Семинар
Гетероскедастичность и нетипичные наблюдения

Задание 1. После оценивания линейной регрессионной модели y на предикторы x, z, d и w по 550 наблюдениям в критерии Уайта, использующемся для проверки гетероскедастичности, было получено значение R^2 из вспомогательной регрессии равное 0.4. Сформулируйте нулевую и альтернативную гипотезы, запишите спецификацию дополнительной (вспомогательной) модели, рассчитайте статистику критерия и сделайте вывод на основании p-value.

Решение

Рассчитаем статистику критерия:

$$n^2 = 550 \times 0.4 = 220$$

Данная статистика критерия при верной H_0 имеет распределение хи-квадрат с $df = k$, где k - это количество предикторов во вспомогательной модели. В данном случае таких предикторов у нас 14, из них – 4 исходных предиктора, 4 предикторов в квадрате, 6 попарных произведения.

Далее посчитаем вероятность того, что статистика критерия превысит значение 220, то есть, p-value. Такая вероятность примерно равна 0, следовательно, отвергаем нулевую гипотезу о гомоскедастичности.

Задание 2. По данным 120 стран исследователь оценил зависимость Y от двух объясняющих переменных: X и Z. Исследователь предполагает, что вариация ошибок зависит пропорционально от X. Чтобы проверить это предположение, он упорядочил наблюдения по возрастанию X, исключил из середины выборки 30% наблюдений и оценил регрессии на первом и втором сегментах выборки с наименьшим и наибольшим X. Величины RSS для этих регрессий равны 2500 и 2125 соответственно. Рассчитайте статистику критерия Голдфелда-Квандта и соответствующее значение p-value. Какой вывод сделает исследователь на основе представленных результатов?

Решение

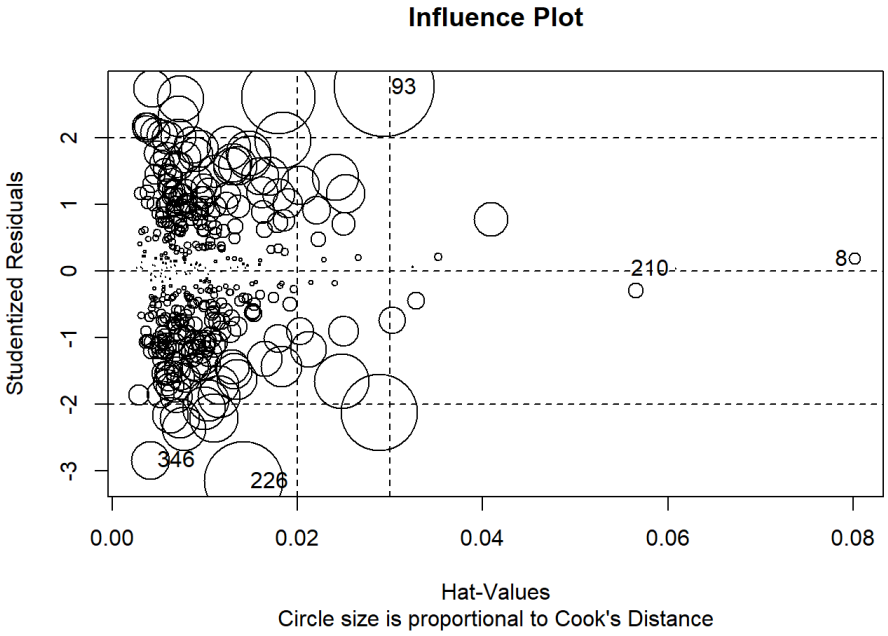
После исключения 30% наблюдений в массиве осталось 84 наблюдения. Значит, в первом и втором сегментах выборки – по 42 наблюдения.

Рассчитаем статистику критерия:

$$\frac{\frac{RSS_1}{n_1-k-1}}{\frac{RSS_2}{n_1-k-1}} = \frac{\frac{2500}{42-2-1}}{\frac{2125}{42-2-1}} \approx 1.176$$

Руководствуясь тем, что статистика критерия имеет при верной H_0 распределение Фишера с df1 = 39; df2 = 39, посчитаем вероятность того, что статистика критерия превысит значение 1.176. Так как p-value составляет примерно 0.3, у нас нет оснований отвергнуть гипотезу о гомоскедастичности. Но при этом не стоит забывать, что тест Голдфелда–Квандта проверяет именно определенный паттерн гетероскедастичности: а именно, что вариация ошибок монотонно возрастает с ростом значений по объясняющей переменной.

Задание 3. Проинтерпретируйте график ниже с точки зрения наличия нетипичных и влиятельных наблюдений.



Ответ

1. Мы можем диагностировать нетипичные значения по X на основе элементов по главной диагонали матрицы-проекции, то есть, по hat-values (на графике - ось абсцисс). Как мы видим, достаточно сильно от среднего значения по X отклоняются наблюдения 210 и 8

2. Нетипичные значения по зависимой переменной мы определяем по стьюдентизированным остаткам. На графике видно, что наблюдения 346 и 226 имеют значения остатков больше 3 по модулю

3. Влиятельные наблюдения могут быть определены при помощи меры Кука. На графике чем больше радиус окружности, тем более влиятельное наблюдение. К примеру, достаточно большой радиус окружности мы видим у 93 наблюдения

Задание 4. По указанным ниже значениям предиктора и отклика (x и y соответственно) в регрессионной модели y на x найдите матрицу проекции (hat matrix). Есть ли в данном случае нетипичные наблюдения по x?

x	2	1	0	-1
y	1	0	0	-3

Решение

$$H = X(X^TX)^{-1}X^T$$

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{pmatrix}$$

$$X^TX = \begin{pmatrix} 4 & 2 \\ 2 & 6 \end{pmatrix}$$

$$(X^TX)^{-1} = \begin{pmatrix} 0.3 & -0.1 \\ -0.1 & 0.2 \end{pmatrix}$$

$$X(X^TX)^{-1} = \begin{pmatrix} 0.1 & 0.3 \\ 0.2 & 0.1 \\ 0.3 & -0.1 \\ 0.4 & -0.3 \end{pmatrix}$$

$$X(X^TX)^{-1}X^T = \begin{pmatrix} 0.7 & 0.4 & 0.1 & -0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.3 & 0.4 \\ -0.2 & 0.1 & 0.4 & 0.7 \end{pmatrix}$$

Определяем порог для нетипичных наблюдений: $\frac{3p}{n} = \frac{3 \times 2}{4} = 1.5$. Нетипичных наблюдений по X в данном случае нет, но принимаем во внимание маленький размер выборки.