

Темы из программы, выносимые на кр 20.03 и задачи по этим темам

Тема XVII

Мультиколлинеарность данных. Идеальная и практическая мультиколлинеарность (квазимультиколлинеарность). Теоретические последствия мультиколлинеарности для оценок параметров регрессионной модели. Нестабильность оценок параметров регрессии и их дисперсий при малых изменениях исходных данных в случае мультиколлинеарности. Признаки наличия мультиколлинеарности. Показатели степени мультиколлинеарности. Вспомогательные регрессии и показатель "вздутия" дисперсии (VIF). Индекс обусловленности информационной матрицы (bad conditioned index - BCI), как показатель степени мультиколлинеарности. Методы борьбы с мультиколлинеарностью. Переспецификация модели (функциональные преобразования переменных). Исключение объясняющей переменной, линейно связанной с остальными. Понятие о методе главных компонент как средстве борьбы с мультиколлинеарностью данных. Понятие о методе LASSO.

Задачи

1) Признаком мультиколлинеарности служит:

1. маленькие t -статистики при R^2 , близком к 1
2. близкое к 0 значение коэффициента множественной детерминации
3. значительные изменения в оценках коэффициентов регрессии при небольших изменениях в данных
4. близкие к 0 значения коэффициентов корреляции регрессоров
5. все ответы верны

2) Оцененная с помощью МНК зависимость заработной платы индивида EARNINGS от его возраста AGE, опыта EXP, пола MALE, длительности обучения S, длительности обучения матери SM имеет вид (в скобках стандартные отклонения коэффициентов):

$$\widehat{EARN} = -24 - 0.099 AGE + 2.49 S + 0.26 SM + 0.46 EXP + 6.23 MALE, R^2 = 0.247$$

(10.6) (0.25) (0.249) (0.24) (0.14) (1.11)

Были оценены также вспомогательные регрессии:

$$\widehat{AGE} = -.007 + 0.53 AGE - 0.6 S + 0.23 SM + 1.23 MALE, R^2 = 0.2,$$

$$\widehat{S} = 8.47 + 0.095 AGE + 0.4 SM - 0.2 EXP + 0.12 MALE, R^2 = 0.25,$$

$$\widehat{SM} = 6.16 - 0.045 AGE + 0.42 S + 0.08 EXP + 0.42 MALE, R^2 = 0.18,$$

$$\widehat{EXP} = -.07 + 0.53 AGE - 0.6 S + 0.23 SM + 1.23 MALE, R^2 = 0.2,$$

VIF для переменной EXP равен ____.

Ответ. 1.25

3) При применении к модели, результаты оценки которой приведены ниже,

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
S	2.578227	.2288185	11.27	0.000	2.128729	3.027726
AGE	-10.70493	9.211662	-1.16	0.246	-28.80062	7.390769
Agesq	.1300605	.1125515	1.16	0.248	-.0910395	.3511605
EXP	.4429137	.1442633	3.07	0.002	.159518	.7263094
ETHHISP	-1.078255	2.268688	-0.48	0.635	-5.534941	3.378432
ETHBLACK	-4.014172	2.152185	-1.87	0.063	-8.241996	.2136528
MALE	6.364055	1.111968	5.72	0.000	4.179668	8.548442
_cons	193.7202	187.6859	1.03	0.302	-174.9761	562.4165

. vif

Variable	VIF	1/VIF
AGE	1411.96	0.000708
Agesq	1411.13	0.000709
EXP	1.29	0.778114
S	1.14	0.875122
ETHBLACK	1.04	0.962602
MALE	1.03	0.966488
ETHHISP	1.02	0.983851
Mean VIF	404.09	

метода последовательного исключения, на ближайшем шаге из уравнения регрессии будет удалена переменная

1) S 2) AGE 3) EXPSQ 4) EXP 5) ETHWHITE 6) ETHHISP 7) FEMALE 8) ни одна из перечисленных

4) Первой главной компонентой системы показателей $X_1, \dots, X_k$ называется такая линейная комбинация этих показателей

1. в которой коэффициент при X_1 равен 1 2. которая обладает наименьшей дисперсией 3. которая обладает наибольшей дисперсией 4. которая ортогональна всем $X_j, j = 1, \dots, k$

5) (Д.А.Борзых, Б.Б.Демешев, часть задачи 7.4)

Пионеры, Крокодил Гена и Чебурашка собирали металлолом несколько дней подряд. В распоряжение иностранной шпионки, гражданки Шапокляк, попали ежедневные данные по количеству собранного металлолома: вектор g – для Крокодила Гены, вектор h – для Чебурашки и вектор x – для пионеров. Гена и Чебурашка собирали вместе, поэтому выборочная корреляция $\hat{c}or(g, h) = -0.9$. Гена и Чебурашка собирали независимо от пионеров, поэтому $\hat{c}or(g, x) = 0$, $\hat{c}or(h, x) = 0$. Если регрессоры g , h , x центрировать и нормировать, то получится матрица $\tilde{X}$. Вычислите одну или две главные компоненты (выразите их через вектор-столбцы матрицы $\tilde{X}$), объясняющие не менее 70% общей выборочной дисперсии регрессоров.

Ответ. $(\tilde{X}_1 - \tilde{X}_2)/\sqrt{2}$; $\tilde{X}_3$

Тема XVIII

Типы ошибок спецификации модели. Пропущенные и излишние переменные. Неправильная функциональная форма модели. Смещение в оценках коэффициентов, вызываемое невключением существенных переменных.

Ухудшение точности оценок (увеличение оценок дисперсий) при включении в модель излишних переменных. Проверка гипотезы о группе излишних переменных. RESET тест Рамсея (Ramsey's RESET test) для проверки гипотезы о существовании упущенных переменных.

- 1) По 150 наблюдениям оценили зависимость почасовой заработной платы от пола (переменная MALE равно 1 для мужчин и 0 для женщин), длительности обучения S и возраста AGE.

$$\hat{Y} = 3.6 + 3.5 MALE + 3.24 S + 0.44 AGE, \text{ RSS} = 7632$$

(3.09) (1.21) (0.53) (0.057)

Используя результаты двух вспомогательных регрессий, приведенных ниже, проведите RESET – тест и ответьте, правильная ли спецификация модели выбрана.

$$\hat{Y} = 12.37 - 0.29 MALE - 0.49 S - 0.08 AGE + 0.0064 \hat{Y}^2, \text{ RSS} = 7154$$

(4.09) (1.7) (1.3) (0.17) (0.002)

$$\hat{Y} = -18.1 + 9.2 MALE + 7.93 S + 1.1 AGE - 0.012 \hat{Y}^2 + 1.75 \cdot 10^{-10} \hat{Y}^3, \text{ RSS} = 6069$$

(4.42) (2.44) (2.05) (0.28) (0.004) (3.45 \cdot 10^{-11})

Ответ. Неправильная

Тема XIX

Метод максимального правдоподобия. Свойства оценок метода максимального правдоподобия. Соотношение между оценками коэффициентов линейной регрессии, полученными методом максимального правдоподобия и методом наименьших квадратов в случае нормально распределенной случайной составляющей. Свойства оценки дисперсии случайной составляющей, полученной методом максимального правдоподобия. Проверка гипотез с помощью теста Вальда, теста отношения правдоподобия, теста множителей Лагранжа

Задачи

- 1) Логарифмическая функция правдоподобия, используемая для оценивания классической линейной регрессионной модели, имеет вид:

$$1) \ln L(\beta, \sigma_\varepsilon^2) = \frac{n}{2} \ln \sqrt{2\pi} - \frac{n}{2} \ln \sigma_\varepsilon^2 - \frac{1}{2} \frac{(Y - X\beta)'(Y - X\beta)}{\sigma_\varepsilon^2};$$

$$2) \ln L(\beta, \sigma_\varepsilon^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma_\varepsilon^2 - \frac{1}{2} \frac{\varepsilon' \varepsilon}{\sigma_\varepsilon^2};$$

$$3) \ln L(\beta, \sigma_\varepsilon^2) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (Y - X\beta)' \Sigma^{-1} (Y - X\beta);$$

$$4) \ln L(\beta, \sigma_\varepsilon^2) = \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} (Y - X\beta)' \Sigma (Y - X\beta);$$

5) ни один из вышеперечисленных.

- 2) Оценки метода максимального правдоподобия:

$$\left(\frac{\hat{C}}{GNP}\right)_t = \frac{25.92}{(2.22)} \frac{1}{GNP_t} + \frac{0.6246}{(0.00597)} - \frac{0.4315}{(0.0597)} \left(\frac{D}{GNP}\right)_t,$$

где C – агрегированные частные потребительские расходы,
 GNP – ВНР,

D – национальные расходы на оборону.

С какой целью оценили второе уравнение? Какое при этом было сделано предположение о дисперсии ошибок?

Ответ. $\sigma_i^2 \sim GNP_i^2$

3) По данным для 45 стран исследователь оценил зависимость инвестиций I от государственных расходов G и валового внутреннего продукта Y (все переменные измеряются в миллиардах долларов США):

$$\hat{I} = 18.1 - 1.07G + 0.36Y, \quad R^2 = 0.98$$

(7.79) (0.14) (0.02)

Исследователь упорядочил наблюдения по увеличению Y и оценил регрессии снова для 15 стран с наименьшим Y и 15 стран с наибольшим Y . Величины RSS для этих регрессий равны 421 и 3219 соответственно. Тестируйте модель на наличие пропорциональной формы гетероскедастичности.

4) По данным для 20 стран были оценены коэффициенты уравнения регрессии

$$\hat{Y}_i = 111.78 - 0.0042 X_{2i} - 0.4898 X_{3i} \quad R^2 = 0.492$$

$t=4.79$ $t=-2.53$ $t=-1.71$

где Y – младенческая смертность (количество в расчете на тысячу рожденных живыми),

X_2 – GNP в расчете на душу населения,

X_3 – процент имеющих начальное образование в определенной возрастной группе.

При проведении теста Уайта была оценена регрессия

$$e_i^2 = 4987 - 0.4718 X_{2i} - 0.8442 X_{3i} + 0.00001 X_{2i}^2 + 0.4435 X_{3i}^2 + 0.0026 X_{2i} X_{3i} \quad R^2 = 0.649$$

$t=4.86$ $t=-0.59$ $t=-2.45$ $t=1.27$ $t=1.62$ $t=0.35$

Проведя тест Уайта, проверьте гипотезу об отсутствии гетероскедастичности.

Решение (верное!):

Рассчитаем тестовую статистику Уайта:

$\chi^2 = n \cdot R^2 = 20 \cdot 0.649 = 12.98$, где R^2 используется из вспомогательной регрессии.

Поскольку $\chi_{0.05, cr}^2(5) = 11.07$, $\chi^2 > \chi_{0.05, cr}^2$, ($5 = 6 - 1$, 6 – число оцениваемых коэффициентов во вспомогательной регрессии) то для модели существует проблема гетероскедастичности возмущений.

5) Предположим, что для модели парной регрессии $\sigma_{ui}^2 = \sigma^2 X_i^4$.

Как избавиться от проблемы гетероскедастичности ошибок?

Решение:

Дисперсия ошибок пропорциональна X_i^4 , а следовательно стандартная ошибка пропорциональна X_i^2 ($\sigma_{ui}^2 \sim X_i^4 \Rightarrow \sigma_{ui} \sim X_i^2$). Решением проблемы будет замена переменных:

$$Y^* = \begin{pmatrix} \frac{Y_1}{X_1^2} \\ \vdots \\ \frac{Y_n}{X_n^2} \end{pmatrix}; \quad 1^* = \begin{pmatrix} \frac{1}{X_1^2} \\ \vdots \\ \frac{1}{X_n^2} \end{pmatrix}; \quad X^* = \begin{pmatrix} \frac{1}{X_1} \\ \vdots \\ \frac{1}{X_n} \end{pmatrix},$$

т.е. необходимо оценить модель

$$Y_i^* = \beta_1^* 1^* + \beta_2^* X^* + u_i^*$$

или

$$\frac{Y}{X^2} = \tilde{\beta}_1 \frac{1}{X^2} + \tilde{\beta}_2 \frac{1}{X} + \tilde{u}_i.$$

6) Тестом, который не только позволяет выявить наличие гетероскедастичности, но и указать способ оценивания параметров σ_i^2 , является

1) тест Уайта 2) тест Глейзера 3) тест Рамсея 4) тест Хаусмана

7) Оценки метода наименьших квадратов коэффициентов регрессии :

$Y = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k + u$ останутся несмещенными при нарушении условий теоремы Гаусса – Маркова

1) $\text{var}(u_i) = \sigma^2$ при всех i

2) $\text{cov}(u_i; u_j) = 0$; $i \neq j$,

3) состоящих во включении в модель лишнего объясняющего фактора Z ,

4) состоящих в невключении в модель необходимого фактора

Тема XXI

Взвешенный метод наименьших квадратов. Обобщенный метод наименьших квадратов

Задачи

1) FGLS-оценка коэффициентов линейной регрессионной модели

$$Y = X\beta + \varepsilon, \quad \hat{V}[\varepsilon] = \Sigma$$

имеет вид:

1) $(X'X)^{-1}(X'Y)$

2) $(X'\hat{\Sigma}^{-1}X)^{-1}(X'Y)$

3) $(X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}Y)$

4) $(X'\hat{\Sigma}^{-1}X)^{-1}(X'\hat{\Sigma}^{-1}Y)$

5) нет верного ответа

2) (Д.А.Борзых, Б.Б.Демешев, задача 8.26)

Пусть $Y_i = \beta X_i + \varepsilon_i$, где $E(\varepsilon_i) = 0$ и известно, что оценка для параметра

$\tilde{\beta} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}$ является наилучшей (в смысле минимума дисперсии) среди всех

линейных несмещенных оценок параметра β . Чему равна в этом случае матрица ковариаций вектора ε с точностью до пропорциональности?

Ответ.

$$V[\varepsilon] = \sigma^2 \begin{pmatrix} X_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & X_n \end{pmatrix}$$