Лабороторная работа 3

Зраева Екатерина

28 Декабрь 2020

В третьей лабораторной мы продолжим работу с примером данных и проверим, удовлетворяют ли остатки четырёх построенных моделей условиям Гаусса-Маркова. Рассматриваются модели:

1. ()
2. ()

Показатели:

* y - ВРП 2015
* x1 - Инвестиции в основной капитал 2014
* x2 - Расходы консолидированных бюджетов субъектов Российской Федерации: на национальную экономику 2014
* x3 - Использование информационных и коммуникационных технологий в организациях: персональные компьютеры 2014
* x4 - Внутренние затраты на научные исследования и разработки 2014

Количество наблюдений: 93

В лабораторной работе рассматривается тесты остатков линейных регрессионных моделей в R. Исходные данные для работы хранятся в файле Example\_model.RData.Построенные нами модели хранятся в списке models.list

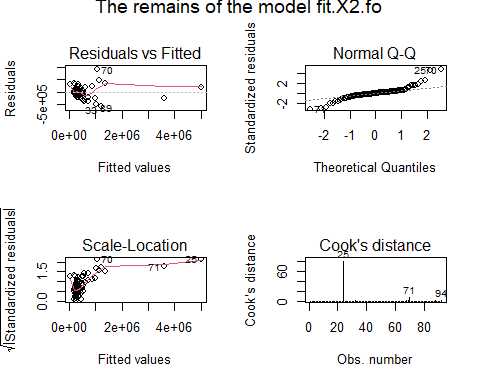
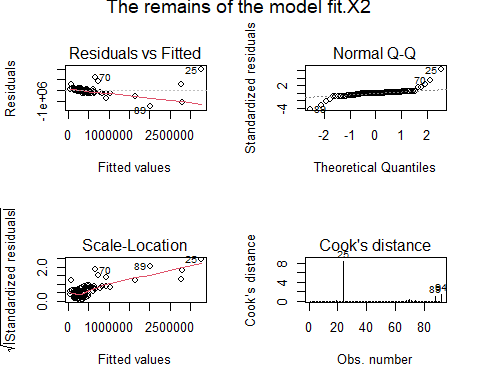
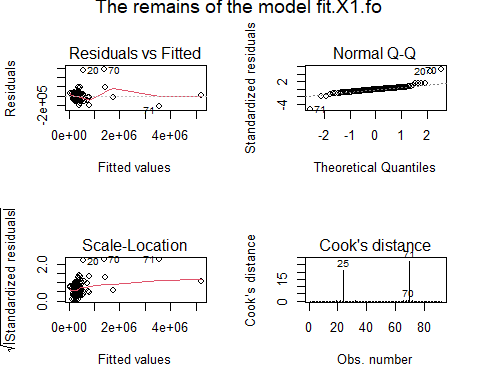
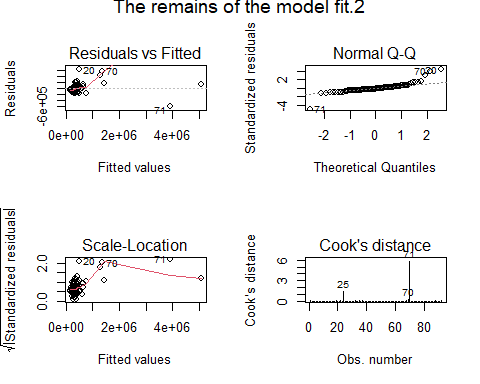
Нам необходимо исследовать остатки этих моделей и определить, не нарушаются ли условия Гаусса-Маркова:

1. Равенство среднего остатков нулю.
2. Постоянство дисперсии остатков (гомоскедастичность).
3. Отсутствие автокорреляции (независимость). Кроме того, модели множественной регрессии тестируют на отсутствие мультиколлинеарности факторов (их значимой взаимосвязи друг с другом).

Графики остатков: “Residuals vs Fitted” и другие:

Представим графически остатки четырёх моделей. Функция plot() от аргумента типа «модель регрессии», построенного функцией lm(), строит шесть графиков остатков (значение второго аргумента – номер графика):

1. «Residuals versus Fitted» – график зависимости остатков модели (по вертикали) от модельных значений (по горизонтали). При наличии явных нарушений условий Гаусса-Маркова они проявятся на этом графике.
2. «Normal Q-Q» – график «квантиль-квантиль», который сопоставляет квантили фактического распределения стандартизованных остатков (по вертикали) с соответствующими теоретическими значениями квантилей нормального распределения (по горизонтали). Чем ближе все точки в прямой, тем «нормальнее» распределение остатков.
3. «Scale-Location» – график, на котором по вертикали отложены корни из стандартизованных значений остатков, а по горизонтали – модельные значения. Этот график служит для поиска гетероскедастичности (непостоянства дисперсии) остатков.
4. «Cook’s Distance» – график расстояний Кука. С помощью расстояния Кука оценивают влияние отдельного наблюдения выборки на модель регрессии. Эта величина показывает разницу между вычисленными коэффициентами уравнения регрессии и значениями, которые получились бы при исключении соответствующего наблюдения

Графики - 1 для моделей fit.2; fit.X1.fo; fit.X2; fit.X2.fo; 

На графиках остатков не видно явных ошибок спецификации. В последней модели обнаруживаются два влияющих наблюдения: регионы с номерами 71 и 25. Это Ненецкий AO и Ямало-Ненецкий АО.Чтобы понять, нужно ли исключать эти наблюдения, нужно установить, насколько сильно они отличаются от всех остальных.

Можно использовать тот факт, что расстояние Кука распределено по закону Фишера и проверить гипотезу:

H0: отличие i-го наблюдения от остальных статистически незначимо.

H1: i-е наблюдение значимо отличается от остальных (и его нужно убрать как влияющее на модель).

Для проверки гипотезы достаточно сравнить расстояния Кука для влияющих наблюдений с квантилем F-распределения, соответствующим заданному уровню значимости. Если расстояния меньше квантиля (критической границы), нулевая гипотеза не отклоняется.

Поскольку расстояния Кука превосходят критическую границу (и, следовательно, соответствующие p-значения меньше 0,05), отвергаем нулевые гипотезы для регионов 25 и 71: их стоит убирать из выборки для построения модели.

Проверим гипотезу о равенстве среднего остатков модели нулю. Эта гипотеза проверяется с помощью t-критерия.

H0: среднее остатков равно 0 в генеральной совокупности.

H1: среднее остатков не равно 0 в генеральной совокупности.

Если p-значение для статистики теста больше уровня значимости (0,05), нулевая гипотеза не отклоняется.

Результатами теста на равенство среднего остатков моделей нулю:

P-value = 1

|  |
| --- |
| x |
| 1 |

Для первой модели нулевая гипотеза не отклоняется. Несложно убедиться, что остатки других моделей также проходят этот тест.

Проверим остатки на постоянство дисперсии тестом Бройша-Пагана

H0: остатки модели гомоскедастичны.

H1: в остатках модели есть гетероскедастичность.

Если p-значение для статистики теста больше уровня значимости (0,05), нулевая гипотеза не отклоняется.

Таблицы с результатами тестов на гетеросткедастичность остатков моделей

P-value test Breush-pagan

|  |  |
| --- | --- |
|  | x |
| BP | 0.0093386 |

Результат первого тест говорит о том, что в остатках первой модели есть гетероскедастичности.

Протестируем остатки моделей на автокорреляцию. Все модели содержат константу, поэтому можно проверить остатки на наличие автокорреляции первого порядка с помощью статистики Дарбина-Уотсона

H0: в остатках отсутствует автокорреляция первого порядка.

H1: автокорреляция первого порядка в остатках есть

Если p-значение > 0,05, нулевая гипотеза не отклоняется.

Таблица с результатами теста на автокорреляцию в остатках моделей

P-value DWtest

|  |
| --- |
| x |
| 0.1788935 |

Таким образом, в остатках моделей автокорреляция отсутствует.

Для проверки модели на мультиколлинеарность используются VIFкоэффициенты, которые в R рассчитывает одноимённая функция vif().

VIF-коэффициенты и выводы по ним

VIF-тест

|  |  |
| --- | --- |
|  | x |
| x1 | 28.08 |
| FOСЗФО | 1.59 |
| FOУФО | 1.92 |
| x1.FOСЗФО | 19.62 |
| x1.FOУФО | 14.66 |
| x1.FOЮФО | 1.03 |

Поскольку не все значения VIF-коэффициентов близки к единице, в модели присутствует мультиколлинеарности факторов (из-за того, что среди факторов всего один количественный показатель, а все остальные – фиктивные переменные, высокая мультиколлинеарность в некоторых случаях неизбежна).

Выводы об остатках лучшей модели, проверка условий Гаусса-Маркова