The General Linear Statistical Model

Study Notes December 28, 2013 Kenneth Chu

Let $Y : \Omega \longrightarrow \mathbb{R}^n$ be an $\mathbb{R}^n$-valued random variable defined on the probability space Ω. We assume that the expected value $E[\,Y\,]$ of Y exists. Then, trivially, we have $E[\,Y\,] \in \mathbb{R}^n$.

1 Assumption on the expected value of the response variable Y

The most fundamental assumption of the General Linear Model is that the expected value of the response variable Y lies in a model-specific subspace of $\mathbb{R}^n$ (this subspace will be called the *estimation space* of the model), in the following sense: One of the "components" of a general linear model is its *model matrix* $X \in \mathbb{R}^{n\times p}$, and the expected value of the response variable Y is assumed to lie in the column space $\mathcal{C}(X) \subset \mathbb{R}^n$.

In other words:

The Estimation Space Assumption

$$E[\,Y\,] \in \mathcal{C}(X);\;\; \text{equivalently,}\;\; E[\,Y\,] = X\beta,\; \text{for some (unknown) } \beta \in \mathbb{R}^p, \tag{1.1}$$

where $\mathcal{C}(X) \subset \mathbb{R}^n$ is the column space of the model matrix $X \in \mathbb{R}^{n\times p}$.

We will call $\mathbb{R}^n$ the *observation space*, and $\mathcal{C}(X)$ the *estimation space* of the model.

2 Assumption of the distribution of the response variable Y

In order to make estimation and hypothesis testing computationally feasible, we need to make certain assumptions on the distribution of the response variable Y.

Assumptions on the distribution of Y:

1. The response variable Y has a multivariate normal distribution.
2. The components of Y are independent $\mathbb{R}$-valued random variables.
3. The variances of the components of Y are all equal.

The assumptions on the expected value and distribution on Y together are equivalent to the following:

$$Y \sim N\big(X\beta, \sigma^2 I_n\big)\,,\; \text{for some (unknown but fixed) } \beta \in \mathbb{R}^p, \text{ and some (unknown but fixed) } \sigma > 0. \tag{2.1}$$

Define $\varepsilon := Y - X\beta$. Then, $\varepsilon : \Omega \longrightarrow \mathbb{R}^n$ is also an $\mathbb{R}^n$-valued random variable, with

$$\varepsilon \sim N\big(0, \sigma^2 I_n\big)\,,\;\; \text{for some}\;\; \sigma > 0. \tag{2.2}$$

Proposition 2.1 (Distribution of the full-model error sum-of-squares)
Let $P_{\mathcal{C}(X)^\perp} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ denote the orthogonal projection operator onto the subspace $\mathcal{C}(X)^\perp$. Then,

$$\frac{\|\, P_{\mathcal{C}(X)^\perp}(Y)\,\|^2}{\sigma^2} \;\sim\; \chi^2\big(\mathrm{rank}\big(\mathcal{C}(X)^\perp\big)\big)$$

3 Testing the hypothesis that $H_0 : E[\,Y\,] \in \mathcal{C}(X_0) \subset \mathcal{C}(X)$

Proposition 3.1
Let $P_{\mathcal{C}(X_0)^\perp\cap\,\mathcal{C}(X)} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ denote the orthogonal projection operator onto the subspace $\mathcal{C}(X_0)^\perp \cap \mathcal{C}(X)$. Then,

$$\frac{\|\, P_{\mathcal{C}(X_0)^\perp\cap\,\mathcal{C}(X)}(Y)\,\|^2}{\sigma^2} \;\sim\; \chi^2\!\left(\mathrm{rank}\big(\mathcal{C}(X_0)^\perp \cap \mathcal{C}(X)\big)\;,\; \frac{\|\, P_{\mathcal{C}(X_0)^\perp\cap\,\mathcal{C}(X)} X\beta\,\|^2}{2\,\sigma^2}\right)$$

Corollary 3.2 (Distribution of F-statistics under validity of full model)

$$\frac{\|\,P_{\mathcal{C}(X_0)^\perp\cap\,\mathcal{C}(X)}(Y)\,\|^2\,/\,\mathrm{rank}\big(\mathcal{C}(X_0)^\perp\,\cap\,\mathcal{C}(X)\big)}{\|\,P_{\mathcal{C}(X)^\perp}(Y)\,\|^2\,/\,\mathrm{rank}(\mathcal{C}(X)^\perp)}\;\sim\;F\!\left(\mathrm{rank}\big(\mathcal{C}(X_0)^\perp\,\cap\,\mathcal{C}(X)\big)\;,\,\mathrm{rank}\big(\mathcal{C}(X)^\perp\big)\;;\;\frac{\|\,P_{\mathcal{C}(X_0)^\perp\cap\,\mathcal{C}(X)}X\beta\,\|^2}{2\,\sigma^2}\right)$$

Corollary 3.3 (Distribution of F-statistics under validity of reduced model)

$$\frac{\|\,P_{\mathcal{C}(X_0)^\perp\cap\,\mathcal{C}(X)}(Y)\,\|^2\,/\,\mathrm{rank}\big(\mathcal{C}(X_0)^\perp\,\cap\,\mathcal{C}(X)\big)}{\|\,P_{\mathcal{C}(X)^\perp}(Y)\,\|^2\,/\,\mathrm{rank}(\mathcal{C}(X)^\perp)}\;\sim\;F\big(\mathrm{rank}\big(\mathcal{C}(X_0)^\perp\,\cap\,\mathcal{C}(X)\big)\;,\,\mathrm{rank}\big(\mathcal{C}(X)^\perp\big)\;;\,0\big)$$

4 Model adequacy checking

Goals for model adequacy checking:

- Linearity of the relationship between the response variable and each of the predictor variables.
- The error terms are independent.
- The error terms follow a common Gaussian distribution with zero mean.

The main graphical tools:

- Partial regression plots.
  - Used to assess the linearity of the relationship between the response variable and each of the predictor variables.
- QQ-plot of standardized residuals against theoretical Gaussian quantiles.
  - outliers and non-normality of error terms.
- Scatter plot of (ordinary) residuals against fitted values for response variable.
  - Validity of model assumptions implies that the fitted values and the ordinary residuals will have zero correlation. Any trends in this scatter plot (i.e. any pattern other than an constant-width horizontal band) may indicate departure from model assumptions.
  - This plot can be used to diagnose presence of outliers, as well as non-constancy of error variance.
- Scatter plot of (ordinary) residuals against each of the predictor variables.
- Time series plot and autocorrelation plot of residuals.
  - These apply if the data (for the response and predictor variables) are collected at different times. If so, these plots can be used to examine whether the residuals show any trend or periodicity with respect to time.

Partial regression plots

Suppose our multiple linear model is:

$$Y\;=\;\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon\;=\;X\cdot\beta+\varepsilon \tag{4.3}$$

We seek a method to assess “correctness” or “adequacy” or “appropriateness” of the above model. To this end, we assume the model is correct, and we derive equations whose validity can be examined visually. We proceed as follows: For each $i = 1, \ldots, p$, let $X_{(i)}$ be the matrix obtained from X by deleting the i^{th} column of X. Let $H_{(i)}$ be the corresponding hat matrix. More concretely, we have:

$$\begin{aligned} X &= [\, X_1 \; X_2 \; \cdots \; X_{i-1} \; X_i \; X_{i+1} \; \cdots \; X_p \,] \\ X_{(i)} &= [\, X_1 \; X_2 \; \cdots \; X_{i-1} \; X_{i+1} \; \cdots \; X_p \,] \\ H_{(i)} &= X_{(i)} \cdot \left(X_{(i)}^t \cdot X_{(i)} \right)^{-1} \cdot X_{(i)}^t \end{aligned}$$

Recall that $H_{(i)}$ is the matrix for the orthogonal projection operator from $\mathbb{R}^n$ onto the column space $\mathcal{C}\big(X_{(i)}\big)$ of $X_{(i)}$, and $\big(I_n - H_{(i)}\big)$ is thus the matrix for the orthogonal projection operator onto residual space $\mathcal{C}\big(X_{(i)}\big)^\perp$ with respect to the model $X_{(i)}$. The model equation (4.3) can be rewritten in terms of $X_{(i)}$ as follows:

$$Y \;=\; X_{(i)} \cdot \beta_{(i)} + \beta_i X_i + \varepsilon\,, \tag{4.4}$$

where $\beta_{(i)} = (\beta_0, \beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots \beta_p)^t$. Left-multiplying through the model equation (4.4) by the matrix $\big(I_n - H_{(i)}\big)$ yields:

$$\big(I_n - H_{(i)}\big)\, Y \;=\; \underbrace{\big(I_n - H_{(i)}\big)\, X_{(i)}}_{0} \cdot \beta_{(i)} + \beta_i \big(I_n - H_{(i)}\big)\, X_i + \big(I_n - H_{(i)}\big)\, \varepsilon$$

which simplifies to:

$$\big(I_n - H_{(i)}\big)\, Y \;=\; \beta_i \big(I_n - H_{(i)}\big)\, X_i + \big(I_n - H_{(i)}\big)\, \varepsilon \tag{4.5}$$

Now, observe that $\big(I_n - H_{(i)}\big)\, Y$ is the residuals resulting from regressing Y to $X_{(i)}$, while $\big(I_n - H_{(i)}\big)\, X_i$ is the residuals resulting from regressing X_i to $X_{(i)}$. In other words, the validity of the model equation (4.3) implies the validity of (4.5), which in turn can be assessed visually: For each $i = 1, \ldots, p$, generate the scatter plot of $\big(I_n - H_{(i)}\big)\, Y$ against $\big(I_n - H_{(i)}\big)\, X_i$ and examine whether the resulting data points roughly lie on a straight line with zero intercept.

(Ordinary) residuals and standardized residuals

Model adequacy checking is large done via examination of the residuals of the model fit. Recall that the least-squares estimator $\widehat{Y} : \Omega \longrightarrow \mathcal{C}(X)$ of the response variable $Y : \Omega \longrightarrow \mathbb{R}^n$ is given by:

$$\widehat{Y} \;=\; X \cdot \left(X^t \cdot X\right)^{-1} \cdot X^t \cdot Y \;=\; H \cdot Y\,,$$

where $H := X \cdot \left(X^t \cdot X\right)^{-1} \cdot X^t$ is called the **hat matrix** of the model. Recall also that, geometrically speaking, the hat matrix H is simply the orthogonal projection operator, defined on $\mathbb{R}^n$ (the observation space), onto the column space $\mathcal{C}(X)$ of X (the estimation space, or the model space). The **residual** $\mathbf{e} : \Omega \longrightarrow \mathcal{C}(X)^\perp$ is defined to be:

$$\mathbf{e} \;:=\; Y - \widehat{Y} \;=\; (I_n - H) \cdot Y\,,$$

where $I_n - H$ is the orthogonal projection operator defined on $\mathbb{R}^n$ (the observation space) onto the orthogonal complement $\mathcal{C}(X)^\perp$ of $\mathcal{C}(X)$. Sometimes, the residuals may be referred to as the **ordinary residuals** in order to clearly distinguish them from other types of residuals under the same discussion. Note that $\mathcal{C}(X)^\perp$ can be regarded as the **residual space** of the model. Recall that our model assumption is:

$$Y \;=\; X \cdot \beta + \varepsilon\,,$$

with $\varepsilon \sim N\big(0, \sigma^2 I_n\big)$; see (2.2). Note that in general, the codomain of the error term $\varepsilon : \Omega \longrightarrow \mathbb{R}^n$ is NOT $\mathcal{C}(X)^\perp$ but all of the observation space $\mathbb{R}^n$. On the other hand, observe that

$$\mathbf{e} \;=\; (I_n - H) \cdot Y \;=\; (I_n - H) \cdot (X \cdot \beta + \varepsilon) \;=\; (I_n - H) \cdot \varepsilon\,,$$

since $I_n - H$ is the orthogonal projection operator onto $\mathcal{C}(X)^\perp$, which maps $X \cdot \beta \in \mathcal{C}(X)$ to zero. We thus see that the residual $\mathbf{e} : \Omega \longrightarrow \mathcal{C}(X)^\perp$ is the orthogonal projection of the error term $\varepsilon : \Omega \longrightarrow \mathbb{R}^n$ onto the residual space $\mathcal{C}(X)^\perp$. Or, more strictly speaking, the residual $\mathbf{e} : \Omega \longrightarrow \mathcal{C}(X)^\perp$ is the composition

$$\mathbf{e} :\; \Omega \;\xrightarrow{\varepsilon}\; \mathbb{R}^n \;\xrightarrow{I_n - H}\; \mathcal{C}(X)^\perp$$

Furthermore, note that

$$\begin{aligned}\mathrm{Var}(\mathbf{e}) &= \mathrm{Var}[\,(I_n - H)\cdot\varepsilon\,] \;=\; (I_n - H)\cdot\mathrm{Var}[\,\varepsilon\,]\cdot(I_n - H)^t \;=\; (I_n - H)\cdot\mathrm{Var}[\,\varepsilon\,]\cdot(I_n - H)\\ &= (I_n - H)\cdot\sigma^2 I_n\cdot(I_n - H) \;=\; \sigma^2\cdot(I_n - H)\cdot(I_n - H)\\ &= \sigma^2\cdot(I_n - H)\;,\end{aligned}$$

where the symmetry and idempotence of the orthogonal projection operator $I_n - H$ is used in the above derivation. The above observations lead to the following "model adequacy checks":

- Generate the scatter plot of the observed residuals $\mathbf{e}$ against the fitted values $\widehat{y}$. Examine this scatter plot for trends between the observed residuals and the fitted values; any trend between the observed residuals and the fitted values may indicate violations of model assumptions.

  This adequacy check is based on the following fact:

  $$\mathrm{Cov}\Big(\,\widehat{Y}\,,\,\mathbf{e}\,\Big) \;=\; \mathrm{Cov}(\,H\cdot Y\,,\,(I_n - H)\cdot Y\,) \;=\; H\cdot\mathrm{Cov}(\,Y\,,\,Y\,)\cdot(I_n - H) \;=\; H\cdot\sigma^2 I_n\cdot(I_n - H) \;=\; 0_{n\times n}$$

- Generate the scatter plot of the observed residuals $\mathbf{e}$ against the observed values of each of the predictor variables (columns of the model matrix X). Any trends in any of these scatter plots may indicate violations of model assumptions.

- Generate the QQ-plot of the **standardized residuals** against the theoretical quantiles of the standard Gaussian distribution, where the standardized residuals are defined as follows:

  $$r_i \;:=\; \frac{e_i}{\sqrt{\mathrm{MS}_{\mathrm{error}}\,(1 - h_{ii})}}$$

  where e_i is the i^{th} component of the observed residual $\mathbf{e}$, h_{ii} is the i^{th} diagonal element of the hat matrix $H := X\cdot(X^t\cdot X)^{-1}\cdot X^t$, and $\mathrm{MS}_{\mathrm{error}}$ is the mean squared error of the model fit, which is defined as follows:

  $$\mathrm{MS}_{\mathrm{error}} \;:=\; \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

  Large deviations of the data points on this QQ-plot from the $y = x$ line may indicate violations of model assumptions. This model adequacy check is based on the observations that (1) $\mathrm{MS}_{\mathrm{error}}$ is an unbiased estimator of σ^2, and (2):

  $$\mathbf{e} \;\sim\; N\big(\,0\,,\,\sigma^2\,(I_n - H)\,\big)\;,$$

  which in turn implies that, for each $i = 1,\ldots,n$,

  $$\frac{e_i}{\sqrt{\sigma^2\,(1 - h_{ii})}} \;\sim\; N(0,1)$$