Прогнозы. Сравнение моделей

Эконометрика. Openedu. Неделя 4

Неделя 4

- Прогнозирование
- Выбор “наилучшей” модели

Прогнозирование

Модель: $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$
Точечный прогноз: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 z_i$

Интервальное прогнозирование

Что мы прогнозируем?

- Средний y_i при данных регрессорах, $E(y_i|x_i, z_i)$:

$E(y_i|x_i, z_i) = \beta_1 + \beta_2 x_i + \beta_3 z_i$

- Конкретный y_i при данных регрессорах:

$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$
Возникает две разных ошибки прогноза!

Ошибка прогнозирования среднего

- условное среднее, $E(y_i|X)$
- ошибка прогноза условного среднего, $\hat{y}_i - E(y_i|X)$
- дисперсия ошибки прогноза:

$$Var(\hat{y}_i - E(y_i|X)|X) = Var(\hat{y}_i|X) = Var(\hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 z_i|X)$$

Ошибка прогноза конкретного значения

- конкретное наблюдение, y_i
- ошибка прогноза, $\hat{y}_i - y_i$
- дисперсия ошибки прогноза:

$$Var(\hat{y}_i - y_i|X) = Var(\hat{y}_i - E(y_i|X) - \varepsilon_i|X) = Var(\hat{y}_i - \varepsilon_i|X) =$$
$$Var(\hat{y}_i|X) + Var(\varepsilon_i|X) = Var(\hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 z_i|X) + Var(\epsilon_i|X)$$

Оценка дисперсии

- $Var(\hat{y}_i|X)$, $Var(\varepsilon_i|X)$ неизвестны, зависят от σ^2
- $\widehat{Var}(\hat{y}_i|X)$, $\widehat{Var}(\varepsilon_i|X)$ известны
- Используем стандартные ошибки: $se(\hat{y}_i) = \sqrt{\widehat{Var}(\hat{y}_i|X)}$

Доверительный интервал для среднего значения

- Асимптотический: $\frac{\hat{y}_i - E(y_i|X)}{se(\hat{y}_i)} \to N(0,1)$

$$E(y_i|X) \in [\hat{y}_i - z_{cr} se(\hat{y}_i); \hat{y}_i + z_{cr} se(\hat{y}_i)]$$

- При предположении о нормальности: $\frac{\hat{y}_i - E(y_i|X)}{se(\hat{y}_i)} \sim t_{n-k}$

$$E(y_i|X) \in [\hat{y}_i - t_{cr} se(\hat{y}_i); \hat{y}_i + t_{cr} se(\hat{y}_i)]$$

Предиктивный интервал для конкретного значения

- Асимптотический: $\frac{\hat{y}_i - y_i}{se(\hat{y}_i - \varepsilon_i)} \to N(0,1)$

$$y_i \in [\hat{y}_i - z_{cr} se(\hat{y}_i - \varepsilon_i); \hat{y}_i + z_{cr} se(\hat{y}_i - \varepsilon_i)]$$

- При предположении о нормальности: $\frac{\hat{y}_i - y_i}{se(\hat{y}_i - \varepsilon_i)} \sim t_{n-k}$

$$y_i \in [\hat{y}_i - t_{cr} se(\hat{y}_i - \varepsilon_i); \hat{y}_i + t_{cr} se(\hat{y}_i - \varepsilon_i)]$$

Пример вычислений [у доски]

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.53539    0.05183  164.68   <2e-16 ***
log(carat)   1.74685    0.07505   23.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2771 on 38 degrees of freedom

vcov(mod)
            (Intercept)  log(carat)
(Intercept) 0.002686470 0.002078281
log(carat)  0.002078281 0.005632675
```

Логарифмирование

Четыре модели:

- $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
- $\ln(y_i) = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$
- $\ln(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_i$
- $y_i = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$

Вывод интерпретации [у доски]

Идея получения интерпретации логарифмических моделей:
$d \ln x = \frac{dx}{x}$ — изменение x в долях

Две популярные версии

- $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. С ростом x на единицу y растет на β_2 единиц.
- $\ln(y_i) = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$. С ростом x на один процент y растет на β_2 процентов.

Полулогарифмические модели

- $\ln(y_i) = \beta_1 + \beta_2 x_i + \varepsilon_i$. С ростом x на единицу y растет на $100\beta_2$ процентов.
- $y_i = \beta_1 + \beta_2 \ln(x_i) + \varepsilon_i$. С ростом x на один процент y растет на $0.01\beta_2$ единиц.

Дамми-переменные

- Объясняющая переменная, принимающая значение 0 или 1, называется дамми-переменной (dummy variable)
- Например, пол респондента в опросе, переменная $male_i$, равная 1 для мужчин и 0 — для женщин.

Дамми-переменные и разные зависимости на подвыборках

Пример 1. Базовая модель.
$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$
Зарплата мужчин и женщин в среднем одинаковая при равном опыте и образовании

Пример 2.

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \beta_4 male_i + \varepsilon_i$

Для мужчин: $wage_i = (\beta_1 + \beta_4) + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$

Для женщин: $wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$

Пример 3.

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \beta_4 male_i + \beta_5 male_i exper_i + \varepsilon_i$
Для мужчин: $wage_i = (\beta_1 + \beta_4) + (\beta_2 + \beta_5) exper_i + \beta_3 educ_i + +\varepsilon_i$
Для женщин: $wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$

Пример 4.

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \beta_4 male_i + \beta_5 male_i educ_i + \varepsilon_i$

Для мужчин: $wage_i = (\beta_1 + \beta_4) + \beta_2 exper_i + (\beta_3 + \beta_5) educ_i + \varepsilon_i$

Для женщин: $wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$

Пример 5.

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \beta_4 male_i + \beta_5 male_i educ_i + \beta_6 male_i exper_i + \varepsilon_i$

Для мужчин: $wage_i = (\beta_1 + \beta_4) + (\beta_2 + \beta_6) exper_i + (\beta_3 + \beta_5) educ_i + \varepsilon_i$

Для женщин: $wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$

Факторная переменная принимает несколько значений

$season_i \in \{$ зима , весна , лето , осень $\}$

1. Выбираем базовое значение факторной переменной: зима.
2. Вводим 3 (четыре сезона минус один базовый) дамми-переменных:

$vesna_i$, $leto_i$, $osen_i$

Вводим дамми-переменные

Наблюдение	Сезон	$vesna_i$	$leto_i$	$osen_i$
1	Зима	0	0	0
2	Весна	1	0	0
3	Лето	0	1	0
4	Осень	0	0	1
5	Зима	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Модели для подвыборок на примере

$icecream_i = \beta_1 + \beta_2 price_i + \beta_3 vesna_i + \beta_4 leto_i + \beta_5 osen_i + \varepsilon_i$
Зима: $icecream_i = \beta_1 + \beta_2 price_i + \varepsilon_i$
Весна: $icecream_i = (\beta_1 + \beta_3) + \beta_2 price_i + \varepsilon_i$
Лето: $icecream_i = (\beta_1 + \beta_4) + \beta_2 price_i + \varepsilon_i$
Осень: $icecream_i = (\beta_1 + \beta_5) + \beta_2 price_i + \varepsilon_i$

Частая ошибка!

Включить дамми-переменные на все значения факторной переменной и константу в регрессию. Благородные доны и дуэньи так не поступают!

Пример с ошибкой (!).

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 male_i + \beta_4 female_i + \varepsilon_i$

Выполнено соотношение $1 = male_i + female_i$.

частая ошибка — нарушение предпосылки

8. С вероятностью 1 среди регрессоров нет линейно зависимых
   - Синонимы в матричном виде: $rank(X) = k$ или $det(X'X) \neq 0$ или $(X'X)^{-1}$ существует

Регрессоры линейно зависимы. Не существует единственных оценок МНК.

Проверка гипотез о нескольких ограничениях сразу

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \beta_4 male_i + \beta_5 male_i educ_i + \varepsilon_i$

Для мужчин: $wage_i = (\beta_1 + \beta_4) + \beta_2 exper_i + (\beta_3 + \beta_5) educ_i + \varepsilon_i$

Для женщин: $wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$

$$H_0 : \begin{cases} \beta_4 = 0 \\ \beta_5 = 0 \end{cases}$$

H_a : хотя бы один коэффициент (β_4 или β_5) отличен от нуля

Проверка гипотезы

1. Оценить неограниченную модель (unrestricted, ur)

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \beta_4 male_i + \beta_5 male_i educ_i + \varepsilon_i$
Посчитать RSS_{UR}

2. Оценить ограниченную модель (restricted, r)

$wage_i = \beta_1 + \beta_2 exper_i + \beta_3 educ_i + \varepsilon_i$
Посчитать RSS_R

Два подхода:

3.1. Асимптотически:

$$\chi^2 = \frac{RSS_R - RSS_{UR}}{RSS_{UR}/(n - k_{UR})} \to \chi^2_r$$

3.2. При нормальности ошибок, $\varepsilon_i | X \sim N(0, \sigma^2)$

$$F = \frac{(RSS_R - RSS_{UR})/r}{RSS_{UR}/(n - k_{UR})} \sim F_{r, n-k_{UR}}$$

r — количество ограничений в H_0

Вывод:

4. Если $F_{obs} > F_{cr}$ или $\chi^2_{obs} > \chi^2_{cr}$, то H_0 отвергается

Пример [у доски]

Проверьте гипотезу о двух ограничениях. RSS для двух моделей:

```
model_1 <- lm(data = h, log(price) ~ log(totsp) + log(livesp) +
              log(kitsp) + brick + metrdist + walk)
model_2 <- lm(data = h, log(price) ~ log(totsp) + log(livesp) +
              log(kitsp)+brick)
deviance(model_1)
```

```
## [1] 62.56589
```

```
deviance(model_2)
```

```
## [1] 69.29502
```

Примечание

RSS ограниченной модели всегда больше:

- $RSS_{UR} = \min_{\hat{\beta}_1, \hat{\beta}_2, \ldots} RSS$
- $RSS_R = \min_{\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_4 = 0} RSS$

TSS в моделях равны, т.к. $TSS = \sum (y_i - \bar{y})^2$
Следовательно, $ESS_{UR} > ESS_R$ и $R^2_{UR} > R^2_R$.

Самый простой случай

Модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

Гипотеза H_0: все наши регрессоры абсолютно бесполезны

$$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \\ \ldots \end{cases}$$

Всего $(k-1)$ ограничение.

Гипотеза о незначимости регрессии.

Доказательство формулы статистики для гипотезы о незначимости регрессии [у доски]

Для гипотезы:

$$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \\ \ldots \\ \beta_k = 0 \end{cases}$$

Статистика приобретает вид:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F_{k-1,n-k}$$

Идея доказательства: ограниченной моделью будет модель $y_i = \beta_1 + \varepsilon_i$.
В ограниченной модели $\hat{\beta}_1 = \bar{y}$ и $RSS_R = TSS_R$, а $ESS_R = 0$.

Проверка гипотезы о незначимости регрессии

Модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \end{cases}$

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F_{k-1,n-k}$$

Пример проверки гипотезы о незначимости регрессии [у доски]

Для регрессии

$$\widehat{wage}_i = -2.5 + 0.6 school_i + 0.157 exper_i$$

Проверьте гипотезу о незначимости регрессии в целом.
$R^2 = 0.09$, $n = 3294$.

Снова БСХС — предпосылки

Если:

1. Истинная зависимость имеет вид $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$
   - В матричном виде: $y = X\beta + \varepsilon$
2. С помощью МНК оценивается регрессия y на константу, x_i, z_i
   - В матричном виде: $\hat{\beta} = (X'X)^{-1}X'y$
3. Наблюдений больше, чем оцениваемых коэффициентов β: $n > k$

БСХС — предположения на ε_i:

4. Строгая экзогенность: $E(\varepsilon_i|$ все регрессоры $) = 0$
   - В матричном виде: $E(\varepsilon_i|X) = 0$
5. Условная гомоскедастичность: $E(\varepsilon_i^2|$ все регрессоры $) = \sigma^2$
   - В матричном виде: $E(\varepsilon_i^2|X) = \sigma^2$
6. $Cov(\varepsilon_i, \varepsilon_j|X) = 0$ при $i \neq j$

БСХС — предпосылки на регрессоры

7. векторы отдельных наблюдений (x_i, z_i, y_i) — независимы и одинаково распределены
8. с вероятностью 1 среди регрессоров нет линейно зависимых

- Синонимы в матричном виде: $rank(X) = k$ или $det(X'X) \neq 0$ или $(X'X)^{-1}$ существует

БСХС — асимптотические свойства (плюс новое)

При $n \to \infty$:

- $\hat{\beta}_j \to \beta_j$ по вероятности
- $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \to N(0,1)$ по распределению
- $\hat{\sigma}^2 \to \sigma^2$ по вероятности
- новое: $\chi^2 = \frac{RSS_R - RSS_{UR}}{RSS_{UR}/(n-k_{UR})} \to \chi^2_r$

$\hat{\sigma}^2 = \frac{RSS}{n-k}$

БСХС — при нормальности

Если дополнительно известно, что $\varepsilon_i \sim N(0, \sigma^2)$:

- Оценки эффективны среди несмещенных
- $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} | X \sim t_{n-k}$, $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k}$
- $RSS/\sigma^2 | X \sim \chi^2_{n-k}$, $RSS/\sigma^2 \sim \chi^2_{n-k}$
- новое: $F = \frac{(RSS_R - RSS_{UR})/r}{RSS_{UR}/(n - k_{UR})} | X \sim F_{r, n-k_{UR}}$

Лишние переменные

- Истина: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
- Оценена регрессия: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 z_i$
- Потеряна: эффективность

Пропущенные переменные

- Истина: $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$
- Оценена регрессия: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$
- Всё плохо!

Мораль:

- Если в теории предполагается зависимость от переменной z, то её лучше включить в модель, даже если она не значима.
- Если переменные значимы, то их лучше оставить в модели, даже если теория говорит, что зависимости от них быть не должно.

Увидеть то, чего нет

- Как проверить не пропущены ли переменные, которых нет?
- RESET-тест Рамсея

$H_0 : y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$
H_a : Есть неизвестные нам пропущенные регрессоры

Алгоритм теста Рамсея:

1. Оценить модель: $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

Получить прогнозы $\hat{y}_i$

2. Оценить модель:
   $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \ldots + \gamma_p \hat{y}_i^{p+1} + \varepsilon_i$
3. Посчитать F-статистику проверяющую гипотезу о равенстве всех γ_i нулю.

Рамсей: при верной H_0 и нормальности остатков $F \sim F_{p,n-k-p}$

Пример теста Рамсея [у доски]

Для регрессии

$$\widehat{wage}_i = -2.5 + 0.6 school_i + 0.157 exper_i$$

Проверьте тест Рамсея, если:

- $R^2 = 0.091$ (в основной регрессии),
- Во вспомогательную регрессию Рамсея включены $\hat{y}_i^2$ и $\hat{y}_i^3$,
- $R^2_{aux} = 0.095$ (во вспомогательной регрессии Рамсея),
- $n = 3294$.

Простые показатели качества

1. R^2. Растет с добавлением регрессоров, $R^2_{ur} > R^2_r$
2. $R^2_{adj} = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2}{TSS/(n-1)}$

Чем больше R^2_{adj} тем меньше $\hat{\sigma}^2$.

Информационные критерии

Модель плохая если:

- плохо предсказывает (RSS большой)
- сложная (много коэффицентов, большое k)

3. Информационные критерии (размер штрафа):

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

Мораль

В этой лекции:

- Прогнозирование
- Гипотезы о нескольких ограничениях
- RESET-тест Рамсея

Далее: о неприятностях :)

Источники мудрости:

- Артамонов Н.В., Введение в эконометрику: главы 2.4, 2.5, 3.2
- Борзых Д.А., Демешев Б.Б. Эконометрика в задачах и упражнениях: глава 3
- Катышев П.К., Пересецкий А. А. Эконометрика. Начальный курс: главы 3.5
- Себер Дж., Линейный регрессионный анализ: главы 4.1, 5.1, 5.2, 5.3