ANOVA I. Decomposition of Variance

Miaoyan Wang

Department of Statistics
UW Madison

Reading: Chapter 4 in R.C. Chapter. 13-14 in J.F.

ANOVA Approach to Regression Analysis

- The idea is to partition the variation into

$$\text{SS Total} = \text{SS Model} + \text{SS Error}$$

- Why partition the variation?
  - Weigh different sources of variation.
  - Hypothesis testing.
- In the linear regression, consider three types of partitions.
  - Deviation for each observation.
  - Total sum of squares.
  - Degrees of freedom.

Partitioning Deviation of Each Observation

$$\underbrace{Y_i - \bar{Y}}_{\text{total dev}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{dev of fitted from mean}} + \underbrace{Y_i - \hat{Y}_i}_{\text{dev of obs from fitted}} .$$

- If $\{\hat{Y}_i - \bar{Y}\}$ are large in relation to $\{Y_i - \hat{Y}_i\}$: then the regression relation explains a large proportion of the total variation in $\{Y_i\}$.

Partitioning Total Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SSTO}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

- The **total sum of squares (SSTO)** is

$$\text{SSTO} \;=\; \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} Y_i\right)^2$$

A measure of total variation in the data (compare to variance).

Partitioning Total Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SSTO}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

- The **regression sum of squares (SSR)** is

$$\text{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \quad \text{where } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

The larger the SSR in relation to SSTO, the larger the proportion of variability in the Y_i's accounted for by the regression relation.

Partitioning Total Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SSTO}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}}.$$

- The **error sum of squares (SSE)** is

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \text{SSTO} - \text{SSR}$$

The greater the variation of the Y_i's around the fitted regression line, the larger the SSE.

Sums of Squares

- Following arguments for SLR in matrix terms, we have sums of squares

$$\begin{aligned}
\text{SSR} &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \boldsymbol{Y}'(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})\boldsymbol{Y} \\
\text{SSE} &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \\
\text{SSTO} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \boldsymbol{Y}'\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J}\right)\boldsymbol{Y}
\end{aligned}$$

- Partitioning of total sum of squares and the corresponding df are

$$\underbrace{\text{SSTO}}_{df=n-1} = \underbrace{\text{SSR}}_{df=p-1} + \underbrace{\text{SSE}}_{df=n-p}.$$

Coefficient of Multiple Determination

- The **coefficient of multiple determination** is denoted by R^2 and is defined as

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}$$

- Interpretation: The proportion of variation in the Y_i's "explained" by the regression relation.

Adjusted Coefficient of Multiple Determination

- What is effect of more explanatory variables on R^2?
- The **adjusted coefficient of multiple determination** is denoted by R_a^2 and is defined as

$$R_a^2 = 1 - \frac{\frac{\text{SSE}}{n-p}}{\frac{\text{SSTO}}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{\text{SSE}}{\text{SSTO}}.$$

- Interpretation:
  The adjusted coefficient of multiple determination R_a^2 may decrease when more explanatory variables are in the model.

Summary: ANOVA table

The ANOVA table is

Source	df	SS	MS	F
Regression	$p-1$	SSR	MSR	$F = \text{MSR}/\text{MSE}$
Error	$n-p$	SSE	MSE	–
Total	$n-1$	SSTO	–	–

What are the last two columns?

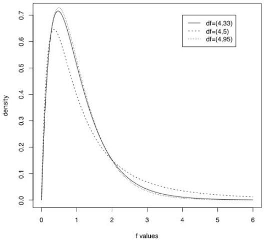
Preparation: Definition of F

F-distribution

An F random variable with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom is

$$F_{\nu_1,\nu_2} = \frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2}$$

where $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ are independent. In particular, $\chi^2_{\nu}\nu = F_{\nu,0}$.



Preparation: Cochran's Theorem

Cochran's Theorem

Consider a linear regression model where SSTO is decomposed into k terms of SS_r, each with degrees of freedom df_r, where $\sum_{r=1}^{k} \mathrm{df}_r = n - 1$. Then under the null $H_0 : \beta_1 = \ldots = \beta_k = 0$,

$$SS_r/\sigma^2 \sim \chi^2_{df_r}, \quad \text{independent w.r.t. } r = 1, \ldots, k.$$

Expectation of quadratic forms

Suppose $\boldsymbol{Y} \sim \mathcal{MVN}(\mu, \boldsymbol{\Sigma})$, and $\boldsymbol{M}$ is a symmetric matrix of constants. Then, $\mathbb{E}(\boldsymbol{Y}^T \boldsymbol{M} \boldsymbol{Y}) = \mathrm{tr}(\boldsymbol{M}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T \boldsymbol{M} \boldsymbol{\mu}$.

Mean Squares

- Define mean squares

$$\text{MSR} = \frac{\text{SSR}}{p-1}, \quad \text{MSE} = \frac{\text{SSE}}{n-p}.$$

- From Cochran's Theorem, we prove that

$$\mathbb{E}(\text{MSE}) = \sigma^2.$$

Proof:

$$\begin{aligned}\mathbb{E}(\text{SSE}) &= (\boldsymbol{X}\beta)'\,(\boldsymbol{I}-\boldsymbol{H})\,(\boldsymbol{X}\beta) + \text{tr}(\sigma^2(\boldsymbol{I}-\boldsymbol{H})) \\ &= 0 + (n-p)\sigma^2\end{aligned}$$

- Thus, we estimate σ^2 by

$$\hat{\sigma}^2 = \text{MSE}.$$

Mean Squares

- Define mean squares

$$\text{MSR} = \frac{\text{SSR}}{p-1}, \quad \text{MSE} = \frac{\text{SSE}}{n-p}.$$

- Similar calculatin shows that

$$\mathbb{E}(\text{MSR}) \quad \begin{cases} = \sigma^2 : & \text{if } \beta_1 = \cdots = \beta_{p-1} = 0; \\ > \sigma^2 : & \text{otherwise} \end{cases}$$

Proof.

$$\begin{aligned} \mathbb{E}(\text{SSR}) &= (\boldsymbol{X}\boldsymbol{\beta})' \left(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J} \right) (\boldsymbol{X}\boldsymbol{\beta}) + \text{tr}(\sigma^2(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J})) \\ &= \sum_{i=1}^{n} \{\beta_1(X_{i1} - \bar{X}_1) + \ldots + \beta_{p-1}(X_{i,p-1} - \bar{X}_{p-1})\}^2 + (p-1)\sigma^2 \end{aligned}$$

Summary: ANOVA table

The ANOVA table is

Source	df	SS	MS	F
Regression	$p-1$	SSR	MSR	$F = \text{MSR}/\text{MSE}$
Error	$n-p$	SSE	MSE	–
Total	$n-1$	SSTO	–	–

What is the last column?

Linear Regression and ANOVA

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

$$Y_i = \beta_0 + \beta_1 X_i^1 + \cdots + \beta_{p-1} X_i^{p-1} + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2)$$

  and obtain SSE(F).

- Consider the **reduced model** (or, **restricted model**) under the $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

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

  and obtain SSE(R).

- It can be shown that SSE(F) $\leq$ SSE(R).

Overall F Test for Regression Relation

- Test

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$$
$$H_A : \text{otherwise, i.e, at least one of the coefficients is non-zero}$$

- Under the $H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$,

$$F^* = \frac{MSR}{MSE} = \frac{\frac{\mathrm{SSE(R)} - \mathrm{SSE(F)}}{\mathrm{df_R} - \mathrm{df_F}}}{\frac{\mathrm{SSE(F)}}{\mathrm{df_F}}} \sim F_{\mathrm{df_R} - \mathrm{df_F}, \mathrm{df_F}}$$

- Thus we can perform an F-test at level α by the decision rule:
  If the observed test statistic $f^* > f_{p-1,n-p,\alpha}$, reject H_0. Otherwise, do not reject H_0.
- Continue to use p-value to gauge the strength of evidence against the H_0.
- Suppose H_0 is rejected, we may further look into which β's are significantly different from 0.

Example: Wetland Species Richness

- For the wetland species richness example,

Source	df	SS	MS	F	p-value
Forest cover	1	49.82	49.817	5.824	0.0191
Error	56	479.03	8.554	–	–
Total	57	528.85	–	–	–

- To test $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$, the observed F test statistic is

$$f^* = \frac{49.817}{8.554} = 5.824$$

- Compared with $F_{1,56}$, the p-value is

$$P(F_{1,56} > 5.824) = 0.0191.$$

- The coefficient of determination is

$$R^2 = \frac{49.82}{528.85} = 0.0942$$

- Interpretation:

Remarks on R^2

- Misunderstandings
  - A high R^2 indicates that accurate predictions can be made.
    not necessary. there may still lack precision in prediction.
  - A high R^2 indicates that the estimated regression line is a good fit.
    not necessary. will see in goodness of fit test
  - An R^2 near zero indicates that X and Y are not related.
    not necessary. small but significant linear coefficient; curvilinear relationship.
- Remarks
  - SSR is "explained variation" and SSE is "unexplained variation" in Y. However, Y does not necessarily depend on X in a causal sense.
  - SLR model does not contain a population parameter for which R^2 estimates.
  - When X are more spread out, R^2 tends to be higher.