_{2,n-4,\alpha}$.

Decomposition of SSR into Extra Sums of Squares

- Begin with

$$\mathrm{SSTO} = \mathrm{SSR}(X_1) + \mathrm{SSE}(X_1).$$

- Since $\mathrm{SSE}(X_1) = \mathrm{SSR}(X_2|X_1) + \mathrm{SSE}(X_1, X_2)$, we have

$$\mathrm{SSTO} = \underbrace{\mathrm{SSR}(X_1) + \mathrm{SSR}(X_2|X_1)}_{\text{explained by regression } \mathrm{SSR}(X_1, X_2)} + \underbrace{\mathrm{SSE}(X_1, X_2)}_{\text{explained by error}}.$$

Sequential SS in ANOVA Table

For $X_1, \ldots, X_{p-1}$ in general, we may summarize the decomposition of SSR into extra sums of squares in an ANOVA table:

Source	SS	df
Regression	$\text{SSR}(X_1, X_2, \ldots, X_{p-1})$	$p-1$
X_1	$\text{SSR}(X_1)$	1
X_2	$\text{SSR}(X_2 \mid X_1)$	1
$\ldots$	$\ldots$	
X_{p-1}	$\text{SSR}(X_{p-1} \mid X_1, \ldots, X_{p-2})$	1
Error	$\text{SSE}(X_1, X_2, \ldots, X_{p-1})$	$n-p$
Total	SSTO	$n-1$

Order of Fitting

- The order of the explanatory variables is arbitrary. For example,

$$
\begin{aligned}
\text{SSTO} &= \text{SSR}(X_1) + \text{SSR}(X_2|X_1) + \text{SSE}(X_1, X_2) \\
\text{SSTO} &= \text{SSR}(X_2) + \text{SSR}(X_1|X_2) + \text{SSE}(X_1, X_2).
\end{aligned}
$$

- Generally, decomposition depends on order of explanatory variables.
- The number of possible orderings becomes large as the number of explanatory variables increases.
- The extra sums of squares in the ANOVA table above are called **sequential SS**).
- When is sequential SS useful?
  when there is a pre-determined order for selecting explanatory variables (e.g. main effect, interaction effect).

Partial SS in ANOVA Table

For $X_1, \ldots, X_{p-1}$ in general, we may summarize the decomposition of SSR into **partial sums of squares** in an ANOVA table:

Source	SS	df
Regression	$SSR(X_1, X_2, \ldots, X_{p-1})$	$p-1$
X_1	$SSR(X_1 \mid X_2, X_3, \ldots, X_{p-1})$	1
X_2	$SSR(X_2 \mid X_1, X_3, \ldots, X_{p-1})$	1
$\ldots$	$\ldots$	
X_{p-1}	$SSR(X_{p-1} \mid X_1, X_2, \ldots, X_{p-2})$	1
Error	$SSE(X_1, X_2, \ldots, X_{p-1})$	$n-p$
Total	SSTO	$n-1$

- The results are independent of the order of the explanatory variables.
- The partial sums of squares do not add up to anything meaningful.

Coefficient of Partial Determination

- Coefficient of partial determination: measures the marginal contribution of one explanatory variable when all others are already included in the regression model.
- For example, with 3 explanatory variables, the coefficients of partial determination are

$$
\begin{aligned}
R^2_{Y1|23} &= \frac{\mathrm{SSR}(X_1|X_2, X_3)}{\mathrm{SSE}(X_2, X_3)} \\
R^2_{Y2|13} &= \frac{\mathrm{SSR}(X_2|X_1, X_3)}{\mathrm{SSE}(X_1, X_3)} \\
R^2_{Y3|12} &= \frac{\mathrm{SSR}(X_3|X_1, X_2)}{\mathrm{SSE}(X_1, X_2)}
\end{aligned}
$$

- Coefficient of partial correlation: square root of a coefficient of partial determination with the same sign as the corresponding fitted regression coefficient.