Сравнение моделей

Сравнение вложенных моделей

Если нужно сравнить модели

$$y = \beta_1 + \beta_2 x^{(1)} + \beta_3 x^{(2)} + \beta_4 z^{(1)} + \beta_5 z^{(2)} + \varepsilon$$

и

$$y = \beta_1 + \beta_2 x^{(1)} + \beta_3 x^{(2)} + \varepsilon$$

То можно использовать тест для сравнения «короткой» и «длинной» регрессии

$H_0: \beta_4 = \beta_5 = 0$

Такие модели называются **вложенными**.

Сравнение НЕ вложенных моделей

(с одинаковой зависимой переменной)

Но что делать, если нужно сравнить модели

$$y = \beta_1 + \beta_2 x^{(1)} + \beta_3 x^{(2)} + \varepsilon \quad (\boldsymbol{A})$$

и

$$y = \beta_1 + \beta_2 z^{(1)} + \beta_3 z^{(2)} + \varepsilon \quad (\boldsymbol{B})$$

?

Такие модели называются **не вложенными** (nonnested)

Сравнение НЕ вложенных моделей

Оцениваем общую модель:

$$y = \beta_1 + \beta_2 x^{(1)} + \beta_3 x^{(2)} + \beta_4 z^{(1)} + \beta_5 z^{(2)} + \varepsilon$$

Проверяем **гипотезу *A***: $H_0: \beta_4 = \beta_5 = 0$

Проверяем **гипотезу *B***: $H_0: \beta_2 = \beta_3 = 0$

A принимается, ***B*** отклоняется => модель А лучше

B принимается, ***A*** отклоняется => модель В лучше

R^2 (сравниваем модели с одинаковой зависимой переменной)

- R^2 всегда растет при включении в модель дополнительных переменных!!! Поэтому лучше использовать R^2_{adj} (см. лекцию 4). Можно сравнивать вложенные модели.

- Если в двух моделях зависимые переменные разные, то их нельзя сравнивать, используя R^2!!!!

Информационные критерии (Акаике, Шварца).

Модель плохая если:

- Плохо предсказывает (*RSS* большой)
- Сложная (много коэффицентов, большое k)

Штрафуем модель за большое k и большую *RSS*

Информационные критерии:

- Акаике $AIC = n\ln(RSS/n) + 2k$
- Шварца $BIC = n\ln(RSS/n) + \ln(n)k$

Если мы хотим сравнить модели с зависимыми переменными y и ln(y)

- Для модели с ln(y) рассчитывается AIC’

$$AIC' = AIC + 2\sum_{t=1}^{T} \log y_t,$$

Если мы хотим сравнить модели с зависимыми переменными y и ln(y)

Обратимся к R. Для нашего примера мы будем использовать данные longley из пакета datasets. Для начала оценим две простые модели (аддитивную и мультипликативную):

```
modelAdditive <- lm(GNP~Employed,data=longley)
modelMultiplicative <- lm(log(GNP)~Employed,data=longley)
```

Теперь посмотрим на информационные критерии:

```
AIC(modelAdditive)
> 142.7824
AIC(modelMultiplicative)
> -44.5661
```

Как видим, значения не сравнимы. Скорректируем второй информационный критерий:

```
AIC(modelMultiplicative)+2*sum(log(longley$GNP))
> 145.118
```

Теперь стало намного лучше! Можем заключить, что по информационному критерию первая модель (аддитивная) лучше второй.