22 October 2020

One categorical variable

Model I (Regression model)

Categorical variable m levels $\Rightarrow$ $(m-1)$ dummy variables (or indicator variables)

$$y_i = \beta_0 + \alpha_1 * g_{i,1} + \ldots + \alpha_{m-1} * g_{i,m-1} + e_i$$

for $i = 1, \ldots, n$, where $g_{i,j} = 1$ if i^{th} observation is in j^{th} level and $g_{i,j} = 0$ otherwise.

Model II (ANOVA model)

$$y_{ij} = \mu_i + e_{ij}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n_i$.

Remarks

1. Model I is the model we normally use if there are both categorical and continuous independent variables.
2. Model I and Model II are equivalent such that $\mu_i = \beta_0 + \alpha_i$ for $i = 1, \ldots, m-1$ and $\mu_m = \beta_0$, i.e., $\beta_0 = \mu_m$ and $\alpha_i = \mu_i - \mu_m$ for $i = 1, \ldots, m-1$. Thus, the last group is called reference group.
3. Model II is good in both interpretation & calculation.

Based on Model II,

$$\begin{aligned}
\hat{\mu}_i &= \bar{y}_{i.} \\
Var(\hat{\mu}_i) &= \frac{\sigma^2}{n_i} \\
Cov(\hat{\mu}_i, \hat{\mu}_j) &= 0 \quad \text{for} \quad i \neq j \\
\text{Res.S.S.} &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \\
\Rightarrow \hat{\sigma}^2 &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}{\sum_{i=1}^{m} n_i - m}
\end{aligned}$$

$\Rightarrow$ $(1-\alpha)\%$ C.I. for μ_i:

$$\bar{y}_{i.} \pm t_{\alpha/2, (\sum_{i=1}^{m} n_i - m)} \hat{\sigma} \sqrt{\frac{1}{n_i}}$$

For testing $H_0 : \mu_i = \mu_{i0}$,

$$t = \frac{\bar{y}_{i.} - \mu_{i0}}{\hat{\sigma}\sqrt{\frac{1}{n_i}}}$$

Reject H_0 if $|t_{obs}| > t_{\alpha/2, (\sum_{i=1}^{m} n_i - m)}$.

$H_o: \ \mu_1 = \mu_2 = \cdots = \mu_m = \mu$ (in Model IIA) is equivalent to $H_0: \ \alpha_1 = \ldots = \alpha_{m-1} = 0$ (in Model I)

$$\begin{aligned}
\text{Res. S. S.} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\
\text{Total S. S.} &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2
\end{aligned}$$

The One-way ANOVA table is given below:

Source of Variation	Sum of Squares	Degrees of freedom	Mean Square	Computed f
Model	$\sum\limits_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$m-1$	$\frac{\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{m-1}$	$\frac{(\sum_{i=1}^{m} n_i - m)\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{(m-1)\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}$
Error	$\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$\sum\limits_{i=1}^{m} n_i - m$	$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}{\sum_{i=1}^{m} n_i - m}$	
Total	$\sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	$\sum\limits_{i=1}^{m} n_i - 1$		

<u>Calculation</u>

1. Total S.S. can be calculated as $\left(\sum\limits_{i=1}^{m} n_i - 1\right) S_T^2$, where S_T^2 is the sample variance for all observations.

2. Res.S.S. can be calculated as $\sum\limits_{i=1}^{m}(n_i - 1)S_i^2$, where S_i^2 is the sample variance for observations in the i^{th} level.

3. Reg.S.S. is equal to Total S.S. - Res.S.S.

4. Or,

$$\begin{aligned}
SSA &= \sum_{i=1}^{m}\frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N} \\
&= \frac{1}{n}\sum_{i=1}^{m}\left(T_{i.} - \frac{T_{..}}{m}\right)^2
\end{aligned}$$

if $n_i = n$ for all i.