NATIONAL RESEARCH UNIVERSITY

Лекции
7.12.20, 14.12.20

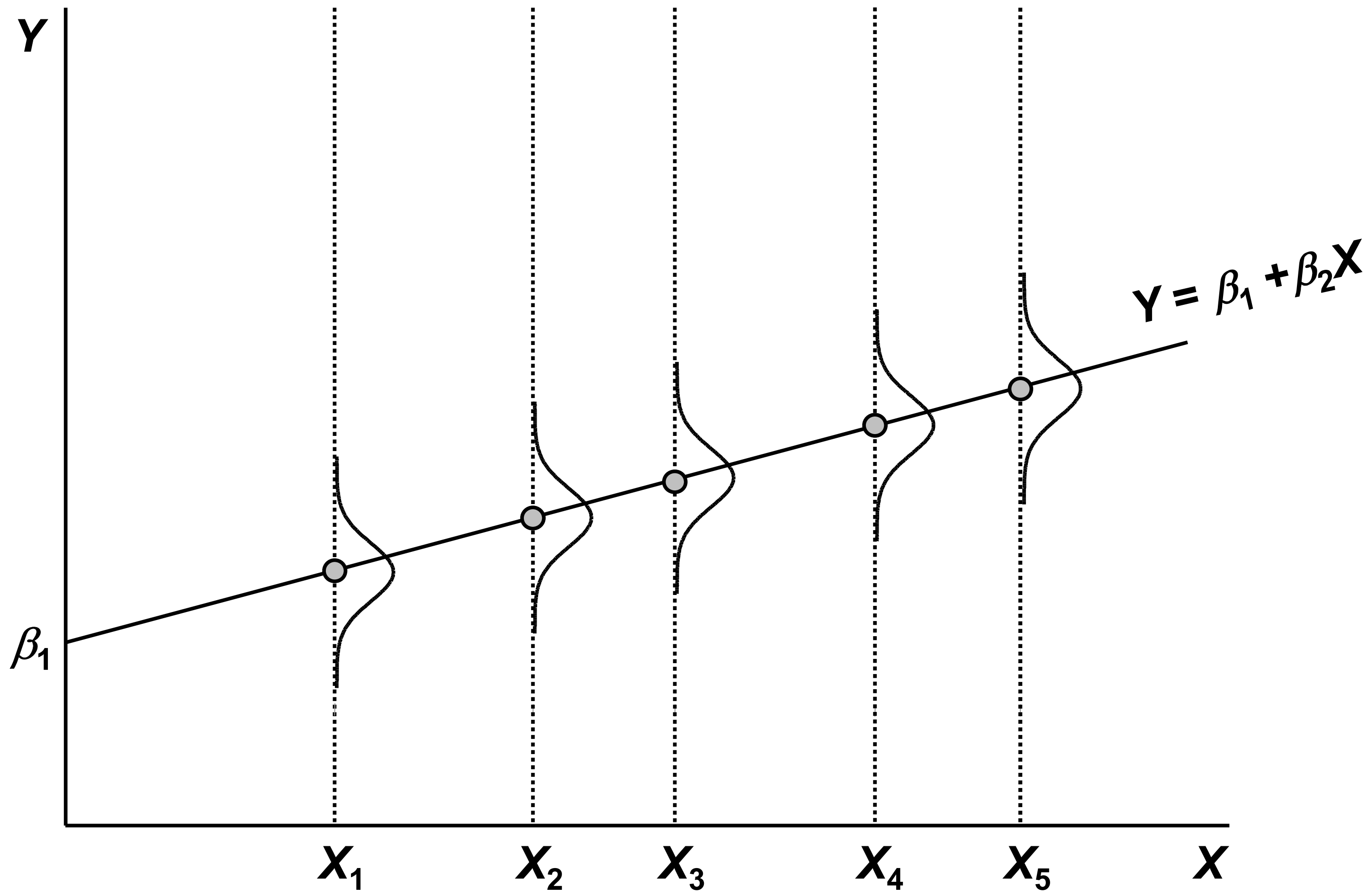
Гетероскедастичность

Демидова О.А.
E-mail:demidova@hse.ru

План лекции

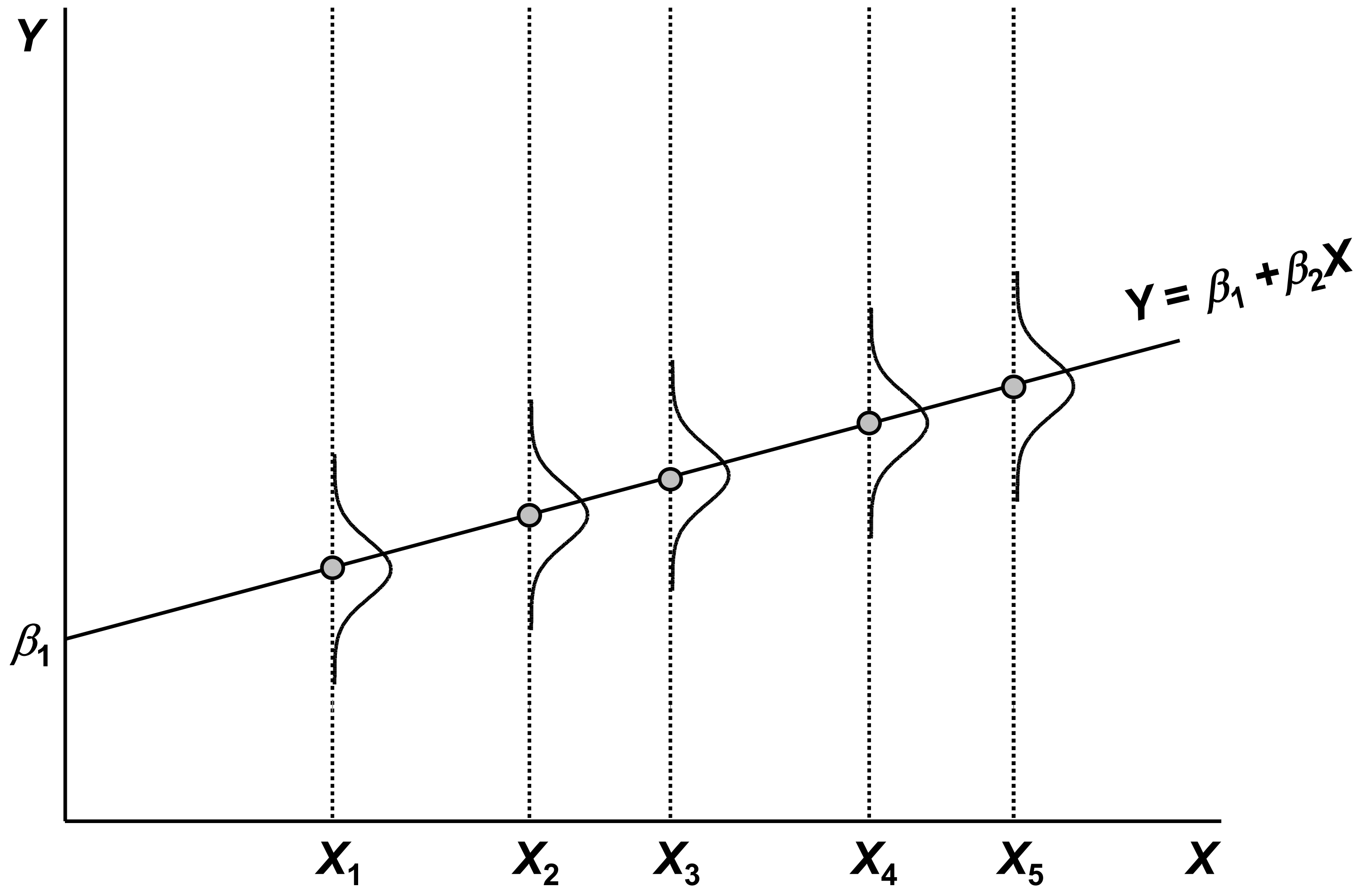
1) Нарушение гипотезы о гомоскедастичности ошибок регрессии. Последствия гетероскедастичности для оценок коэффициентов регрессии методом наименьших квадратов и проверки статистических гипотез.
2) Тесты на выявление гетероскедастичности.
3) Оценивание при наличии гетероскедастичности. Взвешенный метод наименьших квадратов.
4) Робастные стандартные ошибки оценок коэффициентов регрессии в форме Уайта (White).
5) Обобщенный метод наименьших квадратов

Гетероскедастичность



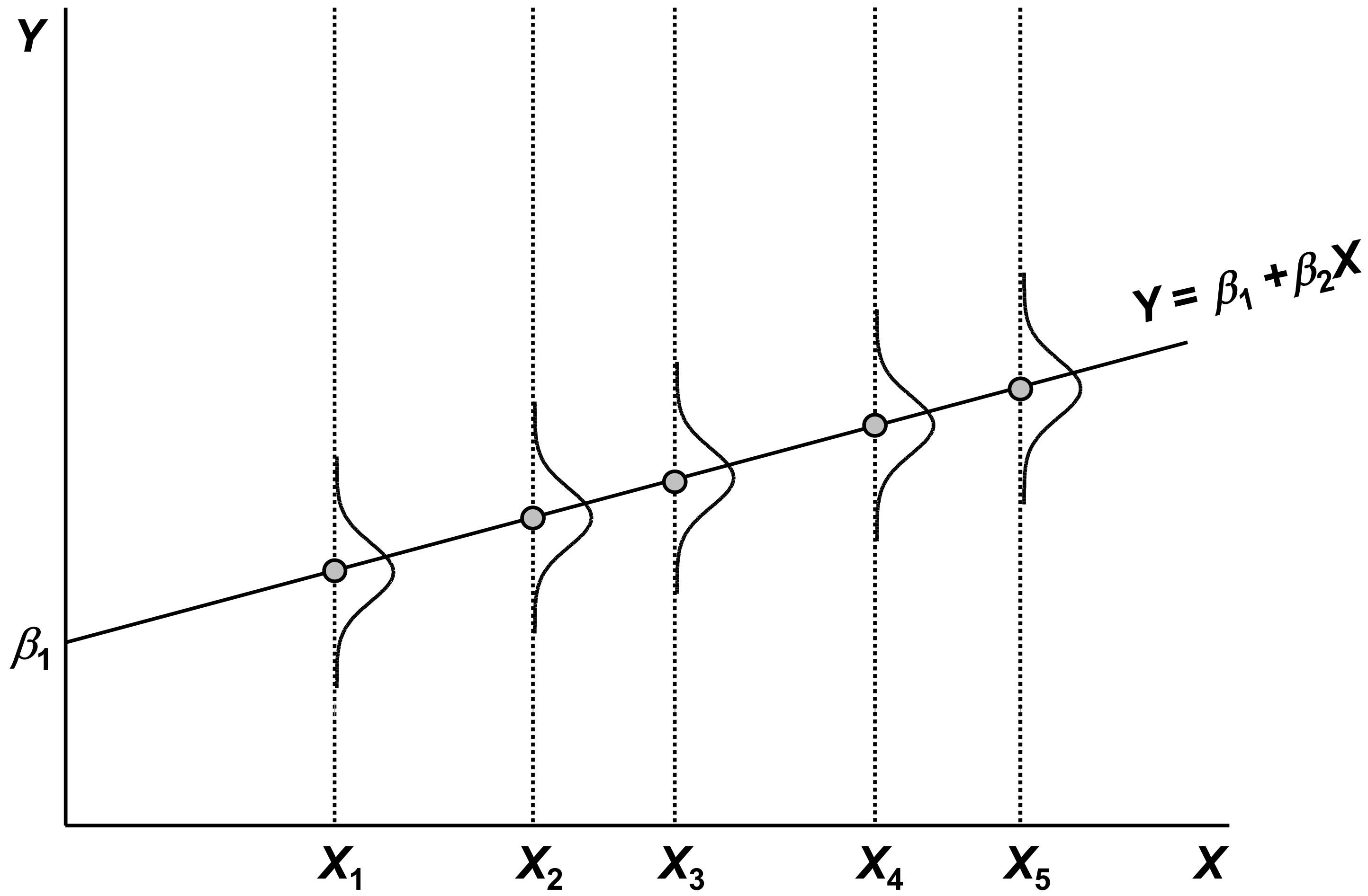
Одно из условий теоремы Гаусса – Маркова состоит в том, что возмущения u имеют нулевое математическое ожидание и одинаковую дисперсию.

Гетероскедастичность



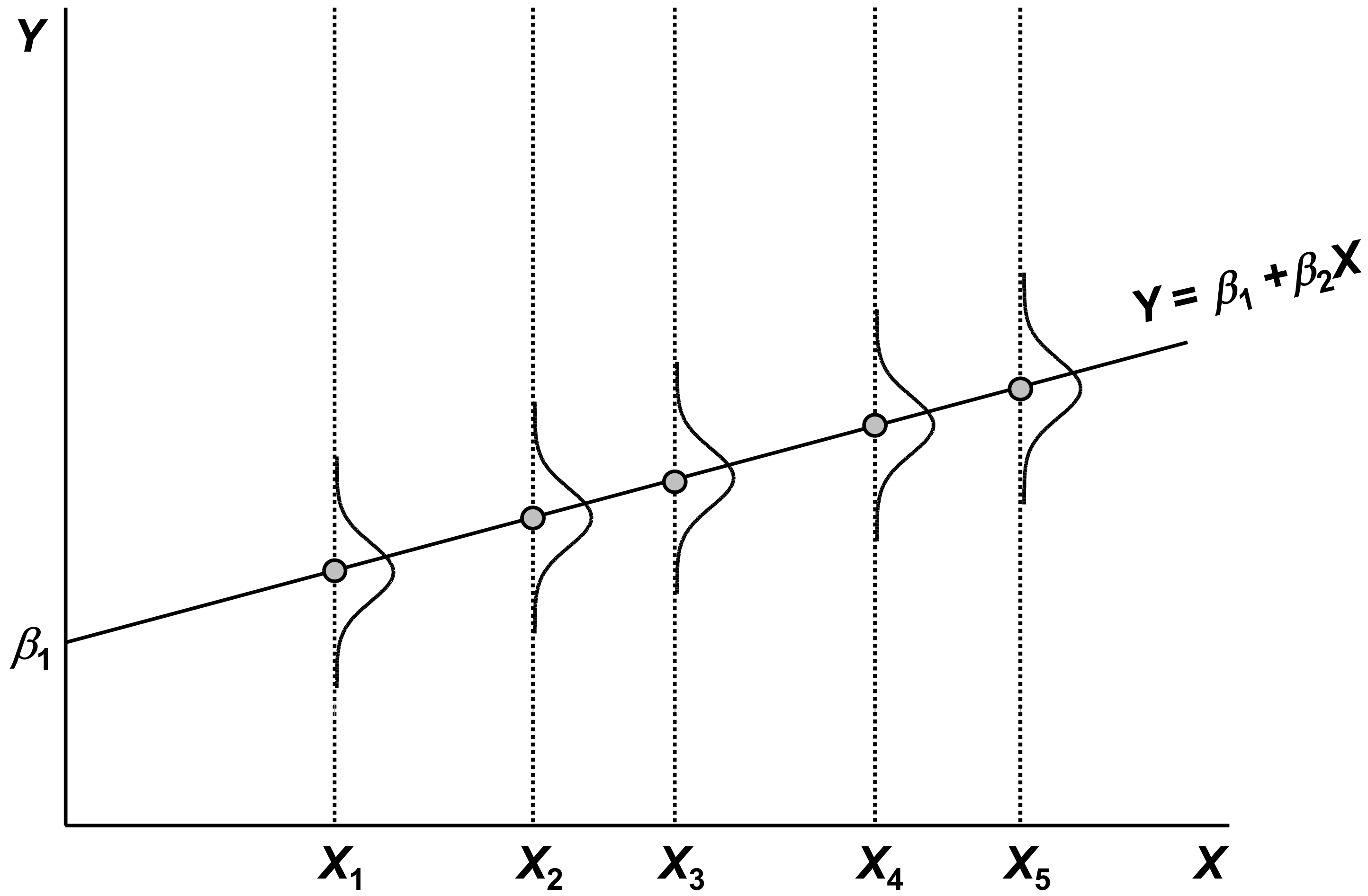
Было сделано также дополнительное предположение о нормальном законе распределения возмущений.

Гетероскедастичность



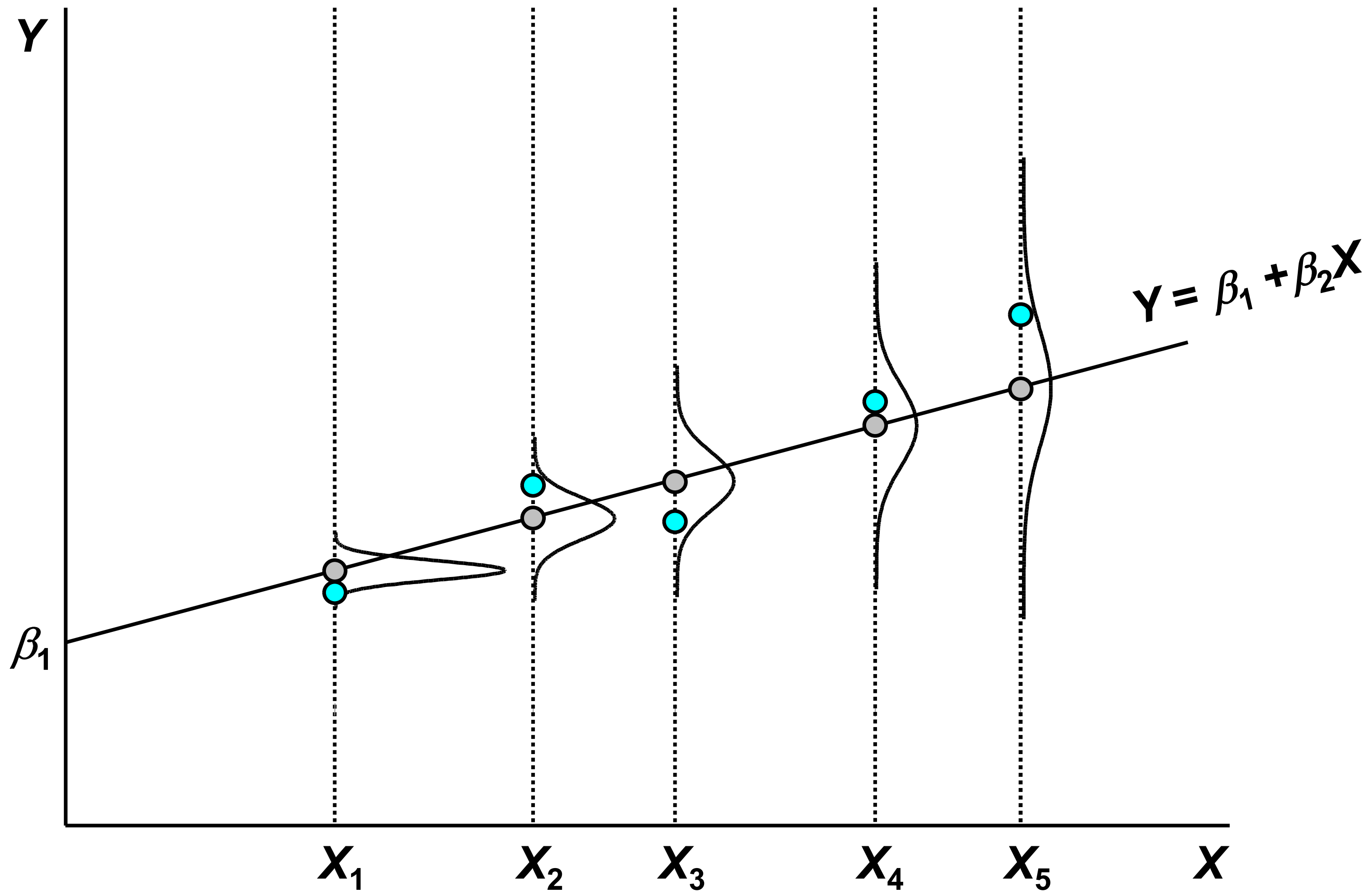
Свойство одинаковой дисперсии возмущений называется гомоскедастичностью.

Гетероскедастичность



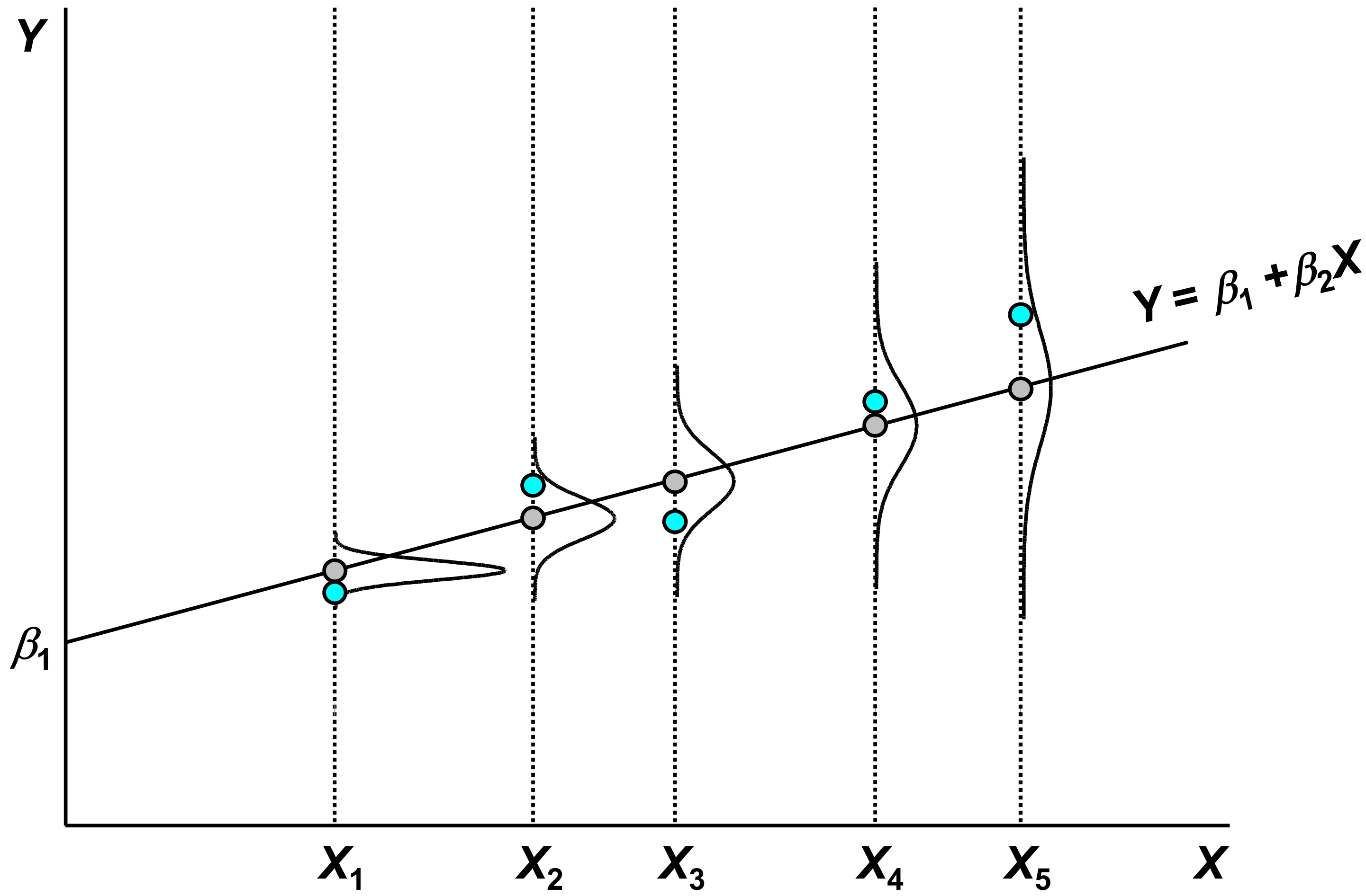
Линия теоретической регрессии $Y = \beta_1 + \beta_2 X$, которую мы не можем провести и проверить, одинаково ли распределены возмущения.

Гетероскедастичность



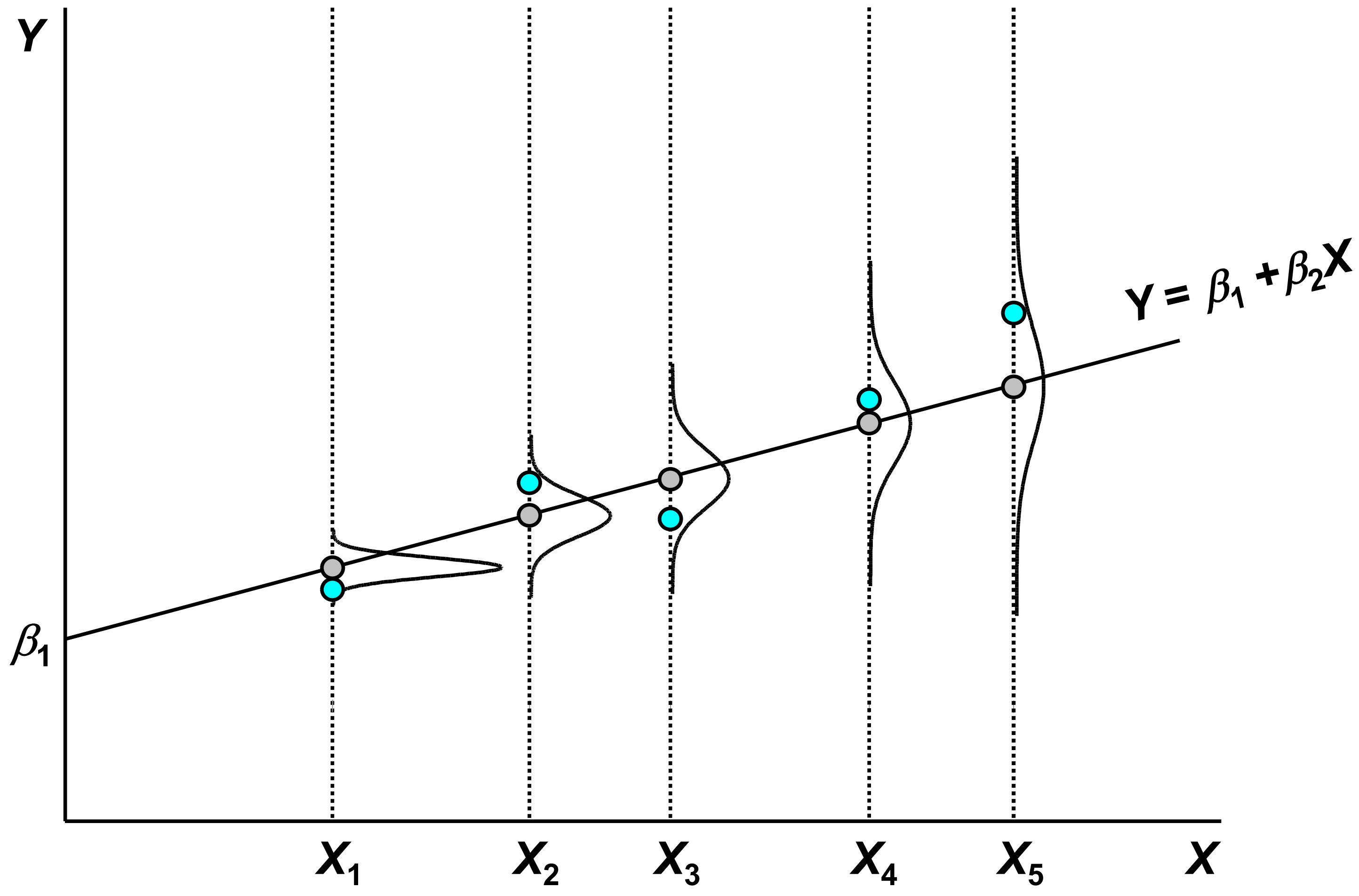
Если дисперсии возмущений различны, то это явление называется гетероскедастичностью.

Гетероскедастичность



Наличие гетероскедастичности можно заподозрить, если отклонение наблюдений от линии выборочной регрессии (остатки) достаточно сильно различаются.

Гетероскедастичность



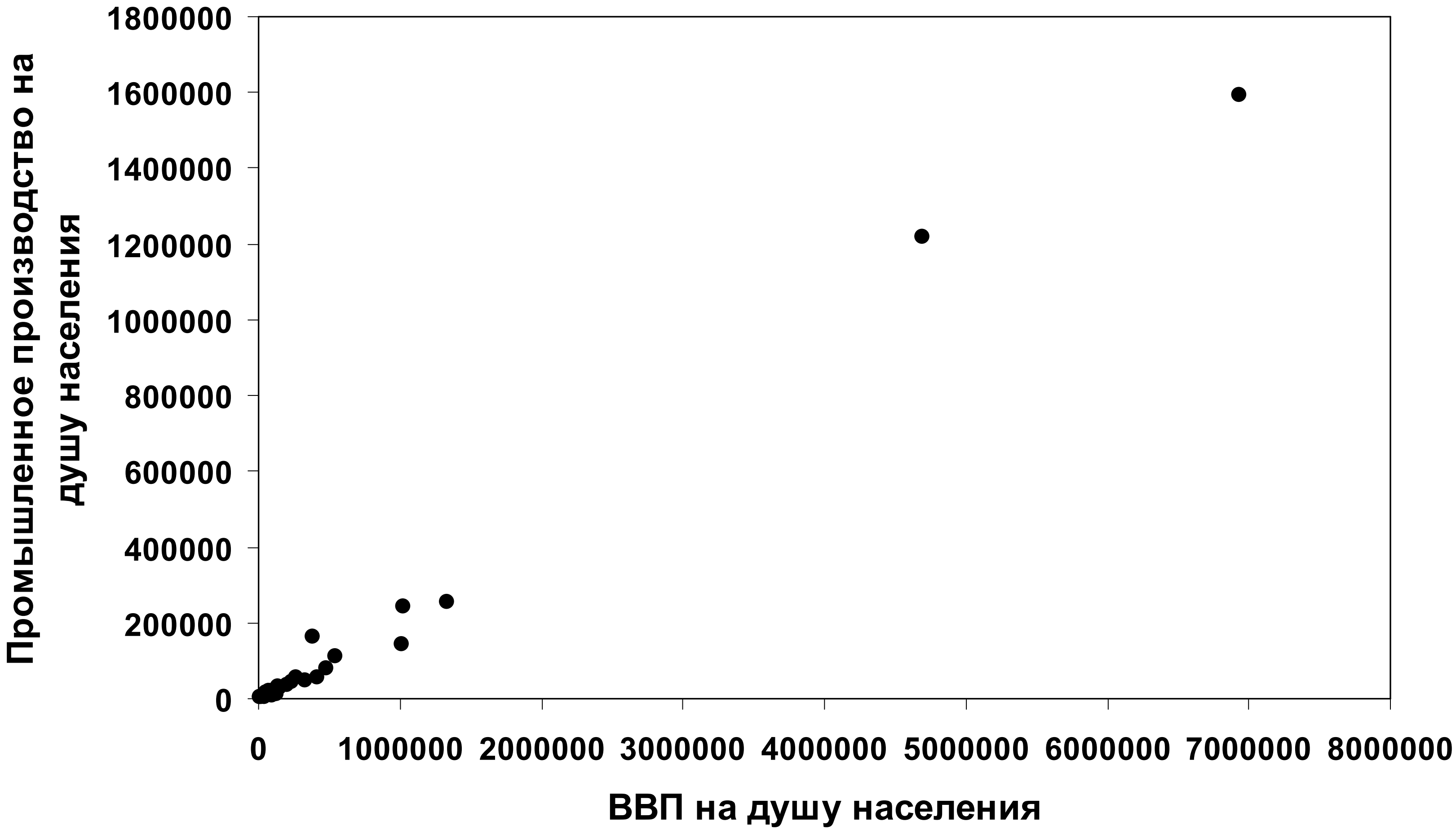
Однако ответ на вопрос, имеет ли место гетероскедастичность, можно получить только с помощью тестов.

Последствия гетероскедастичности

Если предположение об одинаковых дисперсиях возмущений не выполняется, то

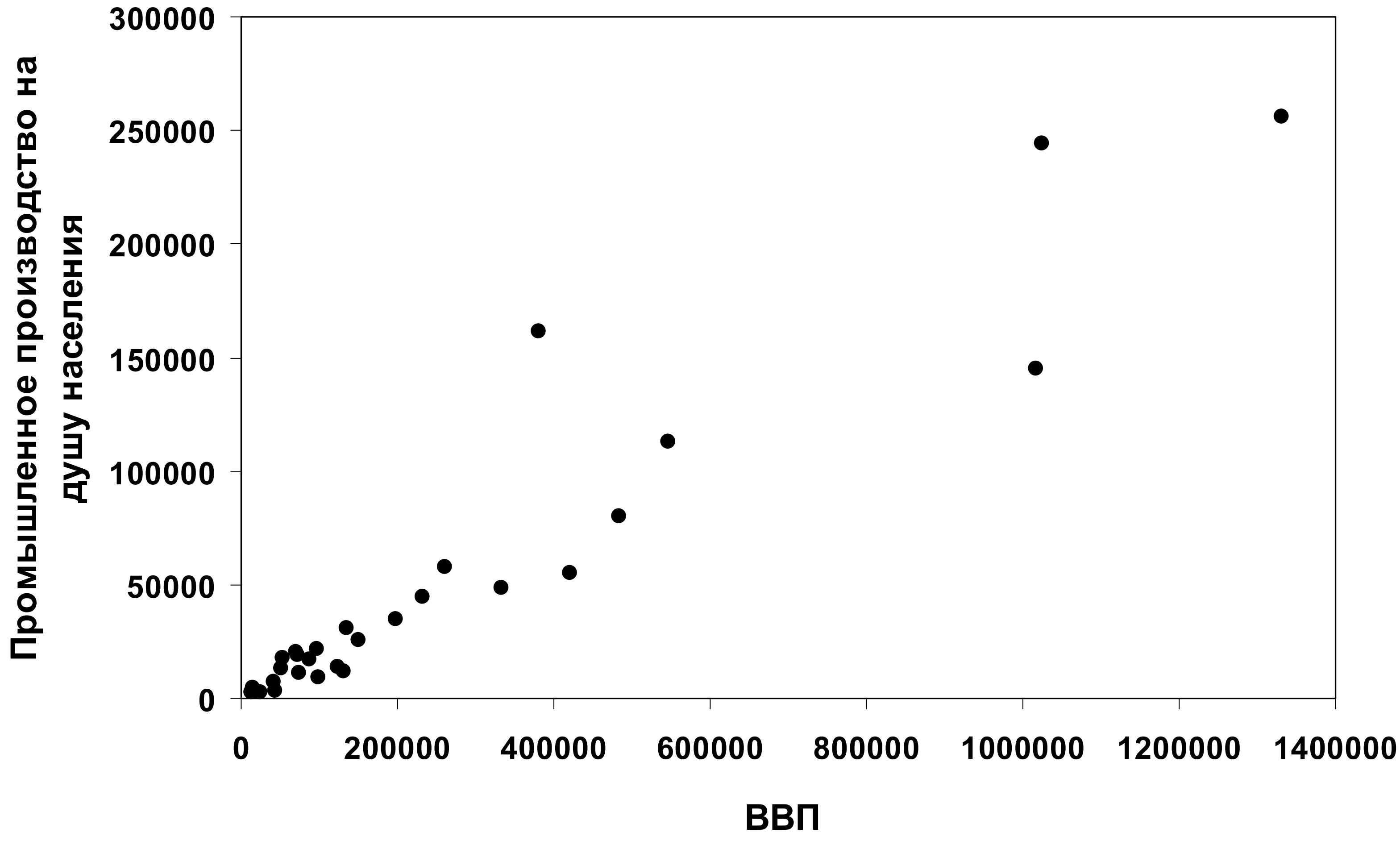
- **стандартные ошибки коэффициентов регрессии вычисляются по неверным формулам**
- **t – тесты для проверки гипотез о конкретных значениях коэффициентов не дают правильных результатов**
- **F – тесты для проверки гипотез о линейных ограничениях на коэффициенты регрессии не дают правильных результатов**
- **Оценки МНК коэффициентов регрессии больше не являются BEST, теряется эффективность оценок .**

Пример



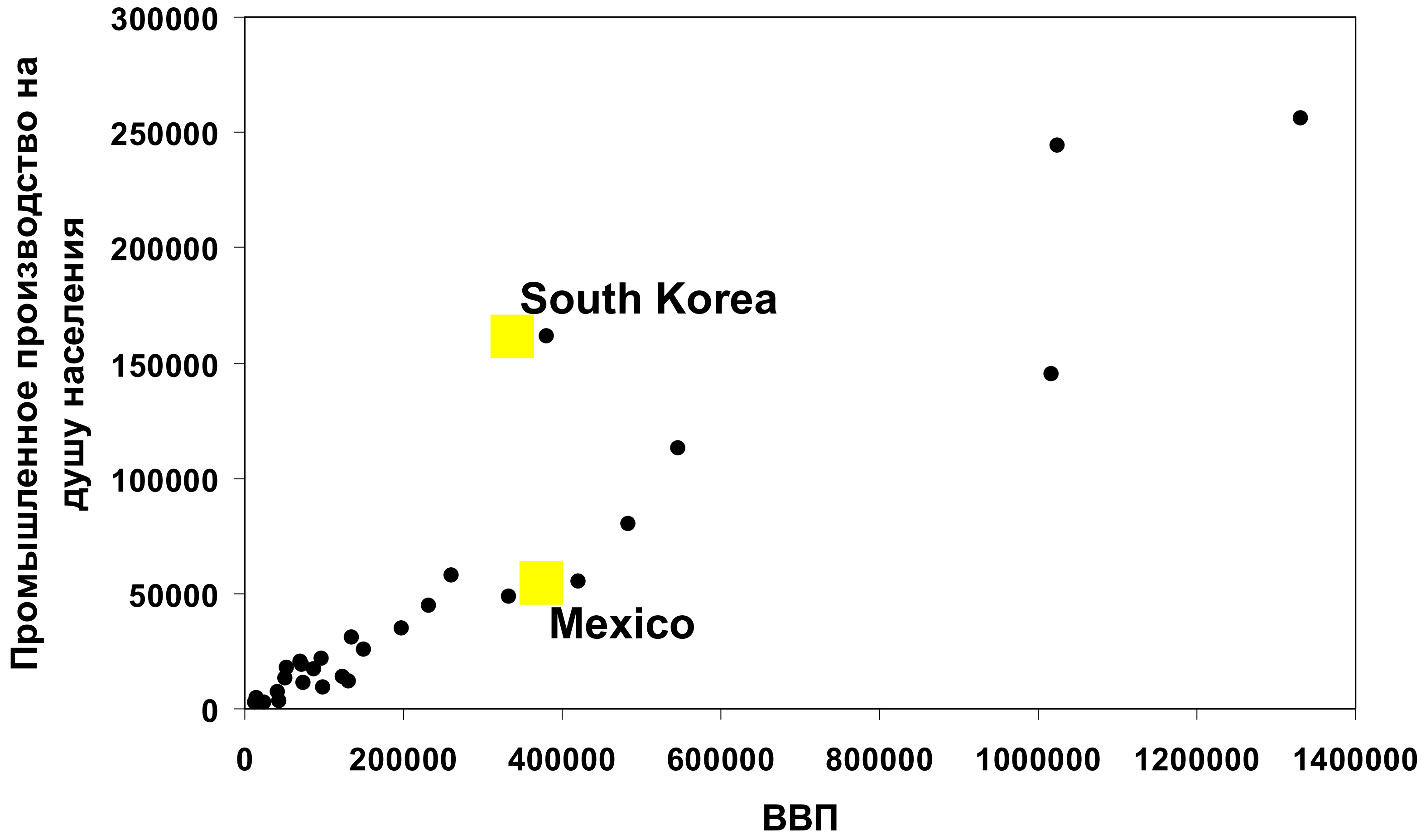
Данные для 30 стран в 1997.

Пример



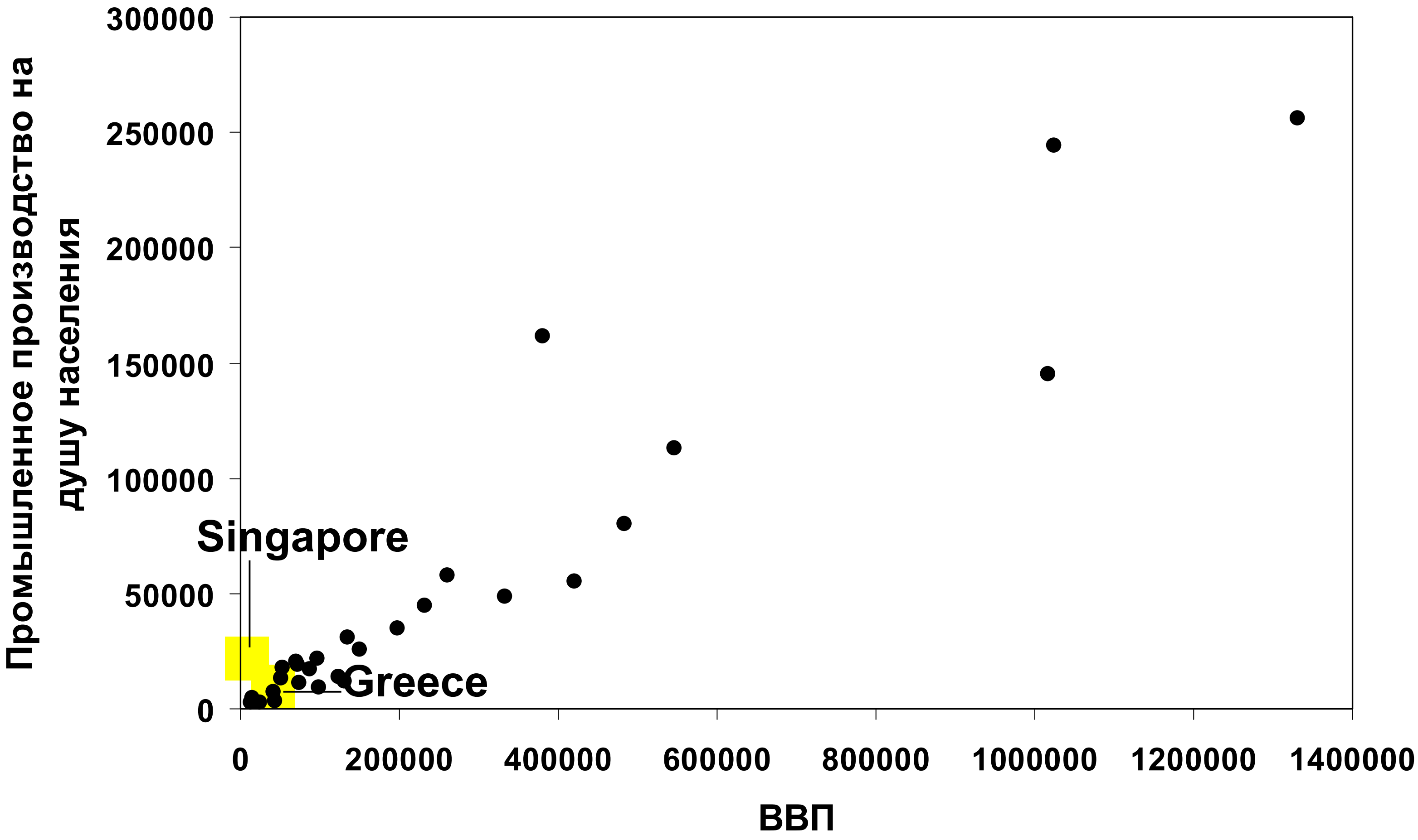
Взглянув на этот рисунок, можно сделать предположение, что с ростом ВВП дисперсия возмущений увеличивается.

Пример



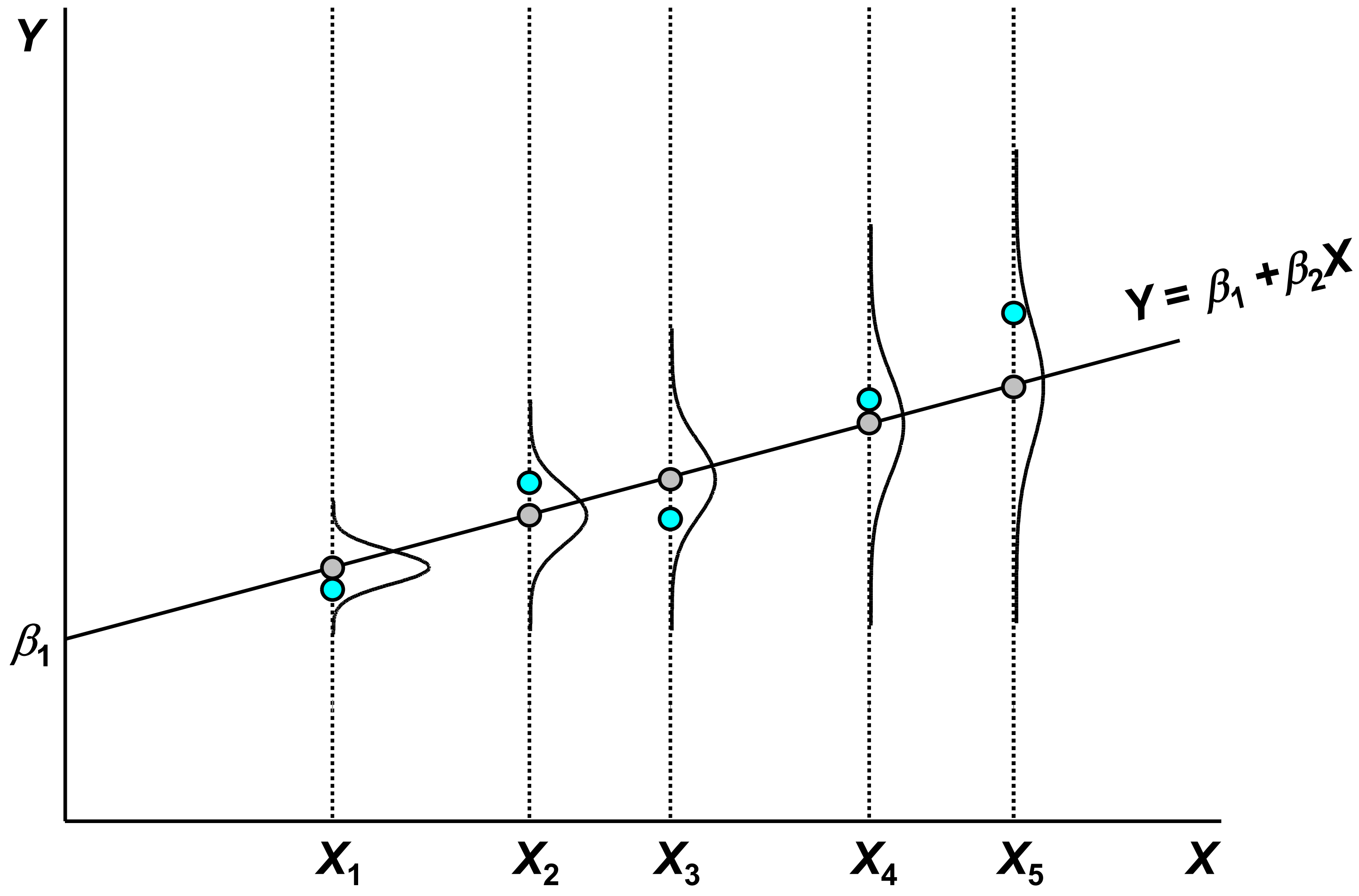
Сравним Южную Корею и Мексику с приблизительно одинаковым уровнем ВВП.

Пример



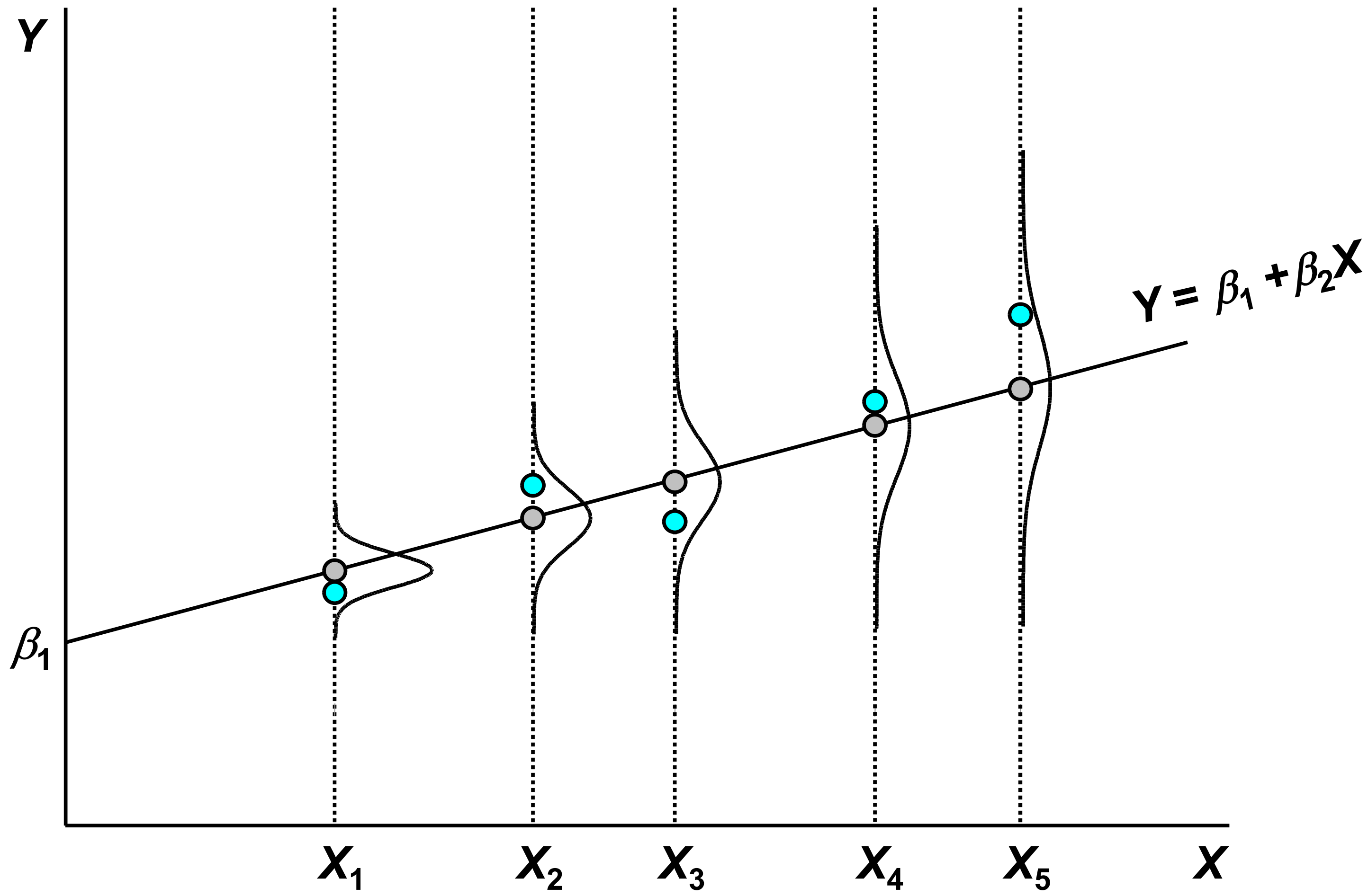
Другая пара для сравнения – Сингапур и Греция, также с почти одинаковым уровнем ВВП. Очевидно, что для первой пары с большим ВВП и разница больше. Можно предположить наличие гетероскедастичности.

Тест Голдфелда – Квандта



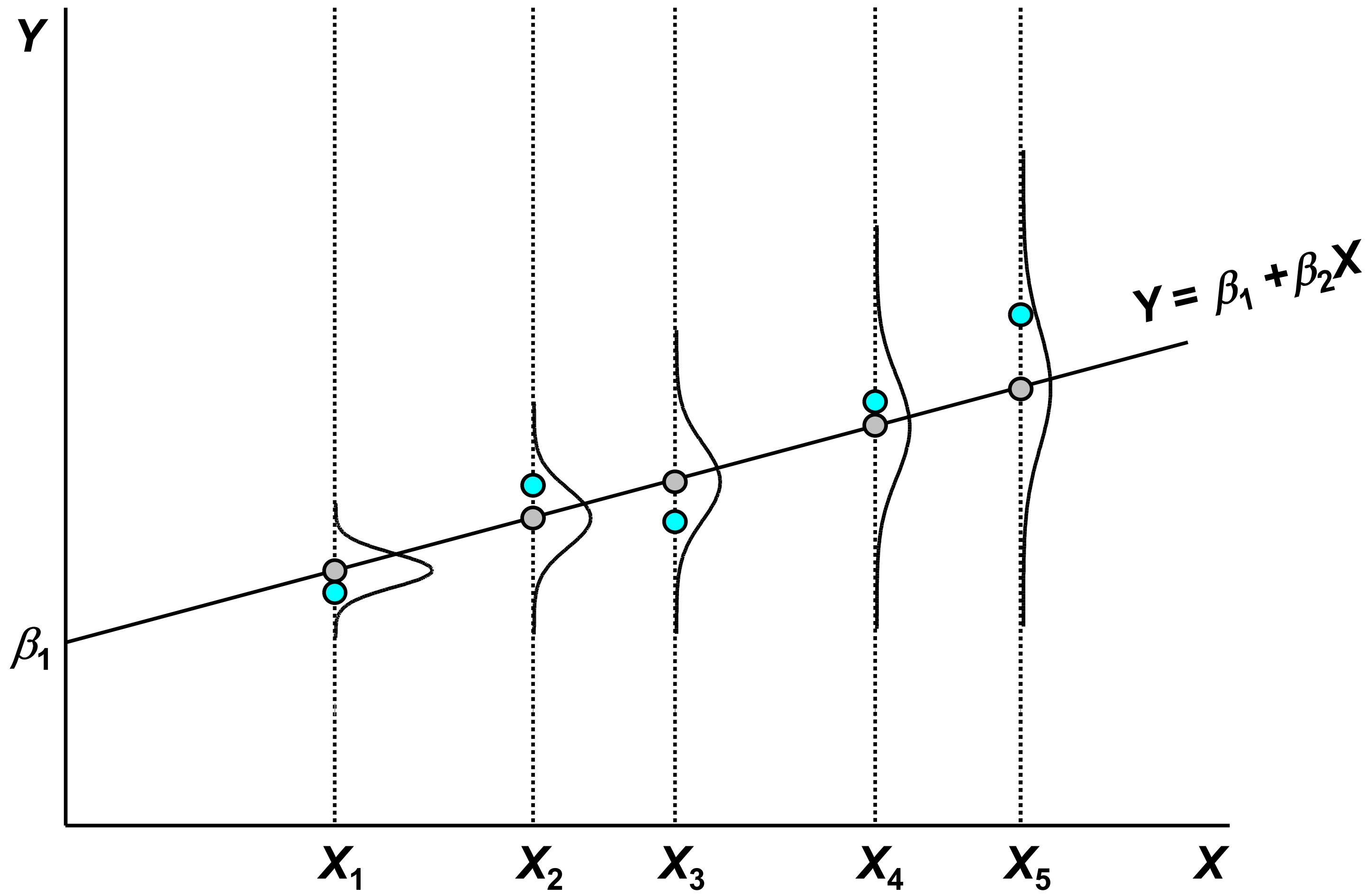
Гетероскедастичность – различие дисперсий возмущений для различных наблюдений. Ясно, что видов гетероскедастичности может быть сколь угодно много.

Тест Голдфелда – Квандта



Однако одним из самых распространенных видов гетероскедастичности является пропорциональность стандартного отклонения возмущений одной из объясняющих переменных.

Тест Голдфелда – Квандта



Этот тип гетероскедастичности иллюстрируется на приведенной диаграмме. Дисперсия возмущений пропорциональна переменной X.

Тест Голдфелда – Квандта

Основная и альтернативная гипотезы в тесте Голфелда – Квандта (и во всех остальных тестах, в которых проверяется, имеет ли место гетероскедастичность) формулируются следующим образом:

H_0: гомоскедастичность

H_1: гетероскедастичность

Однако сам тест зависит от того, какой вид гетероскедастичности мы предполагаем в альтернативной гипотезе.

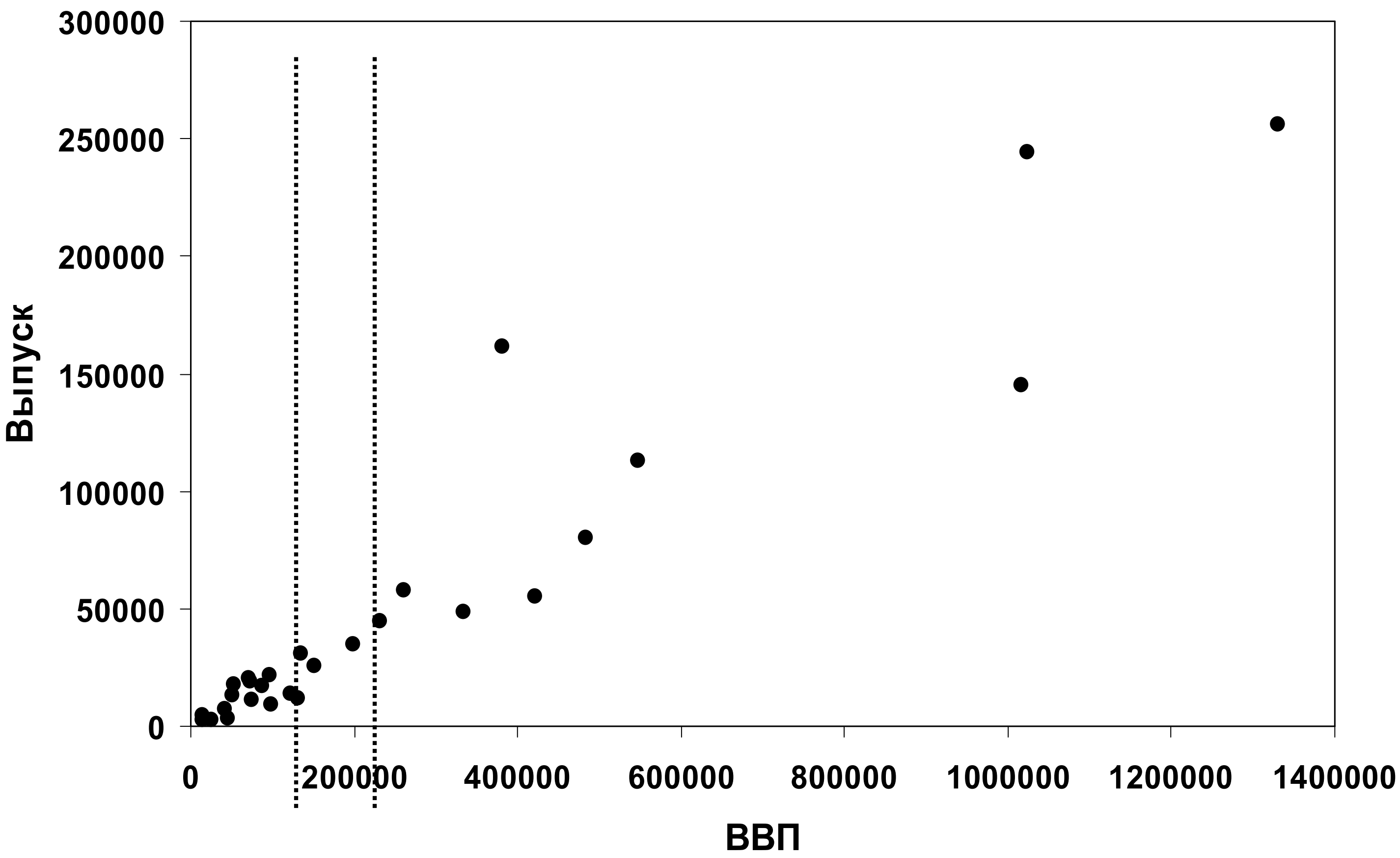
Тест Голдфелда – Квандта

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad D(u_i) = \sigma_i^2, \, i = 1, \ldots, n$$

H$_0$: $\sigma_i^2 = \sigma_u^2 \quad \forall i = 1, \ldots, n$

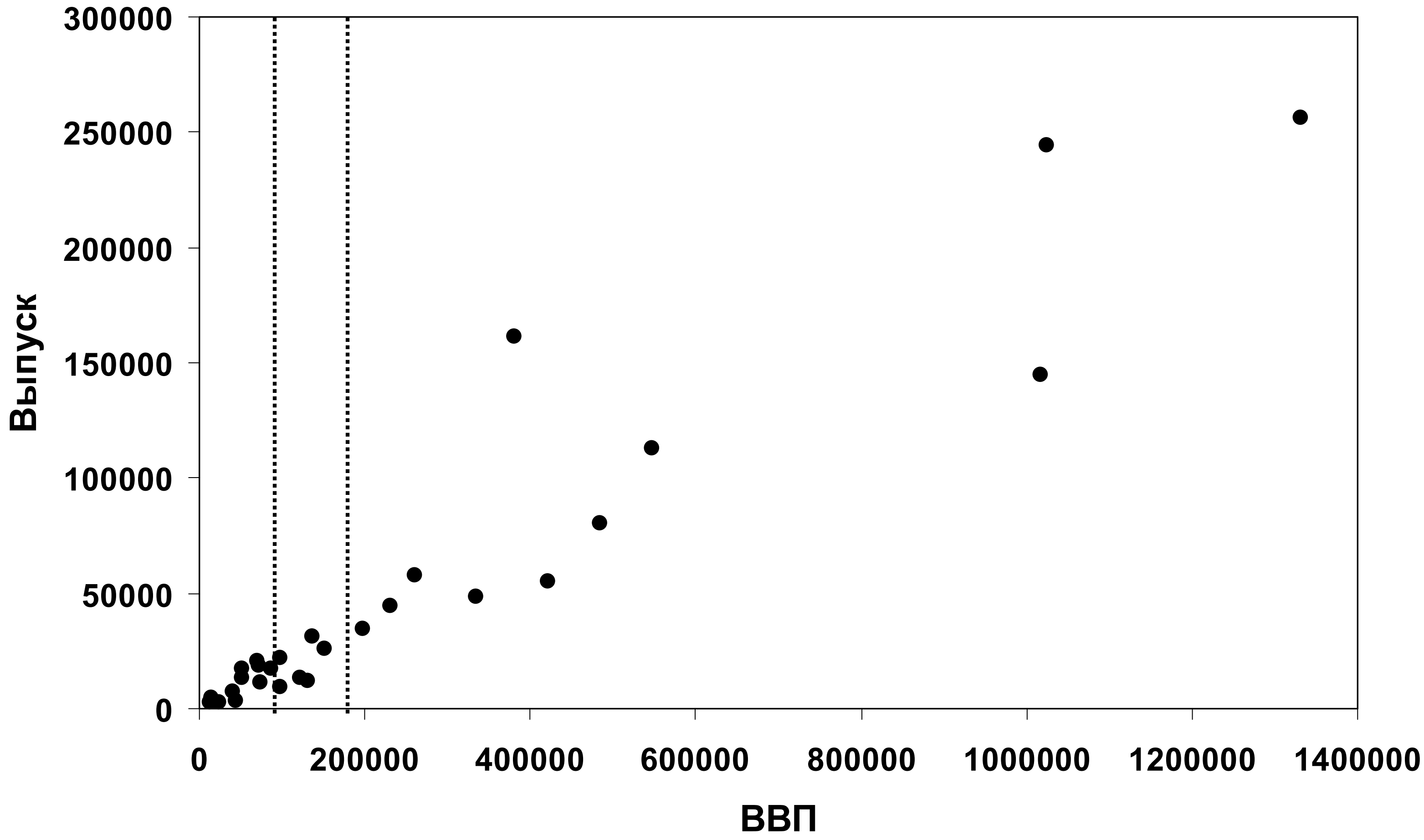
H$_1$: $\sigma_i \sim X_{ji} \quad \text{для} \quad \text{некоторого} \quad X_j, \, i = 1, \ldots, n$

Тест Голдфелда – Квандта



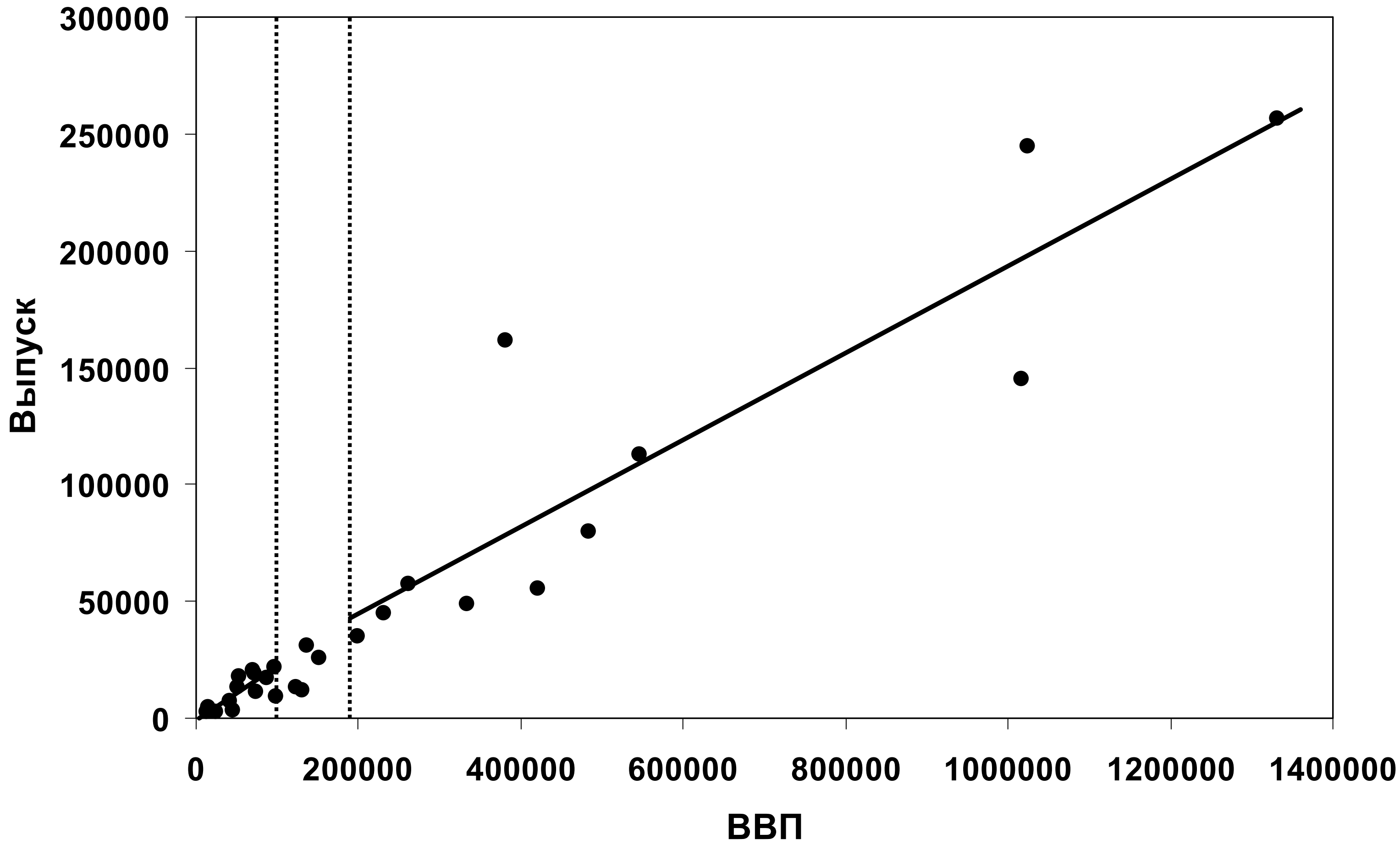
Для проведения теста Голдфелда – Квандта все наблюдения упорядочиваются по Xj и делятся на 3 части. Если выборка небольшая, то выделяют приблизительно 3/8 части всех наблюдений для первой и третьей части и приблизительно 1/4 в середине.

Пример проведения теста Голдфелда – Квандта



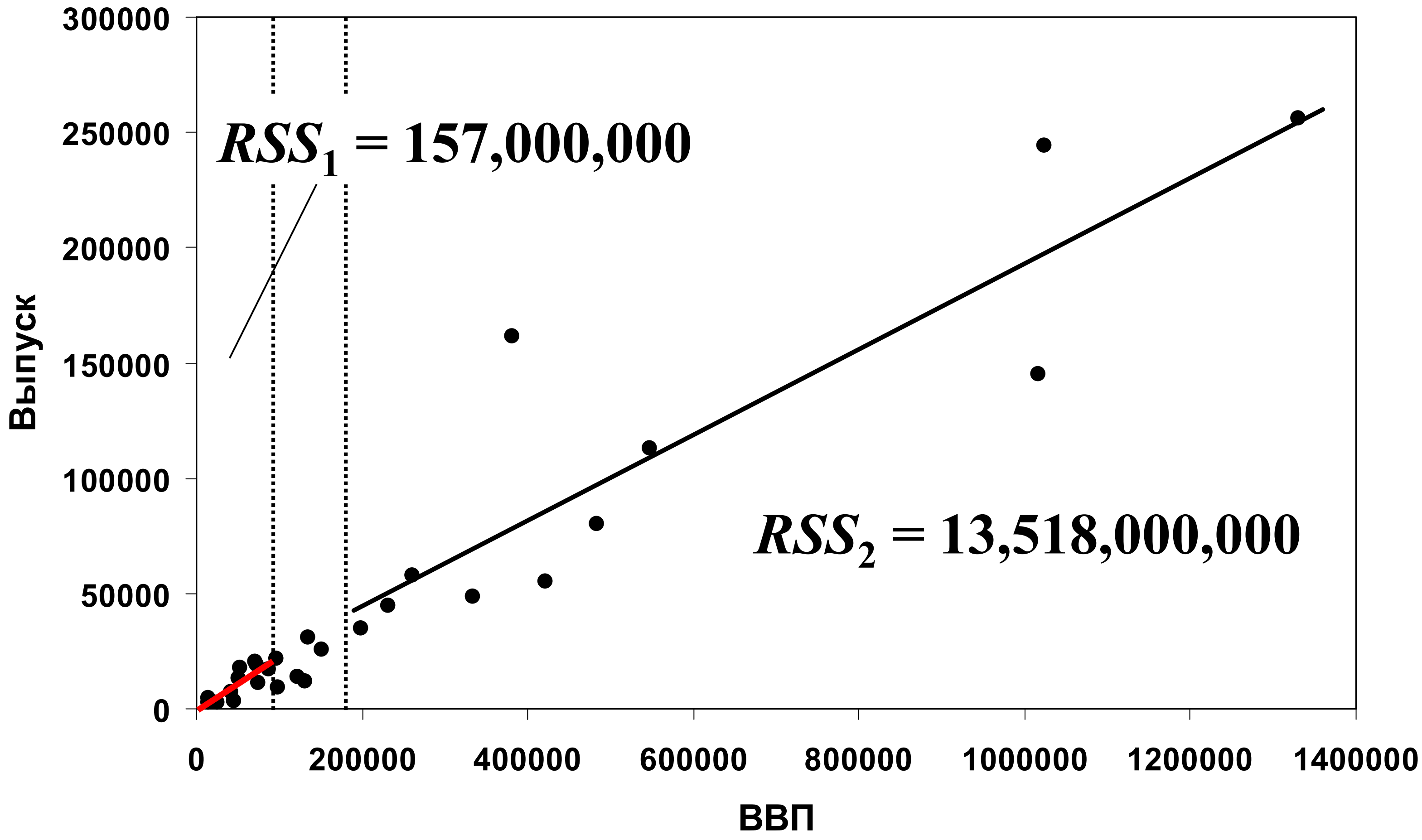
Для 28 стран оценивается зависимость выпуска продукции обрабатывающей промышленности от ВВП. Выделено 11 стран с маленьким ВВП, 6 со средним и 11 с большим.

Пример проведения теста Голдфелда – Квандта



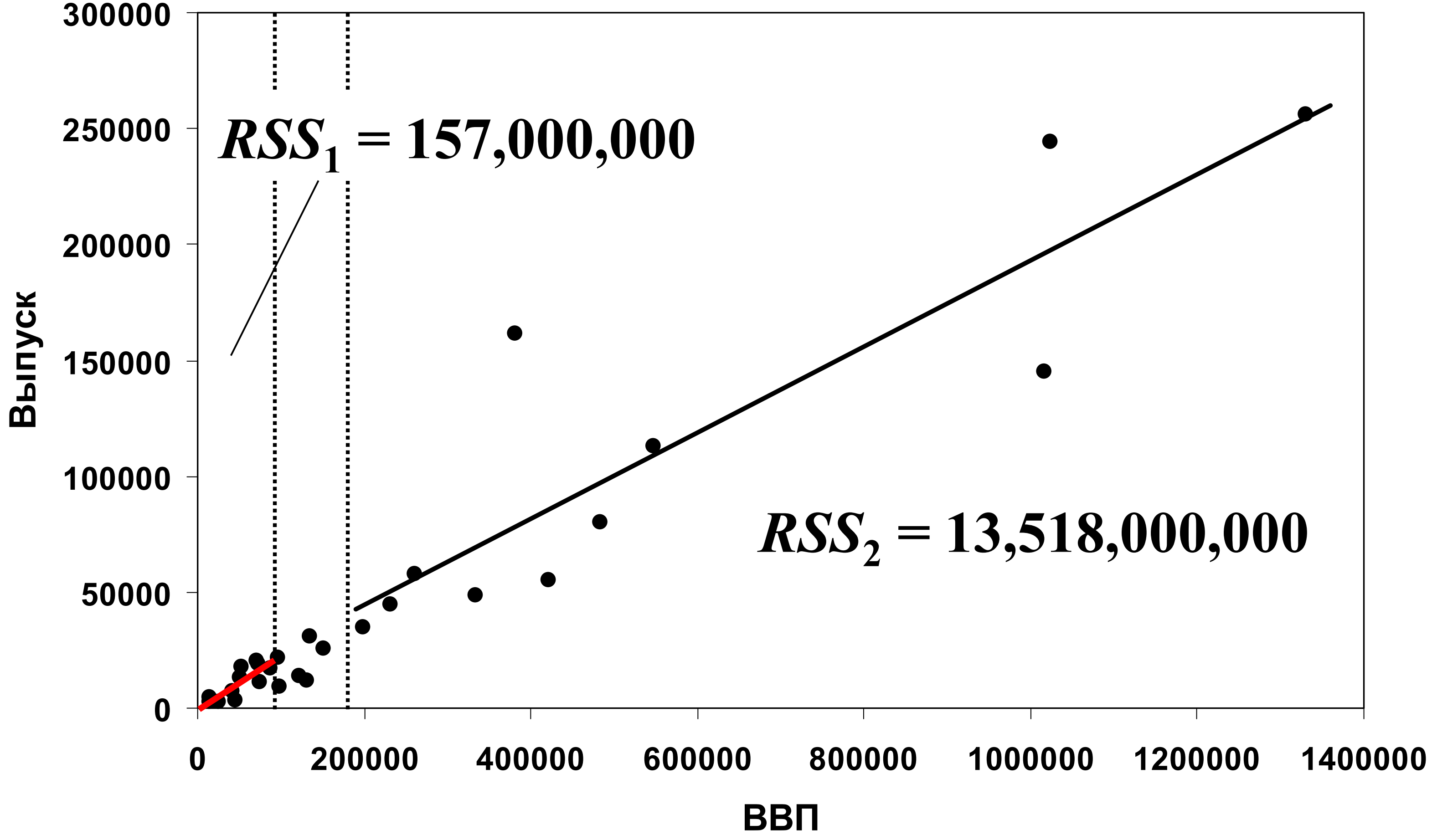
Отдельно оцениваются регрессии для 11 стран с маленьким ВВП и для 11 стран с большим ВВП.

Пример проведения теста Голдфелда – Квандта



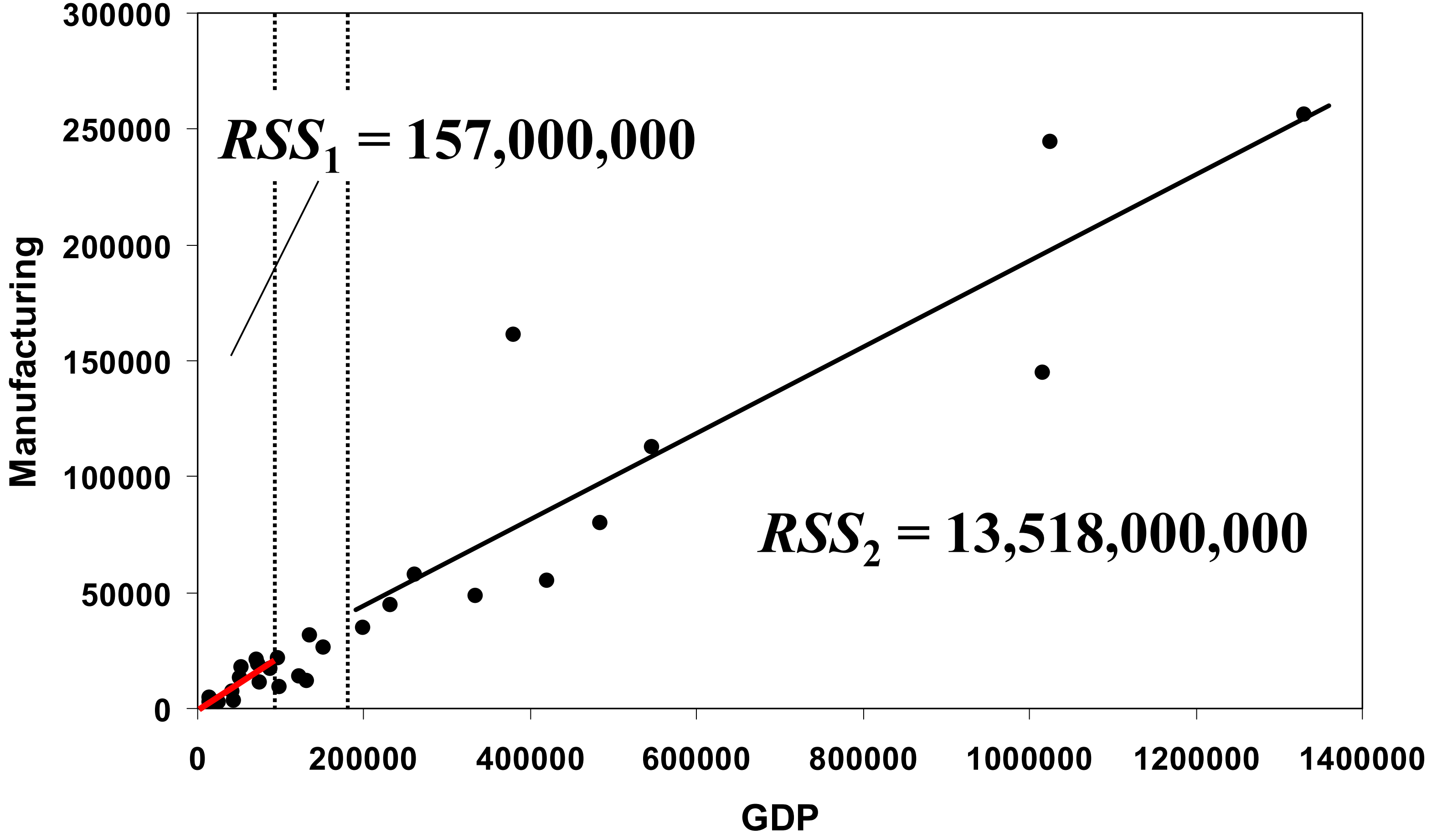
Для каждой регрессии находятся суммы квадратов остатков RSS_1 и RSS_2.

Пример проведения теста Голдфелда – Квандта



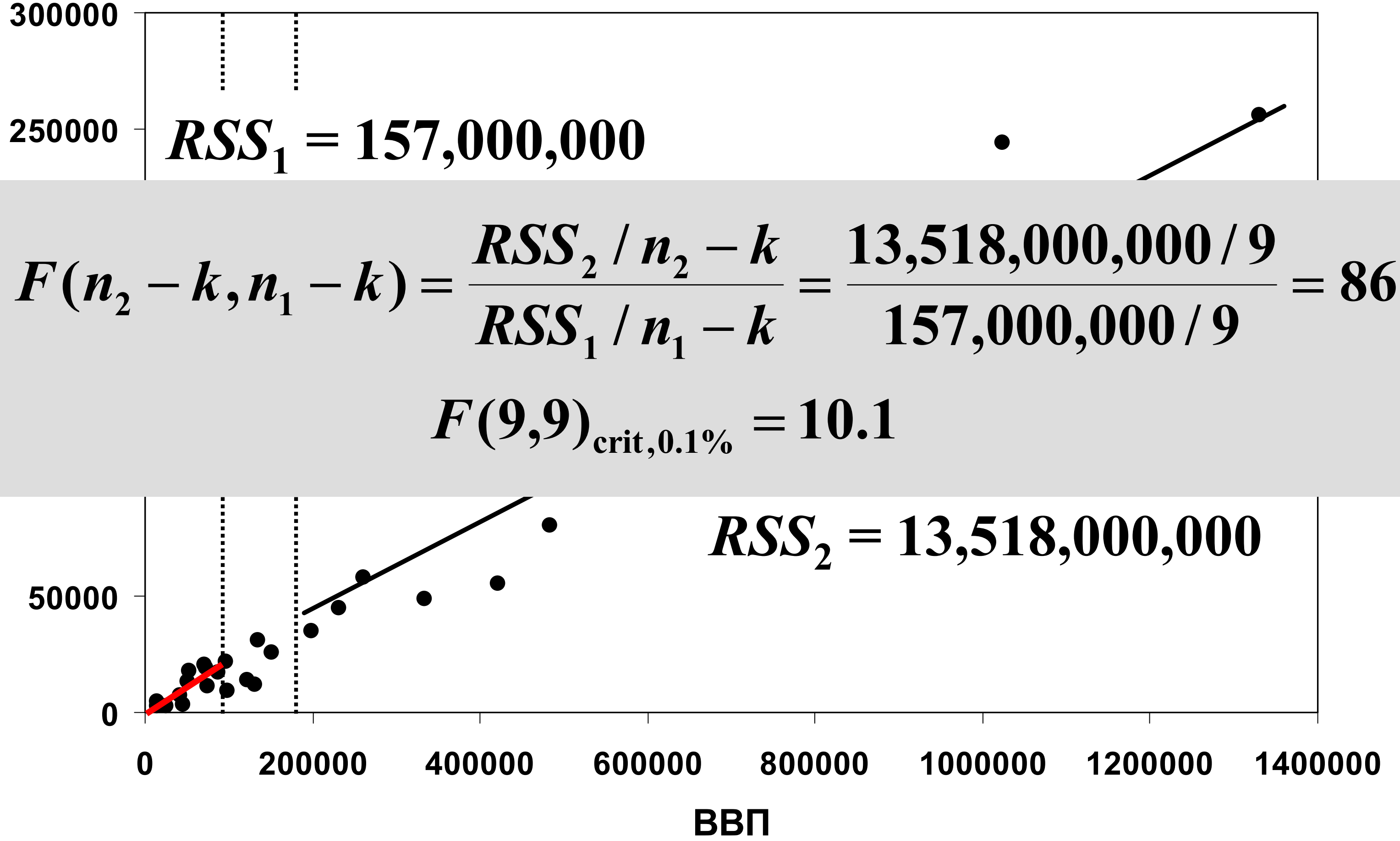
Если имеет место гомоскедастичность, RSS_1 и RSS_2 не должны сильно различаться (если число наблюдений в оцениваемых регрессиях совпадает).

Пример проведения теста Голдфелда – Квандта



Однако в рассматриваемом примере RSS_2 значительно превышает RSS_1.

Тестовая статистика в тесте Голдфелда - Квандта

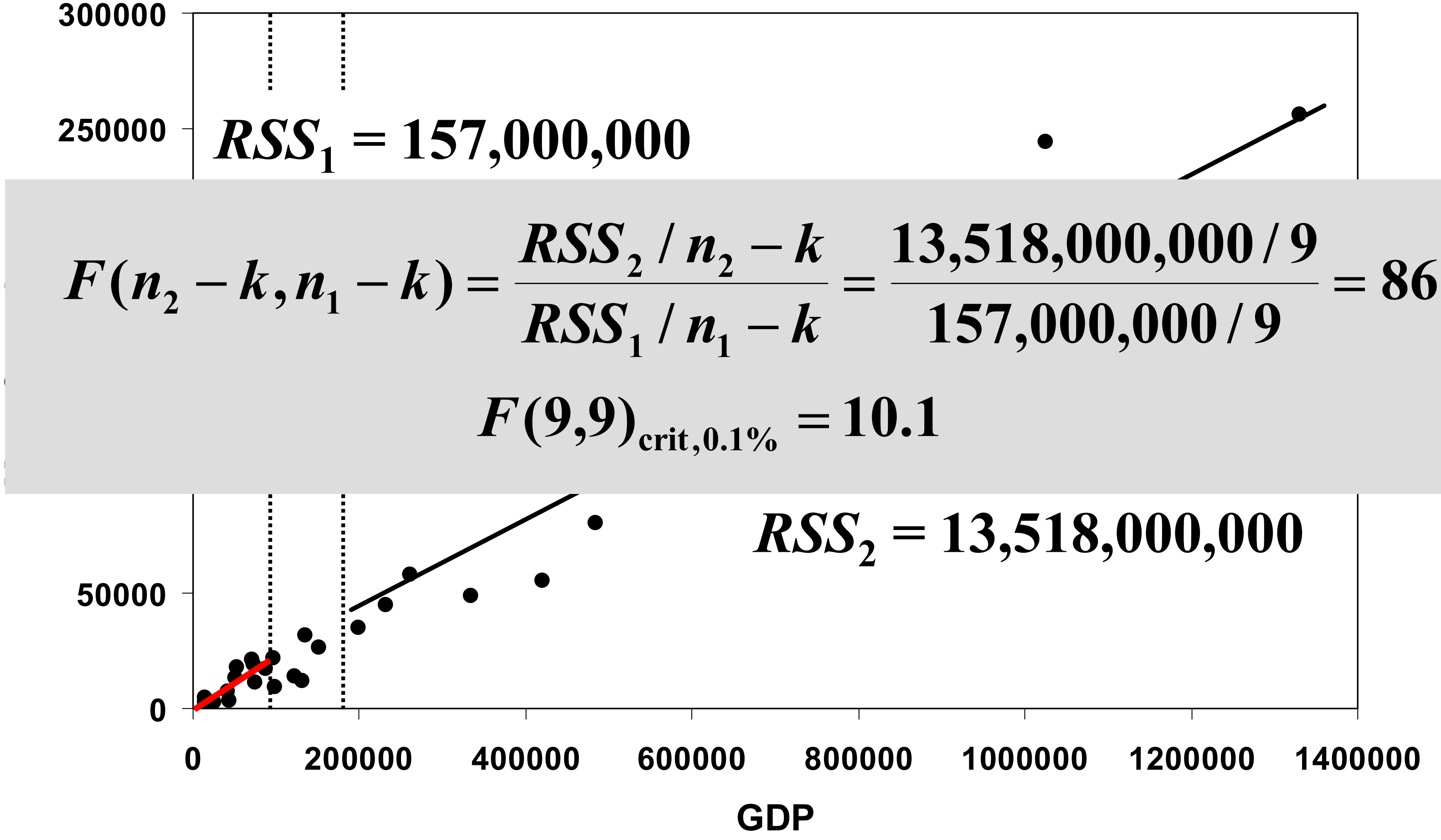


$$F(n_2 - k, n_1 - k) = \frac{RSS_2 / n_2 - k}{RSS_1 / n_1 - k} = \frac{13{,}518{,}000{,}000 / 9}{157{,}000{,}000 / 9} = 86.1$$

$$F(9,9)_{\text{crit},0.1\%} = 10.1$$

Тестовая статистика *F* рассчитывается по приведенной выше формуле. В числителе – оценка дисперсии возмущений по последним n_2 наблюдениям, а в знаменателе - оценка дисперсии возмущений по первым n_1 наблюдениям. *K* – число параметров в модели.

Тестовая статистика в тесте Голдфелда - Квандта



$$F(n_2-k, n_1-k) = \frac{RSS_2 / n_2 - k}{RSS_1 / n_1 - k} = \frac{13{,}518{,}000{,}000 / 9}{157{,}000{,}000 / 9} = 86.1$$

$$F(9,9)_{\text{crit},0.1\%} = 10.1$$

Тестовая F – статистика превышает критическое значение даже при уровне значимости 0.1% . Нулевая гипотеза о гомоскедастичности отвергается.

Формальное описание теста Голфелда - Квандта

- Оценивается регрессия по всем наблюдениям.
- Полезно взглянуть на график остатков. Может появиться предположение, что дисперсия возмущений увеличивается с ростом некоторой переменной.
- Упорядочиваем все наблюдения по модулю подозрительной переменной.
- Делим все наблюдения на три группы (если наблюдений достаточно много, то приблизительно на трети). Удобно, если в первой и третьей группах количество наблюдений одинаково.
- Наблюдениями средней группы пренебрегаем, а по первым n_1 и последним n_2 наблюдениям оцениваем отдельные регрессии.
- Используя суммы квадратов остатков (RSS) в оцененных регрессиях, рассчитываем тестовую статистику по формуле

$$F(n_2 - k, n_1 - k) = \frac{RSS_2/(n_2 - k)}{RSS_1/(n_1 - k)}$$

- Сравниваем полученное значение F – статистики с критическим (при выбранном уровне значимости).
- Если значение F – статистики превышает критическое, нулевая гипотеза о гомоскедастичности отвергается.

Тест Глейзера

Дисперсия возмущений не обязательно пропорциональна какому-либо фактору, может быть и другой вид зависимости, для определения которой используется тест Глейзера.

Тест Глейзера

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad D(u_i) = \sigma_i^2, i = 1, \ldots, n$$

H$_0$: $$\sigma_i^2 = \sigma_u^2 \quad \forall i = 1, \ldots, n$$

H$_1$: $$\sigma_i \sim X^{\gamma} \quad \textit{для некоторого} X_j,$$
$$\gamma = 1 \textit{ или } \gamma = 1/2 \textit{ или } \gamma = -1, \; i = 1, \ldots, n$$

Формальное описание теста Глейзера

- **Оценивается регрессия по всем наблюдениям.**
- **Сохраняются остатки регрессии e_i.**
- **Оцениваются регрессии**

$$|e_i| = \alpha + \beta X_i + u_i, i = 1, \ldots, n$$

$$|e_i| = \alpha + \beta \sqrt{X_i} + u_i, i = 1, \ldots, n$$

$$|e_i| = \alpha + \beta \frac{1}{X_i} + u_i, i = 1, \ldots, n$$

- **Если коэффициент β значим хотя бы в одной из трех регрессий (значимость коэффициента проверяется как обычно с помощью t – статистики), то имеет место гетероскедастичность (соответствующего вида).**

Тест Уайта

H_0: гомоскедастичность

H_1: гетероскедастичность

Вид гетероскедастичности не конкретизируется.

Формальное описание теста Уайта

- Оценивается регрессия по всем наблюдениям.
- Сохраняются остатки регрессии e_i.
- Оцениваются регрессия квадратов остатков на все регрессоры, их квадраты, попарные произведения и константу.
- В последней оцененной регрессии находим коэффициент множественной детерминации R^2
- Вычисляем тестовую статистику по формуле nR^2. Тестовая статистика имеет распределение «хи – квадрат» с k-1 степенями свободы, где k – число оцениваемых коэффициентов.
- Сравниваем полученное значение тестовой статистики с критическим при выбранном уровне значимости. Если значение тестовой статистики превышает критическое, то нулевая гипотеза о гомоскедастичности отвергается.

Тест Уайта

Привлекательной чертой теста Уайта является его универсальность. Однако этот тест не является конструктивным. Если гетероскедастичность выявлена, то тест Уайта не дает указания на функциональную форму гетероскедастичности. Единственным способом коррекции является применение стандартных ошибок в форме Уайта.

Пример применения теста Уайта

```
. reg MANU GDP

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  1,    26) =  210.73
       Model |  1.1600e+11     1  1.1600e+11           Prob > F      =  0.0000
    Residual |  1.4312e+10    26   550462775           R-squared     =  0.8902
-------------+------------------------------           Adj R-squared =  0.8859
       Total |  1.3031e+11    27  4.8264e+09           Root MSE      =   23462

------------------------------------------------------------------------------
        MANU |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         GDP |    .193693   .0133428    14.52   0.000     .1662665    .2211195
       _cons |   603.9453   5699.677     0.11   0.916    -11111.91     12319.8
------------------------------------------------------------------------------

. predict EMANU, resid
. gen EMANUSQ = EMANU*EMANU
```

Оценена регрессия выпуска от ВВП для 28 стран. Остатки регрессии сохранены.

Проведение теста Уайта

```
. gen GDPSQ = GDP*GDP

. reg EMANUSQ GDP GDPSQ

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    25) =    3.35
       Model |  1.3183e+19     2  6.5913e+18           Prob > F      =  0.0514
    Residual |  4.9179e+19    25  1.9671e+18           R-squared     =  0.2114
-------------+------------------------------           Adj R-squared =  0.1483
       Total |  6.2361e+19    27  2.3097e+18           Root MSE      = 1.4e+09

------------------------------------------------------------------------------
     EMANUSQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         GDP |   6271.896   2758.253     2.27   0.032     591.1687    11952.62
       GDPSQ |  -.0041155   .0022626    -1.82   0.081    -.0087754    .0005444
       _cons |  -4.21e+08   4.51e+08    -0.93   0.359    -1.35e+09    5.08e+08
------------------------------------------------------------------------------
```

В качестве зависимой переменной выбраны квадраты остатков, а в качестве независимых переменных – GDP и GDP в квадрате (т.к. переменная одна, то попарных произведений нет). В регрессию включена также константа.

Проведение теста Уайта

```
. gen GDPSQ = GDP*GDP

. reg EMANUSQ GDP GDPSQ

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    25) =    3.35
       Model |  1.3183e+19     2  6.5913e+18           Prob > F      =  0.0514
    Residual |  4.9179e+19    25  1.9671e+18           R-squared     =  0.2114
-------------+------------------------------           Adj R-squared =  0.1483
       Total |  6.2361e+19    27  2.3097e+18           Root MSE      = 1.4e+09

------------------------------------------------------------------------------
     EMANUSQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         GDP |   6271.896   2758.253     2.27   0.032     591.1687    11952.62
       GDPSQ |  -.0041155   .0022626    -1.82   0.081    -.0087754    .0005444
       _cons |  -4.21e+08   4.51e+08    -0.93   0.359    -1.35e+09    5.08e+08
------------------------------------------------------------------------------
```

Вычисляем тестовую статистику по формуле nR^2, беря R^2 из таблицы.

Проведение теста Уайта

```
. gen GDPSQ = GDP*GDP

. reg EMANUSQ GDP GDPSQ

      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  2,    25) =    3.35
       Model |  1.3183e+19     2  6.5913e+18           Prob > F      =  0.0514
    Residual |  4.9179e+19    25  1.9671e+18           R-squared     =  0.2114
-------------+------------------------------           Adj R-squared =  0.1483
       Total |  6.2361e+19    27  2.3097e+18           Root MSE      = 1.4e+09

------------------------------------------------------------------------------
     EMANUSQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         GDP |   6271.896   2758.253     2.27   0.032     591.1687    11952.62
       GDPSQ |  -.0041155   .0022626    -1.82   0.081    -.0087754    .0005444
       _cons |  -4.21e+08   4.51e+08    -0.93   0.359    -1.35e+09    5.08e+08
------------------------------------------------------------------------------
```

$$nR^2 = 28 \times 0.2114 = 5.92 \qquad \chi^2_{5\%}(2) = 5.99$$

R^2 равен 0.2114 и n равно 28. Тестовая статистика равна 5.92. Критическое значение статистики «хи –квадрат» с двумя степенями свободы равно 5.99 при 5 % уровне значимости. Полученное значение тестовой статистики не превышает критическое, следовательно, нулевая гипотеза о гомоскедастичности не отвергается.

Тест Бройша - Пагана

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad D(u_i) = \sigma_i^2, i = 1, \ldots, n$$

H$_0$: $\sigma_i^2 = \sigma_u^2 \quad \forall i = 1, \ldots, n$

H$_1$: $\sigma_i^2 \sim f(\alpha_0 + \alpha_1 Z_{1i} + .. + \alpha_r Z_{ri}),$
для некоторых переменных $Z_1, \ldots, Z_r$
$i = 1, \ldots, n$

Вид функции f может быть любой.

Формальное описание теста Бройша - Пагана

- **Оцениваетcя регрессия**

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \; i = 1, \ldots, n$$

по всем наблюдениям. Сохраняются остатки регрессии e_i,

$i = 1,\ldots,n$. Находится RSS.

- **Находится оценка дисперсии возмущений по формуле**

$$\hat{\sigma}_u^2 = \frac{RSS}{n}$$

- **Оценивается регрессия e^2 на $Z_1,\ldots,Z_r$, находится ESS_0.**

Формальное описание теста Бройша - Пагана

•**Тестовая статистика**

$$\chi^2 = \frac{ESS_0}{2\hat{\sigma}^4}$$

•**Имеет распределение «хи – квадрат» с r степенями свободы.**

•**Если** $\chi^2 > \chi^2_{cr,\alpha}(r)$ **при выбранном уровне значимости,**

то гипотеза H_0 о гомоскедастичности отвергается.

Что делать в случае гетероскедастичности?

$$Y = \beta_1 + \beta_2 X + u$$

Предположим, что нам известны дисперсии возмущений σ_i^2 для всех наблюдений i = 1,…,n.

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $u_i = \sigma_i^2$

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i}$$

Разделим обе части равенства на σ_i для каждого наблюдения.

$$\boldsymbol{Y = \beta_1 + \beta_2 X + u}$$

дисперсия $\quad u_i = \sigma_i^2$

$$\boldsymbol{\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i}}$$

дисперсия новых возмущений $\left\{\frac{u_i}{\sigma_i}\right\} = \frac{1}{\sigma_i^2}$ дисперсии $\quad u_i$

$$= \frac{\sigma_i^2}{\sigma_i^2} = 1$$

Тогда дисперсии возмущений в новой регрессии станут одинаковыми и равными 1.

Преобразование переменных

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $u_i = \sigma_i^2$

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i}$$

$$Y' = \beta_1 H + \beta_2 X' + u' \quad Y' = \frac{Y_i}{\sigma_i}, \quad H = \frac{1}{\sigma_i}, \quad X' = \frac{X_i}{\sigma_i}, \quad u' = \frac{u_i}{\sigma_i}$$

Все сводится к оценке новой регрессии с преобразованными факторами, оцениваем регрессию *Y'* на *X'* и *H*, которые определенны выше. Отметим, что в новой регрессии нет константы. β_1 становится коэффициентом наклона перед переменной $1/\sigma_i$.

Взвешенный метод наименьших квадратов

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $u_i = \sigma_i^2$

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i}$$

$$Y' = \beta_1 H + \beta_2 X' + u' \qquad Y' = \frac{Y_i}{\sigma_i}, \quad H = \frac{1}{\sigma_i}, \quad X' = \frac{X_i}{\sigma_i}, \quad u' = \frac{u_i}{\sigma_i}$$

Указанный метод называется взвешенным методом наименьших квадратов. Наибольший вес 1/σ_i получают наблюдения с наименьшей дисперсией возмущений σ_i.

Взвешенный метод наименьших квадратов

$$Y = \beta_1 + \beta_2 X + u$$

$$\text{variance of } u_i = \sigma_i^2$$

$$\sigma_i = \lambda Z_i$$

Однако на практике стандартные отклонения возмущений обычно неизвестны. Но, оказывается, достаточно знать эти стандартные отклонения с точностью до постоянного множителя. Предположим, что стандартные отклонения возмущений пропорциональны некоторой известной переменной Z_i.

Взвешенный метод наименьших квадратов

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $u_i = \sigma_i^2$

$$\sigma_i = \lambda Z_i$$

$$\frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_i}{Z_i} + \frac{u_i}{Z_i}$$

В этом случае мы достигаем гомоскедастичности остатков, разделив все переменные на Z_i.

Взвешенный метод наименьших квадратов

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $u_i = \sigma_i^2$

$$\sigma_i = \lambda Z_i$$

$$\frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_i}{Z_i} + \frac{u_i}{Z_i}$$

дисперсия $\left\{\frac{u_i}{Z_i}\right\} = \frac{1}{Z_i^2}\sigma_i^2 = \frac{\sigma_i^2}{\sigma_i^2 / \lambda^2} = \lambda^2$

$$Y' = \beta_1 H + \beta_2 X' + u' \qquad Y' = \frac{Y_i}{Z_i},\ H = \frac{1}{Z_i},\ X' = \frac{X_i}{Z_i},\ u' = \frac{u_i}{Z_i}$$

Действительно, как показано выше, дисперсии новых остатков одинаковы и равны λ^2. Нам нет необходимости знать λ^2. Достаточно того, что это константа (т.е. одинаковые дисперсии для всех возмущений, гомоскедастичность) .

Взвешенный метод наименьших квадратов

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $u_i = \sigma_i^2$

$$\sigma_i = \lambda Z_i$$

$$\frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_i}{Z_i} + \frac{u_i}{Z_i}$$

$$Y' = \beta_1 H + \beta_2 X' + u' \qquad Y' = \frac{Y_i}{Z_i},\ H = \frac{1}{Z_i},\ X' = \frac{X_i}{Z_i},\ u' = \frac{u_i}{Z_i}$$

Если после выполнении теста Голдфелда – Квандта гипотеза о гомоскедастичности отвергается, то в качестве Z может быть использована переменная X_j.

Взвешенный метод наименьших квадратов

$$Y = \beta_1 + \beta_2 X + u$$

дисперсия $\quad u_i = \sigma_i^2$

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i}$$

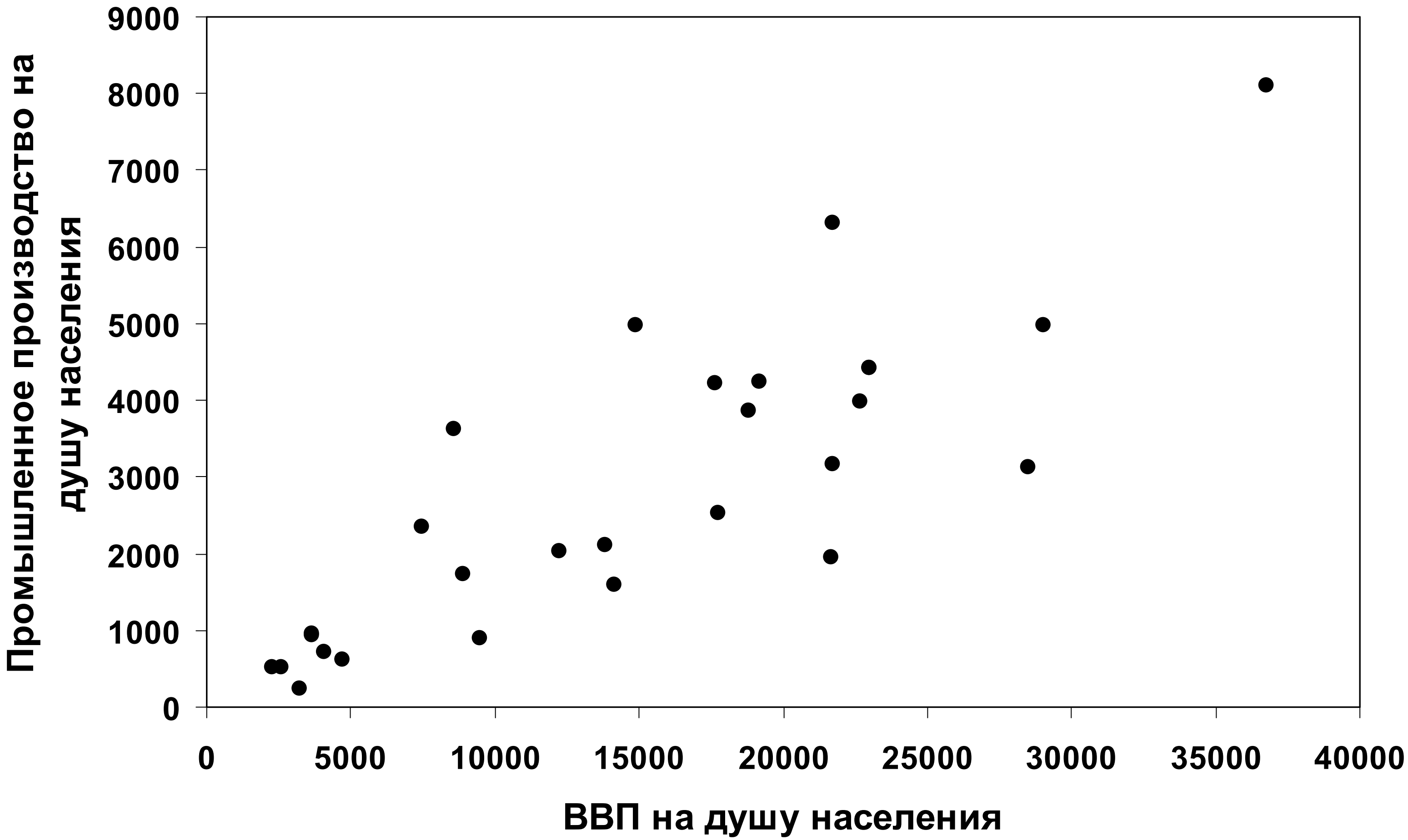
$$Y' = \beta_1 H + \beta_2 X' + u' \qquad Y' = \frac{Y_i}{\sigma_i}, \quad H = \frac{1}{\sigma_i}, \quad X' = \frac{X_i}{\sigma_i}, \quad u' = \frac{u_i}{\sigma_i}$$

На практике вместо σ_i часто используют их оценки. Например, если после проведения теста Глейзера гипотеза о гомоскедастичности была отвергнута, поскольку в регрессии

$$|e_i| = \alpha + \beta X_i + u_i, i = 1, \ldots, n$$

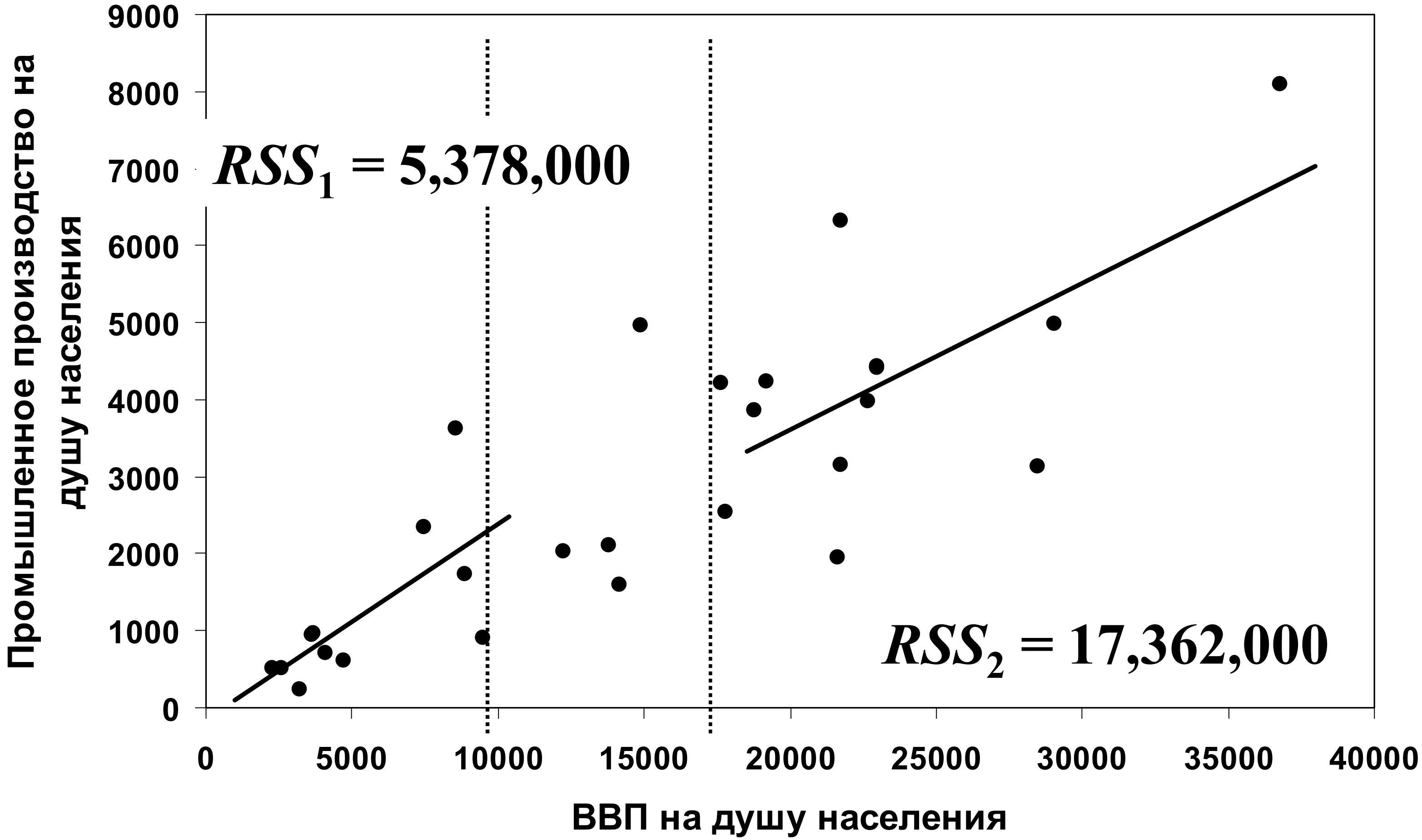
коэффициент β значим, то $\hat{\sigma}_i = |\hat{e}_i|$, i = 1,..,n

Пример



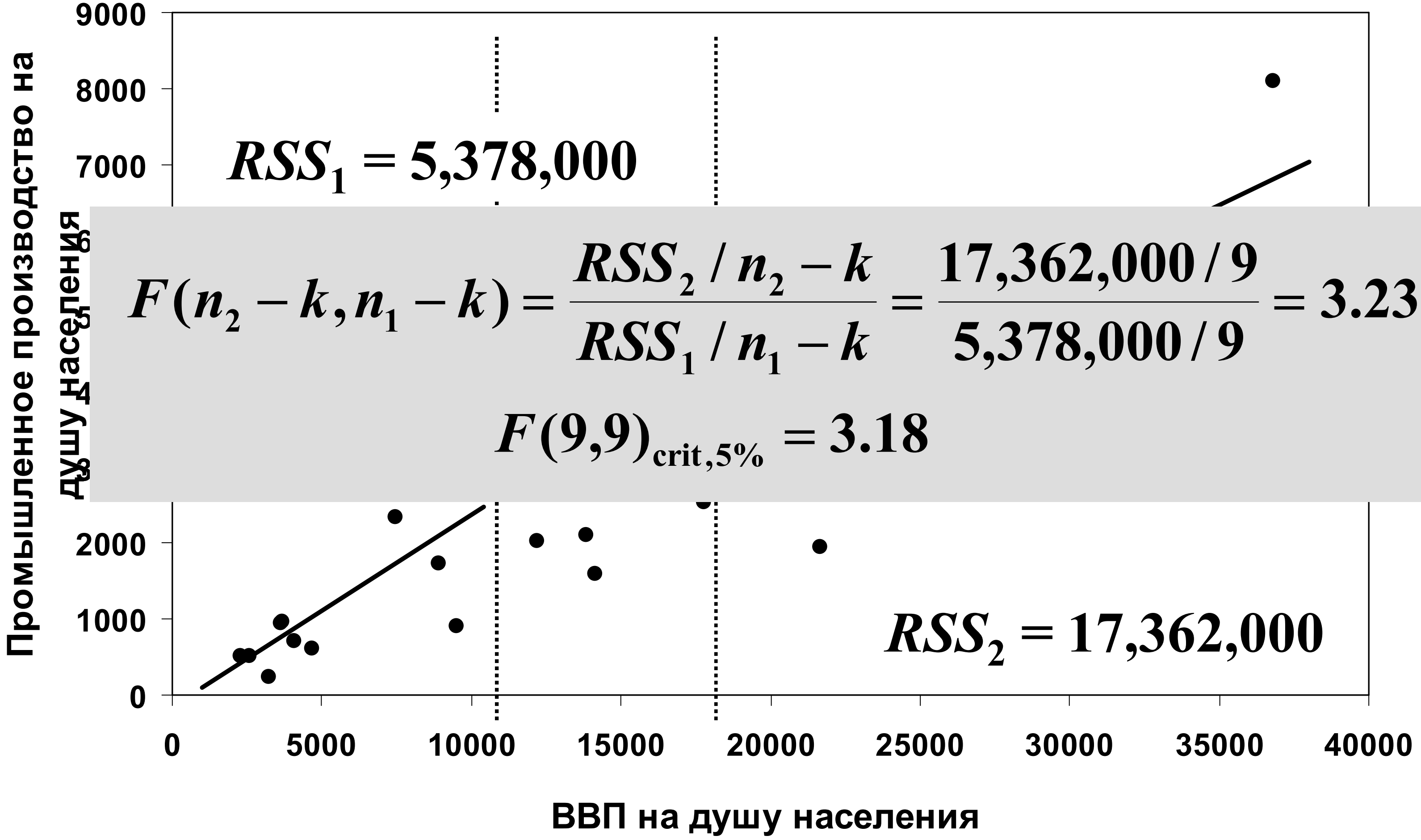
Пример зависимости производства на душу населения от ВВП на душу населения (диаграмма рассеивания).

Пример



Упорядочив страны по возрастанию ВВП на душу населения, разбивает их на три группы, средние наблюдения выкидывает, а для первой и последней группы наблюдений оцениваем регрессии и находим RSS.

Пример



$$F(n_2 - k, n_1 - k) = \frac{RSS_2 / n_2 - k}{RSS_1 / n_1 - k} = \frac{17{,}362{,}000 / 9}{5{,}378{,}000 / 9} = 3.23$$

$$F(9,9)_{\mathrm{crit},5\%} = 3.18$$

Проводим тест Голфелда - Квандта. Поскольку тестовая статистика больше критической при 5% уровне значимости, нулевая гипотеза о гомоскедастичности отвергается.

Пример

$$\boldsymbol{Y} = \beta_1 + \beta_2 \boldsymbol{X} + \boldsymbol{u}$$

$$\text{дисперсия} \quad u_i = \sigma_i^2$$

$$\sigma_i = \lambda \boldsymbol{X}_i$$

Альтернативная гипотеза в тесте Голфелда – Квандта предполагает пропорциональность стандартного отклонения возмущений объясняющей переменной (в данном примере X = GDP) .

Пример

$$\boldsymbol{Y} = \beta_1 + \beta_2 \boldsymbol{X} + \boldsymbol{u}$$

$$\textit{дисперсия} \quad = \sigma_i^2$$

$$\sigma_i = \lambda \boldsymbol{X}_i$$

$$\frac{\boldsymbol{Y}_i}{\boldsymbol{X}_i} = \beta_1 \frac{\mathbf{1}}{\boldsymbol{X}_i} + \beta_2 + \frac{\boldsymbol{u}_i}{\boldsymbol{X}_i}$$

Напомним, что для получения эффективных оценок требуется преобразовать переменные, разделив их на ту переменную, которой пропорционально стандартное отклонение возмущений.

Пример

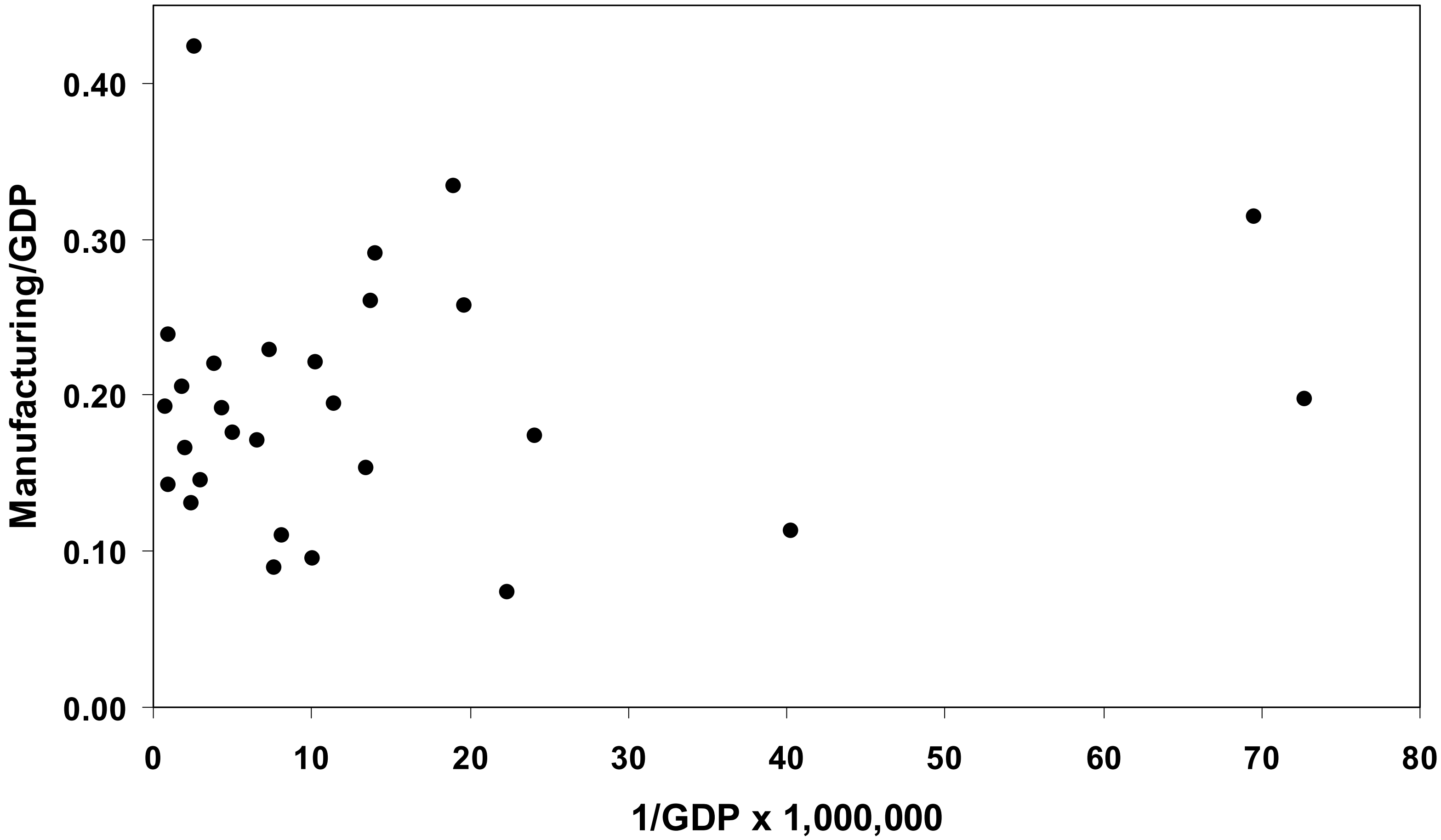
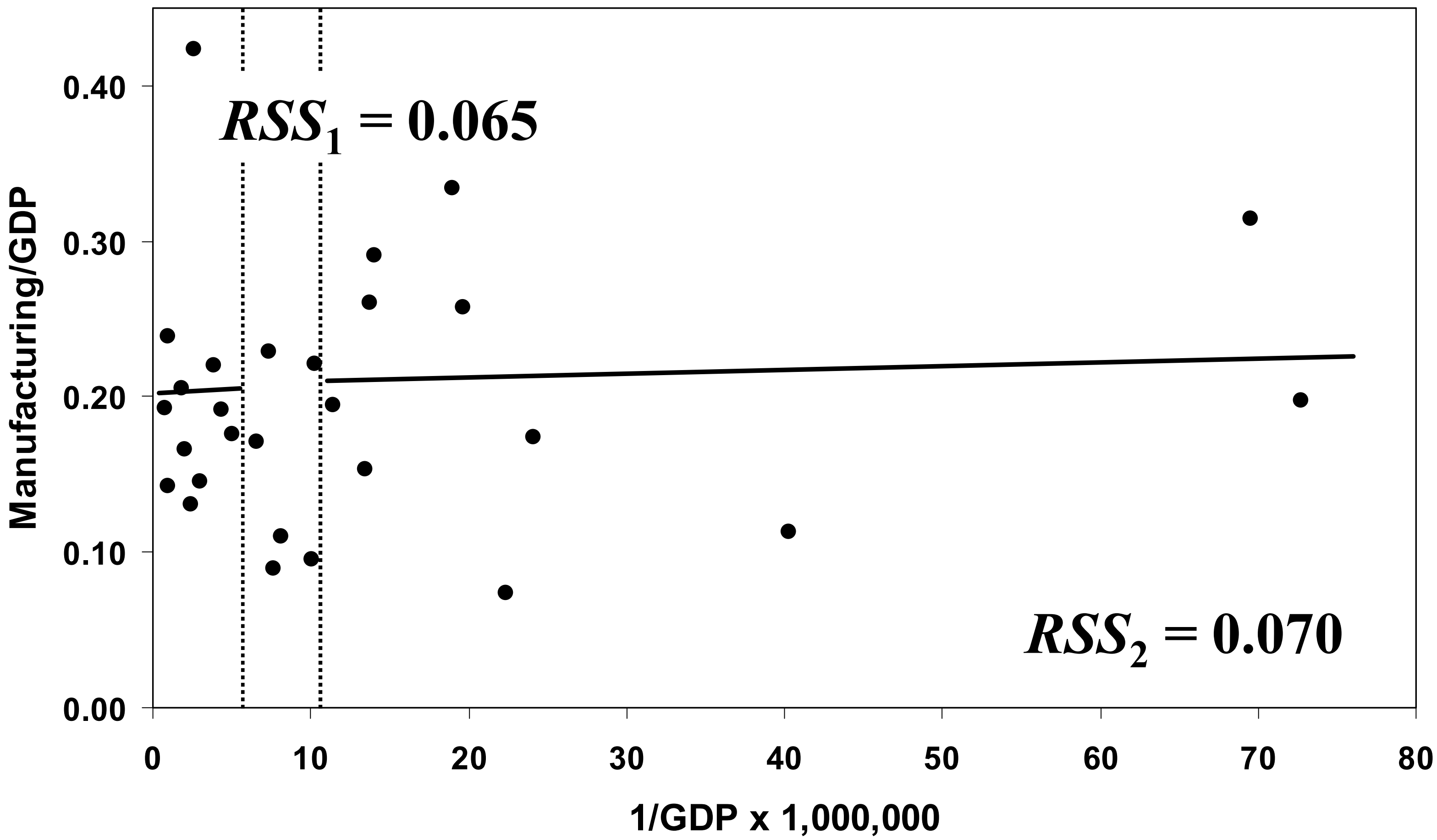


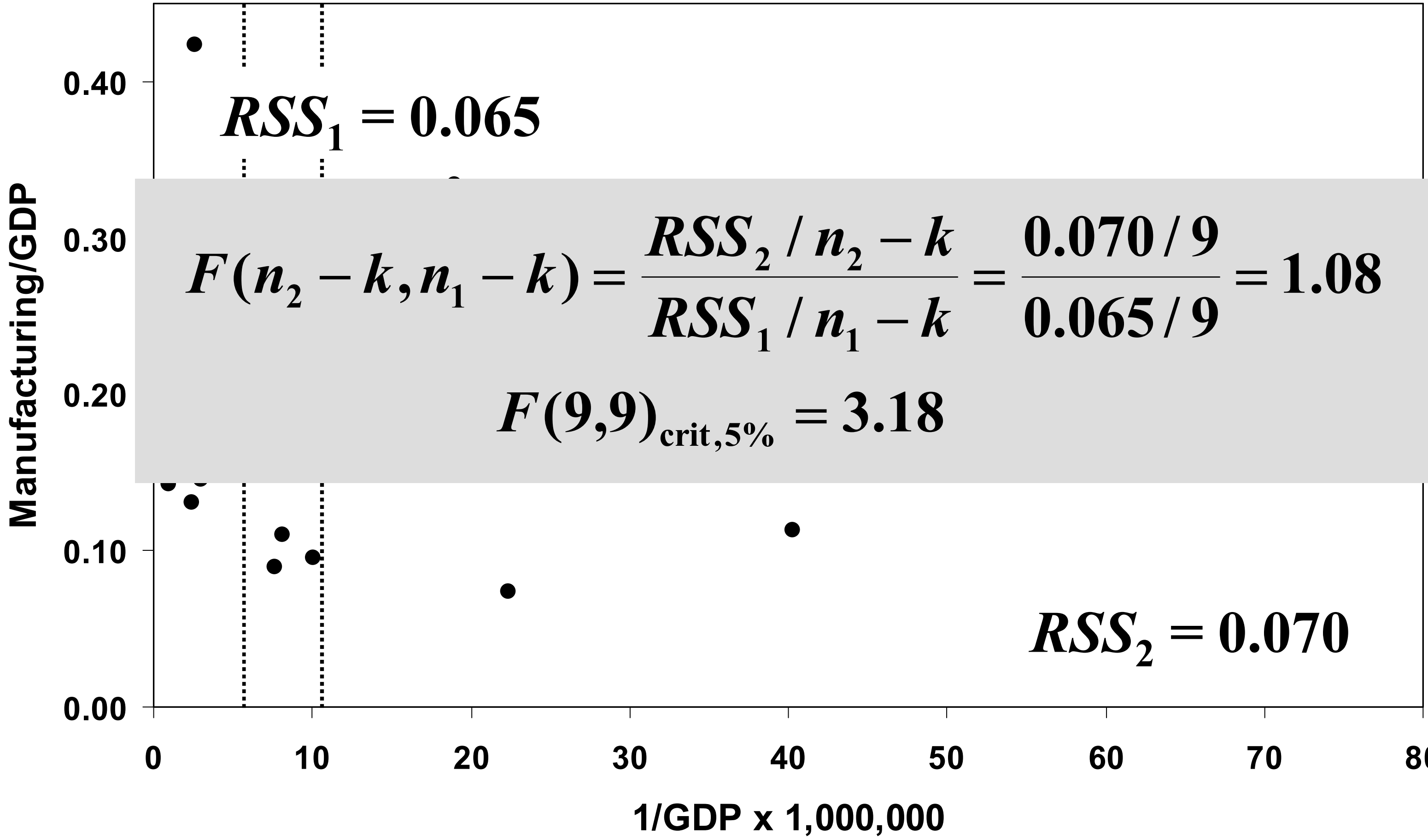
Диаграмма рассеяния в преобразованных переменных.

Пример



Снова проводим тест Голдфелда - Квандта.

Пример

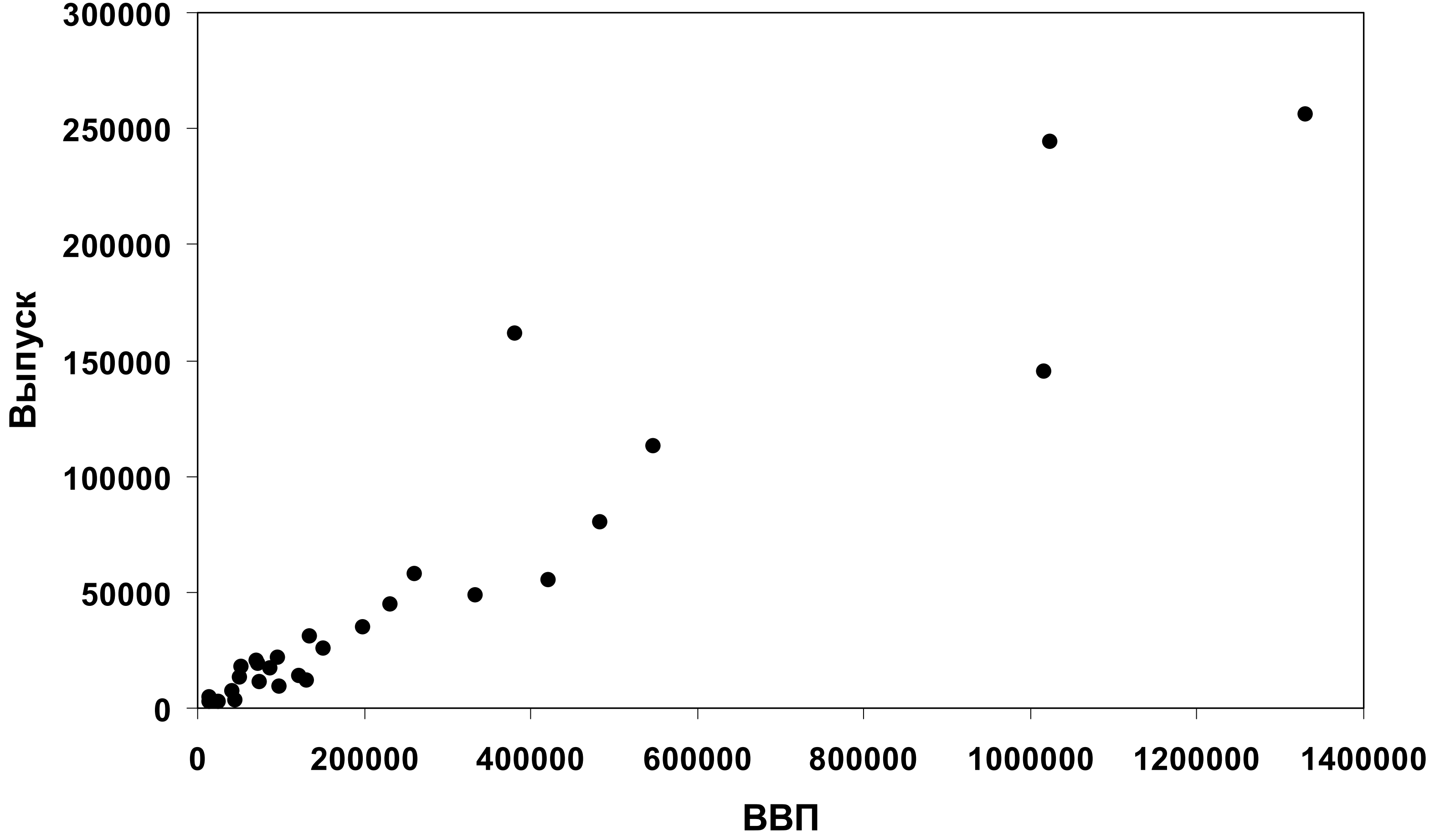


$$F(n_2 - k, n_1 - k) = \frac{RSS_2 / n_2 - k}{RSS_1 / n_1 - k} = \frac{0.070 / 9}{0.065 / 9} = 1.08$$

$$F(9,9)_{\text{crit},5\%} = 3.18$$

На этот раз гипотеза о гомоскедастичности не отвергается. С помощью преобразования гетероскедастичность была устранена.

Второй способ борьбы с гетероскедастичностью



Существует другой способ борьбы с гетероскедастичностью, связанный с выбором другой функциональной формы модели, а именно, линейной в логарифмах.

Логарифмическое преобразование данных

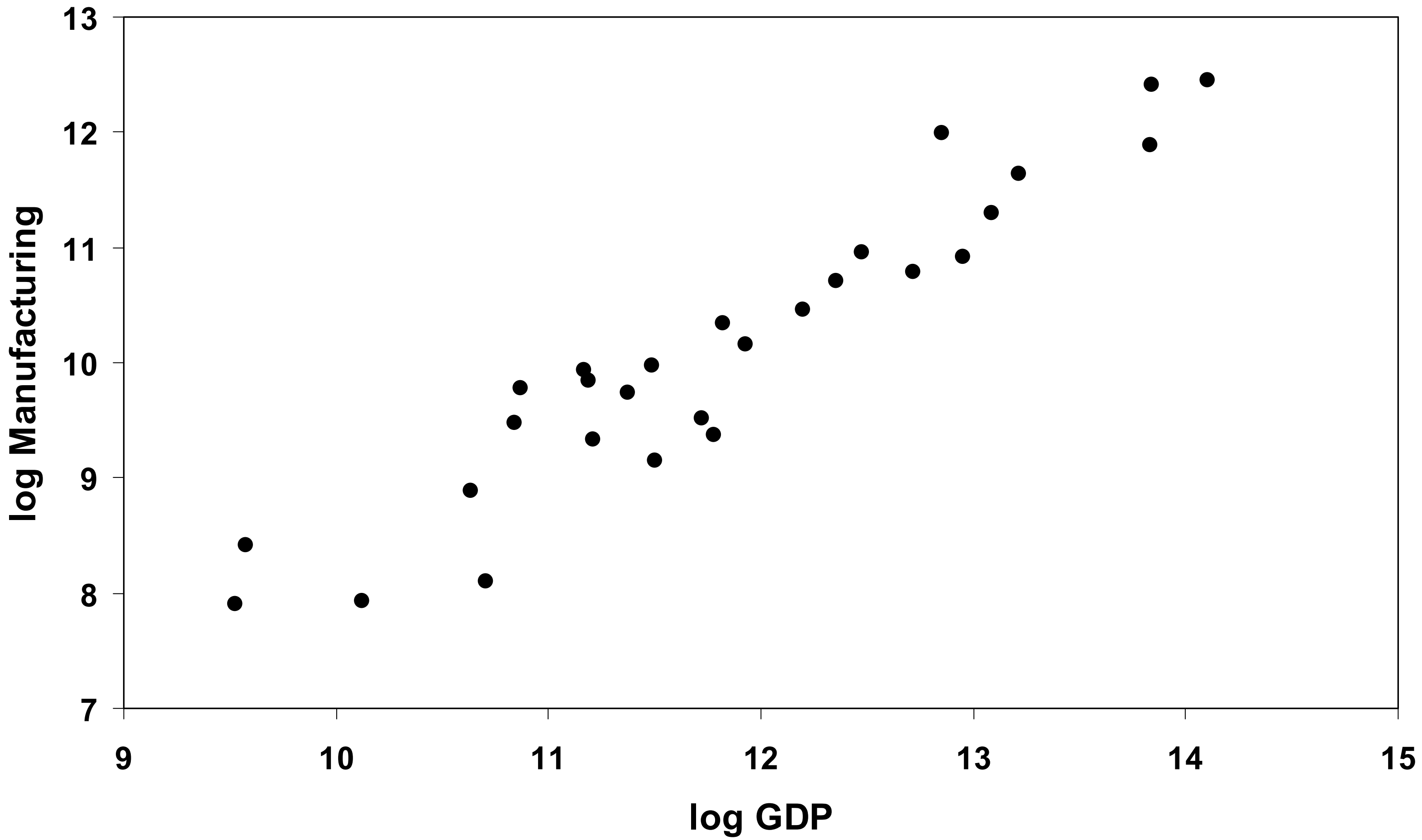
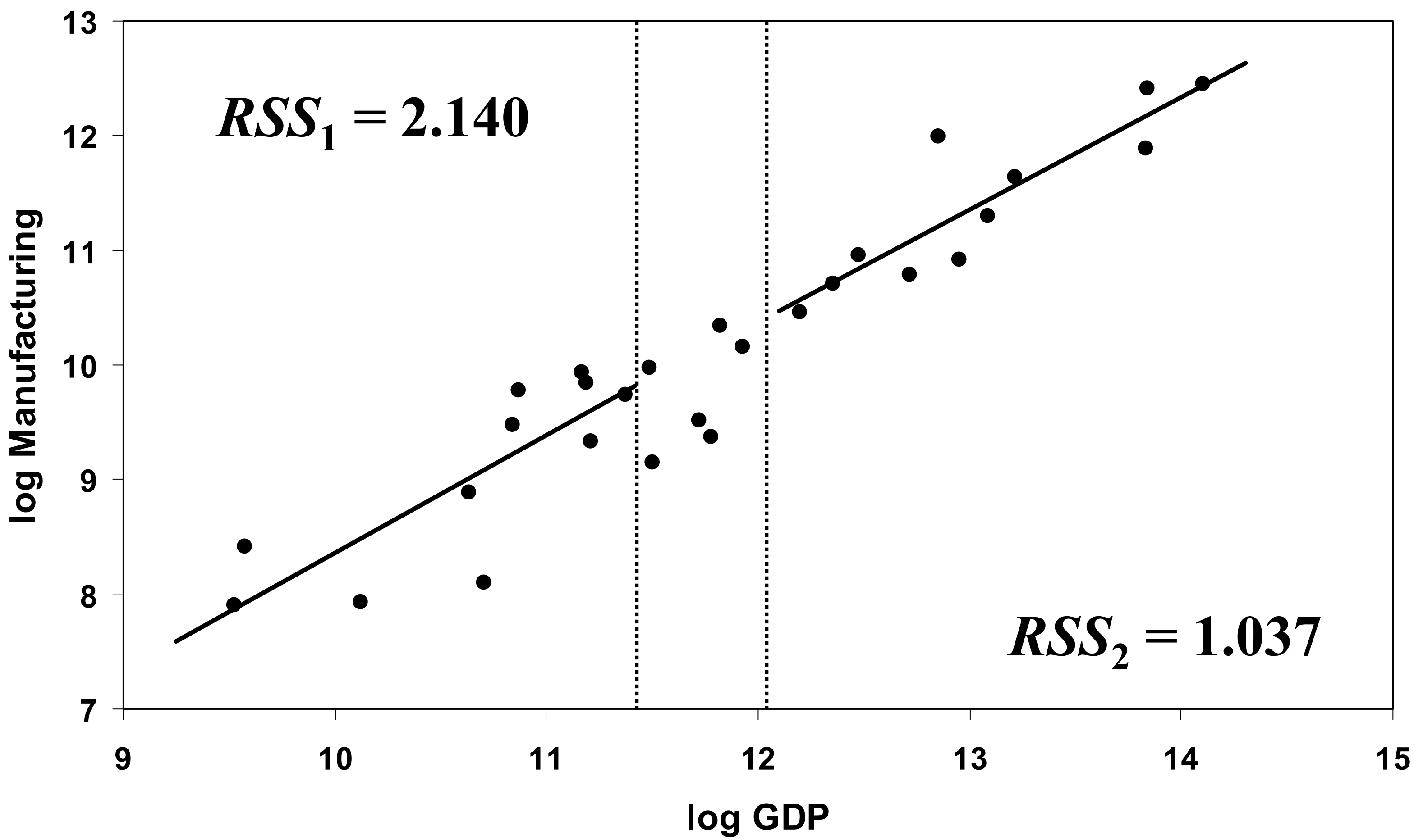


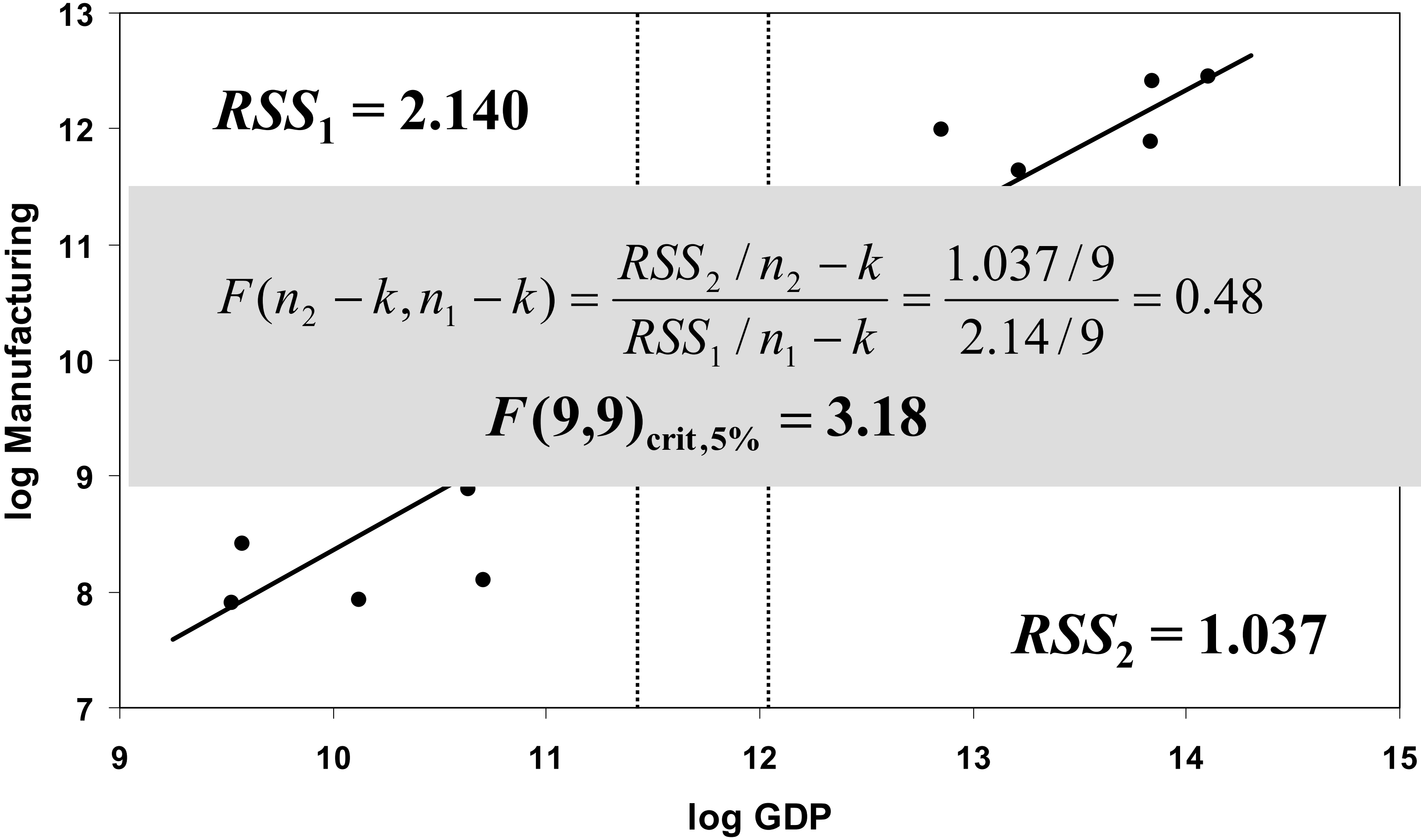
Диаграмма рассеяния для переменных в логарифмическом масштабе.

Линейная в логарифмах модель



Проведем снова тест Голдфелда – Квандта для линейной в логарифмах модели.

HETEROSCEDASTICITY: WEIGHTED AND LOGARITHMIC REGRESSIONS

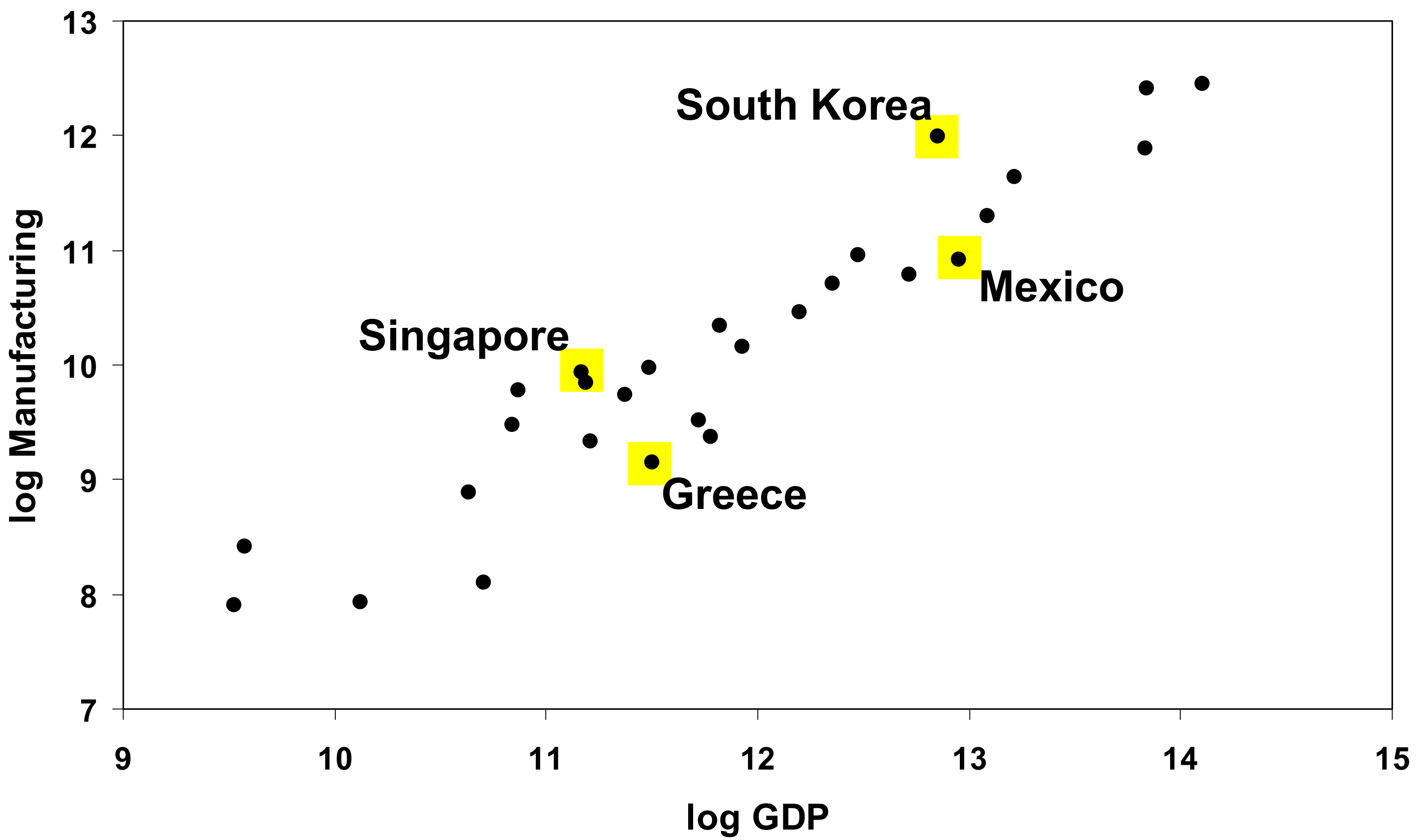


$$F(n_2 - k, n_1 - k) = \frac{RSS_2 / n_2 - k}{RSS_1 / n_1 - k} = \frac{1.037/9}{2.14/9} = 0.48$$

$$F(9,9)_{\text{crit},5\%} = 3.18$$

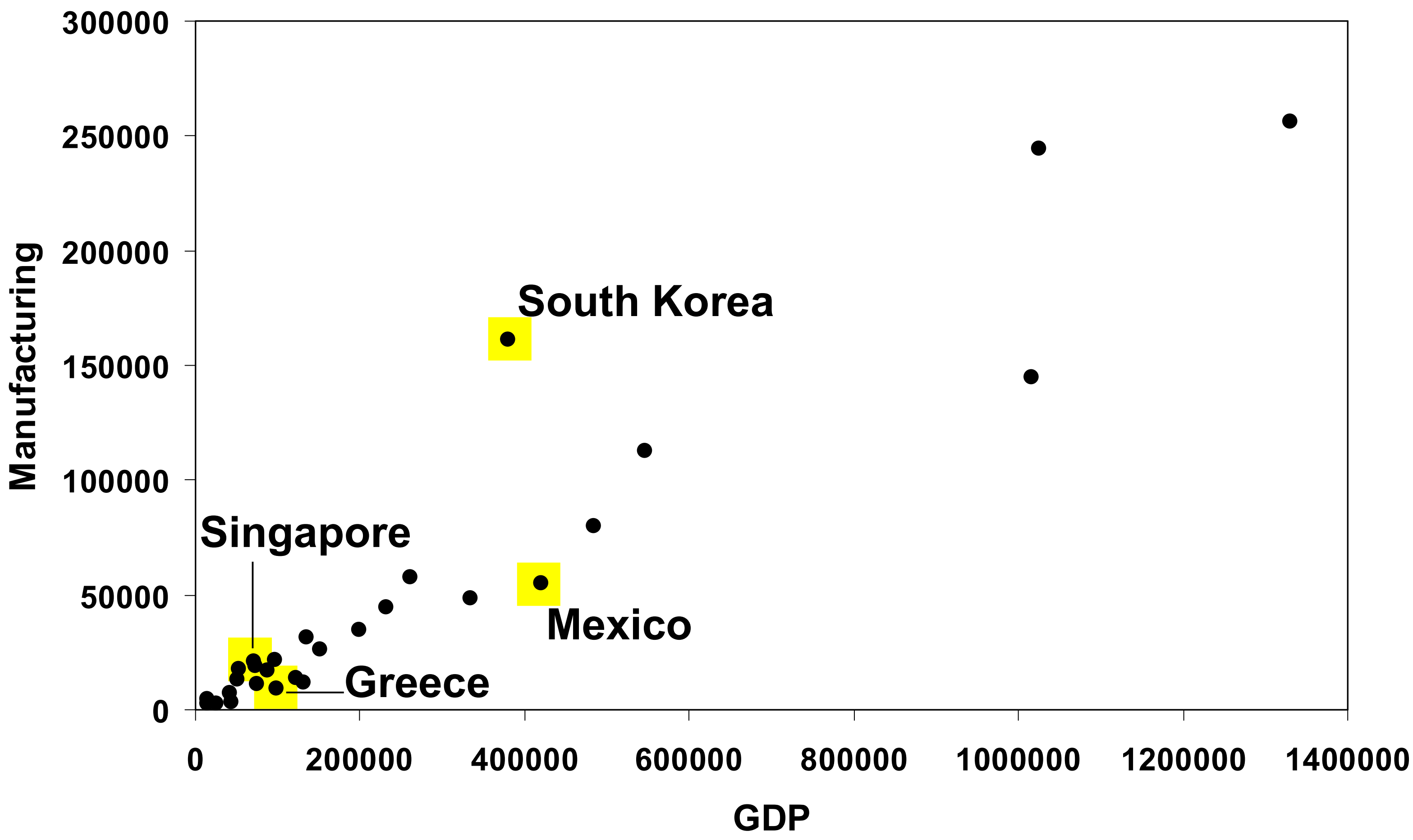
Нулевая гипотеза о гомоскедастичности не отвергается.

Логарифмический масштаб



В логарифмическом масштабе разнице между Южной Кореей и Мексикой не так сильно отличается от разницы для Сингапура и Греции, как при линейном масштабе.

Линейный масштаб



Различные спецификации

$$\underset{(5700)}{\hat{MANU}} = 604 + \underset{(0.013)}{0.194}GDP \qquad R^2 = 0.89$$

$$\frac{\hat{MANU}}{GDP} = \underset{(0.019)}{0.189} + \underset{(841)}{533}\frac{1}{GDP} \qquad R^2 = 0.02$$

$$\log \hat{MANU} = \underset{(0.785)}{-1.694} + \underset{(0.066)}{0.999}\log GDP \qquad R^2 = 0.90$$

По одним и тем же данным оценено несколько моделей.

Состоятельные при гетероскедастичности стандартные ошибки Уайта

$$Y = X\beta + \varepsilon$$

$$\hat{\beta} = \beta + (X'X)^{-1} X'\varepsilon$$

$$\mathrm{Var}[\hat{\beta}] = (X'X)^{-1} X'\mathrm{var}[\varepsilon]X\,(X'X)^{-1}$$

$$\widehat{Var}\,[\hat{\beta}] = (X'X)^{-1} \left[X'\widehat{var[\varepsilon]}X\right] (X'X)^{-1}$$

Оценка Уайта: $\mathrm{n}\widehat{Var}\,[\hat{\beta}] =$

$$= \left(\frac{1}{n}X'X\right)^{-1} \cdot \left(\frac{1}{n}\sum_{s=1}^{n} e_s^2\,(x_s' x_s)\right) \cdot \left(\frac{1}{n}X'X\right)^{-1}$$

x_s - s-ая строка матрицы X.

Пример

```
. reg manu gdp
      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  1,    26) =  210.73
       Model |  1.1600e+11     1  1.1600e+11           Prob > F      =  0.0000
    Residual |  1.4312e+10    26   550462775           R-squared     =  0.8902
-------------+------------------------------           Adj R-squared =  0.8859
       Total |  1.3031e+11    27  4.8264e+09           Root MSE      =   23462
------------------------------------------------------------------------------
        manu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |    .193693   .0133428    14.52   0.000     .1662665    .2211195
       _cons |   603.9453   5699.677     0.11   0.916    -11111.91     12319.8
```

```
. reg manu gdp, robust
Regression with robust standard errors                 Number of obs =      28
                                                       F(  1,    26) =  116.39
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8902
                                                       Root MSE      =   23462
------------------------------------------------------------------------------
             |               Robust
        manu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |    .193693   .0179542    10.79   0.000     .1567877    .2305983
       _cons |   603.9453   3542.388     0.17   0.866    -6677.538    7885.429
```

$\sigma^2_{b_2^{OLS}}$

Пример

```
. reg manu gdp
      Source |       SS       df       MS              Number of obs =      28
-------------+------------------------------           F(  1,    26) =  210.73
       Model |  1.1600e+11     1  1.1600e+11           Prob > F      =  0.0000
    Residual |  1.4312e+10    26   550462775           R-squared     =  0.8902
-------------+------------------------------           Adj R-squared =  0.8859
       Total |  1.3031e+11    27  4.8264e+09           Root MSE      =   23462
------------------------------------------------------------------------------
        manu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |    .193693   .0133428    14.52   0.000     .1662665    .2211195
       _cons |   603.9453   5699.677     0.11   0.916    -11111.91     12319.8
```

```
. reg manu gdp, robust
Regression with robust standard errors                 Number of obs =      28
                                                       F(  1,    26) =  116.39
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.8902
                                                       Root MSE      =   23462
------------------------------------------------------------------------------
             |               Robust
        manu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gdp |    .193693   .0179542    10.79   0.000     .1567877    .2305983
       _cons |   603.9453   3542.388     0.17   0.866    -6677.538    7885.429
```

Пример1. История, произошедшая на Нью-Йорской фондовой бирже

Securities and Exchange Commission vs Antitrust division of the US Department of Justice

Биржевой комитет: Комиссионные брокерам не являются объектом соглашения между брокерами и клиентами, а устанавливаются биржевым комитетом

Подразделение министерства юстиции: Цены комиссионных д.б. либерализованы

История, произошедшая на Нью-йоркской фондовой бирже

Биржевой комитет:

$$\underset{t=\textit{отношение}}{\hat{Y}} = \underset{(2.98)}{476000} + \underset{(40.39)}{31.348}X - \underset{(-6.54)}{1.083 \times 10^{-6}} X^2$$

где Y – доход брокерских компаний, X – количество акций в сделке.

Вывод: естественная монополия, не надо либерализовывать цены.

История, произошедшая на Нью-Йорской фондовой бирже

Подразделение министерства юстиции :

Дисперсия ошибок зависит от объема сделки. Надо поделить все переменные на $\sqrt{X}$. Новое оцененное уравнение:

$$\underset{t=отношение}{\hat{Y}} = \underset{(32.3)}{342000} + \underset{(7.07)}{25.77}X + \underset{(0.503)}{4.34\times10^{-6}}X^2$$

Вывод: это не естественная монополия, надо либерализовать цены.

Пример 2. Оценка функции спроса на макаронные изделия в России

База данных: РМЭЗ - Российский мониторинг экономического положения и здоровья населения

Переменные:

buymacar_c – стоимость макаронных изделий, купленных семьей за последние 7 дней (в руб.),

pr_macar - цена 1 кг макаронных изделий,

inc – доход семьи за месяц

Оценка функции спроса на макаронные изделия в России

reg buymacar_c pr_macar inc

Source	SS	df	MS		
				Number of obs	= 1756
				F(2, 1753)	= 999.28
Model	4914486.1	2	2457243.05	Prob > F	= 0.0000
Residual	4310662.73	1753	2459.02038	R-squared	= 0.5327
				Adj R-squared	= 0.5322
Total	9225148.83	1755	5256.49506	Root MSE	= 49.589

buymacar_c	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
pr_macar	.5073996	.01136	44.67	0.000	.4851191	.5296802
inc	.0000559	.0000642	0.87	0.384	-.00007	.0001818
_cons	19.85896	1.443465	13.76	0.000	17.02786	22.69005

Тест Уайта

estat imtest, white

White's test for Ho: homoskedasticity

against Ha:unrestricted heteroskedasticity

chi2(5) =   624.97

Тест Голдфелда –Квандта (упорядочивание по переменной inc)

sort inc

. reg buymacar_c pr_macar inc in 1/585

Source	SS	df	MS	
				Number of obs = 295 F(2, 292) = 465.68
Model	839154.929	2	419577.464	Prob > F = 0.0000
Residual	263093.755	292	901.006012	R-squared = 0.7613
				Adj R-squared = 0.7597
Total	1102248.68	294	3749.14518	Root MSE = 30.017

buymacar_c	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
pr_macar	2.716744	.0894214	30.38	0.000	2.540752	2.892736
inc	-.0044199	.0023256	-1.90	0.058	-.0089969	.0001572
_cons	-11.64927	5.555395	-2.10	0.037	-22.58296	-.715578

. reg buymacar_c pr_macar inc in 1172/1756

Source	SS	df	MS	
				Number of obs = 293
				F(2, 290) = 2564.69
Model	1861980.84	2	930990.42	Prob > F = 0.0000
Residual	105270.882	290	363.003043	R-squared = 0.9465
				Adj R-squared = 0.9461
Total	1967251.72	292	6737.16343	Root MSE = 19.053

buymacar_c	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
pr_macar	.9843328	.013775	71.46	0.000	.9572211	1.011444
inc	.0021632	.0013107	1.65	0.100	-.0004165	.0047429
_cons	-6.335153	8.961121	-0.71	0.480	-23.97223	11.30193

Тест Голдфелда –Квандта (упорядочивание по переменной inc)

$$F = \frac{RSS_2 /(n_2 - k)}{RSS_1 /(n_1 - k)} < 1 \Rightarrow F < F^{cr}$$

Вывод: Гипотеза H_0 не отвергается, т.е. нет гетероскедастичности (по переменной inc)

Тест Голдфелда –Квандта (упорядочивание по переменной pr_macar)

```
sort pr_macar
. reg  buymacar_c pr_macar       inc  in 1/585
Source          SS           df      MS             Number of obs  =    585
                   F(  2,   582)        =    0.30
Model      309.359477    2    154.679738              Prob > F     =  0.7404
Residual  299328.145    582   514.309528              R-squared  =  0.0010
                                                     Adj R-squared = -0.0024
Total   299637.505      584  513.077919          Root MSE  =  22.678

buymacar_c      Coef.    Std. Err.      t  P>t   [95% Conf.        Interval]
pr_macar    .3833284    .6208139    0.62      0.537 -.8359801   1.602637
inc           -.0000419    .000085     -0.49      0.622 -.000209    .0001251
_cons       21.96761      8.479713    2.59      0.010 5.313038    38.62217

. reg  buymacar_c pr_macar       inc  in 1172/1756
Source           SS           df      MS             Number of obs  =    585
                                                     F(  2,   582)=  364.75
Model       4793973.58    2    2396986.79                    Prob > F    =  0.0000
Residual   3824678.33   582  6571.61226              R-squared =  0.5562
                                                     Adj R-squared =  0.5547
Total   8618651.92      584  14757.9656          Root MSE  =  81.065
buymacar_c      Coef.        Std. Err.     t          P>t       [95% Conf.        Interval]
pr_macar      .5059595    .0187335     27.01   0.000       .469166     .542753
inc              .0000173    .0001254     0.14     0.891       -.0002291   .0002636
_cons          22.21863    3.995971     5.56     0.000       14.37035    30.06691
```

Тест Голдфелда –Квандта (упорядочивание по переменной pr_macar)

$$F = \frac{RSS_2}{RSS_1}^{(*} = \frac{382467.833}{299328.145} = 12.7775 > F^{cr}_{0.05}$$

Вывод: Гипотеза H_0 отвергается, есть гетероскедастичность (по переменной pr_macar)

*** - т.к. число наблюдений в выборках совпадает**

Преобразование данных и оценивание новой регрессии

```
gen buymacar_cnew =  buymacar_c/ pr_macar
(1579 missing values generated)

. gen consnew =1/  pr_macar
(1579 missing values generated)

. gen  incnew= inc/ pr_macar
(1579 missing values generated)

. reg  buymacar_cnew consnew incnew
```

Source	SS	df	MS	
				Number of obs = 1756
				F(2, 1753) = 58.43
Model	241.386174	2	120.693087	Prob > F = 0.0000
Residual	3621.08302	1753	2.06564918	R-squared = 0.0625
				Adj R-squared = 0.0614
Total	3862.46919	1755	2.20083715	Root MSE = 1.4372

buymacar_cnew	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
consnew	20.81843	2.01338	10.34	0.000	16.86955	24.76731
incnew	.0000831	.0000388	2.14	0.032	7.05e-06	.0001592
_cons	.3681707	.1218333	3.02	0.003	.1292168	.6071246

Сравнение результатов начальной и скорректированной регрессий

buymacar_c	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
pr_macar	.5073996	.01136	44.67	0.000	.4851191	.5296802
inc	.0000559	.0000642	0.87	0.384	-.00007	.0001818
_cons	19.85896	1.443465	13.76	0.000	17.02786	22.69005

buymacar_cnew	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
consnew	20.81843	2.01338	10.34	0.000	16.86955	24.76731
incnew	.0000831	.0000388	2.14	0.032	7.05e-06	.0001592
_cons	.3681707	.1218333	3.02	0.003	.1292168	.6071246

Обобщенный метод наименьших квадратов

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Выполнены все условия ТГМ, кроме скалярности ковариационной матрицы ошибок регрессии

Var[ε**]** $=\Omega$ – известная матрица.

Оценка МНК коэффициентов $\boldsymbol{\beta}$ в этом случае будет несмещенной, но не эффективной.

Ω – положительно определенная симметричная матрица ⇨

Ω **=** $C^{-1}\Lambda\, C$, Λ **-** диагональная матрица (на диагонали – собственные числа), $C^{-1} = C'$

Обобщенный метод наименьших квадратов

$$\Omega^{-1/2} = \mathrm{C}^{-1}\ \Lambda^{-1/2}\ \mathrm{C},$$

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \ \boldsymbol{\varepsilon},$$

$$\Omega^{-1/2}\boldsymbol{Y} = \Omega^{-1/2}\boldsymbol{X\beta} + \Omega^{-1/2}\boldsymbol{\varepsilon},$$

$$\boldsymbol{Y}^{*} = \boldsymbol{X}^{*}\boldsymbol{\beta} + \ \boldsymbol{\varepsilon}^{\ *}$$

Обобщенный метод наименьших квадратов

$$\boldsymbol{Y}^{*}=\boldsymbol{X}^{*}\boldsymbol{\beta}+\ \boldsymbol{\varepsilon}^{\ *},$$

$$\widehat{\boldsymbol{\beta}}_{GLS}=(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*})^{-\mathbf{1}}\ (\boldsymbol{X}^{*\prime}\boldsymbol{Y}^{*})$$ **=…=**

$$=(\mathrm{X}'\Omega^{-\mathbf{1}}\mathrm{X})^{-\mathbf{1}}\ (\mathrm{X}'\Omega^{-\mathbf{1}}\mathrm{Y})$$ **– эффективная оценка.**

Матрицу Ω достаточно знать с точностью до пропорциональности.

NATIONAL RESEARCH
UNIVERSITY

Thank you for your attention!

20, Myasnitskaya str., Moscow, Russia, 101000
Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931
www.hse.ru