

Эконометрика, 2017-2018, 3 модуль

Семинар 1

15.01.18 для

Группы Э_Б2015_Э_3

Семинарист О.А.Демидова

1) Упражнение 9.5. (Учебник Демидова, Малахов)

В файле `rlms14.dta` приведены данные о заработной плате, возрасте, образовании, профессиональной принадлежности россиян из базы данных RLMS, 14 раунд опроса, проводившийся в 2005 г. (описание переменных дано в Приложении 1).

1) Оцените зависимость заработной платы от возраста, продолжительности рабочей недели, длительность работы на последнем месте. Выберите в качестве зависимой переменной `income` или `wage`, а в качестве независимых – возраст (переменную необходимо предварительно рассчитать по формуле $age = 2005 - birth_year$), длительность работы на последнем месте (переменную необходимо предварительно рассчитать по формуле $tenure = 2005 - beginning$), и т.д.

2) Адекватна ли оцененная регрессия? Все ли оценки коэффициентов имеют ожидаемый знак? Правильно ли выбрана спецификация модели, исходя из полученных результатов?

3) Проведите тест Рамсея. Если гипотеза H_0 о правильной спецификации модели будет отвергнута, переходите к следующему пункту.

4) Оцените регрессию снова, добавив в уравнение регрессии квадрат одной из независимых переменных, например, возраста. Если новая оцененная регрессия является адекватной, дайте экономическую интерпретацию полученным результатам.

5) Поэкспериментируйте с включением в модель других переменных.

Методические рекомендации

1) Создайте новые переменные `age` и `tenure` с помощью команд:

```
gen age = 2005 - birth_year,  
gen tenure=2005 - beginning
```

2) Оцените коэффициенты уравнения регрессии $income = \beta_0 + \beta_1 age + \beta_2 tenure + u$,

набрав в командном окне

```
reg income age tenure
```

3) Проведите тест Рамсея с помощью команды

```
ovtest
```

Мультиколлинеарность. Метод пошагового выбора независимых переменных

Для регрессии $Y = \beta_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$ имеет место проблема мультиколлинеарности, если между факторами X_1, \dots, X_k существует линейная или близкая к линейной зависимость.

Признаки мультиколлинеарности:

- Адекватность регрессии в целом при незначимости коэффициентов при многих факторах
- неадекватные знаки и величина коэффициентов;
- неустойчивость оценок коэффициентов по отношению к увеличению объема выборки или добавлению новых регрессоров.

Последствия мультиколлинеарности:

- высокие стандартные ошибки оценок коэффициентов регрессии;
- возможная незначимость коэффициентов регрессии;
- невозможность оценить влияние регрессоров по отдельности.

Способы диагностики мультиколлинеарности:

- изучение корреляционной матрицы регрессоров, если коэффициент корреляции двух факторов близок к 1, то это свидетельствует о наличии мультиколлинеарности;

- вычисление параметра обусловленности матрицы $X'X$: $CI = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$ (где $\lambda_{\max}, \lambda_{\min}$ -

соответственно наибольшее и наименьшее собственные значения матрицы $X'X$, CI – conditional index), если значение CI достаточно велико (больше 4), то это свидетельствует о наличии мультиколлинеарности;

- вычисление VIF_j (Variance inflation factor), $j \in \{2, \dots, k\}$ по формуле:

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j \in \{2, \dots, k\}, \quad \text{где } R_j^2 - \text{коэффициент множественной детерминации в}$$

регрессии фактора X_j на все остальные факторы. Если хотя бы один из VIF_j , $j \in \{2, \dots, k\}$ достаточно велик (больше 8), то это свидетельствует о наличии мультиколлинеарности.

Некоторые способы устранения мультиколлинеарности:

- добавление новых наблюдений;
- исключение некоторых регрессоров;
- выбор другой функциональной формы модели.

2) Проблемы мультиколлинеарности при моделировании продаж одежды

В файле clothing (STATA) содержатся данные о продажах одежды в 400 немецких магазинах одежды.

Переменные:

tsales – среднегодовые продажи в гюльденах,

sales - продажи в расчете на квадратный метр,
 margin – маржинальная валовая прибыль,
 nown – количество собственников (менеджеров),
 nfull – количество полностью занятых,
 npart - количество частично занятых,
 naux – количество временно работающих,
 hoursw – общее число отработанных часов,
 hourspw – количество отработанных часов в расчете на одного работающего,
 inv1 – капиталовложения в помещения,
 inv2 - капиталовложения в автоматизацию,
 ssize – размер магазина в м²,
 start – год открытия магазина.

- 1) Оцените зависимость среднегодовых продаж (переменная tsales) или продаж в расчете на квадратный метр (sales) или маржинальной валовой прибыли (margin) от всех остальных переменных.
- 2) Проверьте адекватность регрессии. Если регрессия адекватна, то переходите к следующим пунктам.
- 3) Рассчитайте VIF-ы. Существует ли для построенной регрессии проблема мультиколлинеарности?
- 4) Выберите факторы, которые должны быть исключены из уравнения регрессии, используя метод пошагового исключения незначимых переменных.
- 5) Выберите факторы, которые должны быть включены в уравнение регрессии, используя метод пошагового включения переменных.
- 6) Сравните результаты, полученные в пунктах 1, 4, 5. В качестве показателя качества подгонки регрессии используйте коэффициент множественной детерминации, скорректированный на число степеней свободы. Дайте экономическую интерпретацию полученным результатам.

Методические рекомендации по выполнению упражнения в пакете STATA

1) Для оценки, например, параметров уравнения регрессии

$$\begin{aligned}
 \text{sales} = & \beta_1 + \beta_2 \text{nown} + \beta_3 \text{nfull} + \beta_4 \text{npart} + \beta_5 \text{naux} + \beta_6 \text{hoursw} + \beta_7 \text{hourspw} + \beta_8 \text{inv1} + \beta_9 \text{inv2} + \\
 & + \beta_{10} \text{ssize} + \beta_{11} \text{start} + \varepsilon
 \end{aligned}$$

наберите в командном окне

reg sales nown nfull npart naux hoursw hourspw inv1 inv2 ssize start

2) Для нахождения VIFов после оценки уравнения регрессии необходимо набрать команду *vif*

и воспользоваться выданной таблицей..

3) Для применения метода пошагового исключения незначимых переменных из уравнения регрессии, наберите в командном окне

stepwise, pr(0.1): reg имя зависимой переменной имена независимых переменных

Выбранный в скобках уровень значимости 0.1 можно изменить.

4) Для применения метода пошагового включения переменных в уравнение регрессии, наберите в командном окне

stepwise, pr(0.1): reg имя зависимой переменной имена независимых переменных

Выбранный в скобках уровень значимости 0.1 можно изменить.

4) (Демешев, Борzych, 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 $s\text{Corr}(g, h) = -0.9$. Гена и Чебурашка собирали независимо от Пионеров, поэтому выборочные корреляции $s\text{Corr}(g, x) = 0$, $s\text{Corr}(h, x) = 0$. Если регрессоры g , h и x центрировать и нормировать, то получится матрица \tilde{X} .

1. Найдите параметр обусловленности матрицы $(\tilde{X}'\tilde{X})$.
2. Вычислите одну или две главные компоненты (выразите их через вектор-столбцы матрицы \tilde{X}), объясняющие не менее 70% общей выборочной дисперсии регрессоров.
3. Шпионка Шапокляк пытается смоделировать ежедневный выпуск танков, y . Выразите оценки коэффициентов регрессии $y = \beta_1 + \beta_2 g + \beta_3 h + \beta_4 x + \varepsilon$ через оценки коэффициентов регрессии на главные компоненты, объясняющие не менее 70% общей выборочной дисперсии.

5) (Демешев, Борzych, 7.11)

Эконометресса Алевтина перешла от исходных регрессоров к трём главным компонентам, z_1 , z_2 и z_3 . И далее посчитала коэффициенты вздутия дисперсии, VIF_j , для главных компонент. Чему они оказались равны?

6) (Демешев, Борzych, 7.13)

Известно, что выборочная корреляция между переменными x и z равна 0.9.

1. Найдите коэффициенты VIF для x и z в регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$.
2. В каких пределах могут лежать коэффициенты VIF для x и z в регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 w_i + \varepsilon_i$?