Тема 4. Мультиколлинеарность.

Что такое мультиколлинеарность.

Одна из предпосылок теоремы Гаусса-Маркова для множественной регрессии состояла в том, что $rang(X) = k$, это означает, что объясняющие переменные линейно не зависимы. В случае, если это условие не выполняется, говорят о том, что имеет место точная мультиколлинеарность.

Пусть у нас есть теоретическая модель регрессии

(4.01) $$Y_i = \beta_1 + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i3} + \beta_4 \cdot X_{i4} + \varepsilon_i .$$

Если $X_{i2} = X_{i3} + X_{i4}$ (это и есть точная мультиколлинеарность), то мы не сможем оценить уравнение (4.01) метом наименьших квадратов. Уточним почему: $rang(X) = 3$, а не 4, соответственно $rang(X^T X) = 3 \Rightarrow \det(X^T X) = 0$, матрица $(X^T X)^{-1}$- не будет определена. Поэтому вектор оценок $\hat{\beta} = (X^T X)^{-1} X^T Y$ не может быть вычислен. Проблема возникает из-за того, что не возможно отделить влияние X_2, от влияния X_3, X_4.

Точная мультиколлинеарность, как правило, возникает вследствие неправильной спецификации модели. В нашем случае нужно просто исключить одну из переменных из уравнения регрессии.

Однако обычно под мультиколлинеарностью подразумевают не точную мультиколлинеарность, а ситуацию, когда точной линейной связи между переменными нет, но они все равно меняются похожим образом. Формально в этом случае нужно говорить о квазимультиколлинеарности, но обычно ее тоже называют просто мультиколлинеарностью. Например, для уравнения (4.01) это может выглядеть, так $X_{i2} = X_{i3} + X_{i4} + \xi_i$, где ξ_i- это некоторая величина, которая принимает небольшие произвольные значения. В этом случае, мы можем вычислить МНК оценки $\hat{\beta} = (X^T X)^{-1} X^T Y$, более того, **они по-прежнему будут линейными несмещенными и эффективными оценками** параметров регрессии. Проблема состоит в том, что, если переменные почти линейно зависимы, то $\det(X^T X) \to 0$, соответственно элементы матрицы $(X^T X)^{-1}$ будут стремиться к бесконечности, то же самое будет происходить с элементами ковариационной матрицы $\mathrm{cov}(\hat{\beta}) = (X^T X)^{-1} \sigma^2$. Большие дисперсии будут означать, что точность оценки коэффициентов будет низкая, и они будут незначимы. Это происходит потому что, если факторы связаны, то трудно **отдельно** оценить влияние каждого из них.

Экономическая теория предполагает зависимость потребительских расходов от богатства и дохода. В таблице 4.1 приведены результаты оценки такой регрессии в логарифмической форме (данные смотри в приложении 1). Как мы видим, коэффициенты при объясняющих переменных не значимы. В тоже время регрессия адекватна, причем на любом уровне значимости. Да и коэффициент детерминации достаточно большой. Происходит это так как доход и богатство сильно коррелированны (коэффициент корреляции =0.99), да и наблюдений у нас довольно мало.

Еще раз обратим внимание, с точки зрения теории у нас все хорошо. Оценки не смещенные. Что означает свойство на практике? Оно означает, что если у нас есть много выборок и по каждой из них мы оценим параметры регрессии, а потом посчитаем среднюю оценку, то она совпадет с истинным значением. Но нам от этого не легче, у нас есть только одна выборка и по ней получаются плохие результаты.

Оценки эффективные, они обладают наименьшей дисперсией среди несмещенных оценок. Но опять нам от этого не легче, так как в нашем случае эта наименьшая дисперсия, достаточно велика, чтобы сделать точность оценок неприемлемой.

Более того, оценки еще и состоятельны. Об этом не говориться в теореме Гаусса-Маркова, но это так. Это означает, что, если размер выборки будет стремиться к бесконечности, то оценки будут стремиться к истинным значениям параметров, которые мы оцениваем. Это хорошо, но у нас размер выборки задан, и изменить его мы не можем.

Таблица 4.1

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/27/07 Time: 16:42
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.608456	4.899733	0.124181	0.9047
LOG(INCOME)	0.643406	2.137046	0.301073	0.7721
LOG(WEALTH)	0.108048	2.126388	0.050813	0.9609

R-squared	0.947998	Mean dependent var	4.670518
Adjusted R-squared	0.933140	S.D. dependent var	0.300991
S.E. of regression	0.077828	Akaike info criterion	-2.025296
Sum squared resid	0.042401	Schwarz criterion	-1.934521
Log likelihood	13.12648	F-statistic	63.80453
Durbin-Watson stat	2.811065	Prob(F-statistic)	0.000032

Итак, с точки зрения теории все нормально, а с точки зрения практики мы не можем решить стоящую перед нами задачу. В этом проблема мультиколлинеарности очень похожа на ситуацию, когда у нас мало наблюдений в выборке. Не случайно некоторые методы борьбы с мультиколлинеарностью применяются и для случая малых выборок.

Так как проблема не теоретическая, а практическая, то методы ее решения будут носить кустарный характер. Не случайно в большинстве современных продвинутых учебников по эконометрике темы мультиколлинеарность просто нет.

Признаки мультиколлинеарности.

Начнем по порядку. Характерными признаками, при которых обычно возникает подозрение, что есть мультиколлинеарность, являются

1) Незначимые коэффициенты и при этом большой R^2 и адекватная регрессия
2) Чувствительность оценок коэффициентов и оценок дисперсий коэффициентов к добавлению и исключению наблюдению из выборки.

Например, чтобы проиллюстрировать 2 признак исключим одно из наблюдений из уравнения для потребительских расходов. Сравнив таблицы 4.1 и 4.2 можно видеть, что оценки коэффициентов изменились в разы.

Таблица 4.2

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/27/07 Time: 17:23
Sample: 1 9
Included observations: 9

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.345546	5.649167	-0.061168	0.9532
LOG(INCOME)	0.276879	2.422325	0.114303	0.9127
LOG(WEALTH)	0.489127	2.423483	0.201828	0.8467

R-squared	0.940169	Mean dependent var	4.632727
Adjusted R-squared	0.920226	S.D. dependent var	0.293008
S.E. of regression	0.082758	Akaike info criterion	-1.884587
Sum squared resid	0.041093	Schwarz criterion	-1.818845
Log likelihood	11.48064	F-statistic	47.14140

.

Численные измерители мультиколлинеарности.

VIF

Как мы уже сказали, проблема возникает из-за того, что объясняющие переменные сильно взаимосвязаны. Наиболее информативным показателем для нас является коэффициент детерминации в регрессии, которая представляет собой зависимость одной из объясняющих переменных от всех остальных. Обозначим R_j^2- коэффициент детерминации регрессии X_j на оставшиеся объясняющие переменные. Тогда, будет справедлива формула

$$\text{var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\sum\left(X_{ij} - \bar{X}_j\right)^2} \cdot \frac{1}{1 - R_j^2} \qquad (4.02)$$

Формула (4.02) показывает, что дисперсию коэффициента в множественной регрессии можно разложить на две составляющие. Во-первых, это дисперсия, коэффициента, которая была бы, если бы регрессия была однофакторная, то есть если бы в качестве объясняющей переменной использовалась только X_j. В главе 2 рассматривалась однофакторная модель $Y_i = \alpha + \beta \cdot X_i + \varepsilon_i$, и там $\text{var}\left(\hat{\beta}\right) = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}$, а в формуле (4.02) как раз присутствует множитель $\frac{\sigma^2}{\sum\left(X_{ij} - \bar{X}_j\right)^2}$. Вторая составляющая это $\frac{1}{1 - R_j^2}$, она всегда будет больше единицы. Поэтому дисперсия всегда будет больше, за счет того, что регрессия множественная и в ней есть другие объясняющие переменные. Введем обозначение

$$VIF_j = \frac{1}{1 - R_j^2} \qquad (4.03)$$

VIF- это сокращение от variance inflation factor, что и означает множитель, который увеличивает дисперсию. Имеется в виду увеличение дисперсии по сравнению с тем, какой бы она была в случае однофакторной регрессии. Если переменные ортогональны $R_j^2 = 0$, соответственно $VIF_j = 1$ и переменные не будут снижать значимость друг друга. Если

наоборот имеет место линейная зависимость или точная мультиколлинеарность $R_j^2 = 1$, $VIF_j = +\infty$ дисперсия коэффициентов тоже будет бесконечной. Мы уже отмечали, что оценить регрессию в этом случае невозможно.

Среди эконометристов существует убеждение, что если для одной из переменных $VIF > 10$, то в регрессии есть мультиколлинеарность. Эту цифру можно даже найти в учебнике Gudgarati. Некоторые скептики, говорят, что квазимультиколлинеарность не является теоретической проблемой. И поэтому нет механизмов для получения точных критических значений или выработки процедур для проверки гипотез о ее наличии или отсутствии, но они ничего не понимают в эконометрике. =)

Проверим, чему в нашем примере с уравнением для потребительских расходов равны VIF и выполняется ли соотношение (4.03).

Оценим вспомогательную регрессию

Таблица 4.3

Dependent Variable: LOG(INCOME)
Method: Least Squares
Date: 02/27/07 Time: 18:40
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.281251	0.081117	-28.12281	0.0000
LOG(WEALTH)	0.994530	0.010958	90.76102	0.0000
R-squared	0.999030	Mean dependent var		5.071773
Adjusted R-squared	0.998909	S.D. dependent var		0.389734
S.E. of regression	0.012876	Akaike info criterion		-5.690056
Sum squared resid	0.001326	Schwarz criterion		-5.629539
Log likelihood	30.45028	F-statistic		8237.563
Durbin-Watson stat	1.907461	Prob(F-statistic)		0.000000

По данным таблицы 4.3 можно вычислить $VIF_2 = \dfrac{1}{1-0.999030} = 1030.928 > 10$. Очевидно, что мультиколлинеарность есть. Проверим, выполняется ли формула (4.02), оценим однофакторную регрессию.

Таблица 4.4

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/27/07 Time: 18:46
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.856832	0.316697	2.705527	0.0268
LOG(INCOME)	0.751943	0.062278	12.07403	0.0000
R-squared	0.947978	Mean dependent var		4.670518
Adjusted R-squared	0.941476	S.D. dependent var		0.300991
S.E. of regression	0.072815	Akaike info criterion		-2.224928
Sum squared resid	0.042416	Schwarz criterion		-2.164411
Log likelihood	13.12464	F-statistic		145.7822
Durbin-Watson stat	2.829128	Prob(F-statistic)		0.000002

Как видно из таблицы 4.4, стоило только исключить из уравнения одну из двух коррелированных переменных, и все стало нормально. Все коэффициенты значимы и т.д. Как мы видим, формула 4.02 справедлива, а именно

$$\left(2.137046\right)^2 = \left(0.062278\right)^2 \cdot \frac{1}{1-0.999030}$$

Индекс обусловленности

Мы упоминали, что большие дисперсии коэффициентов получаются, так как $\det\left(X^T X\right) \to 0$. Еще мы знаем, что определитель матрицы равен произведению собственных чисел. Важную информацию о них нам может дать индекс обусловленности.

(4.04) $$\kappa = \sqrt{\frac{\max\left(\lambda_i\right)}{\min\left(\lambda_i\right)}}$$

где λ_i - это собственные числа матрицы $X^T X$.

Идея состоит в том, что число не нулевых собственных чисел – это и есть число независимых объясняющих факторов, которые можно выделить из иксов. Если есть собственные числа близкие к нулю, значит этих факторов меньше чем число переменных в регрессии, поэтому можно утверждать, что есть мультиколлинеарность. Если $\kappa < 10$, то мультиколлинеарности нет. Если $10 < \kappa < 30$, то есть, но терпимая. Если $\kappa > 30$, то - сильная мультиколлинеарность.

Для нашего примера (регрессия из таблицы 4.1) собственные числа будут равны $\lambda_1 = 6.513 \cdot 10^{-3};\ \lambda_2 = 0.124;\ \lambda_3 = 816.489$ $\kappa = \sqrt{\frac{\lambda_3}{\lambda_1}} = \sqrt{\frac{816.489}{6.513 \cdot 10^{-3}}} = 354.066 > 30$ - как и следовало ожидать, мультиколлинеарность сильная.

Что делать с мультиколлинеарностью.

Важно осознавать, что в результате мультиколлинеарности мы не можем оценить точное значение каждого из параметров **в отдельности**. Однако если мы, например, в регрессии из таблицы 4.1 посмотрим ковариационную матрицу оценок коэффициентов:

24.00739	10.41840	-10.39384
10.41840	4.566967	-4.541984
-10.39384	-4.541984	4.521524

Можно обратить внимание, что оценка дисперсии суммы коэффициентов, которые мы не можем точно оценить

$\hat{D}\left(\hat{\beta}_1 + \hat{\beta}_2\right) = \hat{D}\left(\hat{\beta}_1\right) + \hat{D}\left(\hat{\beta}_2\right) + 2 \cdot \widetilde{\mathrm{cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) = 4.57 + 4.52 + 2 \cdot (-4.54) = 0.01$ - получается, что сумму этих параметров мы можем оценить достаточно точно (дисперсия очень маленькая). Итак, мы не можем оценить оба коэффициента, но можем оценить либо один из коэффициентов, либо их линейную комбинацию. Большая часть методов борьбы с мультиколлинеарностью на этом и основана.

Что делать, если есть мультиколлинеарность?

1) **Ничего.** Ряд исследователей считает, что ничего страшного в том, что коэффициенты неточно оценены, нет. Например, если мы собираемся использовать модель для прогнозирования, то прогноз мы получим хороший и по модели с мультиколлинеарностью.

2) **Изменить спецификацию модели**. Часто мультиколлинеарность возникает из-за того, что модель неправильно построена. Можно (а) просто выкинуть переменные, которые зависят от остальных. (б) Переопределить переменные, перейти к логарифмам/первым разностям и т.д. Например, как видно из нашего примера (таблица 4.4), когда мы оставляем только одну переменную из двух, модель только улучшается (скорректированный коэффициент детерминации возрастает).

3) **Увеличит выборку.** Добавить данных с другой структурой, где объясняющие переменные не так сильно зависят друг от друга. (Хорошая идея, но откуда их взять =)

3) **Перейти к линейной комбинации переменных.** Это можно сделать по-разному.

3.1) **Использовать априорные предположения о взаимосвязи между коэффициентами или о значениях некоторых параметров.** Эти предположения заимствуются из других исследований или просто из теории. Возвращаясь к нашему примеру, пусть ранее установлено, что эластичность потребления по доходу должна быть в два раза больше чем по богатству, поэтому будем оценивать уравнение с ограничениями $\beta_2 = 2 \cdot \beta_3$

Получаем результат. Таблица 4.5.

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/27/07 Time: 23:49
Sample: 1 10
Included observations: 10
LOG(CONSUMP) = C(1) + C(2)*LOG(INCOME) + 0.5*C(2) *LOG(WEALTH)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.281794	0.364265	0.773597	0.4614
C(2)	0.500510	0.041459	12.07231	0.0000

R-squared	0.947964	Mean dependent var	4.670518
Adjusted R-squared	0.941460	S.D. dependent var	0.300991
S.E. of regression	0.072825	Akaike info criterion	-2.224658
Sum squared resid	0.042428	Schwarz criterion	-2.164141
Log likelihood	13.12329	Durbin-Watson stat	2.786224

Как мы видим, в этой модели уже все нормально.

Либо, может быть, априорное предположение о значении одного из параметров. Например, известно, что эластичность по богатству равна 15%, поэтому оцениваем при ограничении, что $\beta_3 = 0.15$. Тоже получаем хорошее уравнение.

Таблица 4.6

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 02/28/07 Time: 00:09
Sample: 1 10
Included observations: 10
LOG(CONSUMP) = C(1) + C(2)*LOG(INCOME) + 0.15*LOG(WEALTH)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.512020	0.316647	1.617004	0.1445
C(2)	0.601265	0.062268	9.656078	0.0000

R-squared	0.947995	Mean dependent var	4.670518
Adjusted R-squared	0.941494	S.D. dependent var	0.300991
S.E. of regression	0.072804	Akaike info criterion	-2.225241
Sum squared resid	0.042403	Schwarz criterion	-2.164724
Log likelihood	13.12620	Durbin-Watson stat	2.803871

3.2) **Использовать методы снижения числа объясняющих факторов, путем перехода к линейным комбинациям факторов со специальным способом подобранными коэффициентами.** Например, метод главных компонент. Этот метод широко используется не только в борьбе с мультиколлинеарностью, но и во многих приложениях поэтому о нем будет сказано отдельно.

Метод главных компонент.

Алгоритм вычисления главных компонент.

Пусть у нас есть k факторов: $X_1, X_2,$ ..... $, X_k$ и по каждому из них n наблюдений, тогда, главные компоненты вычисляются следующим образом.

1) Перейдем к центрированным и нормированным переменным

$$x_{ji} = \frac{X_{ji} - \bar{X}_j}{\sqrt{\sum_{i=1}^{n}\left(X_{ji} - \bar{X}_j\right)^2}} \quad i = 1.....n,\ \ j = 1.....k$$

2) Вычислим корреляционную матрицу, она будет иметь размерность $k \times k$, ее элементы будут вычисляться по формуле $corr_{jl} = \sum_{i=1}^{n} x_{ji} \cdot x_{li}$. На главной диагонали получаться единицы, остальные элементы будут оценками коэффициентов корреляции факторов.

3) Вычислим собственные числа $\lambda_1 > \lambda_2 > ..... > \lambda_k$ (упорядочим их по убыванию) и собственные векторы $C_1 = \begin{pmatrix} c_{11} \\ c_{12} \\ \vdots \\ c_{1k} \end{pmatrix}$ $C_2 = \begin{pmatrix} c_{21} \\ c_{22} \\ \vdots \\ c_{2k} \end{pmatrix}$ $\cdots C_k = \begin{pmatrix} c_{k1} \\ c_{k2} \\ \vdots \\ c_{kk} \end{pmatrix}$ корреляционной матрицы.

При этом должно выполняться условие нормировки $\sum_{j=1}^{k} c_{jl}^2 = 1 \ \ \forall l = 1.....k$

4) Определим главные компоненты $Z_1, Z_2, \ldots \ , Z_k$ - это новые факторы, они являются линейными комбинациями исходных факторов. Тоже будут включать по n наблюдений каждая. Вычисляются по формуле:

$$z_{1i} = c_{11} \cdot x_{1i} + c_{12} \cdot x_{2i} + \ldots + c_{1k} \cdot x_{ki}$$
$$z_{2i} = c_{21} \cdot x_{1i} + c_{22} \cdot x_{2i} + \ldots + c_{2k} \cdot x_{ki}$$
$$\vdots$$
$$z_{ki} = c_{k1} \cdot x_{1i} + c_{k2} \cdot x_{2i} + \ldots + c_{kk} \cdot x_{ki} \quad \forall i = 1....n$$

Примечание. Вместо корреляционной матрицы для вычисления главных компонент можно использовать ковариационную матрицу, но тогда результат будет зависеть от единиц измерения исходных переменных.

Свойства главных компонент.

1) Главные компоненты ортогональны, то есть их корреляции равны нулю.
2) Дисперсии главных компонент равны собственным числам матрицы, по которой они вычислены $\text{var}\left(Z_j\right) = \dfrac{\sum_{i=1}^{n}\left(z_{ji} - \bar{z}_j\right)^2}{n} = \lambda_j$
3) Сумма дисперсий главных компонент $\sum \lambda_j = k$. Если бы мы считали их по ковариационной матрице, а не по корреляционной, то $\sum \lambda_j = \sum \text{var}\left(X_j\right)$.
4) Первая главная компонента Z_1, является решением задачи поиска линейной комбинации исходных факторов обладающей наибольшей дисперсией
$$\begin{cases} \text{var}\left(\sum_{j=1}^{k} a_j \cdot x_j\right) \Rightarrow \max\limits_{a_j} \\ \sum a_j^2 = 1 \end{cases}$$
5) Если нам нужно найти $p < k$ линейных комбинаций, которые лучше всего описывают исходные k факторов (в частности дают наибольший R^2 при построении регрессионной зависимости каждого из исходных факторов от этих линейных комбинаций) решением этой задачи будут $Z_1, Z_2, \ldots \ , Z_p$.

6) Главные компоненты обладают свойством наименьшего искажения геометрической структуры исходных данных при переходе в пространство меньшей размерности. Если мы, соответственно, переходим от k факторов к $p < k$ и берем для этого первые p главных компонент.

Более подробно и математически строго свойства главных компонент описаны, например, в учебнике Айвазян С.А., Мхитарян В.С. том 1 «Теория вероятности и прикладная статистика».

Суть в том, что главные компоненты позволяют из большого числа исходных факторов выделить меньшее число факторов, при этом потеря информации при снижении размерности будет минимальной. При переходе от k исходных переменных, к p главным компонентам, долю дисперсии исходного признака, которая при этом сохраняется можно измерить, как $\sum_{j=1}^{p}\lambda_j \Big/ \sum_{j=1}^{k}\lambda_j$. Обычно нескольких первых компонент бывает достаточно, чтобы эта величина достигла 80%-90%.

Применение главных компонент.

Главные компоненты используются во многих статических и эконометрических процедурах (например, в двухшаговом методе наименьших квадратов). Они решают задачу снижения размерности. Эта задача актуальна, когда у нас большое число переменных (относительно числа наблюдений), либо, когда есть проблема мультиколлинеарности. Часто эти две ситуации сложно отличить друг от друга.

С точки зрения регрессионного анализа первая главная компонента представляет собой линейную комбинацию переменных, МНК оценка коэффициента при которой будет иметь наименьшую дисперсию. Действительно, вспомним формулу (4.02) $\operatorname{var}\left(\hat{\beta}_j\right) = \dfrac{\sigma^2}{\sum\left(X_{ij} - \bar{X}_j\right)^2} \cdot \dfrac{1}{1 - R_j^2}$, для главных компонент $R_j^2 = 0$, так как они ортогональны, а $\sum\left(X_{ij} - \bar{X}_j\right)^2$ - будет максимальна в силу свойства 4 главных компонент.

Если мы включим в регрессию все главные компоненты, и после этого, раскрыв скобки, преобразуем уравнение, так, чтобы оно зависело от исходных переменных (главные компоненты это же линейные комбинации исходных переменных), то мы получим то же самое, что мы могли бы получить, оценив МНК регрессию с изначальными переменными. Поэтому использовать главные компоненты разумно, когда мы (а) берем только несколько первых главных компонент (б) главные компоненты имеют самостоятельный смысл: их можно как-то интерпретировать.

Итак, при использовании главных компонент возникает ряд проблем. Во-первых, часто непонятно в чем смысл этой линейной комбинации. Если модель используется для прогнозирования, то это не страшно, так как содержательный смысл переменных, в нее включенных, нас мало волнует. Бывают ситуации, когда смысл есть. Например, посчитаны главные компоненты по нескольким тысячам цен акций на бирже. В первой компоненте наибольшие веса у цен акций одной отрасли, а остальные близки к нулю, во второй у другой отрасли и т.д.- в этом случае можно рассматривать их как отраслевые индексы. Но так бывает редко. Во-вторых, главные компоненты строятся без учета, той переменной, которую мы собираемся с их помощью объяснять. И совершенно не обязательно, что первая главная компонента будет той линейной комбинацией объясняющих переменных, которая позволит получить наибольший R^2. Поэтому есть методы, которые позволяют строить ортогональные линейные комбинации с учетом объясняемой переменной, например Partial Least Squares. Технически это метод даже проще чем метод главных компонент, поэтому тут есть хорошая возможность для доклада.

Закончим эту главу применением метода главных компонент к нашему примеру. Результаты приведены ниже.

Таблица 4.7

Date: 02/28/07 Time: 13:02
Sample: 1 10
Included observations: 10
Correlation of LOG(INCOME) LOG(WEALTH)

	Comp 1	Comp 2
Eigenvalue	1.999515	0.000485
Variance Prop.	0.999757	0.000243
Cumulative Prop.	0.999757	1.000000

Eigenvectors:

Variable	Vector 1	Vector 2
LOG(INCOME)	0.707107	0.707107
LOG(WEALTH)	0.707107	-0.707107

Как видно из таблицы, собственные числа корреляционной матрицы получились равны 1.999515 и 0.000485, как и должно быть, в сумме они равны 2. Первое собственное число составляет 99.98% от общей суммы собственных чисел, это означает, что первая главная компонента будет включать в себя почти всю дисперсию двух исходных факторов. Собственные векторы получились (0.707; 0.707) и (0.707; -0.707), это означает, что первая компонента это будет сумма логарифмов дохода и богатства, вторая разность этих логарифмов. Вывод можно сделать следующий: реально за этими двумя факторами стоит один ненаблюдаемый фактор, который мы можем условно назвать благосостоянием человека. Это фактор представляет собой сумму (или среднее) двух исходных факторов. Если быть точным, при вычислении компонент каждая из переменных еще будут делиться на корень из своей дисперсии. В нашем случае эти величины примерно одинаковые для обоих переменных, поэтому интерпретация от этого не меняется. Оценим модель с первой главной компонентой:

Таблица 4.8

Dependent Variable: LOG(CONSUMP)
Method: Least Squares
Date: 03/02/07 Time: 22:23
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.670518	0.023048	202.6459	0.0000
COMP1	0.196603	0.016299	12.06215	0.0000

R-squared	0.947881	Mean dependent var	4.670518
Adjusted R-squared	0.941366	S.D. dependent var	0.300991
S.E. of regression	0.072883	Akaike info criterion	-2.223062
Sum squared resid	0.042496	Schwarz criterion	-2.162545
Log likelihood	13.11531	F-statistic	145.4955
Durbin-Watson stat	2.763665	Prob(F-statistic)	0.000002

В этой модели нет мультиколинеарности, но по величине коэффициента детерминации, она уступает, например модели из таблицы 4.7. Так как мы сказали до этого, первая главная компонента не обязательно является наилучшим объясняющим фактором.

Приложение. 1 Данные, использованные для построения численного примера.

Источник Gujarati D. N. Basic Econometrics p.356										
consumption $	70	65	90	95	110	115	120	140	155	150
income $	80	100	120	140	160	180	200	220	240	260
wealth $	810	1009	1237	1425	1633	1876	2025	2201	2435	2686

Рекомендуемая литература по теме мультиколлинеарность.

1 К. Доугерти. “Введение в эконометрику” М., ИНФРА-М, 2000 глава 5 (или глава 4 в новом издании)./ Сам не читал, поэтому не могу ничего сказать.

2 D. Gujarati. “Basic econometrics” McGraw-Hill 1995 Ch.10 pp.341-386.

3 G. S. Maddala. Introduction to econometrics. Macmillan Publishing Co. 1992 Ch.7 pp.269-296.

4 Johnston, J. Econometric methods. New York McGraw-Hill, 3e 1991 . Ch.6 pp239-259 (это не Johnston J., DiNardo J.J. Econometric methods, а более старая версия).

Выдумкин Платон
email platonhse@mail.ru