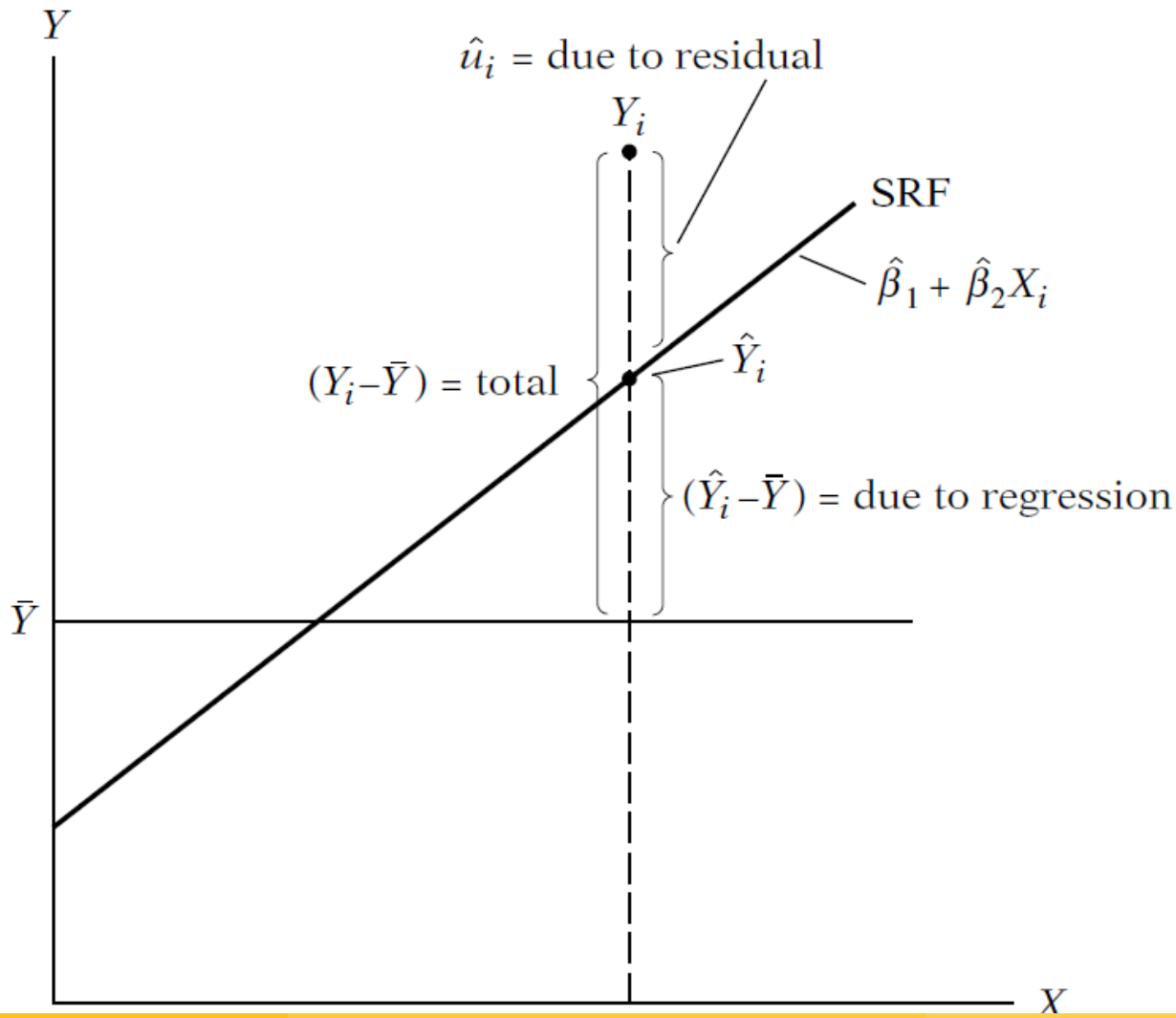
НИС: регрессионный анализ

Меры качества линейной регрессионной модели
Сравнение альтернативных спецификаций

19 ноября 2025

Разложение дисперсии в линейной регрессии: визуализация

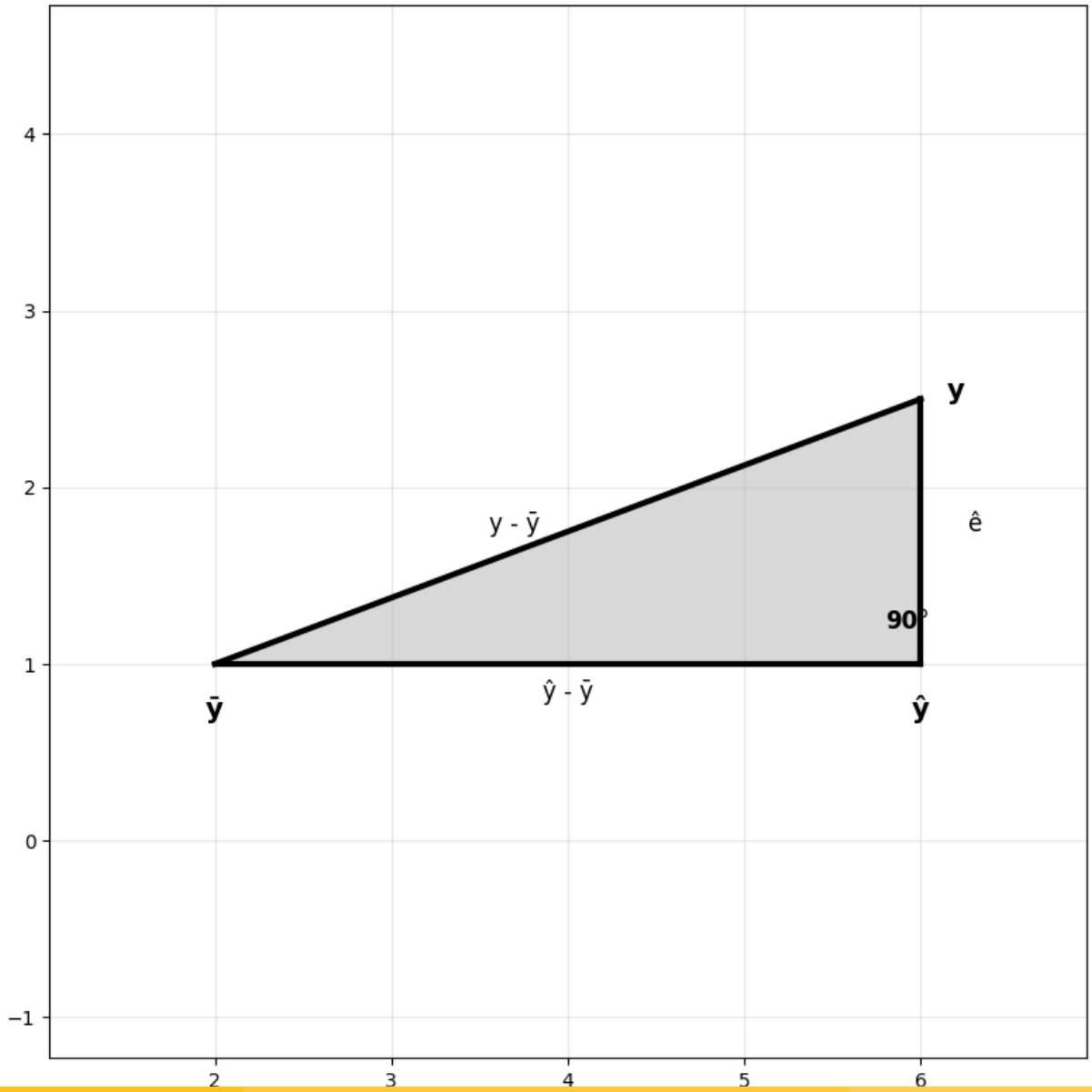


ANOVA-таблица

	df	SS	MSS	F
x	k	ESS	$\dfrac{\text{ESS}}{k}$	$\dfrac{\text{ESS}/k}{\text{RSS}/(n-k-1)}$
Residual	$n-k-1$	RSS	$\dfrac{\text{RSS}}{n-k-1}$	
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 $\vec{\hat{y}} - \bar{y} \times \vec{1}$

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

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

$$= cor(y_i, \hat{y}_i)$$

R^2 скорректированный

При добавлении новых параметров в модель, вне зависимости от того, значимы они или нет для объяснения дисперсии зависимой переменной, коэффициент детерминации R^2 будет увеличиваться. 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)
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 (при сравнении вложенных моделей выберем более экономную спецификацию)

F-тест для вложенных моделей

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?