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

Lecture 5: Linear Models with Categorical predictors

Lecturer: Prof. Jingyi Jessica Li *Subscribers: Narek Manoukian, Mina Shahi*

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.*

Tips for Homework 1

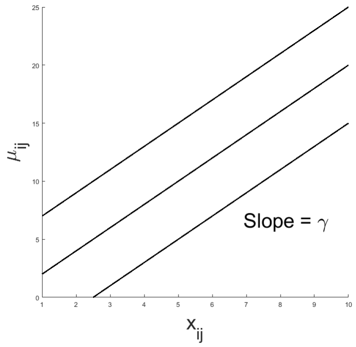
Generalized form of Wald Test:
If A is known to be an $n \times p$ matrix
$H_0 \rightarrow A\beta = 0$
$H_1 \rightarrow A\beta \neq 0$
Test statistic: $W = (A\hat{\beta})^T \cdot (\text{Var}(A\hat{\beta}))^{-1} \cdot (A\hat{\beta}) \xrightarrow{n\rightarrow\infty} \chi^2_m$

Distribution of RSS(SSE)
$RSS = \sum_{i=1}^{n} r_i^2 = r^T r$
where $r = (I - H)Y$ and $Y = X\beta + \epsilon$
$RSS = Y^T(I - H)Y = \epsilon^T(I - H)\epsilon$
$\frac{RSS}{\sigma^2} = (\epsilon/\sigma)^T(I - H)(\epsilon/\sigma) \sim \chi^2_{n-p}$ where $(\epsilon/\sigma) \sim N(0, I_n)$ and rank $(I - H) = (n - p)$

5.1 Analysis of Covariance Models

- Combination of categorical factors and continuous variables.
- x continuous with 1 degree of freedom, z categorical with I levels and $I - 1$ degrees of freedom.
- n_i observations in level i of z. $n = \sum_{i=1}^{I} n_i$.
- Random structure: $Y_{ij} \sim N(\mu_{ij}, \sigma^2)$, $i = 1, \ldots, I$; $j = 1, \ldots, n_i$
- Systematic structure: $\mu_{ij} = \mu + \alpha_i + \gamma x_{ij}$ (i.e., additive model). Impose $\alpha_1 = 0$ for identifiability.

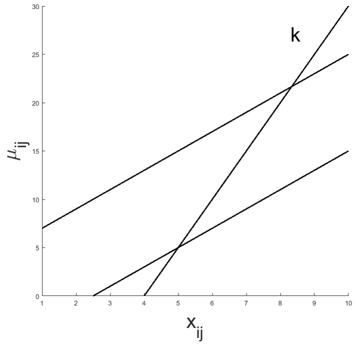
The parameters are $\beta = (\mu, \alpha_2, \ldots, \alpha_I, \gamma)^T$.



Then this model represents I parallel lines, one for each group. The X matrix will look like

$$
\begin{array}{c}
\begin{array}{cccccc} 1 & \alpha_2 & \alpha_3 & \dots & \alpha_I & \gamma \end{array} \\
\begin{pmatrix}
1 & 0 & 0 & \dots & 0 & x_{11} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & \dots & 0 & x_{1n_1} \\
1 & 1 & 0 & \dots & 0 & x_{21} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 1 & 0 & \dots & 0 & x_{2n_2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & \dots & 1 & x_{I1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & \dots & 1 & x_{In_I}
\end{pmatrix}
\end{array}
$$

We can drop the parallel lines assumption. Then $\mu_{ij} = \mu + \alpha_i + (\gamma + \eta_i)x_{ij}$.



- Idenfiability conditions: $\alpha_1 = \eta_1 = 0$
- Design matrix X? Homework question.
- Can test $H_0 : n_2 = \cdots n_I = 0$ by Wald test, LRT, or hierarchical ANOVA.

5.2 Regression Diagnostics

Statistical modeling has three stages:

1. Formulate a model - include random-ness and assumption
2. Fit the model to data - Optimization procedure
3. Check the model - Run model diagnostics

5.2.1 Residual Diagnostics

Residuals may also be expressed as:

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X(X^TX)^{-1}X^TY \\ &= Y - HY \\ &= (I-H)Y. \end{aligned}$$

It can be easily shown that the matrix H is idempotent, i.e. $H^2 = H$ and symmetric, so that $H^T = H$. Also, $(I-H)^T = (I-H)$ and $(I-H)^2 = (I-H)$.

Furthermore, linearity of $\mathbb{E}\left[.\right]$ implies that

$$\begin{aligned} \mathbb{E}\left[e\right] &= \mathbb{E}\left[(I-H)Y\right] \\ &= (I-H)\mathbb{E}\left[Y\right] \\ &= (I-H)X\beta \\ &= X\beta - X(X^TX)^{-1}X^TX\beta \\ &= 0. \end{aligned}$$

It also follows that:

$$\begin{aligned} \mathrm{Var}\left[e\right] &= \mathrm{Var}\left[(I-H)Y\right] \\ &= \mathbb{E}\left[(I-H)YY^T(I-H)^T\right] - \mathbb{E}^2\left[(I-H)Y\right] \\ &= (I-H)(\mathbb{E}\left[YY^T\right](I-H)^T - 0 \\ &= (I-H)\mathrm{Var}\left[Y\right](I-H) \\ &= (I-H)(\sigma^2 I)(I-H) \\ &= \sigma^2(I-H) \end{aligned}$$

Hence, $\text{Var}(e_i) = (1 - h_{ii})\sigma^2$, so residuals do not have constant variance. However, note that $\text{Var}(\varepsilon_i) = \sigma^2$, the error terms have the same variance (constant variance assumption). Furthermore, we can also show that:

$$\begin{aligned} \text{tr}(H) &= \text{tr}(X(X^TX)^{-1}X^T) \\ &= \text{tr}(X^TX(X^TX)^{-1}) \\ &= \text{tr}(I_p) \\ &= p \end{aligned}$$

where $p = \text{rank}(X)$. Now by symmetry we have $h_{ij} = h_{ji}$ which in combination with idempotency implies:

$$\begin{aligned} h_{ii} &= \sum_{j=1}^{n} h_{ij}^2 \\ &= h_{i1}^2 + \cdots + h_{ii}^2 + \cdots + h^2 s_{in} \\ &\geq h_{ii}^2 \end{aligned}$$

which is only possible if $h_{ii} \in [0, 1]$.

Hence, $\text{Var}(e_i) = (1 - h_{ii})\sigma^2$, so residuals do not have the same variance. Thus, $\text{Var}(e_i) > 0$, as $h_{ii} < 1$. And we can also see that larger the h_{ii}, smaller the $\text{Var}(e_i)$.

We can also show $h_{ii} > 0$, thus $\text{Var}(e_i) < \text{Var}(\epsilon_i) = \sigma^2$ given the error terms have the same variance (constant variance assumption).

Ex When $p = 1$ (simple linear model):

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \overline{x})^2}, \text{ where } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Setting $x_i = \bar{x}$ minimizes h_{ii} and maximizes $\text{Var}(e_i)$, here the minimum value of h_{ii} is $\frac{1}{n}$

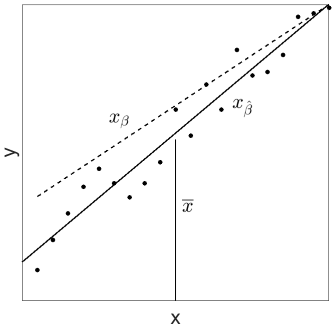
When x_i is far from $\bar{x}$, $\text{Var}(e_i)$ is small since h_{ii} is large. That is, a point farther away from the mean value has more impact on fitting the line, so the variance of its residual will be smaller, but variance of a residual around $\bar{x}$ is larger because the data point has smaller impact on the fitted line.

Definition 5.1 *A standardized residual is defined as* $s_i = \dfrac{e_i}{se(e_i)} = \dfrac{e_i}{\sqrt{1 - h_{ii}}\hat{\sigma}} \sim N(0, 1)$.

where $\hat{\sigma}^2 = \frac{RSS}{n-p}$

Look for observations that have $|s_i| \geq 2$. This is a rule of thumb for detecting potential outliers, but $\hat{\sigma}$ itself may be influenced by outliers. This limits the use of standardized residuals.

<u>Solutions</u>



1. Jack knifed residual

   Estimate σ from $(n-1)$ data points without using the i^{th} observation.

	Jack knifed Residuals	Bootstrap
	(Tokey)	(Efron)
Computer	X	✓
Sample-size	n-1	n
Sampling	without replacement	with replacement
No. of samples	n	∞

   Book for reference: Jackknife and Bootstrap (1993) by Jon Shao

   Definition 5.2 *The jackknifed residual is defined as* $t_i = \dfrac{e_i}{\sqrt{1-h_{ii}}\hat{\sigma}_{(i)}}$

   $$\textit{where } \hat{\sigma}^2_{(i)} = \tfrac{RSS_{(i)}}{n-p-1}$$
   $$RSS_{(i)} \rightarrow \textit{residual sum of squares after leaving out the i-th observation.}$$
   $$\textit{That is, } \hat{\sigma}^2_{(i)} = \frac{\sum_{j\neq i}^{n} e^2_{(i)j}}{(n-1)-p}$$

   To calculate all t_i's, we don't need to do n regressions. By Weisberg (1985, p293),

   $$t_i = s_i\sqrt{\frac{n-p-1}{n-p-s_i^2}},$$

   where $s_i = \frac{r_i}{\sqrt{1-h_{ii}}\hat{\sigma}}$. As a result, one regression is enough and t_i is monotonic in s_i. Hence, ordering observations by s_i or by t_i will give the same rank.

2. Predictive residual $(Y_i - \hat{Y}_{(i)})$

   Let $y_i - \hat{y}_{(i)}$ be the predicted value of the i^{th} observation without using the i^{th} observation in estimation of the regression line. The statistic $y_i - \hat{y}_{(i)}$ is known as the predictive residual. Note that y_i and $\hat{y}_{(i)}$ are independent, hence Cov $\left[y_i, \hat{y}_{(i)}\right] = 0$. It follows that:

$$\begin{aligned}
\mathrm{Var}\left[y_i - \hat{y}_{(i)}\right] &= \mathrm{Var}\left[y_i\right] + \mathrm{Var}\left[\hat{y}_{(i)}\right] \\
&= \sigma^2 + \mathrm{Var}\left[x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}\right] \\
&= \sigma^2 + x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T \mathrm{Var}\left[Y_{(i)}\right] (x_i^T (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T)^T \\
&= \sigma^2 + \sigma^2 (x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i)
\end{aligned}$$