Oct. 30

1. One categorical variable.

1) Models

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

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.

2) Inference. (model II)

Unknown param: μ_i, $i=1,\cdots,m$; σ^2

① Point est.

$$\hat{\mu}_i = \bar{y}_{i\cdot} \quad i\text{th group.} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2}{\sum_{i=1}^{m} n_i - m}$$

Properties:

$$\mathbb{E}\hat{\mu}_i = \mu_i \qquad Var\,\hat{\mu}_i = \frac{\sigma^2}{n_i} \qquad Cov(\hat{\mu}_i, \hat{\mu}_j) = 0, \; i \neq j.$$

$$\mathbb{E}\,\hat{\sigma}^2 = \sigma^2.$$

② $(1-\alpha)$ C.I. for μ_i:

$$\bar{y}_{i\cdot} \mp t_{\frac{\alpha}{2}}\left(\sum_{i=1}^{m} n_i - m\right) \hat{\sigma}\sqrt{\frac{1}{n_i}}.$$

③ HT. (single param)

$H_0: \mu_i = \mu_{i0}$ (given)

Test stat $\quad t = \dfrac{\bar{y}_{i\cdot} - \mu_{i0}}{\hat{\sigma}/\sqrt{n_i}}.$

Reject H_0 if $|t_{obs}| > t_{\frac{\alpha}{2}}\left(\sum_{i=1}^{m} n_i - m\right)$

④ HT (multi-param)

- $H_0: \mu_1 = \cdots = \mu_m$ ($\Leftrightarrow H_0: \alpha_1 = \cdots = \alpha_m = 0$ in Model I).

- Sum of Squares.

partitioning: $$\sum_{i=1}^{m}\sum_{j=1}^{n_i}[(y_{ij}-\bar{y}_{..})^2] = \sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y}_{..})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2$$

SST (total) SSA (treatment/reg.) SSE (error/res.)

D.F. $\sum_{i=1}^{m} n_i - 1$ $m-1$ $\sum_{i=1}^{m} n_i - m$

- One-way ANOVA:

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

- Computation:

$$SST = \left(\sum_{i=1}^{m} n_i - 1\right) S_T^2$$

$$RSS = SSE = \sum_{i=1}^{m}(n_i - 1) S_i^2$$

$$SSA = SST - SSE$$

- Reject H_0 if $F_{obs} > F_\alpha(m-1, \sum_{i=1}^{n} n_i - m)$

⑤ Single-DF Comparison.

- Contrast $\omega = \sum_{i=1}^{m} c_i \mu_i$ s.t. $\sum_{i=1}^{m} c_i = 0$.

Est. $\hat{\omega} = \sum_{i=1}^{m} c_i \bar{y}_{i.}$

- $H_0: \sum_{i=1}^{m} c_i \mu_i = 0$. $\overset{H_0}{\sim} F(1, \sum_{i=1}^{m} n_i - m)$.

Test stat $F = \frac{SSW}{\hat{\sigma}^2}$, where $SSW = \frac{(\sum_{i=1}^{m} c_i \bar{y}_{i.})^2}{\sum_{i=1}^{m} c_i^2/n_i}$.

1. Standard statistical inference.
   $\Rightarrow$ interpretation
2. Prediction

Correlation　　　causation　　　$X \to Y$

$Y = \beta_0 + \beta_1 X + \varepsilon$　　　$H_0: \beta_0 = \beta_1 = 0.$　　　$Y \to X$

causal inference

$$\begin{array}{ccc} & Z & \\ \swarrow & & \searrow \\ X & & Y \end{array}$$

3. Other analysis beyond correlation.

$X \perp\!\!\!\perp Y \mid Z$