Гетероскедастичность

Гетероскедастичность

1. Что такое гетероскедастичность?
2. Чем она плоха?
3. Как понять, есть ли эта проблема в модели или нет?
4. Что можно сделать в случае гетероскедастичности?
5. Ложная гетероскедастичность

Предпосылки МНК, связанные с ошибками

(3) Математическое ожидание случайных ошибок равно нулю: $E(\varepsilon_i) = 0$

(4) случайные ошибки имеют постоянную дисперсию: $V(\varepsilon_i) = \sigma^2 = const$

Нарушение этой предпосылки и называется гетероскедастичностью

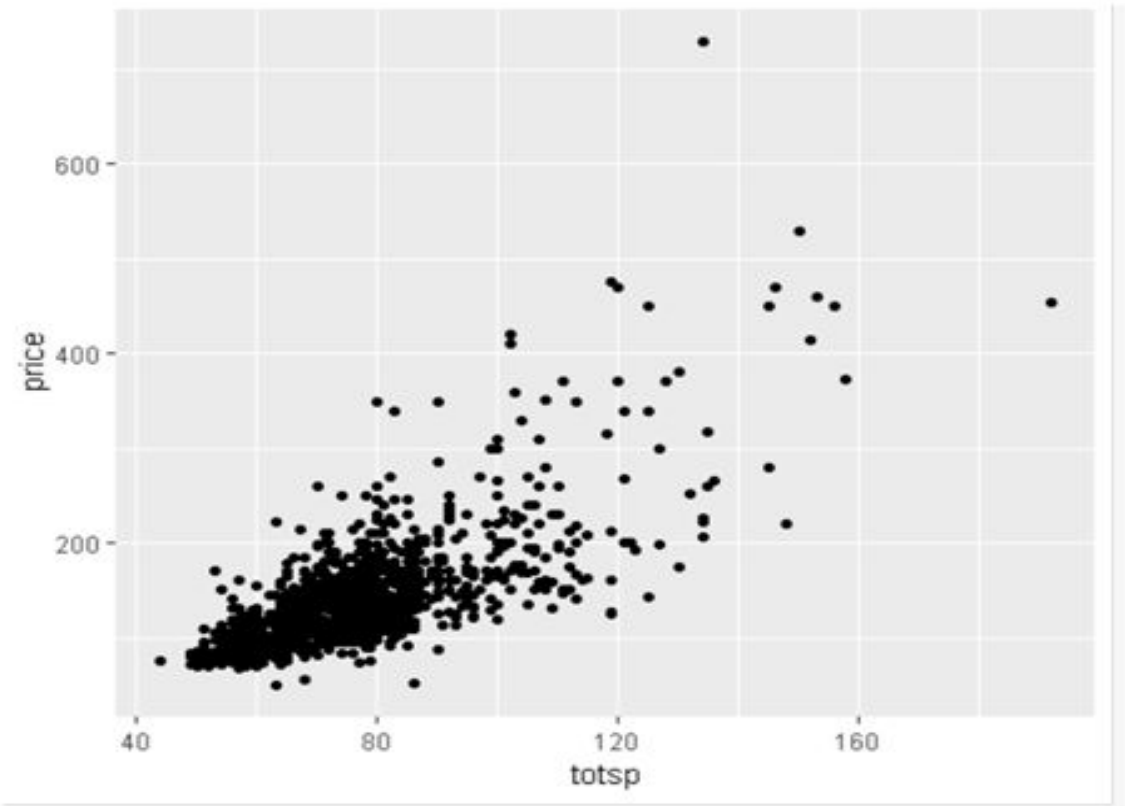
(5) Случайные ошибки, соответствующие разным наблюдениям, не зависят друг от друга (не коррелированны) $Cov(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$

Гетероскедастичность

Гомоскедастичность (условие №4 теоремы Гаусса — Маркова) — случайные ошибки имеют постоянную дисперсию: $V(\varepsilon_i) = \sigma^2 = const$

Гетероскедастичность — случайные ошибки имеют **не**постоянную дисперсию: $V(\varepsilon_i) = \sigma_i^2 \neq const$

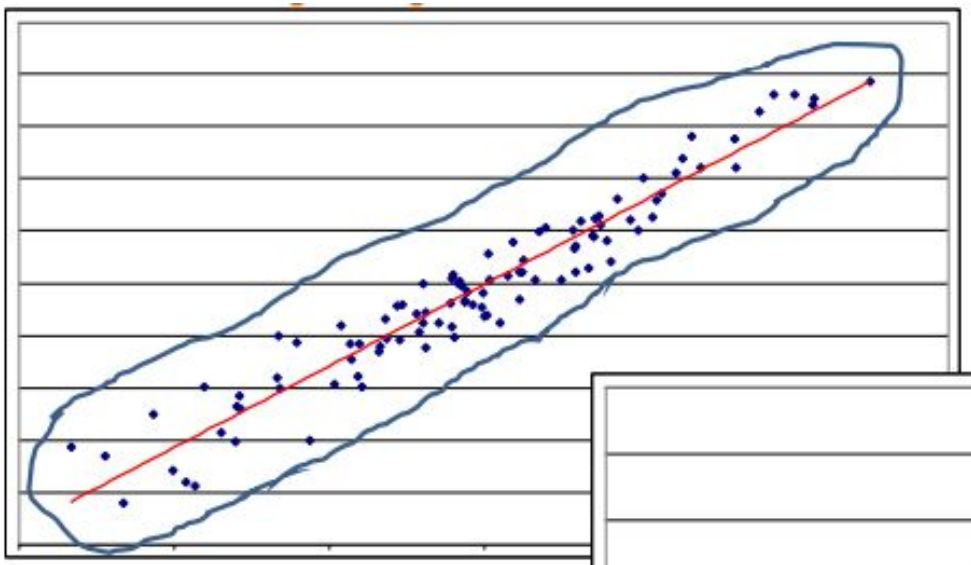
Пример гетероскедастичности



Price – цена квартиры, тыс.$; **totsp** – общая площадь квартиры, м.кв.

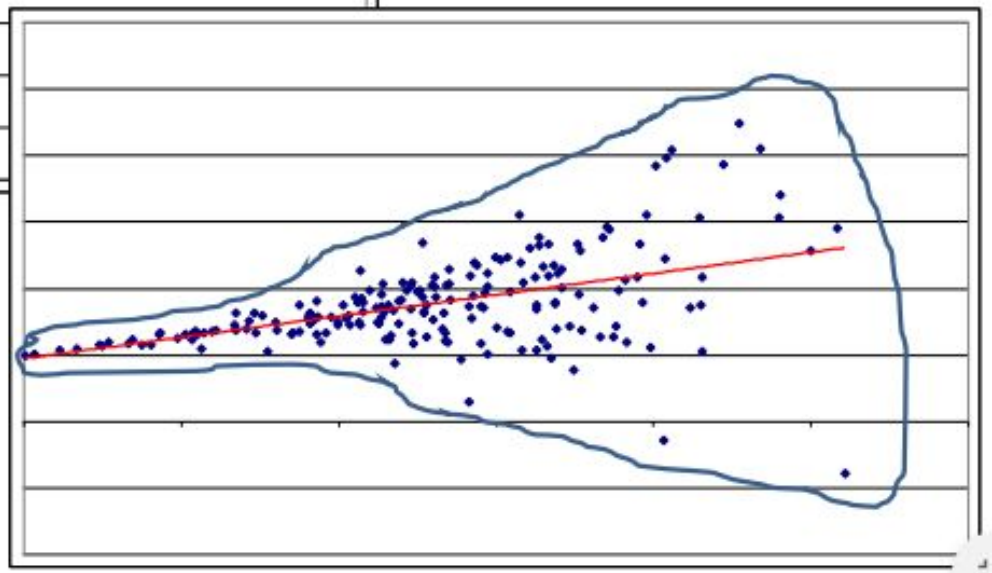
Смотрим на разброс **price** по вертикали. При маленьких значениях **totsp** разброс **price** небольшой, при увеличении **totsp** разброс **price** увеличивается.

Пример гетероскедастичности



Нет гетероскедастичности

<=



Есть гетероскедастичность

=>

Когда появляется гетероскедастичность

Если в выборку включены разнородные объекты (большие и малые предприятия, домохозяйства с разным уровнем дохода, регионы с различной численностью населения и т.п.)

Последствия гетероскедастичности

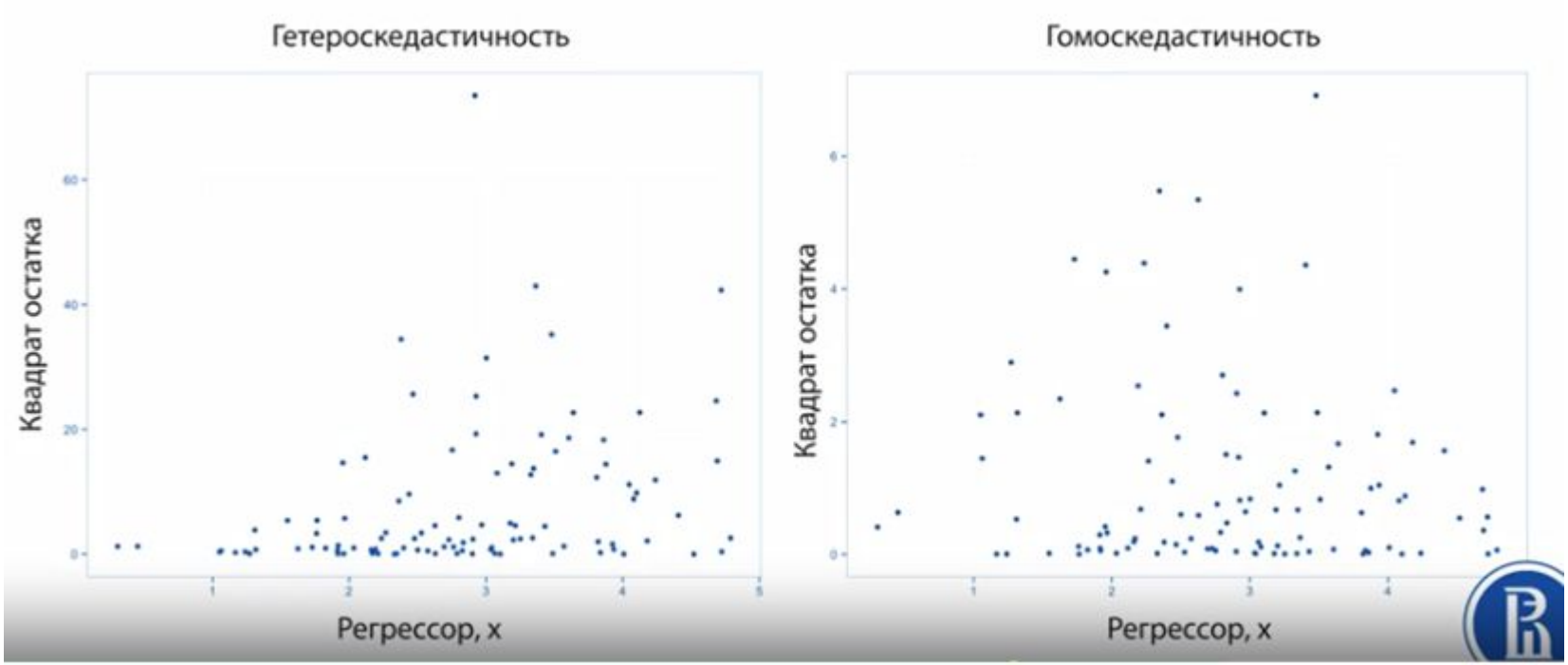
1) + МНК-оценки коэффициентов $\hat{\beta}$ остаются несмещенными, т.е. мы можем их спокойно интерпретировать и использовать;
2) – дисперсии ошибок разные и уже не имеют общей формулы для оценки $\hat{\sigma}_{\varepsilon}^{2} = \frac{RSS}{n-k-1}$, а значит формулы для стандартных отклонений коэффициентов неприменимы $\Rightarrow$ мы не можем тестировать гипотезы с помощью t - статистики и F – статистики, не можем строить доверительные интервалы.

Обнаружение гетероскедастичности

- оцениваем интересующую нас модель с помощью МНК;

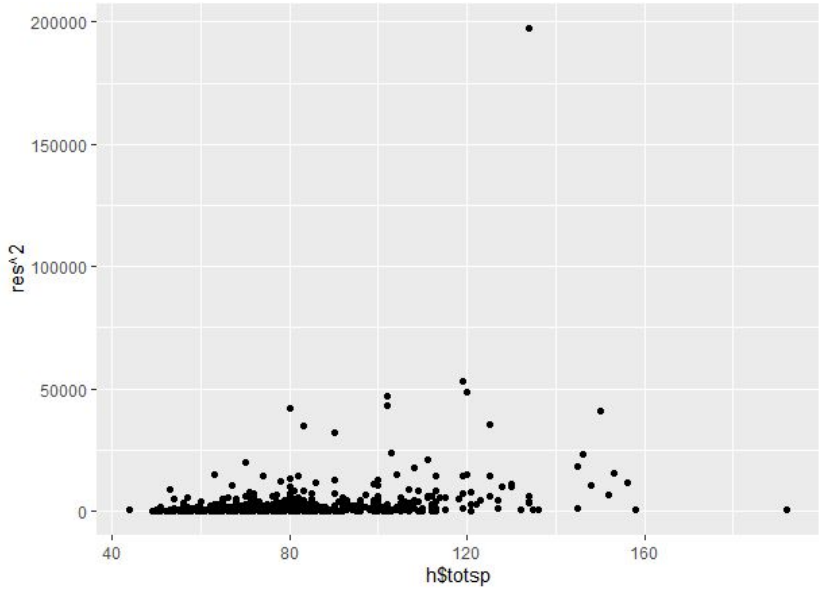
- строим график зависимой переменной (y) от регрессора (x_j);

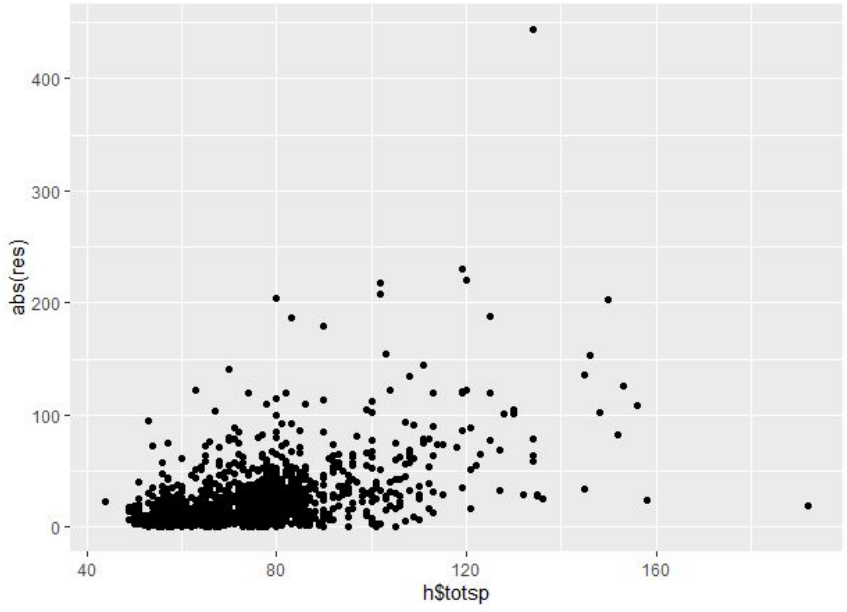
- строим график квадратов (или модулей) остатков в зависимости от регрессора (x_j).



Как построить график квадратов или модулей остатков (Excel, R-studio)

```
h= read.csv("flats_moscow.txt", sep="\t", header=TRUE, dec=".")#выгружаем файл
model=lm(data=h,price~totsp)#строим регрессионную модель
qplot(data=h,totsp,price)#строим поле корреляции
res=residuals(model)#достаем остатки
res
qplot(h$totsp,res^2)#график зависимости квадратов остатков от регрессора
qplot(h$totsp,abs(res))#график зависимости модулей остатков от регрессора
```





Тесты на наличие гетероскедастичности

- - тест Бройша-Пагана (BP test)
- - тест Уайта
- - тест Гольдфелда-Квандта
- -тест Глейзера

Тест Бройша-Пагана

H_0 : $Var(\varepsilon | x_1, x_2, \ldots x_k) = \sigma^2$ (имеет место гомоскедастичность)

Вспомним, что M(ε)=0, поэтому можем сформулировать ее по-другому:

H_0 : $M(\varepsilon^2 | x_1, x_2, \ldots x_k) = M(\varepsilon^2) = \sigma^2$

Для тестирования этой гипотезы нужно проверить, зависят ли квадраты ошибок от исходных регрессоров. Для этого составим регрессионное уравнение:

$$\varepsilon^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + error$$

Тогда:

H_0: $\delta_1 = \delta_2 = \ldots\ldots = \delta_k = 0$ (гомоскедастичность)

H_1: хотя бы один из коэффициентов $\delta_i \neq 0$ (гетероскедастичность)

Тест Бройша-Пагана

Оценим это регрессионное уравнение:

$$\hat{\varepsilon}^2 = \hat{\delta}_0 + \hat{\delta}_1 x_1 + \hat{\delta}_2 x_2 + \ldots + \hat{\delta}_k x_k + \hat{e}$$

Тестовая статистика:

$$F_{расч} = \frac{R^2_{\hat{\varepsilon}^2} / k}{(1 - R^2_{\hat{\varepsilon}^2}) / (n - k - 1)}$$

n - число наблюдений; k - число факторов

Или:

$LM = n \cdot R^2_{\hat{\varepsilon}^2}$ **- распределение** χ^2 с k степенями свободы

Пример (Excel, R-studio)

```
model = lm(data=Rost,Ves~Rost+Dohod)
bptest(model)
```

```
> model = lm(data=Rost,Ves~Rost+Dohod)
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 4.4539, df = 2, p-value = 0.1079
```

- H0 принимается, гетероскедастичности нет

Тест Уайта

Проверяется, зависят ли квадраты ошибок от исходных регрессоров, их квадратов и попарных произведений. Для этого составляется регрессионное уравнение (например, для k = 3):

$$\varepsilon^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 +$$

$$+\delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error$$

Тогда:

H_0: $\delta_1 = \delta_2 = ... = \delta_k = 0$ **(гомоскедастичность)**

H_1: хотя бы один из коэффициентов $\delta_i \neq 0$
(гетероскедастичность)

Тест Уайта

Тестовая статистика:

$$F_{расч} = \frac{R^2_{\hat{\varepsilon}^2} / k}{(1 - R^2_{\hat{\varepsilon}^2}) / (n - k - 1)}$$

n - число наблюдений; k - число факторов

Или:

$LM = n \cdot R^2_{\hat{\varepsilon}^2}$ - **распределение** χ^2 с k степенями свободы

Пример (Excel, R-studio)

```
model = lm(data=Rost,Ves~Rost+Dohod)
bptest(model, data = Rost, varformula = ~poly(Rost,Dohod,degree=2))
```

```
> bptest(model, data = Rost, varformula = ~poly(Rost,Dohod,degree=2))

        studentized Breusch-Pagan test

data:  model
BP = 7.6522, df = 5, p-value = 0.1765
```

- H0 принимается, гетероскедастичности нет

Пример (Эконометрика. Демидова, Малахов)

Задача 10.5. По данным для 20 стран были оценены коэффициенты уравнения регрессии (под оценками коэффициентов даны их стандартные ошибки):

$$\hat{Y}_i = \underset{t=4,79}{111,78} - \underset{t=-2,53}{0,0042}\,X_{2i} - \underset{t=-1,71}{0,4898}\,X_{3i},\ R^2 = 0,492,$$

где Y — младенческая смертность (количество умерших в расчете на тысячу рожденных живыми); X_2 — валовой национальный продукт в расчете на душу населения; X_3 — процент имеющих начальное образование в определенной возрастной группе.

Для проведении теста Уайта была оценена регрессия

$$e_i^2 = \underset{t=4,86}{4987} - \underset{t=-0,59}{0,4718}\,X_{2i} - \underset{t=-2,45}{0,8442}\,X_{3i} + \underset{t=1,27}{0,00001}\,X_{2i}^2 + \underset{t=1,62}{0,4435}\,X_{3i}^2 + \underset{t=0,35}{0,0026}\,X_{2i}X_{3i},\ R^2 = 0,649.$$

Проверим гипотезу об отсутствии гетероскедастичности с помощью теста Уайта.

H0: гетероскедастичности нет

H1: гетероскедастичность есть

LMрасч = 20*0,649 = 12,98

LM табл (5, 0,05) = 11,071

LMрасч > LMтабл, значит H0 отвергается, гетероскедастичность есть

Тест Голдфелда-Квандта

Если есть подозрение, что дисперсии ошибок пропорциональны некоторой переменной X_j. Тест подходит для малых выборок.

H_0 : $Var(\varepsilon | x_1, x_2, \ldots x_k) = \sigma^2$ (имеет место гомоскедастичность)

H_1: $\sigma_i \sim X_{ij}$ для некоторого X_j, i = 1…n (стандартные отклонения ошибок пропорциональны некоторой переменной).

Процедура теста Голдфелда-Квандта

1. Оцениваем коэффициенты основной регрессии

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

2. Сохраняем остатки регрессии. После анализа графиков остатков может появиться предположение о том, с ростом какой переменной увеличивается дисперсия ошибок.
3. Упорядочиваем все наблюдения по увеличению выбранной в п.2 переменной.
4. Делим все наблюдения на 3 группы (в средней группе обычно оставляют 20% наблюдений).
5. Убираем все наблюдения второй группы.
6. Оцениваем исходную модель отдельно по наблюдениям первой и третьей групп.

Процедура теста Голдфелда-Квандта

7. Проверка гипотезы H_0 сводится к проверке гипотезы о равенстве дисперсий первых n_1 и последних n_2 наблюдений с помощью F-статистики:

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

где RSS_1 и RSS_2 – суммы квадратов остатков в регрессиях, оцененных по первым n_1 и последним n_2 наблюдениям.

8. Если значение тестовой статистики F превышает критическое $F_{\alpha,(n2-k,n1-k)}$ (при выбранном уровне значимости α), то гипотеза H_0 отвергается.

Пример (Excel, R-studio)

```
model=lm(data=f,rashod~dohod)
gqtest(model, order.by =~dohod, data = f, fraction = 0.2)
```

```
> gqtest(model, order.by =~dohod, data = f, fraction = 0.2)

        Goldfeld-Quandt test

data:  model
GQ = 24.124, df1 = 6, df2 = 6, p-value = 0.0005935
alternative hypothesis: variance increases from segment 1 to 2
```

Пример

(Эконометрика, Демидова, Малахов)

Задача 10.1[3]. Исследователь рассматривает вопрос о том, происходит ли вытеснение инвестиций государственными расходами, оценивая регрессию по данным о государственных расходах G, инвестициях I, валовом внутреннем продукте Y и численности населения P для 30 стран в 1997 г. (под оценками коэффициентов приведены их стандартные ошибки):

$$\hat{I} = \underset{7,79}{18,1} - \underset{0,14}{1,07}G + \underset{0,02}{0,36}Y,\ R^2 = 0,99.$$

Исследователь также упорядочил наблюдения по увеличению Y и оценил регрессии снова для 11 стран с наименьшим уровнем дохода и для 11 стран с наибольшим уровнем дохода. Величины RSS для этих регрессий равны 321 и 28 101 соответственно.

Выполним тест Голдфелда — Квандта на гетероскедастичность.

- Fрасч = 28101/321=87,5
- Fтабл = 3,23
- Fрасч> Fтабл, т.е. H0 не принимается, в модели есть гетероскедастичность

Тест Глейзера

Однако зависимость дисперсии ошибок от X_j может быть необязательно линейной. Более разнообразные формы функциональной зависимости проверяются в ***тесте Глейзера*** с альтернативной гипотезой

$$H_1: \sigma_i \sim X_{ji}^{\gamma} \text{ для некоторого } X_j,\ i = 1, ..., n,\ \gamma \in \{1; 0{,}5; -1\}.$$

Тест Глейзера состоит из следующих шагов (первые два шага в тесте Глейзера такие же, что и в тесте Голдфелда — Квандта).

3. Оцениваем вспомогательные регрессии:

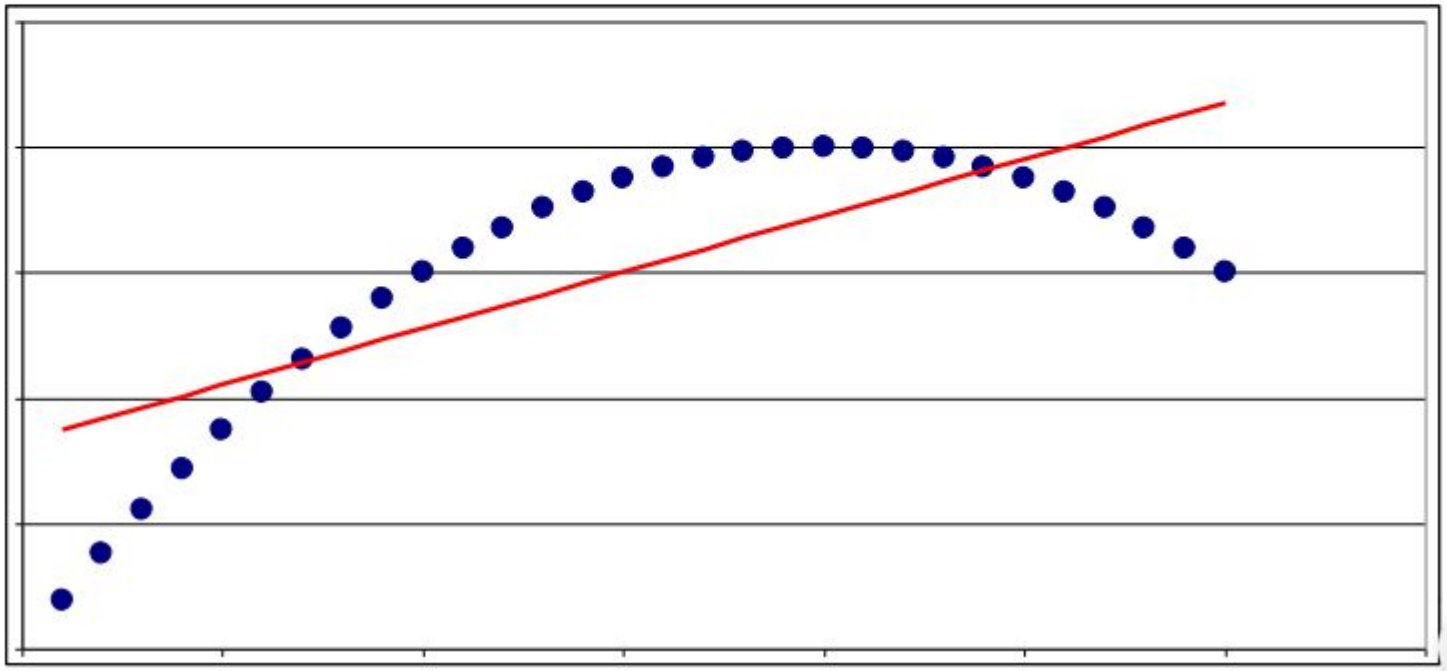
$$|e| = \alpha + \beta X_j + \varepsilon, \quad |e| = \alpha + \beta\sqrt{X_j} + \varepsilon, \quad |e| = \alpha + \frac{\beta}{X_j} + \varepsilon.$$

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

Ложная гетероскедастичность

Возникает в результате неправильной спецификации уравнения регрессии

Пример ситуации, в которой формальные тесты могут показать наличие г/с, хотя на самом деле проблема в другом:



Что делать?

1. Вместо обычных стандартных ошибок $se(\hat{\beta}_j)$ использовать стандартные ошибки, устойчивые к гетероскедастичности $se_{HC}(\hat{\beta}_j)$ - робастные стандартные ошибки

HC – heteroskedascity consistent – устойчивые к гетероскедастичности

*Их следует использовать, если есть случайная выборка и объекты могут быть разного размера

Робастные стандартные ошибки

Утверждение 5.1[2]. Если выполнены условия теоремы Гаусса — Маркова, то

$$\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-k-1} \tag{5.11}$$

является несмещенной оценкой дисперсии ошибок σ_ε^2.

Замечание 5.2. Оценки дисперсий МНК-оценок вычисляются по формуле

$$\hat{\sigma}_{\hat{\beta}_j}^2 = \hat{\sigma}_\varepsilon^2 (\mathbf{X}^T\mathbf{X})_{jj}^{-1},\ j = 0, 1, \ldots, k,$$

где через $(\mathbf{A})_{jj}$ обозначен элемент, находящийся на пересечении j-й строки и j-го столбца матрицы $\mathbf{A}$.

В выдачах статистических пакетов $\hat{\sigma}_{\hat{\beta}_j}$ обычно обозначают $s.e.(\hat{\beta}_j)$, где *s.e.* — стандартные ошибки.

Наиболее распространенным способом коррекции гетероскедастичности в общем виде является использование оценок Уайта[1] для дисперсий коэффициентов:

$$\hat{\mathbf{V}}(\hat{\beta}) = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \mathbf{x}_i \mathbf{x}_i^T\right)\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1},$$

где $\mathbf{x}_i$ — i-я строка матрицы $\mathbf{X}$, $i = 1, \ldots, n$.

Напомним, что диагональными элементами в матрице $\hat{\mathbf{V}}(\hat{\beta})$ являются оценки дисперсий оценок коэффициентов $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

---

[1] См. статью: *White H.* A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity // Econometrica. 1980. P. 817—838.

Робастные стандартные ошибки

Оценки Уайта являются состоятельными[1], но на конечных выборках даже они не полностью корректируют смещение оценок стандартных ошибок коэффициентов. В современных статистических пакетах опция, выполняющая коррекцию по Уайту, называется HC0 (*heteroskedasticity consistent*, состоятельные при гетероскедастичности), однако рекомендуется использовать опции HC2 или (при наличии в выборке выбросов — наблюдений, сильно отличающихся от остальных) HC3: исследования показывают, что на конечных выборках они дают самые точные результаты[2].

```
vcovHC(m1)
vcovHC(m1, type="HC0")
vcovHC(m1, type="HC2")
coeftest(m1, vcov. =vcovHC(m1))
coeftest(m1, vcov. =vcovHC(m1, type="HC0"))
coeftest(m1, vcov. =vcovHC(m1, type="HC2"))
```

Робастные стандартные ошибки

```
> h= read.csv("flats_moscow.txt", sep="\t", header=TRUE, dec=".")
> m1=lm(data=h,price~totsp)
> vcov(m1)
             (Intercept)        totsp
(Intercept)  13.7772925 -0.180775186
totsp        -0.1807752  0.002473516
> vcovHC(m1,type = "HC0")
             (Intercept)       totsp
(Intercept)  60.7744942 -0.88074075
totsp        -0.8807408  0.01282962
>
```

Робастные стандартные ошибки

```
> coeftest(m1)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -62.044844   3.711778 -16.716 < 2.2e-16 ***
totsp         2.593462   0.049734  52.146 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> coeftest(m1, vcov. = vcovHC(m1,type = "HC0"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -62.04484    7.79580 -7.9588 2.858e-15 ***
totsp         2.59346    0.11327 22.8967 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Робастные стандартные ошибки

Было:

```
> confint(model)
                 2.5 %     97.5 %
(Intercept) -69.324117 -54.765570
totsp         2.495927   2.690998
```

Стало:

```
> ci = mutate(ci,left_ci=estimate-1.96*se_hc,right_ci=estimate+1,96*se_hc)
> ci
    estimate     se_hc    left_ci   right_ci 96 * se_hc
1 -62.044844 7.8591534 -77.448784 -61.044844  754.47873
2   2.593462 0.1141737   2.369682   3.593462   10.96067
```

Доверительный интервал при переменной totsp увеличился.
Часто это приводит к изменению значимости коэффициента регрессии.

Робастные стандартные ошибки

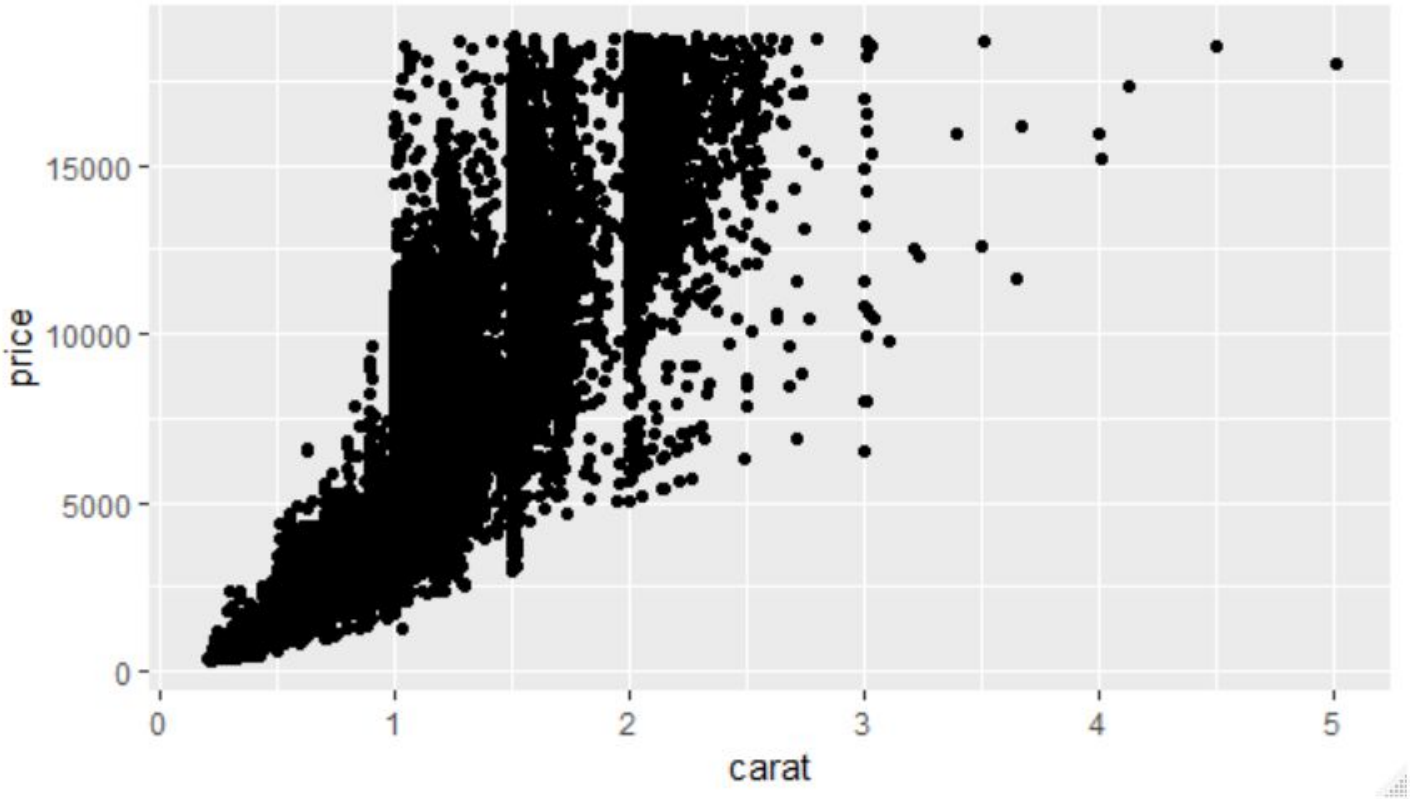
Стандартные ошибки в форме Уайта не устраняют саму гетероскедастичность, но устраняют одно из главных ее негативных последствий.

На практике в пространственных данных гетероскедастичность есть почти всегда.

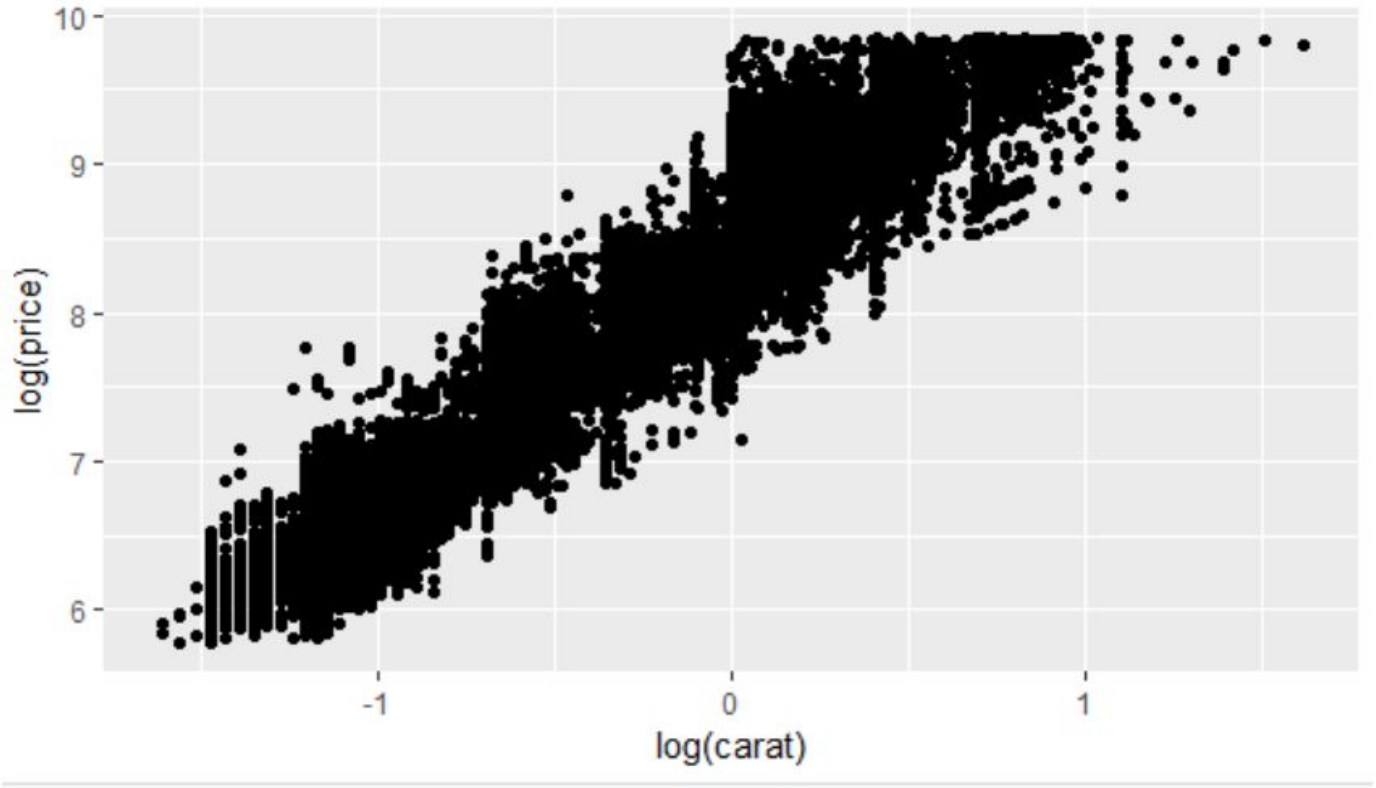
=> Если у вас нет веских оснований говорить об отсутствии гетероскедастчиности в вашей модели, то следует использовать робастные стандартные ошибки.

Что делать?

2. Переход к логарифмам.



Переход к логарифмам



Пример(Wooldridge)

We use the data in HPRICE1.RAW to test for heteroskedasticity in a simple housing price equation. The estimated equation using the levels of all variables is

$$\widehat{price} = -21.77 + .00207\, lotsize + .123\, sqrft + 13.85\, bdrms \quad [8.17]$$
$$(29.48)\ (.00064)\qquad (.013)\qquad (9.01)$$
$$n = 88,\ R^2 = .672.$$

This equation tells us *nothing* about whether the error in the population model is heteroskedastic. We need to regress the squared OLS residuals on the independent variables. The R-squared from the regression of $\hat{u}^2$ on *lotsize*, *sqrft*, and *bdrms* is $R^2_{\hat{u}^2} = .1601$. With $n = 88$ and $k = 3$, this produces an F statistic for significance of the independent variables of $F = [.1601/(1 - .1601)](84/3) \approx 5.34$. The associated p-value is .002, which is strong evidence against the null. The LM statistic is $88(.1601) \approx 14.09$; this gives a p-value $\approx$ .0028 (using the χ^2_3 distribution), giving essentially the same conclusion as the F statistic. This means that the usual standard errors reported in (8.17) are not reliable.

In Chapter 6, we mentioned that one benefit of using the logarithmic functional form for the dependent variable is that heteroskedasticity is often reduced. In the current application, let us put *price*, *lotsize*, and *sqrft* in logarithmic form, so that the elasticities of *price*, with respect to *lotsize* and *sqrft*, are constant. The estimated equation is

$$\widehat{\log(price)} = -1.30 + .168 \log(lotsize) + .700 \log(sqrft) + .037\, bdrms \quad [8.18]$$
$$(.65)\ (.038)\qquad (.093)\qquad (.028)$$
$$n = 88,\ R^2 = .643.$$

Regressing the squared OLS residuals from this regression on log(*lotsize*), log(*sqrft*), and *bdrms* gives $R^2_{\hat{u}^2} = .0480$. Thus, $F = 1.41$ (p-value = .245), and $LM = 4.22$ (p-value = .239). Therefore, we fail to reject the null hypothesis of homoskedasticity in the model with the logarithmic functional forms. The occurrence of less heteroskedasticity with the dependent variable in logarithmic form has been noticed in many empirical applications.

Пример R-studio

```
m2<-lm(data=df2, price~lotsize+sqrft+bdrms)
bptest(m2)
m3<-lm(data=df2,log(price)~log(lotsize)+log(sqrft)+bdrms)
bptest(m3)
```

```
> m2<-lm(data=df2, price~lotsize+sqrft+bdrms)
> bptest(m2)

        studentized Breusch-Pagan test

data:  m2
BP = 14.092, df = 3, p-value = 0.002782

> m3<-lm(data=df2,log(price)~log(lotsize)+log(sqrft)+bdrms)
> bptest(m3)

        studentized Breusch-Pagan test

data:  m3
BP = 4.2232, df = 3, p-value = 0.2383
```

Что делать?

3. **Вместо Метода Наименьших Квадратов (OLS – Ordinary Least Squares) использовать Взвешенный Метод Наименьших Квадратов (WLS – Weighted Least Squares)**

Взвешенный МНК

Как и раньше, будем предполагать, что среднее значение остаточных величин равно нулю. А вот дисперсия их не остается неизменной для разных значений фактора, а пропорциональна величине K_i, т. е.

$$\sigma^2_{\varepsilon_i} = \sigma^2 \cdot K_i,$$

где $\sigma^2_{\varepsilon_i}$ – дисперсия ошибки при конкретном i-м значении фактора;
σ^2 – постоянная дисперсия ошибки при соблюдении предпосылки о гомоскедастичности остатков;
K_i – коэффициент пропорциональности, меняющийся с изменением величины фактора, что и обусловливает неоднородность дисперсии.

При этом предполагается, что σ^2 неизвестна, а в отношении величины K выдвигаются определенные гипотезы, характеризующие структуру гетероскедастичности.

Взвешенный МНК

В общем виде для уравнения

$$y_i = a + b \cdot x_i + \varepsilon_i \text{ при } \sigma^2_{\varepsilon_i} = \sigma^2 \cdot K_i,$$

модель примет вид: $y_i = \alpha + \beta_i \cdot x_i + \sqrt{K_i} \cdot \varepsilon_i$.

В ней остаточные величины гетероскедастичны. Предполагая в них отсутствие автокорреляции, можно перейти к уравнению с гомоскедастичными остатками, поделив все переменные, зафиксированные в ходе i-го наблюдения на $\sqrt{K_i}$. Тогда дисперсия остатков будет величиной постоянной, т. е. $\sigma^2_{\varepsilon_i} = \sigma^2$.

Иными словами, от регрессии y по x мы перейдем к регрессии на новых переменных: $y/\sqrt{K}$ и $x/\sqrt{K}$.

Взвешенный МНК

Уравнение регрессии примет вид:

$$\frac{y_i}{\sqrt{K_i}} = \frac{\alpha}{\sqrt{K_i}} + \beta \cdot \frac{x_i}{\sqrt{K_i}} + \varepsilon_i .$$

Взвешенный МНК

Исходные данные для данного уравнения будут иметь вид:

$$y = \begin{vmatrix} \frac{y_1}{\sqrt{K_1}} \\ \frac{y_2}{\sqrt{K_2}} \\ \cdots\cdots \\ \frac{y_n}{\sqrt{K_n}} \end{vmatrix}, \qquad x = \begin{vmatrix} \frac{x_1}{\sqrt{K_1}} \\ \frac{x_2}{\sqrt{K_2}} \\ \cdots\cdots \\ \frac{x_n}{\sqrt{K_n}} \end{vmatrix}.$$

По отношению к обычной регрессии уравнение с новыми, преобразованными переменными представляет собой взвешенную регрессию, в которой переменные y и x взяты с весами $1/\sqrt{K}$.

Взвешенный МНК

Оценка параметров нового уравнения с преобразованными переменными приводит к взвешенному методу наименьших квадратов, для которого необходимо минимизировать сумму квадратов отклонений вида

$$S = \sum \frac{1}{K_i} \cdot (y_i - a - b \cdot x_i)^2.$$

Взвешенный МНК

Соответственно получим следующую систему нормальных уравнений:

$$\begin{cases} \sum \frac{y_i}{K_i} = a \cdot \sum \frac{1}{K_i} + b \cdot \sum \frac{x_i}{K_i}, \\ \sum \frac{y_i \cdot x_i}{K_i} = a \cdot \sum \frac{x_i}{K_i} + b \cdot \sum \frac{x_i^2}{K_i}. \end{cases}$$

Если преобразованные переменные x и y взять в отклонениях от средних уровней, то коэффициент регрессии b можно определить как

$$b = \frac{\sum \frac{1}{K} \cdot x \cdot y}{\sum \frac{1}{K} \cdot x^2}.$$

Взвешенный МНК

Если преобразованные переменные x и y взять в отклонениях от средних уровней, то коэффициент регрессии b можно определить как

$$b = \frac{\sum \frac{1}{K} \cdot x \cdot y}{\sum \frac{1}{K} \cdot x^2}.$$

При обычном применении метода наименьших квадратов к уравнению линейной регрессии для переменных в отклонениях от средних уровней коэффициент регрессии b определяется по формуле

$$b = \frac{\sum (x \cdot y)}{\sum x^2}.$$

Взвешенный МНК

- Обычно в качестве величины K рассматривают фактор (квадрат фактора), который отвечает за гетероскедастичность (предварительно проводят тест Гольдфельда-Квандта или Глейзера)

Что делать?

$$sav_i = \beta_0 + \beta_1 inc_i + u_i$$

$$\mathrm{Var}(u_i|inc_i) = \sigma^2 inc_i.$$

$$sav_i/\sqrt{inc_i} = \beta_0(1/\sqrt{inc_i}) + \beta_1\sqrt{inc_i} + u_i^*,$$

OLS vs. WLS (учебник Wooldridge)

We now estimate equations that explain net total financial wealth (*nettfa*, measured in \$1,000s) in terms of income (*inc*, also measured in \$1,000s) and some other variables, including age, gender, and an indicator for whether the person is eligible for a 401(k) pension plan. We use the data on single people (*fsize* = 1) in 401KSUBS.RAW. In Computer Exercise C12 in Chapter 6, it was found that a specific quadratic function in *age*, namely $(age - 25)^2$, fit the data just as well as an unrestricted quadratic. Plus, the restricted form gives a simplified intepretation because the minimum age in the sample is 25: *nettfa* is an increasing function of *age* after *age* = 25.

The results are reported in Table 8.1. Because we suspect heteroskedasticity, we report the heteroskedasticity-robust standard errors for OLS. The weighted least squares estimates, and their standard errors, are obtained under the assumption $\text{Var}(u|inc) = \sigma^2 inc$.

OLS vs. WLS (учебник Wooldridge)

TABLE 8.1 Dependent Variable: *nettfa*

Independent Variables	(1) OLS	(2) WLS	(3) OLS	(4) WLS
inc	.821 (.104)	.787 (.063)	.771 (.100)	.740 (.064)
$(age - 25)^2$	—	—	.0251 (.0043)	.0175 (.0019)
male	—	—	2.48 (2.06)	1.84 (1.56)
e401k	—	—	6.89 (2.29)	5.19 (1.70)
intercept	−10.57 (2.53)	−9.58 (1.65)	−20.98 (3.50)	−16.70 (1.96)
Observations	2,017	2,017	2,017	2,017
R-squared	.0827	.0709	.1279	.1115

OLS vs. WLS (учебник Wooldridge)

```
library(wooldridge)
library(psych)
library(sandwich)
library(lmtest)
library(car)

df4<-k401ksubs
m6 <- lm(nettfa ~ inc, data = df4, subset = (fsize == 1))
summary(m6)

m6.2 <- lm(nettfa ~ inc, weights = 1/inc, data = df4, subset= (fsize == 1))
summary(m6.2)
```

Пример

Задача 10.4[1]. По данным с 1946 по 1975 г. Е. А. Ханушек и Дж. Е. Джексон[2] оценили коэффициенты уравнений регрессий (под оценками коэффициентов указаны их стандартные ошибки):

$$\hat{C}_t = \underset{2,73}{26,19} + \underset{0,006}{0,6248}\,GNP_t - \underset{0,0736}{0,4398}\,D_t;$$

$$\widehat{\left(\frac{C}{GNP}\right)}_t = \underset{2,22}{25,92}\frac{1}{GNP_t} + \underset{0,00397}{0,6246} - \underset{0,0597}{0,4315}\frac{D_t}{GNP_t},$$

где C — агрегированные частные потребительские расходы; GNP (*gross national product*) — валовой национальный продукт; D — национальные расходы на оборону.

С какой целью оценили второе уравнение? Какое при этом было сделано предположение о дисперсии ошибок? Можно ли сравнивать $R2$ в двух регрессиях? Ответ следует обосновать и дать интерпретацию полученным результатам.

- Скорее всего, с помощью тестов (Гольдфельдта-Кванта или Глейзера) была выявлена гетероскедастичность по переменной GNP. Второе уравнение оценивалось с целью устранения гетероскедастичности в модели.
- При этом делалось предположение, что дисперсия ошибок пропорциональна квадрату переменной GNP.
- Сравнивать R2 в этих регрессиях мы не можем, т.к. зависимые переменные разные.

Таблица значений F-критерия Фишера на уровне значимости a = 0,05

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.84	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.99	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314