Семинар №8.

Мультиколлинеарность. Метод главных компонент.

1. Теоретическая регрессионная зависимость и корреляционная матрица регрессоров имеют вид: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$, $Corr(X) = \begin{pmatrix} 1 & r & 0 \\ r & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, где $r=0.95$.
   a. Найдите параметр обусловленности для матрицы $(\tilde{x}'\tilde{x})$, где $\tilde{x}$ - матрица центрированных и нормированных значений регрессоров.
   b. Вычислить одну или две главные компоненты (т.е. выразить их через линейные комбинации столбцов $\tilde{x}$), объясняющие не менее 70% общей дисперсии.
   c. Выразить коэффициенты исходной регрессии через коэффициенты регрессии на главные компоненты, объясняющие не менее 70% общей дисперсии.

Задание для выполнения на компьютерах. Часть 1.

1. Откройте файл **housing.dta.** Описание переменных содержится в файле **housing.txt.** Посмотрите на описательные статистики переменных.

2. Постройте уравнение регрессии вида: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{11} X_{11} + \varepsilon$, где Y -цена продажи дома (в канадских долларах), $X_1, ..., X_{11}$ все остальные переменные, которые есть в файле.

3. Проверьте адекватность регрессии в целом и значимость коэффициентов по отдельности. Дайте экономическую интерпретацию полученным результатам.

4. Попробуйте исключить из модели какую-нибудь значимую переменную. Что изменилось? Что произошло с коэффициентами модели?

5. Проверьте гипотезу о совместной незначимости группы переменных.

6. Исключите из модели незначимые переменные. Что изменилось?

7. Проведите тест Рамсея для проверки гипотезы о существовании упущенных переменных для вашей модели. Сделайте вывод.

8. Создайте логарифм переменной price (цена продажи дома). Оцените полулогарифмическую модель. $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{11} X_{11} + \varepsilon$. Проинтерпретируйте полученные результаты. Что показывают коэффициенты в такой модели?

9. На основании теста Бокса-Кокса сделайте вывод о том, какая модель лучше: линейная или полулогарифмическая.

10. Теперь на основании теста Дэвидсона-Маккиннена (Davidson, R., and J. G. MacKinnon) сделайте вывод о том, какая модель лучше: линейная или полулогарифмическая.

11. Постройте прогноз $\hat{Y}$ для выбранной вами модели. Получите ряд остатков. Постройте графики для прогноза и для остатков. Какой можно сделать вывод на основании этих графиков.

12. Протестируйте наличие мультиколлинеарности в выбранной регрессии. Примите какие-то действия на основании сделанных результатов.

Задание для выполнения на компьютерах. Часть 2.

1. Сгенерируйте три зависимые нормально распределенные случайные величины. Число наблюдений равно 50. Назовите их x_1, x_2 и x_3. Со следующей ковариационной матрицей:

$$\text{var}[X] = \begin{pmatrix} 10 & 7 & 5 \\ 7 & 6 & 4 \\ 5 & 4 & 2.73 \end{pmatrix}$$

2. Сгенерируйте стандартную нормальную случайную величину. Назовите её ε (epsilon).
3. Постройте гистограмму распределения для ε.
4. Создайте переменную $y = -5 + 3x_1 - 8x_2 + \varepsilon$.
5. Найдите описательные статистики переменных x_1, x_2, x_3 и y. Постройте для них корреляционную матрицу. Найдите определитель этой матрицы.
6. Оцените линейную регрессию $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ или в матричной форме $Y = X\beta + \varepsilon$.
7. Прокомментировать содержимое таблицы результатов.
8. Рассчитайте значения VIF.
9. Выведите на экран оценку ковариационной матрицы коэффициентов регрессии.
10. Исключите из регрессии сильно коррелируемые регрессоры и оцените её. Что изменилось?
11. Найдите главные компоненты. Оцените регрессию на главных компонентах. Как можно проинтерпретировать полученный результат?
12. Сгенерируйте снова три зависимые нормально распределенные случайные величины x_1, x_2 и x_3 с той же ковариационной матрицей, как и в пункте 1, однако число наблюдений теперь пусть будет равно 1500.

13. Проделайте шаги со второго по 10. Какие можно сделать выводы относительно влияния мультиколлинеарности в модели.