$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(}}$$

Эконометрическое моделирование

Лабораторная работа № 4

Парная нелинейная регрессия

Оглавление

Простое преобразование переменных 3

Задание 1. Простое преобразование переменных 4

Логарифмические преобразования 4

Задание 2. Логарифмические преобразования 5

Полулогарифмические модели 5

Задание 3. Логарифмически-линейная модель 6

Задание 4. Линейно-логарифмическая модель 6

Выбор функции 6

Задание 5. Выбор функции 8

Простое преобразование переменных

Нелинейные соотношения гораздо лучше подходят для описания многих экономических процессов, чем линейные. Одним из недостатков линейного регрессионного анализа, как это следует из самого названия, является то, что он может быть применен только к линейным уравнениям, где каждый объясняющий элемент (за исключением свободного члена) записывается как произведение коэффициента и переменной:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (1)$$

Уравнения вида:

$$Y = \beta_0 + \frac{\beta_1}{X} \quad (2)$$

и

$$Y = \beta_0 X^{\beta_1} \quad (3)$$

являются нелинейными. Тем не менее, именно зависимости (2) и (3) считаются приемлемыми для описания кривых Энгеля, характеризующих соотношение между спросом на определенный товар Y и общей суммой дохода X.

Как можно определить параметры β_0 и β_1 в каждом уравнении, имея данные о значениях Y и X? В действительности в обоих случаях можно в конце концов применить линейный регрессионный анализ. Во-первых, заметим, что уравнение (1) является *линейным в двух смыслах*. Правая часть его линейна *по переменным*, если определить их в представленном виде, а не как функции. Следовательно, она состоит из взвешенной суммы переменных, а параметры являются весами. Правая часть также линейна *по параметрам*, так как она состоит из взвешенной суммы параметров, а переменные в данном случае являются весами.

Для целей линейного регрессионного анализа важное значение имеет только второй тип линейности. Нелинейность по переменным всегда можно обойти путем использования соответствующих определений. Например, предположим, что соотношение имеет вид:

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 log X_3 \quad (4)$$

Если определить $Z_1 = X_1^2$, $Z_2 = \sqrt{X_2}$, $Z_3 = logX_3$ и т.д., то соотношение примет следующий вид:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 \quad (5)$$

и теперь оно является линейным как по переменным, так и по параметрам. Такой тип преобразований является лишь косметическим, и обычно уравнения регрессии записываются с нелинейными выражениями относительно переменных. Это позволяет избежать лишних обозначений.

С другой стороны, уравнение типа (3) является нелинейным как по параметрам, так и по переменным, и его нельзя преобразовать только путем замены определений. (Не следует думать, что его можно преобразовать в линейное, если определить Z = X^{β_1} и подставить X^{β_1} вместо Z; поскольку β_2 неизвестно, мы не сможем рассчитать выборочные

значения Z.) Проблема преобразования нелинейных по параметрам соотношений будет рассмотрена позже. В случае (2), однако, единственное, что нам нужно сделать, – это определить Z= (1/X). Тогда уравнение (2) примет вид

$$Y = \beta_0 + \beta_1 Z \quad (6)$$

и оно будет линейным, и мы оцениваем регрессию У на Z. Постоянный член в уравнении регрессии будет представлять собой оценку β_0, а коэффициент при Z – оценку β_1.

Задание 1. Простое преобразование переменных

В данной лабораторной работе используйте данные из лабораторной работы №3. Построй регрессионную модель вида (2) отражающую зависимость между налоговыми поступлениями и численностью занятых. Для построения модели любой удобный для вас способ в MS Excel.

Запишите полученное уравнение модели, проверьте значимость полученных параметров с помощью t-критерия Стьюдента, оцените значимость полученного уравнения в целом, найдите R^2 и сумму квадратов отклонений (СКО), дайте словесную интерпретацию полученной взаимосвязи.

Логарифмические преобразования

Рассмотрим далее функции вида (3), которые являются нелинейными как по параметрам, так и по переменным:

$$Y = \beta_0 X^{\beta_1} \quad (7)$$

Когда вы видите такую функцию, вы можете сразу сказать, что эластичность Y по X постоянна и равна β_1. Это можно легко продемонстрировать. Независимо от математической формулы связи между Y и X или определения величин Y и X, эластичность Y по X определяется как относительное изменение Y на единицу относительного изменения X:

$$\text{Эластичность} = \frac{dY/Y}{dX/X} \quad (8)$$

Таким образом, если, например, Y – это спрос, а X – доход, то данное выражение определяет эластичность спроса на данный товар по доходу. Выражение для эластичности можно переписать в следующем виде:

$$\text{Эластичность} = \frac{dY/dX}{Y/X} \quad (9)$$

Если соотношение между Y и X имеет вид (7), то

$$\frac{dY}{dX} = \beta_0\beta_1 X^{\beta_1 - 1} = \beta_1 \frac{Y}{X} \quad . \quad (10)$$

Следовательно,

$$\text{Эластичность} = \frac{dY/dX}{Y/X} = \frac{\beta_1 Y/X}{Y/X} = \beta_1 \quad (11)$$

Таким образом, если имеется кривая Энегеля вида $Y = 0{,}01X^{0{,}3}$, то это означает, что эластичность спроса по доходу равна 0,3. Если вы хотите объяснить это другим способом, то наиболее просто будет сказать, что изменение X (дохода) на 1% вызывает изменение У (спроса) на 0,3%.

Функция описанного типа может быть преобразована в линейную путем использования логарифмов. Уравнение (7) можно преобразовать в линейное как

$$lnY = ln\beta_0 + \beta_1 lnX \quad (12)$$

Если обозначить Y’ = lnY, Z = lnX и β_0’=lnβ_0, то уравнение можно переписать в следующем виде:

$$Y' = \beta_0' + \beta_1 Z. \quad (13)$$

Процедура оценивания регрессии теперь будет следующей. Сначала вычислим Y' и Z для каждого наблюдения путем логарифмирования исходных значений. Затем оценивается регрессию Y' на Z. Коэффициент при Z будет представлять собой непосредственно оценку β_1. Постоянный член является оценкой β_0’, т.е. lnβ_0. Для получения оценки необходимо взять антилогарифм, т. е. вычислить exp(β_0’).

Задание 2. Логарифмические преобразования

Построй регрессионную модель вида (7) отражающую зависимость между налоговыми поступлениями и численностью занятых. Для построения модели любой удобный для вас способ в MS Excel.

Запишите полученное уравнение модели, проверьте значимость полученных параметров с помощью t-критерия Стьюдента, оцените значимость полученного уравнения в целом, найдите R^2 и сумму квадратов отклонений (СКО), дайте словесную интерпретацию полученной взаимосвязи.

Полулогарифмические модели

Еще одна широко распространенная функциональная форма представлена уравнением вида:

$$Y = \beta_0 e^{\beta_1 X} \quad (14)$$

Здесь β_1 можно интерпретировать как относительное изменение У в расчете на единицу (абсолютного) изменения X. Это снова легко показать. Дифференцируя, получаем:

$$\frac{dY}{dX} = \beta_0 \beta_1 e^{\beta_1 X} = \beta_1 Y \quad (15)$$

Следовательно,

$$\frac{dY/dX}{Y} = \beta_1 \quad (16)$$

На практике более естественно говорить не об относительном, а о процентном изменении Y в расчете на единицу изменения X, и в этом случае оценку коэффициента β_1 нужно умножить на 100. Данная функция может быть трансформирована в модель, линейную по параметрам, путем логарифмирования обеих частей:

$$lnY = ln\beta_0 e^{\beta_1 X} = ln\beta_0 + lne^{\beta_1 X} = ln\beta_0 + \beta_1 X lne = ln\beta_0 + \beta_1 X \quad (17)$$

Заметим, что только левая часть является логарифмической по переменным, и поэтому сама модель названа полулогарифмической.

Интерпретация коэффициента β_1 как относительного изменения Y на единицу изменения X правомерна лишь для малых β_1 (<0.1). Если β_1 велико, то интерпретация становится несколько более сложной.

Обозначим прирост зависимой переменной как

$$\Delta y = \frac{y_1 - y_0}{y_0} = \frac{y_1}{y_0} - 1 = \frac{\beta_0 e^{\beta_1 x_1}}{\beta_0 e^{\beta_1 x_0}} = e^{\beta_1 \Delta x} - 1$$

Таким образом, при изменении X на единицу Y изменится на $\left(e^{\beta_1} - 1\right) \cdot 100\%$.

Задание 3. Логарифмически-линейная модель

Построй регрессионную модель вида (14) отражающую зависимость между налоговыми поступлениями и численностью занятых. Для построения модели любой удобный для вас способ в MS Excel.

Запишите полученное уравнение модели, проверьте значимость полученных параметров с помощью t-критерия Стьюдента, оцените значимость полученного уравнения в целом, найдите R^2 и сумму квадратов отклонений (СКО), дайте словесную интерпретацию полученной взаимосвязи.

Задание 4. Линейно-логарифмическая модель

Построй регрессионную модель, отражающую зависимость между налоговыми поступлениями и численностью занятых, следующего вида

$Y = \beta_0 + \beta_1 lnX$

Параметр β_1 в данной модели имеет следующую интерпретацию: при увеличении x на 1 % y увеличится на (β_1/100) единиц.

Для построения модели любой удобный для вас способ в MS Excel.

Запишите полученное уравнение модели, проверьте значимость полученных параметров с помощью t-критерия Стьюдента, оцените значимость полученного уравнения в целом, найдите R^2 и сумму квадратов отклонений (СКО), дайте словесную интерпретацию полученной взаимосвязи.

Выбор функции

Возможность построения нелинейных моделей, как с помощью приведения к линейному виду, так и путем использования нелинейной регрессии, значительно повышает универсальность регрессионного анализа, но и усложняет задачу исследователя. Часто несколько разных нелинейных функций приблизительно соответствуют наблюдениям, если они лежат на некоторой кривой.

При рассмотрении альтернативных моделей с одним и тем же определением зависимой переменной процедура выбора достаточно проста. Наиболее разумным является оценивание регрессии на основе всех вероятных функций, которые вы можете вообразить, и выбор функции, в наибольшей степени объясняющей изменения зависимой переменной. Если две или более функций подходят примерно одинаково, то вы должны представить результаты для каждой из них. Если получилось, что линейная функция объясняет 69% дисперсии Y, а гиперболическая функция – 97%, следует без колебаний выбрать последнюю. Если, однако, разные модели используют разные функциональные формы для зависимой переменной, то проблема выбора модели становится более сложной, так как нельзя непосредственно сравнить коэффициенты R^2 и суммы квадратов отклонений. В частности – и это наиболее общий пример для данной

проблемы, – нельзя сравнить эти статистики для линейного и логарифмического вариантов модели.

Пример. Допустим, что линейная регрессия между Y и числом X имела коэффициент R^2 = 0,104, а сумма квадратов отклонений (RSS, или СКО) была равна 34420. Для полулогарифмического варианта модели (формула 14, 17) соответствующие значения были равны 0,141 и 132. Во втором случае СКО существенно меньше, но это ничего не решает. Значения переменной lnY значительно меньше соответствующих значений Y, поэтому неудивительно, что остатки также значительно меньше. Величина коэффициента R^2 безразмерна, однако в двух уравнениях она относится к разным понятиям. В одном уравнении она измеряет объясненную регрессией долю дисперсии Y, а в другом – объясненную регрессией долю дисперсии логарифма Y. Если для одной модели коэффициент R^2 значительно больше, чем для другой, то вы сможете сделать оправданный выбор без особых раздумий. Однако если значения R^2 для двух моделей близки друг к другу, то проблема выбора существенно усложняется.

В этом случае следует использовать стандартную процедуру, известную под названием *теста Бокса-Кокса* (Box, Cox, 1964). Если вы хотите сравнить модели, с использованием Y и lnY в качестве зависимой переменной, то можно использовать вариант теста, разработанный П. Зарембкой (Zarembka, 1968). Данный тест предполагает такое преобразование масштаба наблюдений Y, при котором обеспечивалась бы возможность непосредственного сравнения СКО в линейной и логарифмической моделях. Процедура включает следующие шаги:

1. Вычисляется *среднее геометрическое значений* Y в выборке. Оно совпадает с экспонентой среднего арифметического lnY, которое можно рассчитать:

$$e^{\frac{1}{n}\sum \ln Y_i} = e^{\frac{1}{n}ln\,(Y_1\times\ldots\times Y_n)} = e^{ln\,(Y_1\times\ldots\times Y_n)^{\frac{1}{n}}} = (Y_1 \times \ldots \times Y_n)^{\frac{1}{n}}$$

2. Пересчитываются наблюдения Y, они делятся на это значение, т. е.

$$Y_i^* = Y_i/\text{Среднее геометрическое } Y$$

3. Оценивается регрессия для линейной модели с использованием Y^* вместо Y в качестве зависимой переменной и для логарифмической модели с использованием $\ln Y^*$ вместо lnY; во всех других отношениях модели должны оставаться неизменными. Теперь значения СКО для двух регрессий сравнимы, и, следовательно, модель с меньшей суммой квадратов отклонений обеспечивает лучшее соответствие.

4. Для того чтобы проверить, обеспечивает ли одна из моделей значимо лучшее соответствие, можно вычислить величину (n/2)lnZ, где Z – отношение значений СКО в пересчитанных регрессиях, а n – число наблюдений, и взять ее абсолютное значение (т. е. игнорировать знак «минус», если он имеется). Эта статистика имеет распределение χ^2 (хи-квадрат) с одной степенью свободы. Если она превышает критическое значение χ^2 при выбранном уровне значимости, то делается вывод о наличии значимой разницы в качестве оценивания.

Пример. Допустим выборка состоит из 570 значений и среднее значение переменной lnY=2,43. Тогда масштабирующий множитель равен exp(2,43) = 11,36. Сумма квадратов отклонений в регрессии между масштабированным по методу Зарембки Y ($Y_i^*=Y_i/11,36$) и X равняется 266,7; сумма квадратов в регрессии логарифма того же

масштабированного по методу Зарембки Y получилась равной 132,1. Следовательно тестовая статистика равна:

570/2*ln(266,7/132,1)=200,2

Критическое значение χ^2 с одной степенью своды при 0,1-процентном уровне значимости равно 10,8. Следовательно, полулогарифмическая спецификация обеспечивает лучшие оценки.

Замечание. Масштабированные по методу Зарембки регрессии могут использоваться исключительно для принятия решения о том, какую модель предпочесть. Не следует обращать внимание на их коэффициенты, рассматривая только их суммы квадратов отклонений. Коэффициенты регрессии получают непосредственно из немасштабированного варианта выбранной моедели.

Задание 5. Выбор функции

Выберите функциональную зависимость, которая наилучшим образом описывает взаимосвязь между налоговыми поступлениями и численностью занятых. Свой выбор обоснуйте.