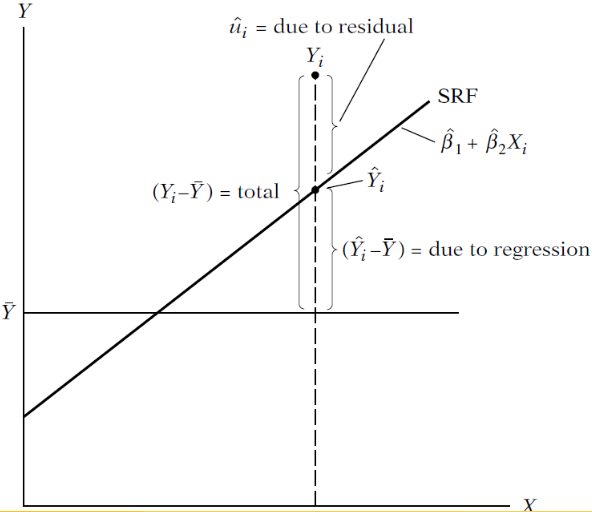
Регрессионный анализ социально-экономических процессов

Меры качества линейной регрессионной модели
Кросс-валидация
Сравнение альтернативных спецификаций

26 января 2026

Разложение дисперсии в линейной регрессии: визуализация



ANOVA-таблица

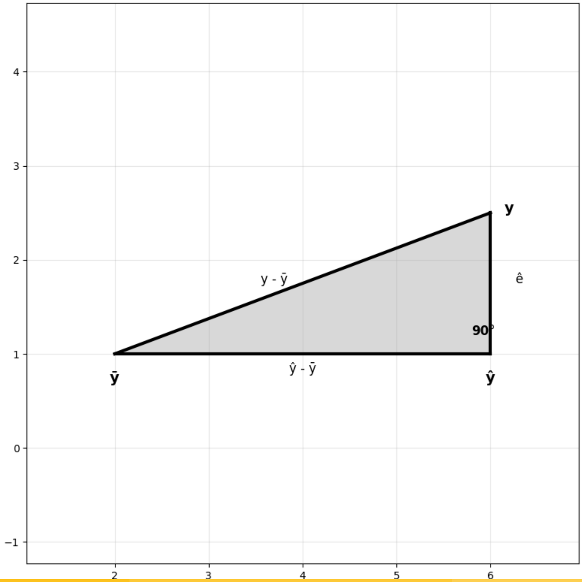
	df	SS	MSS	F
x	k	ESS	$\frac{\text{ESS}}{k}$	$\frac{\text{ESS}/k}{\text{RSS}/(n-k-1)}$
Residual	$n-k-1$	RSS	$\frac{\text{RSS}}{n-k-1}$	
Total	$n-1$	TSS		

Примечание:

k – количество предикторов

n – количество наблюдений

Геометрическая интерпретация R^2



Геометрическая интерпретация R^2

Пусть α – это угол между векторами $\vec{y} - \bar{y} \times \vec{1}$ и $\vec{\hat{y}} - \bar{y} \times \vec{1}$

$$R^2 = \frac{ESS}{TSS} = \frac{\|\vec{\hat{y}} - \bar{y} \times \vec{1}\|^2}{\|\vec{y} - \bar{y} \times \vec{1}\|^2} = cos^2\alpha = cor^2(y_i, \hat{y}_i)$$

$$cos\alpha = \frac{(\vec{y} - \bar{y} \times \vec{1}) \cdot (\vec{\hat{y}} - \bar{y} \times \vec{1})}{\|\vec{y} - \bar{y} \times \vec{1}\| \|\vec{\hat{y}} - \bar{y} \times \vec{1}\|} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}$$
$$= cor(y_i, \hat{y}_i)$$

Проблема переобучения

При переобучении наблюдается существенная разница в объяснительной силе модели на обучающей и тестовой выборке. Модель по сути перестает быть моделью, а просто описывает обучающие данные, подстраиваясь под особенности обучающей выборки. Таким образом, при переобучении

- обобщающая способность модели (возможность перенести результаты на другую выборку) низкая
- результаты оценивания модели неустойчивы

В частности, сильная мультиколлинеарность может способствовать переобучению

$Var(\hat{\beta}_j)$ и VIF_j (1)

1. Пусть дана исходная модель:
   $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki}$
2. По теореме Фриша-Во-Ловелла получим $\hat{\beta}_j$: для этого очистим отклик от всех предикторов кроме j-го, а также очистим j-ый предиктор от всех остальных объясняющих переменных:
   $$\vec{\hat{y}}_i = X_{-j}\vec{\hat{\alpha}}$$
   Сохраняем $\hat{w}_i$ – остатки из этой модели (очищенный y)
   $$\vec{\hat{x}}_j = X_{-j}\vec{\hat{\gamma}}$$
   Сохраняем $\hat{u}_i$ – остатки из этой модели (очищенный x_j)
3. Оцениваем регрессию $\hat{w}_i$ на $\hat{u}_i$: $\widehat{\beta_j} = \frac{Cov(\hat{w}_i, \hat{u}_i)}{Var(\hat{u}_i)} = \frac{\hat{u}^T \hat{w}}{\hat{u}^T \hat{u}}$

$Var(\hat{\beta}_j)$ и VIF_j (2)

$$\frac{\widehat{Cov}(\hat{w}_i, \hat{u}_i)}{\widehat{Var}(\hat{u}_i)} = \frac{\hat{u}^T \hat{w}}{\hat{u}^T \hat{u}} = \frac{\hat{u}^T (y_i - \hat{y}_{(-j)i})}{\hat{u}^T \hat{u}} = \frac{\hat{u}^T y}{\hat{u}^T \hat{u}}$$

$$Var(\hat{\beta}_j) = Var(\frac{\hat{u}^T y}{\hat{u}^T \hat{u}}) = \frac{\hat{u}^T \sigma^2 I \hat{u}}{(\hat{u}^T \hat{u})^2} = \frac{\sigma^2 \hat{u}^T \hat{u}}{(\hat{u}^T \hat{u})^2} = \frac{\sigma^2}{\hat{u}^T \hat{u}}$$

$$\hat{u}^T \hat{u} = RSS_j = \sum_{i=1}^{n} (x_{ji} - \hat{x}_{ji})^2 = TSS_j(1 - R_j^2) = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2 (1 - R_j^2)$$

$$VIF_j = \frac{1}{1 - R_j^2}$$

показывает, во сколько раз увеличивается дисперсия $\hat{\beta}_j$ при мультиколлинеарности по сравнению со случаем, когда j-ый предиктор ортогонален другим предикторам

Регуляризация: метод гребневой регрессии

Гребневая регрессия накладывает штраф на большие по модулю коэффициенты:

$$\min_{\beta}\left(\text{RSS} + \alpha\sum_{j=1}^{k}\beta_j^2\right)$$

Мы изменяем матрицу $X^T X$, добавляя константу α к её диагональным элементам

Получается, что

$$\hat{\beta}_{ridge} = (X^T X + \alpha I)^{-1} X^T Y$$

, где α – это параметр регуляризации

- Матрица $X^T X + \alpha I$ обратима
- Получаем более устойчивые результаты
- Оценки коэффициентов получаются смещенными, однако уменьшается их дисперсия

Гребневая регрессия: поиск параметра α

k-блочная кросс-валидация:

1. массив разбивается на k равных подвыборок
2. для каждого заданного параметра α k итераций: j-ая подвыборка выступает тестовой, остальные подвыборки составляют обучающую
3. считаем среднее MSE по k итерациям для каждого значения α
4. выбираем то значение α, при котором усредненное MSE принимает минимальное значение

R^2 скорректированный

Вне зависимости от того, добавили ли мы в модель значимые или незначимые для объяснения дисперсии зависимой переменной предикторы, коэффициент детерминации R^2 не убывает. Поэтому для того, чтобы учесть случай, когда некоторые параметры могут оказаться бесполезными в нашей спецификации модели, мы будем использовать альтернативу - R^2_{adj} (R^2 скорректированный не превышает R^2: $R^2_{adj} \leq R^2$)
В R^2 скорректированном мы вводим корректировку на количество степеней свободы (см. таблицу разложения дисперсии ANOVA):

$$R^2_{\text{adj}} = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-1}} = 1 - \frac{RSS(n-1)}{TSS(n-k-1)} = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

Вложенные VS Невложенные модели

Вложенные модели (Nested Models)

Мы можем получить спецификацию одной из другой посредством только добавления или только исключения параметров. Пример:

$$M_1 : \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$
$$M_2 : \hat{y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}$$

Невложенные модели (Non-nested Models)

Получить одну из другой только исключением или только добавлением параметров невозможно. Пример:

$$M_1 : \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$
$$M_3 : \hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_2 x_{3i}$$

Информационные критерии

$$IC = -2lnL + c$$

Чем лучше модель описывает данные, тем больше функция правдоподобия: $-2ln(L)$ — это мера «плохости» подгонки. Чем меньше значение $-2ln(L)$, тем лучше модель

c – штраф на добавление параметров

AIC – Akaike Information Criterion

BIC – Bayesian Information Criterion

$$AIC = -2lnL + 2p$$

$$BIC = -2lnL + p\ln(N)$$

p – количество параметров в регрессионной модели;

N – размер выборки;

lnL – натуральный логарифм функции правдоподобия модели

Какой вывод делаем по IC?

Мы рассчитываем информационные критерии (AIC / BIC) для альтернативных моделей, оцененных на одном наборе данных, и сравниваем значения AIC / BIC. Само по себе значение IC ничего не говорит, важно только в сравнительной перспективе

Чем меньше значение рассчитанного информационного критерия (AIC/BIC), тем модель лучше

Эмпирическое правило: если разница между информационными критериями оцененных моделей превышает 10, то одна модель считается существенно лучше, чем другая. Если разница меньше, то значимой разницы между моделями нет

F-тест для вложенных моделей

$$M\_short : \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$$
$$M\_full : \hat{y}_i = \hat{\gamma}_0 + \hat{\gamma}_2 x_{1i} + \ldots + \hat{\gamma}_k x_{ki} + \hat{\gamma}_{k+1} x_{(k+1)i} + \ldots + \hat{\gamma}_{k+j} x_{(k+j)i}$$

$$H_0 : \gamma_{k+1} = \ldots = \gamma_{k+j} = 0$$

При верной H_0 справедливо, что

$$F = \frac{(RSS_{short} - RSS_{long})/\Delta df}{RSS_{long}/df_{long}} \sim F(\Delta df, df_{long})$$

Если p-value достаточно велико, то делаем выбор в пользу модели более экономной (с меньшим количеством параметров)

ВОПРОСЫ?