STATS 205: Hierarchical Linear Model **Fall 2018**

Lecture 3: Analysis of Variance

Lecturer: Prof. Jingyi Jessica Li *Subscribers: Dong Wang*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Recap

3.1.1 Testing Beta ($\hat{\beta}$): t-test

Model: $Y_{(n\times 1)} = X_{(n\times p)}\beta_{(p\times 1)} + \epsilon_{(n\times 1)}$ and $\beta \sim \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix}$

1. Null Hypothesis:
$$H_0 : \beta_j = 0; \quad H_A : \beta_j \neq 0$$
2. The t-statistics:
$$t = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim T - distribution$$
where:
$$\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$$
3. Estimator for σ^2:
$$\widehat{\sigma^2_{MLE}} = \frac{1}{n} \underset{(1\times n)}{(y - X\hat{\beta})^T} \underset{(n\times 1)}{(y - X\hat{\beta})} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
Since RSS(Residual Sum of Squares) equals to
$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 = \text{SSE (Sum of Squared Errors)}$$
$\widehat{\sigma^2_{MLE}}$ can be expressed as
$$\widehat{\sigma^2_{MLE}} = \frac{\text{RSS}}{n}(or\ \frac{\text{SSE}}{n})$$
As $\frac{\text{RSS}}{\sigma^2} \sim \chi^2{}_{(n-p)}$, we know that $E[\frac{\text{RSS}}{\sigma^2}] = n - p$, so $E[\widehat{\sigma^2_{MLE}}] = \frac{n-p}{n}\sigma^2$. To make it unbiased, we can use $\hat{\sigma}^2_{unbiased} = \frac{RSS}{n-p}$. Under the null hypothesis(H_0), we can calculate a t-statistic which will follow a t-distribution with $(n-p)$ degree of freedom:
$$t = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2_{\text{unbiased}} \times j^{th} \text{ diagonal element of } (X^TX)^{-1}}}$$

and this t-distribution converges to normal distribution as $n \to \infty$:

$$t_{n-p} \overset{n\to\infty}{\longrightarrow} N(0,1)$$

The t-distribution is heavy-tailed relative to the normal distribution, and therefore, it is a more conservative test than the z test.

3.1.2 Wald Test

Assume: $\beta = \begin{pmatrix} \underset{(p_1\times 1)}{\beta_1} \\ \underset{(p_2\times 1)}{\beta_2} \end{pmatrix} \quad X = [\underset{(p_1\times 1)}{X_1}, \underset{(p_2\times 1)}{X_2}]$

1. Null Hypothesis:
$$H_0 : \beta_2 = 0 \quad H_a : \beta_2 \neq 0$$

2. Wald test statistic:
$$W = \hat{\beta_2}^T \widehat{var(\hat{\beta_2})}^{-1} \hat{\beta_2}, \qquad \text{if } p_2 = 1, W = t^2$$
Under H_0: $W/p_2 \sim F(p_2, n-p)$ (exact, under the normality assumption) or $W \overset{\text{approx}}{\sim} \chi^2_{p_2}$(when n is large)

3. the smaller model: $Y = X_1\beta_1 + \epsilon$, which has p_1 predictors
the larger model: $Y = X\beta + \epsilon$, which has p predictors

4. In practice, we use the larger model to calculate $\hat{\sigma}^2 = \frac{1}{n-p}\text{RSS}_{large}$, and $\widehat{var(\hat{\beta_2})} = \hat{\sigma}^2\times$ the lower block diagonal $p_2 \times p_2$ matrix of $(X^TX)^{-1}$

3.1.3 Likelihood Ratio Test

1. Null Hypothesis:
$$H_0 : \beta = 0 \quad H_a : \beta \neq 0$$

2. Likelihood Ratio Statistic:
$$\Lambda = \frac{\max L(\beta)_{H_0}}{\max L(\beta)}$$
$$\log\Lambda = \max\ l(\beta)_{H_0} - \max\ l(\beta)$$
$$\Rightarrow \log\Lambda = -\frac{1}{2\sigma^2}[RSS_{H_0} - RSS]$$
$$-2\log\Lambda = \frac{RSS_{H_0} - RSS}{\sigma^2}$$
plug in the $\hat{\sigma^2}_{unbiased}$, we have that: $-2\log\hat{\Lambda} = \frac{RSS_{H_0} - RSS}{RSS/(n-p)} \overset{\text{approx}}{\sim} \chi^2_{p_2}$

3.2 ANOVA

3.2.1 Extreme Linear Models

1. Null Model: (no predictors, just intercept)

$$Y_i = \mu + \epsilon_i$$

   If $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, then $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$

2. Saturate Model:

$$Y_i = \mu_i + \epsilon_i \qquad i = 1, 2, \cdots, n$$

   where $\mu = (\mu_1, \cdots, \mu_n)^T$ are the parameters and $\hat{\mu_i} = Y_i$

3.2.2 Basic ANOVA

Model: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

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

Source of Variation (SV)	Sum of Squares (SS)	Degrees of Freedom (DF)	Mean Square (MS)	F Ratios
Regression(Model)	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$p-1$	$MSR = \frac{SSR}{p-1}$	$F = \frac{MSR}{MSE}$ $\sim F(p-1, n-p)$
Residuals	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = \frac{SSE}{n-p}$	H_0: $\beta_1 = 0$
Total	$SST = SSR + SSE$ $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$		

Note: $MSR \sim \chi^2_{p-1}$ as $n \to \infty$, $MSE \sim \chi^2_{n-p}$,so $F \sim F_{p-1,n-p}$, so we can do hypothesis testing using F distribution.

3.2.3 Hierarchical ANOVA

Original Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Small Model: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

SV	SS	DF	MS	F
X_1	SSR_1	1	$MSR_1 = \frac{SSR_1}{1}$	
$X_2 \vert X_1$	$SSR_{12} - SSR_1 = SSE_1 - SSE_{12}$	1	$MSR_{2\vert 1} = \frac{SSR_{2\vert 1}}{1}$	$\frac{MSR_{2\vert 1}}{MSE} \sim F(1, n-3)$
Residual	SSE_{12}	$n-3$	$MSE_{12} = \frac{SSE_{12}}{(n-3)}$	H_0: $\beta_2 = 0$
Total		$n-1$		

Remark: In $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, it's possible that H_0: $\beta_1 = \beta_2 = 0$ (ANOVA F test) is rejected at $\alpha = 0.05$, but neither H_0: $\beta_1 = 0$ nor H_0: $\beta_2 = 0$ (t test) is rejected.

In R, `summary(lm(`$Y \sim X_1 + X_2$`))` shows the following:

Estimate	Std. error	t
$\hat{\beta}_1$	$\sqrt{\widehat{Var}(\hat{\beta}_1)}$	$t_1 = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-3}$
$\hat{\beta}_2$	$\sqrt{\widehat{Var}(\hat{\beta}_2)}$	$t_2 = \frac{\hat{\beta}_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} \sim t_{n-3}$

Note:

1. $\frac{MSR_{X_2|X_1}}{MSE} \sim F(1, n-3) \Longleftrightarrow \hat{\beta}_2 \sim t_{n-3}$.
2. If using `summary(lm(`$Y \sim X_2 + X_1$`))`, then $\frac{MSR_{X_1|X_2}}{MSE} \sim F(1, n-3) \Longleftrightarrow \hat{\beta}_1 \sim t_{n-3}$.

3.3 Coefficient of determination: R^2

1. In a simple linear model $Y = \beta_0 + \beta_1 X_1 + \epsilon$, the coefficient of determination $R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = r^2$, where

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \hat{cor}^2(Y, X)$$

   is the Pearson correlation coefficient and is symmetric between X_1 and Y.

2. In a multiple linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, the coefficient of determination $R^2 = 1 - \frac{SSE}{SST}$.

3.3.1 Partial correlation

For the linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ with two variables, partial correlation between Y and X_2 conditional on X_1 ($r_{YX_2|X_1}$) can be calculated as follows:

1. Regress Y onto X_1 to obtain residuals $Y \sim X_1 \xrightarrow{\text{linear regression}} e_{Y|X_1}$
2. Regress X_2 onto X_1 to obtain residuals $X_2 \sim X_1 \xrightarrow{\text{linear regression}} e_{X_2|X_1}$
3. Compute Pearson correlation of the residuals $r_{YX_2|X_1} = \text{Cor}(e_{Y|X_1}, e_{X_2|X_1})$

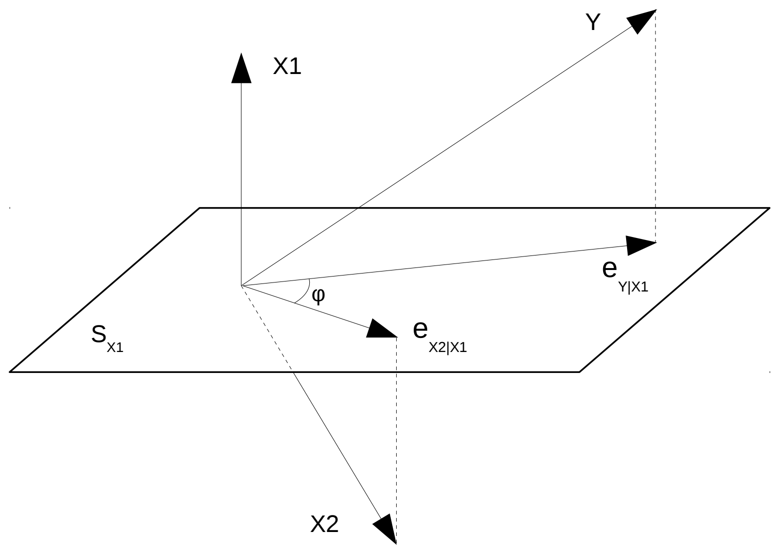

Figure 3.1: $\cos\psi$ gives the partial correlation between Y and X_2 conditional on X_1.

Partial correlation can be interpreted as the cosine of the angle between the projection of Y on to the plane (S_{X_1}) orthogonal to X_1 and the projection of X_2 on to S_{X_1}. More specifically, let $e_{Y|X_1} = \text{proj}_{S_{X_1}} Y$, $e_{X_2|X_1} = \text{proj}_{S_{X_1}} X_2$, and $\psi = \angle(e_{Y|X_1}, e_{X_2|X_1})$, then $r_{YX_2|X_1} = \cos\psi$.

The sample partial correlation can be calculated as

$$r_{YX_2|X_1} = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{\sqrt{(1 - r_{YX_1})^2 (1 - r_{X_1X_2})^2}}.$$

As a recap, for the linear model $Y = \beta_0 + \beta X + \epsilon$, the estimate of β is $\hat{\beta} = r_{XY} \frac{\text{sd}(Y)}{\text{sd}(X)}$. The estimate of β_0 is $\hat{\beta}_0 = \bar{Y} - \hat{\beta}\bar{X}$. And the residual is $e_{Y|X} = Y - \hat{\beta}_0 - \hat{\beta}X$.

Partial correlation is widely used in network analysis. In a network with n nodes $X = \{X_1, \cdots, X_n\}$, the partial correlation between two nodes condtional on the rest of the nodes, $r_{X_iX_j|X\setminus\{X_i,X_j\}}$, can be used to detect the association between two nodes.