Спецкурс "Регрессия"

Шкляев А.В.

15 января 2019 г.

1 Нормальная модель

1.1 Проверка гипотез в линейной модели

Будем записывать наше предположение $\vec{Y} = X\vec{a} + \vec{\varepsilon}$, $\vec{\varepsilon} \sim \mathcal{N}(0, \sigma^2 E)$ в терминах $Y - \vec{\varepsilon} \in L$, где

$$L = \{X\vec{a}, \vec{a} \in \mathbb{R}^n\}.$$

Будем проверять гипотезу $H_0 : Y - \vec{\varepsilon} \in L_0$, где L_0 — линейное подпространство L с общей альтернативой $H_1 : Y - \vec{\varepsilon} \in L \setminus L_0$.

Воспользуемся обобщенным критерием правдоподобий, который предлагает для проверки H_0 против H_1 использовать статистику отношения правдоподобий

$$\frac{L(X_1, ..., X_n; \widehat{\theta})}{L(X_1, ..., X_n; \widetilde{\theta})} > c,$$

где $\widetilde{\theta}$ — ОМП при выполненной гипотезе, а $\widehat{\theta}$ — ОМП в общей модели $H_0 \cup H_1$.

Как мы уже обсуждали выше, правдоподобие в модели имеет вид

$$L(y_1, ..., y_n; \vec{a}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}||y - X\vec{a}||^2\right).$$

При этом ОМП для σ^2 при H_0 будет иметь вид

$$\widehat{\sigma}^2 = ||y - X\vec{\widehat{a}}||^2/n = ||y - \mathrm{proj}_{L_0} y||^2/n,$$

а правдоподобие приобретет вид

$$L(y_1, ..., y_n; \vec{\widehat{a}}, \widehat{\sigma}^2) = \left(\frac{\sqrt{n}}{\sqrt{2\pi}||y - \mathrm{proj}_{L_0} y||}\right)^n \exp\left(-\frac{n}{2}\right).$$

Аналогичным образом при $H_0 \cup H_1$ омп для σ^2 и правдоподобие приобретут вид

$$\widetilde{\sigma}^2 = ||y - \mathrm{proj}_L y||^2/n,\ L(y_1, ..., y_n; \vec{\widetilde{a}}, \widetilde{\sigma}^2) = \left(\frac{\sqrt{n}}{\sqrt{2\pi}||y - \mathrm{proj}_L y||}\right)^n \exp\left(-\frac{n}{2}\right).$$

Тогда отношение правдоподобий имеет вид

$$\frac{L(y_1, ..., y_n; \vec{\widehat{a}}, \widehat{\sigma}^2)}{L(y_1, ..., y_n; \vec{\widetilde{a}}, \widetilde{\sigma}^2)} = \left(\frac{||y - \mathrm{proj}_L y||}{||y - \mathrm{proj}_{L_0} y||}\right)^{n/2}.$$

Следовательно, критическое множество критерия обобщенного отношения правдоподобий имеет вид

$$\left\{\vec{y}: \frac{||\vec{y}-\text{proj}_{L_0}\vec{y}||}{||\vec{y}-\text{proj}_L\vec{y}||} > c\right\} = \left\{\vec{y}: \frac{||\vec{y}-\text{proj}_{L_0}\vec{y}||^2}{||\vec{y}-\text{proj}_L\vec{y}||^2} > c^2\right\} = \left\{\frac{||\vec{y}-\text{proj}_{L_0}\vec{y}||^2 - ||\vec{y}-\text{proj}_L\vec{y}||^2}{||\vec{y}-\text{proj}_L\vec{y}||^2} > \widetilde{c}\right\}$$

для некоторого $\widetilde{c}$. В числителе стоит разница квадратов длин перпендикуляров из $\vec{y}$, опущенных на L_0 и L, равная квадрату длины $\text{proj}_L\vec{Y} - \text{proj}_{L_0}\vec{Y}$ по теореме Пифагора. Следовательно, мы должны найти $\widetilde{c}$ такое что

$$\mathbf{P}_{H_0}\left(\frac{||\,\text{proj}_L\vec{Y} - \text{proj}_{L_0}\vec{Y}||^2}{||\vec{Y}-\text{proj}_L\vec{Y}||^2} > \widetilde{c}\right) = 1-\alpha.$$

При верной гипотезе H_0 величины $\text{proj}_L\vec{Y} - \text{proj}_{L_0}\vec{Y}$ и $\vec{Y} - \text{proj}_L\vec{Y}$ представляют собой проекции вектора $\vec{Y} - X\vec{a} \sim \mathcal{N}(0,\sigma^2)$ на два ортогональных пространства $L^\perp$ и $L_1 \cap L_0^\perp$. Следовательно,

$$\mathbf{P}_{H_0}\left(\frac{||\,\text{proj}_L\vec{Y} - \text{proj}_{L_0}\vec{Y}||^2/(m-m_0)}{||\vec{Y}-\text{proj}_L\vec{Y}||^2/(n-m)} > \widehat{c}\right) = 1 - F_{f_{m-m_0,n-m}}(\widehat{c}),$$

где $\dim L = m$, $\dim L_0 = m_0$, $f_{k,l}$ — распределение Фишера-Снедекора с k и l степенями свободы.

Отсюда получаем критерий с критическим множеством

$$\frac{D_{0,1}/(m-m_0)}{D_1/(n-m)} = \frac{(D_0-D_1)/(m-m_0)}{D_1/(n-m)} > f_{1-\alpha},\ D_{0,1} = ||\,\text{proj}_L\vec{y} - \text{proj}_{L_0}\vec{y}||^2,\ D_1 = ||\vec{y}-\text{proj}_L\vec{y}||^2,$$

где f — квантиль распределения Фишера-Снедекора с $m-m_0, n-m$ степенями свободы.

1.2 One-way ANOVA

Применим полученный критерий к задаче однофакторного дисперсионного анализа (ANalysis Of VAriance или ANOVA).
Пусть $Z_{i,j} \sim \mathcal{N}(\mu_i,\sigma^2), j \le n_i,\ i \le k$ — независимые наблюдения из k различных групп, $n_1 + ... + n_k = n$. Мы хотим проверить гипотезу о том, что группы однородны: $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ с общей альтернативой.

Представим наши данные соответственно линейной модели:

$$\vec{Y} = X\vec{a} + \vec{\varepsilon},$$

где $\vec{Y} = (Z_{1,1}, ..., Z_{1,n_1}, ..., Z_{k,1}, ..., Z_{k,n_k})$, $\vec{a} = (\mu_1, ..., \mu_k)$, ε_i — н.о.р. $\mathcal{N}(0,\sigma^2)$ и

$$X = \begin{pmatrix} 1 & 0 & ... & 0 \\ 1 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 1 & 0 & ... & 0 \\ 0 & 1 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 1 \end{pmatrix}$$

При этом

$$\widehat{a} = (X^tX)^{-1}X^t\vec{Y} = \begin{pmatrix} n_1 & 0 & ... & 0 \\ 0 & n_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & n_k \end{pmatrix}^{-1} X^tY = \begin{pmatrix} \overline{Z}_{1,\cdot} \\ \overline{Z}_{2,\cdot} \\ ... \\ \overline{Z}_{k,\cdot} \end{pmatrix}.$$

Гипотеза утверждает, что $\vec{Y} = a\vec{e} + \vec{\varepsilon}$, где $\vec{e} = (1, ..., 1)$. При этом $\widetilde{a} = \overline{Y} = \overline{Z}_{\cdot,\cdot}$. Таким образом,

$$RSS = ||\vec{Y} - \text{proj}_L \vec{Y}||^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{i,j} - \overline{Z}_{i,\cdot})^2, SS_{Reg} = ||\,\text{proj}_{L_0} \vec{Y} - \text{proj}_L \vec{Y}||^2 = \sum_{i=1}^{k} n_i(\overline{Z}_{i,\cdot} - \overline{Z}_{\cdot,\cdot})^2.$$

Критерий тем самым приобретает вид

$$F = \frac{SS_{Reg}/(k-1)}{RSS/(N-k)} > f_{1-\alpha},$$

где $f_{1-\alpha}$ — квантиль распределения Фишера с $k-1, N-k$ степенями свободы

1.3 Проверка адекватности линейной модели

Для частного случая проверки гипотезы $H_0 : \vec{a} = 0$, $H_1 : \vec{a} \neq 0$, где $\vec{Y} = X\vec{a} + b\vec{e}$ мы получаем критерий

$$1 - \frac{D_{01}/(k-1)}{D_1/(N-k)} > f_{1-\alpha}.$$

Можно охарактеризовать статистику величиной

$$R^2 = 1 - \frac{D_1}{D_0} \in [0, 1].$$

Чем ближе эта величина к 1, тем ближе мы к отклонению гипотезы неадекватности модели H_0, то есть тем мы увереннее, что X помогают в оценивании $\vec{Y}$.
Величина R^2 называется коэффициентом детерминации.
Отметим, что величина R^2 увеличивается при добавлении в модель предикторов, поскольку длина проекции D_1 убывает, а D_0 не меняется.

1.4 Снижение количества предикторов

Forward selection.
Начнем с модели $\vec{Y} = \vec{\varepsilon}$ и будем добавлять переменные.
Выбирая, какую переменную добавить в модель, мы будем смотреть на

$$F = \frac{(D_1 - D_0)/(k-1)}{D_1/(N-k)},$$

где D_0 — RSS текущей модели, D_1 — следующей. Максимальное значение эта величина принимает при максимальном

$$R^2 = 1 - \frac{D_1}{D_{00}},$$

где D_{00} рассчитывается по сравнению с моделью $H_0 : \vec{Y} = \vec{\varepsilon}$. Тем самым, мы получаем процедуру добавления переменных — на каждом шаге выбираем переменную с максимальным R^2. Остановиться можно либо когда R^2 стал достаточно близок к 1 (больше данного порога), либо когда статистика F оказалась малой.

Backward selection.
В этом случае мы напротив начинаем со всех предикторов и избавляемся от них один за другим. При этом аналогичным образом можно выбирать тот предиктор, у которого наименьшая статистика R^2 и избавляться от него. Критерием остановки можно считать либо момент, когда наименьший из R^2 достаточно отдалится от 1 (на заданное расстояния), либо когда статистика F оказалось большой.