Эконометрика, НИУ ВШЭ

Составили эконометрессы Максимовская Анастасия, Перевышина Татьяна, Ситникова Арина, а также Пилипейко Роман

1.0 Линейная регрессия с одной объясняющей переменной (парная регрессия)

Линейная регрессионная модель имеет следующий вид:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, ..., n$$

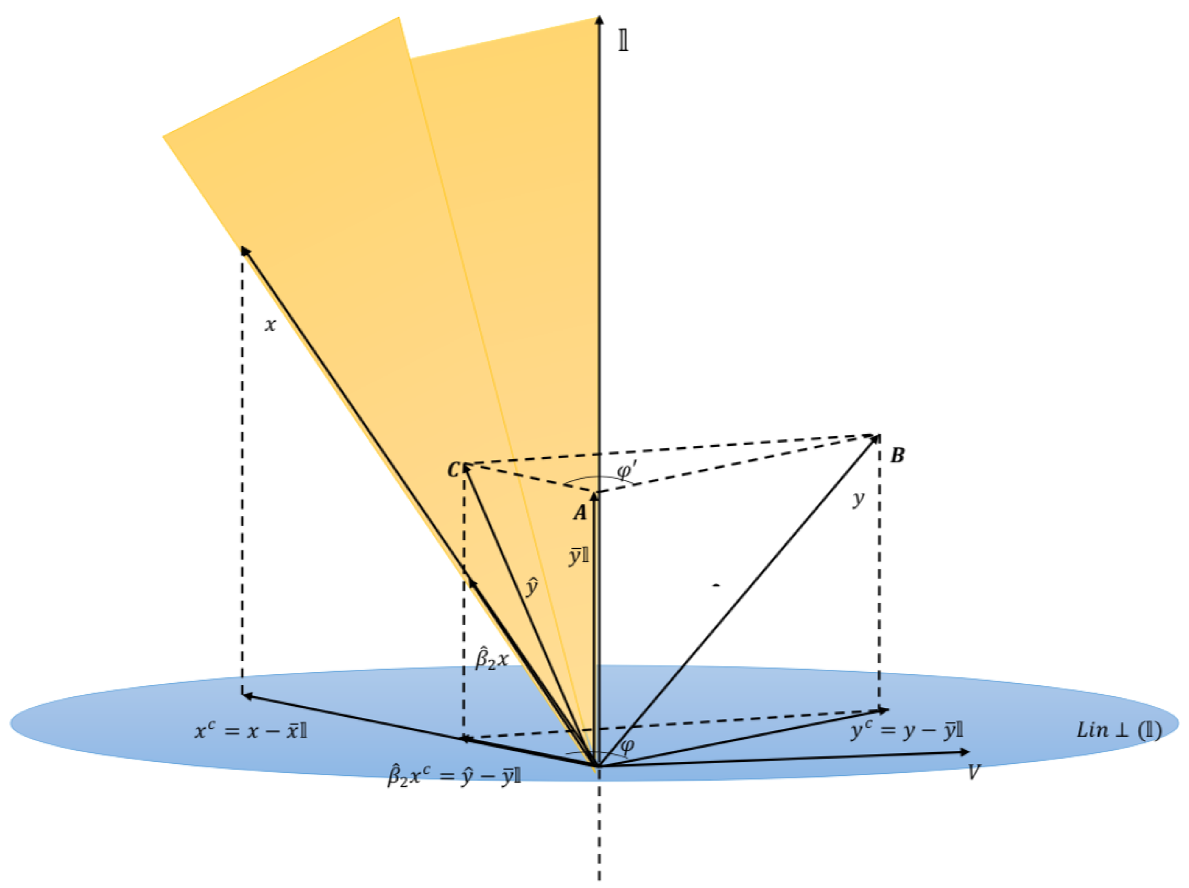
1.1 Метод наименьших квадратов (OLS) для нахождения оценок коэффициентов парной регрессии

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

Зачем он нужен? Для нахождения оценок коэффициентов $\hat{\beta}_1, \hat{\beta}_2$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\text{sCov}(x,y)}{\text{sVar}(x)}$$



$$\text{sCorr}(x;y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Проекция y на $Lin(\mathbb{1}; x)$: $\hat{y} = \hat{\beta}_1 \mathbb{1} + \hat{\beta}_2 x^c$

Проекция x и y на $Lin\perp(\mathbb{1})$: $x^c = x - \bar{x}\mathbb{1} = \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}, y^c = y - \bar{y}\mathbb{1} = \sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

Проекция x и y на $Lin\perp(\mathbb{1})$: $x^c = x - \bar{x}\mathbb{1} = \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Проекция y^c на $Lin(x^c)$: $\hat{\beta}_2 x^c = \hat{y} - \bar{y}\mathbb{1}$

Проекция x^c на $Lin(\mathbb{1})$: 0

Проекция x^c на $Lin\perp(\mathbb{1})$: x^c

$$\text{TSS} = \sum_{i=1}^{n}(y_i-\bar{y})^2 = ESS + RSS = |y - \bar{y}\mathbb{1}|^2 = AB^2 = y^{c^2}$$

$$\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2 = AC^2$$

$$\text{RSS} = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = BC^2$$

$$R^2 = \frac{ESS}{TSS} = \frac{\text{sVar}(\hat{y})}{\text{sVar}(y)} = \frac{AC^2}{AB^2} = \cos(\varphi') = \text{sCorr}^2(y;\hat{y}) = \text{sCorr}^2(y;x)$$

1.2 Теорема Гаусса - Маркова для случая парной регрессии

Зачем она нужна? Для получения дополнительной информации о характеристиках распределения оценок $\hat{\beta}_1, \hat{\beta}_2$

Теорема 1.2 Гаусса - Маркова для случая парной регрессии .
Если для модели $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \varepsilon_i, i = 1, ..., n$ выполняются следующие условия:
1) Строится регрессия y на x
2) Не все x_i равны между собой
3) Математическое ожидание от любой ошибки регрессии равно 0
4) Дисперсия от любой ошибки регрессии равна σ_ε^2
5) $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ - отсутствие автокорреляции,
то МНК - оценки $\hat{\beta}_1$ и $\hat{\beta}_2$ являются *BLUE*

Best (наилучшие)	для любых $\tilde{\beta}_i$ верно: $MSE(\hat{\beta}_i) \leq MSE(\tilde{\beta}_i)$
Linear (линейные)	Оценки линейные по y
Unbiased (несмещенные) **E**stimator (оценки)	$\text{E}(\hat{\beta}_i) = \beta_i$

Утверждение 1.2 Если условия теоремы Гаусса - Маркова выполнены, то дисперсии МНК оценок:

$$\text{Var}(\hat{\beta}_1) = \sigma_\varepsilon^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \text{Var}(\hat{\beta}_2) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

а их ковариация:

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\bar{x}\sigma_\varepsilon^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

1.3 Предположение о нормальном распределении случайной ошибки в рамках классической линейной регрессии

Утверждение 1.3.1 Если верно, что $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $i = 1...n$, то МНК-оценки коэффициентов парной регрессии также имеют нормальное распределение, причём $\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$, $\hat{\beta}_2 \sim N(\beta_2, Var(\hat{\beta}_2))$.

Утверждение 1.3.2 $\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-2}$ является несмещённой оценкой для σ_ε^2.

Утверждение 1.3.3 $\frac{RSS}{\sigma_\varepsilon^2} \sim \chi_{n-2}^2$ для парной регрессии.

Утверждение 1.3.4 Оценки $\hat{\beta}_1$ и $\hat{\sigma}_\varepsilon^2$, $\hat{\beta}_2$ и $\hat{\sigma}_\varepsilon^2$ являются независимыми.

1.3.1 Проверка гипотезы о конкретном значении коэффициентов парной регрессии

$$H_0 : \beta_2 = \beta_2^0$$
$$H_a : \beta_2 \neq \beta_2^0$$

Если $|t = \frac{\hat{\beta}_2 - \beta_2^0}{\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}}| > t^{cr}$ (при уровне значимости $\alpha/2$ и $n-2$ степенях свободы), то гипотеза H_0 отвергается в пользу гипотезы H_a.
Если значение $|t = \frac{\hat{\beta}_2 - \beta_2^0}{\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}}| < t^{cr}$ (при уровне значимости $\alpha/2$ и $n-2$ степенях свободы), то гипотеза H_0 не отвергается.

1.3.2 Проверка гипотезы о значимости коэффициентов парной регрессии

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 \neq 0$$

Если значение $|t = \frac{\hat{\beta}_2}{\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}}| > t^{cr}$ (при уровне значимости $\alpha/2$ и $n-2$ степенях свободы), то гипотеза H_0 отвергается, значит, β_2 значим.

Если значение $|t = \frac{\hat{\beta}_i}{\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}}| < t^{cr}$ (при уровне значимости $\alpha/2$ и $n-2$ степенях свободы), то гипотеза H_0 не отвергается, значит, β_2 не значим.

Доверительный интервал

$$[\hat{\beta}_2 - t^{cr}\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}; \hat{\beta}_2 + t^{cr}\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}]$$

Если доверительный интервал включает в себя нуль, то этот коэффициент не значим, при уровне значимости α.
В обоих случаях t^{cr} рассчитывается на уровне значимости $\alpha/2$ с $n-2$ степенями свободы.

1.3.3 Прогнозирование по модели парной регрессии и его точность

- **Доверительный интервал для среднего прогноза**

$$\hat{\text{Var}}(\hat{y}_{n+1}) = \hat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}) = \hat{\sigma}_\varepsilon^2[\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]$$

$$P(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1} - t_{\alpha/2,n-2}\sqrt{\hat{\text{Var}}(\hat{y}_{n+1})} \leq E(y_{n+1}|x = x_{n+1}) \leq \hat{\beta}_1 + \hat{\beta}_2 x_{n+1} + t_{\alpha/2,n-2}\sqrt{\hat{\text{Var}}(\hat{y}_{n+1})}) = 1 - \alpha$$

- **Доверительный интервал для индивидуального прогноза**

$$\hat{\text{Var}}(y_{n+1}) = \hat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1} + \varepsilon_{n+1}) = \hat{\sigma}_\varepsilon^2[1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}]$$

$$P(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1} - t_{\alpha/2,n-2}\sqrt{\hat{\text{Var}}(y_{n+1})} \leq y_{n+1} \leq \hat{\beta}_1 + \hat{\beta}_2 x_{n+1} + t_{\alpha/2,n-2}\sqrt{\hat{\text{Var}}(y_{n+1})}) = 1 - \alpha$$

- **Доверительный интервал дисперсии ошибок регрессии**

$$P(\frac{RSS}{\chi^2_{\alpha/2,n-2}} \leq \sigma_\varepsilon^2 \leq \frac{RSS}{\chi^2_{1-\alpha/2,n-2}}) = 1 - \alpha$$

1.3.4 Проверка гипотезы о нормальности распределения

- **Критерий Колмогорова - Смирнова**

Когда используется? Выборка состоит из достаточно большого количества наблюдений n

$$H_0 : F(t) = G(t) \text{ для любого t}$$
$$H_a : F(t) \neq G(t) \text{ при некотором t}$$

С использованием таблиц:

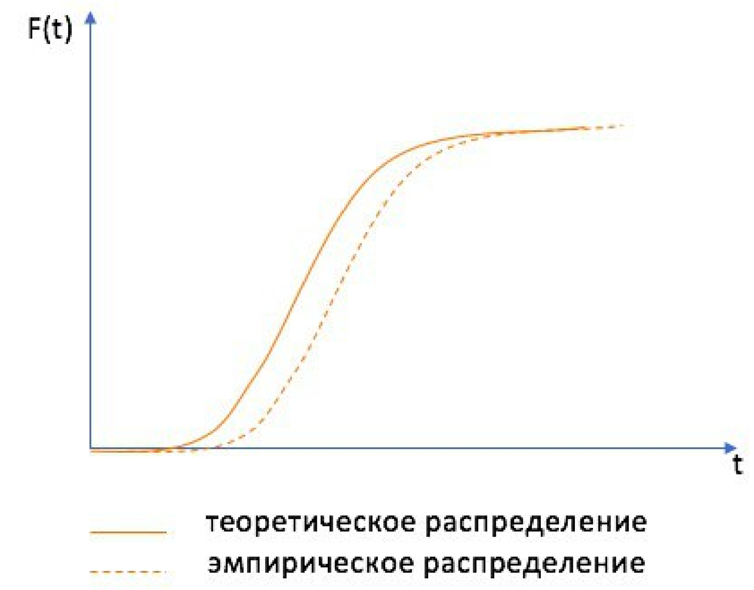
$$J = \frac{mn}{d} \max_{-\infty<t<\infty} |F_m(t) - G_n(t)|$$

где $F_m(t)$ и $G_m(t)$ – эмпирические функции распределения для выборок $x_1, ..., x_m$ и $y_1, ..., y_n$, а именно:

$$F_m(t) = \frac{\text{Число элементов в первой выборке, не превышающих } t}{m}$$
$$G_n(t) = \frac{\text{Число элементов во второй выборке, не превышающих } t}{n}$$

d – НОД(m, n), m и n — общее количество элементов в первой и второй выборке соответственно.

Статистика имеет специальное распределение D при верной H_0:



- **Тест Харке - Бера** *Когда используется? Выборка состоит из достаточно большого количества наблюдений n*

$$H_0 : s = 0, k = 3$$
$$H_a : s \neq 0, k \neq 3$$

Тестовая статистика:

$$JB = \frac{n}{6}(s^2 + \frac{1}{4}(k-3)^2) \sim \chi_2^2$$

Коэффициент асимметрии: $s = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^3}{\hat{\sigma}^3}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$

Для симметричных распределений, в том числе нормального, этот показатель равен нулю.

Коэффициент эксцесса: $k = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i-\bar{x})^4}{\hat{\sigma}^4}$

Для нормального распределения этот показатель равен 3.

Если $JB > \chi_2^2$ при выбранном уровне значимости, то гипотеза отвергается.

- **Тест Шапиро - Уилка**

Когда используется? Выборка состоит из небольшого количества наблюдений n

$$H_0 : x_1, ..., x_n \text{– выборка из нормального распределения}$$
$$H_a : x_1, ..., x_n \text{– выборка не из нормального распределения}$$

Тестовая статистика:

$$W = \frac{(\sum_{i=1}^{n}\alpha_{(i)}x_i)^2}{\sum_{i=1}^{n}(x_{(i)} - \bar{x})^2} \sim W_\alpha$$

Вектор коэффициентов: $\alpha_{(i)} = \frac{m^T V^{-1}}{m^T V^{-1} V^{-1} m}$

m – вектор математических ожиданий порядковых статистик нормального распределения;
V – ковариационная матрица этих порядковых статистик размера n*n.

Для небольших n коэффициенты $\alpha_{(i)}$ берутся из таблиц.

Если $W < W_\alpha$, то гипотеза H_0 отвергается.

2.0 Линейная регрессия с несколькими объясняющими переменными (множественная регрессия)

Модель **множественной линейной регрессии** имеет следующий вид:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, ..., n$$

где i – номер наблюдения, а ε_i – ошибки регрессии

2.1 Формула для МНК - оценок коэффициентов множественной линейной регрессии

Зачем она нужна? Для нахождения оценки коэффициентов $\hat{\beta}_i$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{\beta}_m = \frac{\det\left(\text{sCov}\left(\begin{pmatrix} x_{1i} & x_{1i} \\ x_{2i} & x_{2i} \\ ... \;\; ; & ... \\ x_{mi} & y_i \\ x_{ki} & x_{ki} \end{pmatrix}\right)\right)}{\det\left(\text{sVar}\left(\begin{pmatrix} x_{1i} \\ x_{2i} \\ ... \\ x_{mi} \\ x_{ki} \end{pmatrix}\right)\right)}$$

Показатель качества подгонки множественной регрессии:

$$1) R^2 = \frac{ESS}{TSS} = \hat{r}^2_{y,\hat{y}}, R^2 \in [0;1]; \qquad 2) R^2_{adj} = 1 - \frac{RSS(n-1)}{TSS(n-k)}$$

R^2_{adj} – показатель качества, штрафующий за включение слишком большого количества переменных в модель.

2.2 Теорема Гаусса - Маркова для случая множественной регрессии

Теорема 2.2 Гаусса - Маркова для случая множественной регрессии
Если для модели $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, i = 1, ..., n$ выполнены условия:
1) Модель правильно специфицирована, то есть в нее включен только необходимый x, объясняющий y
2) Нет линейной зависимости между факторами
3) Математическое ожидание от любой ошибки регрессии равно 0
4) Дисперсия от любой ошибки регрессии равна σ^2_ε
5) Cov $(\varepsilon_i, \varepsilon_j) = 0$ – отсутствие автокорреляции,
то оценки МНК $\hat{\beta} = (X^T X)^{-1} X^T y$ являются BLUE.

Если выполнены условия ТГМ, то дисперсии МНК оценок следующие:

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

2.3 Предположение о нормальном распределении случайной ошибки в рамках классической линейной регрессии

Теорема 2.3 Фишера
Если выполнены условия Теоремы Гаусса - Маркова и $\varepsilon \sim N(0, \sigma^2_\varepsilon I)$, то
1) Оценки МНК $\hat{\beta} \sim N(\beta, \sigma^2_\varepsilon (X^T X)^{-1})$
2) $\frac{RSS}{\sigma^2_\varepsilon} \sim \chi^2_{n-k}$
3) $\hat{\beta}$ и RSS независимы

Утверждение 2.3.1 $\hat{\sigma}^2_\varepsilon = \frac{RSS}{n-k}$ является несмещённой оценкой для σ^2_ε.

2.4 Проверка гипотез в случае множественной регрессии для модели $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$

2.4.1 Проверка гипотезы о конкретном значении коэффициентов множественной регрессии

$$H_0 : \beta_j = \beta_j^0$$
$$H_a : \beta_j \neq \beta_j^0$$

Тестовая статистика рассчитывается как:

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\hat{\sigma}^2_\varepsilon (X^T X)^{-1}_{jj}}} \sim t_{n-k}$$

Если $|t| > t^{cr}$(при уровне значимости $\alpha/2$ и $n-k$ степеней свободы), то гипотеза H_0 отвергается.

Для проверки *гипотезы о значимости* используют $\beta_j^0 = 0$.

Доверительный интервал для проверки гипотезы:

$$[\hat{\beta}_j - t^{cr}\sqrt{\hat{\sigma}^2_\varepsilon (X^T X)^{-1}_{jj}}; \hat{\beta}_j + t^{cr}\sqrt{\hat{\sigma}^2_\varepsilon (X^T X)^{-1}_{jj}}]$$

t^{cr} рассчитывается на уровне доверия $\alpha/2$ с $n-k$ степенями свободы. Если в интервал входит нуль, то основная гипотеза не отвергается.

2.4.2 Проверка общей линейной гипотезы о коэффициентах множественной регрессии

$$H_0 : Q\beta = q$$
$$H_a : Q\beta \neq q$$

Инкорпорируем ограничение в регрессию и рассчитываем R^2_r и RSS_r
Также оцениваем регрессию без ограничений и получаем R^2_{ur} и RSS_{ur}
Тестовая статистика рассчитывается как:

$$F = \frac{(RSS_r - RSS_{ur})/q}{RSS_{ur}/(n-k)} = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k)} \sim F_{q,n-k}$$

Если при выбранном уровне значимости α верно, что $F > F^{cr}$, то основная гипотеза H_0 отвергается.

2.4.3 Проверка гипотезы об адекватности регрессии

$$H_0 : \beta_2 = ... = \beta_k = 0$$
$$H_a : \text{хотя бы один коэффициент } \beta_j \text{ отличен от нуля}, j = 2, .., k$$

Тестовая статистика рассчитывается как:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}$$

Если при выбранном уровне значимости α верно, что $F > F^{cr}$, то основная гипотеза H_0 отвергается, регрессия адекватна.

2.5 Фиктивные (dummy) переменные

Фиктивные (dummy) переменные – переменные которые принимают только два значения, обычно 0 и 1.

2.5.1 Проверка гипотез о структурной устойчивости регрессии (о единой зависимости)

- **Тест Чоу**

При $D_i = 1, i = 1, .., n_1$ уравнение регрессии имеет вид:

$$y_i = (\beta_1 + \delta_1) + (\beta_2 + \delta_2) x_{i2} + ... + \varepsilon_i$$

При $D_i = 0, i = n_1 + 1, .., n$ уравнение регрессии имеет вид:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

Пусть $\beta_j' = \beta_j + \delta_j, j = 0, .., k$. Тогда гипотеза будет следующая:

$$H_0 : \beta_i = \beta_i', i = 0, .., k$$
$$H_a : \text{существует такое j, что } \beta_j \neq \beta_j'$$

Тестовая статистика рассчитывается как:

$$F = \frac{(RSS_{pooled} - RSS_1 - RSS_2)/k}{(RSS_1 + RSS_2)/(n-2k)} \sim F_{k,n-2k}$$

где RSS_{pooled} – RSS в регрессии, оцененной по всем n наблюдениям;
RSS_1 – RSS в регрессии, оцененной по наблюдениям $i = 1, ..., n_1$;
RSS_2 – RSS в регрессии, оцененной по наблюдениям $i = n_1 + 1, ..., n_1 + n_2$.

- **С помощью dummy переменных**

$$H_0 : \delta_1 = ... = \delta_k = 0 \text{ – зависимость едина}$$
$$H_a : \delta_1^2 + ... + \delta_k^2 \neq 0 \text{ – зависимость не едина}$$

Регрессия без ограничений: $y_i = \beta_1 + \delta_1 D_i + \beta_2 x_{i2} + \delta_2 D_i x_{i2} + ... + \varepsilon_i$
Регрессия с ограничениями: $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$

Тестовая статистика рассчитывается как:

$$F = \frac{(RSS_r - RSS_{ur})/k}{RSS_{ur}/(n-2k)} \sim F_{k,n-2k}$$

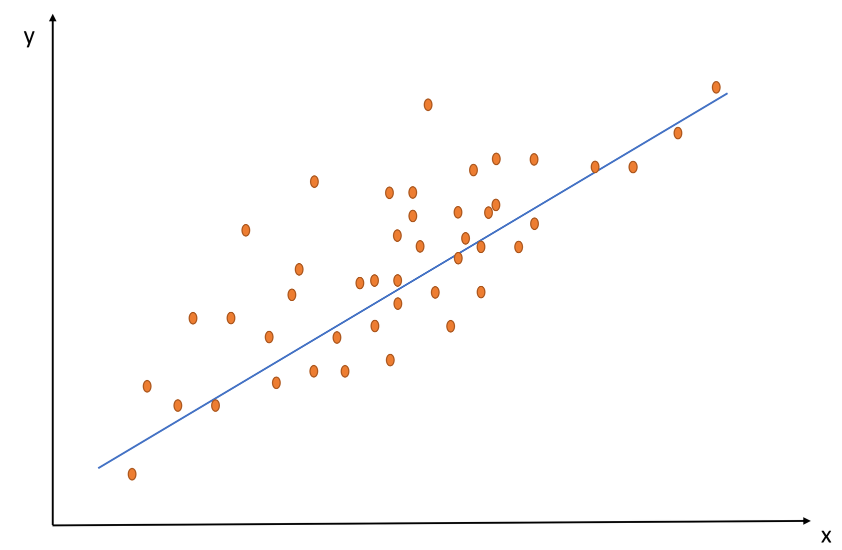
В обоих случаях, если при выбранном уровне значимости α верно, что $F > F^{cr}$, то основная гипотеза H_0 отвергается, и зависимость не едина.

3.0 Выбор функциональной формы модели

3.0.1 Формы модели

- **Линейная модель**

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$



Интерпретация оценки коэффициентов линейной регрессии:
Если коэффициент β_j значим, то при изменении x_j на одну единицу y изменится на $\hat{\beta}_j$ единиц (в сторону увеличения при положительной оценке и в сторону уменьшения при отрицательной).

- **Полулогарифмическая модель (модель с постоянным темпом роста)**

$$\ln y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

Интерпретация оценки коэффициентов полулогарифмической регрессии: Если коэффициент β_j значим, то при изменении x_j на одну единицу y изменится на $\hat{\beta}_j \cdot 100\%$

- **Линейная в логарифмах модель (модель с постоянной эластичностью)**

$$\ln y_i = \beta_1 + \beta_2 \ln x_{i2} + ... + \beta_k \ln x_{ik} + \varepsilon_i$$

Интерпретация оценки коэффициентов логарифмической регрессии:
Если β_j значим, то при изменении x_j на 1% y изменится на $\hat{\beta}_j$%

3.0.2 Выбор между моделями

- **Тест Бера и МакАлера**

Когда используется? Для выбора между полулогарифмической и линейной моделями
Шаг 1. Нахождение оцененных значений зависимой переменной в каждой из этих моделей ($\widehat{\ln y}$ и $\hat{y}$).
Шаг 2. Оценка вспомогательных регрессий

$$\exp(\widehat{\ln y_i}) = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + v_1$$

$$\ln \hat{y}_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ik} + v_2$$

и сохранение остатков регрессий $\hat{v}_1$ и $\hat{v}_2$
Шаг 3. Оценка моделей

$$\ln y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ik} + \theta_1 \hat{v}_1 + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \theta_2 \hat{v}_2 + \varepsilon_i$$

и проверка следующих гипотез о значимости с помощью t-тестов:

$$H_0 : \theta_1 = 0 \qquad H_0 : \theta_2 = 0$$
$$H_a : \theta_1 \neq 0 \qquad H_a : \theta_2 \neq 0$$

Если коэффициент θ_1 не значим, то выбирается полулогарифмическая модель. Если коэффициент θ_2 не значим, то выбирается линейная модель. Если оба коэффициента θ_1 и θ_2 одновременно значимы или не значимы, модель выбрать нельзя.

- **МакКиннона, Уайта и Дэвидсона**

Когда используется? Для выбора между полулогарифмической и линейной моделями
Шаг 1. Нахождение оцененных значений зависимой переменной в каждой из этих моделей ($\widehat{\ln y}$ и $\hat{y}$).
Шаг 2. Оценка вспомогательных регрессий

$$\ln y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \gamma_1 (\hat{y}_i - \exp(\widehat{\ln y_i})) + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \gamma_2 (\ln \hat{y}_i - \widehat{\ln y_i}) + \varepsilon_i$$

и проверка следующих гипотез о значимости с помощью t-тестов:

$$H_0 : \gamma_1 = 0 \qquad H_0 : \gamma_2 = 0$$
$$H_a : \gamma_1 \neq 0 \qquad H_a : \gamma_2 \neq 0$$

Если коэффициент γ_1 не значим, то выбирается полулогарифмическая модель. Если коэффициент γ_2 незначим, то выбирается линейная модель. Если оба коэффициента γ_1 и γ_2 одновременно значимы или не значимы, модель выбрать нельзя.

Также данный тест используется для выбора между линейной и линейной в логарифмах моделями.

Шаг 1. Нахождение оценённых значений зависимых переменных:

$$\widehat{\ln y_i} = \hat{\beta}_1 + \hat{\beta}_2 \ln x_{i2} + ... + \hat{\beta}_k \ln x_{ik}$$

Шаг 2. Оценка вспомогательных регрессий

$$\ln y_i = \beta_1 + \beta_2 \ln x_{i2} + ... + \beta_k \ln x_{ik} + \delta_1 (\hat{y}_i - \exp(\widehat{\ln y_i})) + \varepsilon_i$$

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \delta_2 (\ln \hat{y}_i - \widehat{\ln y_i}) + \varepsilon_i$$

и проверка следующих гипотез о значимости с помощью t-тестов:

$$H_0 : \delta_1 = 0 \qquad H_0 : \delta_2 = 0$$
$$H_a : \delta_1 \neq 0 \qquad H_a : \delta_2 \neq 0$$

Если коэффициент δ_1 не значим, то выбирается линейная в логарифмах модель. Если коэффициент δ_2 не значим, то выбирается линейная модель. Если оба коэффициента δ_1 и δ_2 одновременно значимы или не значимы, модель выбрать нельзя.

- **Тест Бокса - Кокса**

Когда используется? Для выбора функциональной формы модели

$$y_i^{(\lambda)} = \beta_1 + \beta_2 x_{i2}^{(\theta)} + \beta_k x_{ik}^{(\theta)} + \varepsilon_i$$

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln y_i, & \lambda = 0 \end{cases}$$

$$x_j^{(\theta)} = \begin{cases} \dfrac{x_j^{(\theta)} - 1}{\theta}, & \lambda \neq 0 \\ \ln x_j, & \theta = 0 \end{cases}$$

Шаг 1. Оценить коэффициенты $\beta_1, \beta_2, ..., \beta_k$
Шаг 2. Оценить параметры λ и θ с помощью метода максимального правдоподобия.
Для параметров λ и θ можно проверять гипотезы с помощью теста отношения правдоподобия. В случае, когда в модели:

1. $\lambda = \theta$ (зависимая и независимая переменные преобразуются одинаково), можно проверить гипотезу:

$$H_0 : \lambda = \theta = 1 (\text{линейная модель})$$

$$H_a : \lambda = \theta \neq 1$$

а также гипотезу:

$$H_0 : \lambda = \theta = 0 (\text{линейная в логарифмах модель})$$

$$H_a : \lambda = \theta \neq 0$$

2. $\lambda = 1$ (преобразуются только независимые переменные), можно проверить гипотезу:

$$H_0 : \theta = 1 (\text{линейная модель})$$

$$H_a : \theta \neq 1$$

3. $\theta = 1$ (преобразуется только зависимая переменная), можно проверить гипотезу:

$$H_0 : \lambda = 0 (\text{полулогарифмическая модель})$$

$$H_a : \lambda \neq 0$$

- **Метод Зарембки**

Когда используется? Для выбора между полулогарифмической и линейной моделями
Шаг 1.

$$y_{mg} = (\prod_{i=1}^{n} y_i)^{1/n}$$

Шаг 2.

$$y_i* = \frac{y_i}{y_{mg}}$$

Шаг 3. Оценить новые регрессии:

$$y_i^* = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$
$$ln(y_i^*) = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$

И получить RSS_1 и RSS_2 соответственно.
Шаг 4. Проверить гипотезу:

$$H_0 : \text{между двумя моделями нет статистической разницы}$$

$$H_a : \text{между моделями есть статистическая разница}$$

Статистика рассчитывается как:

$$\chi^2 = \frac{n}{2} |ln \frac{RSS_1}{RSS_2}| \sim \chi_1^2$$

При $\chi^2 > \chi^2_{cr}$ гипотеза H_0 отвергается. Тогда из двух построенных моделей следует выбрать ту, для которой RSS будет меньше.

4.0 Ошибки спецификации модели

4.1 Невключение существенной переменной

Предположим, что истинная модель – $y = \beta_1 + \beta_2 x + \beta_3 z + \varepsilon$.
При этом, оценивается модель $y = \beta_1 + \beta_2 x + \varepsilon$.
Тогда оценка МНК коэффициента β_2 будет смещенной во всех случаях, кроме:
1)$\beta_3 = 0$;
2)$\text{sCov}(x, z) = 0$, т.е. когда пропущенный и включенный фактор не коррелируют.
Величина смещения равна $\beta_3 \frac{\text{sCov}(x,z)}{\text{sVar}(x)}$

В матричном виде:
Если истинная модель – $y = X\beta + Z\gamma + \varepsilon$, а оценивается модель $y = X\beta + \varepsilon$, то оценка МНК вектора β будет смещенной во всех случаях, кроме:
1)$\gamma = 0$;
2)$X^T Z = O$, где O – нулевая матрица, т.е. невключенные факторы ортогональны включенным.
Величина смещения равна $(X^T X)^{-1} X^T Z \gamma$

Если же **в модель включены излишние переменные**, то уменьшается эффективность оценок (дисперсия оценок коэффициентов увеличивается, t-статистики уменьшаются), однако они будут несмещенными.

4.2 Тест Рамсея для выявления пропущенных переменных

Исходная модель: $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$

$$H_0 : \text{модель правильно специфицирована}$$
$$H_a : \text{модель неправильно специфицирована}$$

Шаг 1. Оцениваем исходную модель, сохраняем $\hat{y}$ и RSS_r
Шаг 2. Оцениваем вспомогательную модель:

$$y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \alpha_2 \hat{y_i}^2 + ... + \alpha_m \hat{y_i}^m + \varepsilon_i$$

и находим RSS_{ur}
Шаг 3. Проверяем гипотезу:

$$H_0 : \alpha_2 = ... = \alpha_m = 0$$
$$H_a : \alpha_2^2 + ... + \alpha_m^2 > 0$$

Статистика рассчитывается как:

$$F = \frac{(RSS_r - RSS_{ur})/(m-1)}{RSS_{ur}/[n-k-(m-1)]} \sim F_{m-1, n-k-(m-1)}$$

Если при выбранном уровне значимости $F > F^{cr}$, то H_0 отвергается, модель специфицирована неверно.

5.0 Мультиколлинеарность

Теоретическая мультиколлинеарность – линейная зависимость между столбцами матрицы X (например, включение всего набора dummy переменных).
Практическая мультиколлинеарность – "почти линейная зависимость" между столбцами матрицы X.

5.1 Признаки и последствия мультиколлинеарности:

1. Высокие стандартные ошибки и малые t-статистики коэффициентов влекут за собой незначимость коэффициентов; соответственно, невозможно разграничить влияние факторов.
2. Небольшие изменения в данных могут привести к большим изменениям в оценках коэффициентов.

5.2 Индикаторы наличия мультиколлинеарности:

1. Анализ матрицы парных коэффициентов корреляции объясняющих факторов. Если по модулю элемент выше, чем 0.75, то это может свидетельствовать о м/к.
2. $VIF(x_j) = \frac{1}{1-R_j^2}$, где R_j^2 – относится к регрессии фактора x_j на все остальные факторы. Значение больше 6 может свидетельствовать о м/к.
3. $CN = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$, где λ – собственные значения матрицы $x^T x$. Значение больше 30 может свидетельствовать о м/к.

5.3 Методы борьбы с мультиколлинеарностью:

1. Переход к другой функциональной форме модели (например, модель в логарифмах);
2. Метод пошагового исключения (с максимальным P-значением) или включения (с минимальным P-значением) по одной переменной, пока все коэффициенты не останутся значимыми;
3. Метод исключения нескольких переменных одновременно. При этом следует проверить гипотезу:

$$H_0 : \beta_{i_1} = ... = \beta_{i_r} = 0$$
$$H_a : \beta_{i_1}^2 + ... + \beta_{i_r}^2 \neq 0$$

4. PCA (метод главных компонент):

- Каждый вектор $x_1, .., x_k$ центрируется: $\tilde{x}_{ij} = x_{ij} - \bar{x}_j$
- Находятся собственные значения $\lambda_1 > ... > \lambda_k > 0$ и собственные векторы $v_1, .., v_k$ матрицы $\tilde{X}^T \tilde{X}$. Собственные векторы и являются главными компонентами.
- v_1 объясняет изменчивость исходных факторов на $\frac{\lambda_1}{\sum_{i=1}^{k} \lambda_i} \times 100\%$, v_2 – на $\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{k} \lambda_i} \times 100\%$ и т.д.
- В конце оценивается регрессия y на выбранные компоненты.

5. Гребневая (Ридж) регрессия: использование вместо МНК-оценок Ridge-оценок, которые находятся путём минимизации $RSS + \lambda \sum_{i=1}^{n} \beta_i^2$
6. LASSO регрессия: использование вместо МНК–оценок Lasso-оценок, которые находятся путём минимизации $RSS + \lambda \sum_{i=1}^{n} |\beta_i|$

6.0 Прогнозирование по регрессионной модели, оцененной по перекрестным данным

- Доверительный интервал для индивидуального прогноза:

$$[\hat{y}_{n+1} - t^{cr} \sqrt{\hat{\sigma}_\varepsilon^2 (1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})}; \hat{y}_{n+1} + t^{cr} \sqrt{\hat{\sigma}_\varepsilon^2 (1 + x_{n+1}^T (X^T X)^{-1} x_{n+1})}]$$

- Доверительный интервал для среднего прогноза:

$$[\hat{y}_{n+1} - t^{cr} \sqrt{\hat{\sigma}_\varepsilon^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}}; \hat{y}_{n+1} + t^{cr} \sqrt{\hat{\sigma}_\varepsilon^2 x_{n+1}^T (X^T X)^{-1} x_{n+1}}]$$

t^{cr} рассчитывается на уровне доверия $\alpha/2$ с $n-k$ степенями свободы.

7.0 Гетероскедастичность

Гетероскедастичность – ситуация, при которой предположение о гомоскедастичности в теореме Гаусса - Маркова нарушается, т.е. дисперсии ошибок σ_i^2 не совпадают.

Гетероскедастичность не влияет на МНК-оценки коэффициентов регрессии $\hat{\beta}$, они остаются несмещенными. Однако, МНК-оценки уже не являются наилучшими в классе линейных несмещенных оценок, т.е. их дисперсии не являются минимальными.
Стоит отметить, что дисперсии у ошибок при гетероскедастичности разные и не будут иметь общую оценку $\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-k}$.
Также формулы для стандартных отклонений коэффициентов $\hat{\text{Var}}(\hat{\beta}_j) = \hat{\sigma}_\varepsilon^2 (X^T X)_{jj}^{-1}, j = 0, 1, ..., k$, не имеют места.

7.1 Тесты Голдфелда - Квандта, Глейзера, Уайта, Бройша - Пагана для диагностирования гетероскедастичности ошибок

Во всех тестах в данном разделе основная гипотеза H_0 одинакова:

$$H_0 : \sigma_i^2 = \sigma_\varepsilon^2, i = 1, ..., n,$$

т.е. дисперсии всех ошибок одинаковы, имеет место гомоскедастичность ошибок.

- **Тест Голдфелда - Квандта**

Проводится в случае, если есть подозрение, что дисперсии (или стандартные отклонения) ошибок пропорциональны некоторой переменной x_j.
Предпосылки:
– Предполагается, что есть переменная, от которой предположительно монотонно зависит условная дисперсия ошибок
– Требуется нормальнось ошибок
– Тест подходит для малых выборок

Альтернативная гипотеза для некоторого x_j выглядит следующим образом:

$$H_a : \sigma_i \sim x_{ji}, i = 1, ..., n,$$

Для проведения теста необходимо выполнить следующие шаги:

1. Оцениваем коэффициенты основной регрессии $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$.
2. Сохраняем остатки регрессии $e_i, i = 1, ..., n$. После анализа графика остатков может появиться предположение о том, что дисперсия ошибок увеличивается с ростом некоторой переменной x_j.
3. Упорядочиваем все наблюдения по модулю этой переменной.
4. Делим все наблюдения на три группы (если наблюдений достаточно много, то приблизительно на трети). Удобно, если в первой и третьей группах количество наблюдений одинаково.
5. Наблюдениями, входящими в среднюю группу, пренебрегаем, а по первым n_1 и последним n_2 наблюдениям оцениваем отдельные регрессии.
6. Гипотеза H_0 сводится к проверке гипотезы о равенстве дисперсий первых n_1 и последних n_2 наблюдений с помощью F-статистики:
$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{RSS_2/(n_2 - k)}{RSS_1/(n_1 - k)} \sim F_{n_2 - k, n_1 - k}$$
где RSS_1 и RSS_2 - суммы квадратов остатков в регрессиях, оцененным по первым n_1 и последним n_2 наблюдениям.
7. Если значение тестовой статистики F превышает F^{cr} при выбранном уровне значимости α, то гипотеза H_0 отвергается.

- **Тест Глейзера**

Зависимость дисперсии ошибок от x_j может быть необязательно линейной. В тесте Глейзера проверяются более разнообразные формы функциональной зависимости.

Альтернативная гипотеза для некоторого x_j:

$$H_a : \sigma_i \sim x_{ij}^{\gamma}, i = 1, ..., n, \gamma \in \{1; 0.5; -1\}.$$

Тест Глейзера состоит из следующих шагов (первые два шага такие же, как и в тесте Голдфелда - Квандта):

3. Оцениваем вспомогательные регрессии
$$|e| = \alpha + \beta x_j + \varepsilon, |e| = \alpha + \beta \sqrt{x_j} + \varepsilon, |e| = \alpha + \frac{\beta}{x_j} + \varepsilon.$$
4. Если коэффициент β значим хотя бы в одной из трех регрессий (значимость коэффициента проверяется с помощью t-статистики), то имеет место гетероскедастичность.

- **Тест Уайта**

С помощью теста Уайта проверяется наличие гетероскедастичности в случае более сложной зависимости, причем не от одного, а от нескольких факторов. Альтернативная гипотеза выражается как:

$$H_a : \text{имеет место гетероскедастичность.}$$

Первые два шага в тесте Уайта совпадают с двумя предыдущими тестами. Остальные представлены ниже:

3. Оцениваем вспомогательную регрессию:
$$e_i^2 = \alpha_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \sum_{j=2}^{k} \gamma_j x_{ji}^2 + \sum_{\substack{j,m=2 \\ j<m}}^{k} \delta_{jm} x_{ji} x_{mi} + \varepsilon_i, \quad i = 1, ..., n$$
и находим вспомогательный коэффициент множественной детерминации R_e^2.

4. Вычисляем тестовую статистику:
$$\chi^2 = nR_e^2 \sim \chi^2_{m-1}$$
где m – число оцениваемых во вспомогательной регрессии коэффициентов.
5. Если P-значение $< \alpha$, гипотеза H_0 о гомоскедастичности отвергается.

Эта же гипотеза (о равенстве нулю всех коэффициентов $\alpha_i, \gamma_i, \delta_{ij}$) может быть проверена с помощью F-статистики. Если значение тестовой статистики превышает критическое, то нулевая гипотеза о гомоскедастичности отвергается.

- **Тест Бройша - Пагана**

Привлекательной чертой теста Уайта является его универсальность, однако он не дает указания на функциональную форму гетероскедастичности. Если может иметь место и более специфический вид зависимости, в том числе и от не включенных в модель факторов $z_1, ..., z_p$, используют тест Бройша - Пагана. Альтернативная гипотеза для него имеет вид:
$$H_a : \sigma_i^2 \sim f(z_1, ..., z_p)$$
причем f предполагается некоторой гладкой функцией, но ее вид не конкретизируется.
Тест Бройша - Пагана состоит из следующих шагов (первые два шага совпадают с предыдущими тестами):

3. Строим оценку для дальнейшей нормировки квадратов остатков регрессии:
$$\hat{\sigma_i}^2 = \frac{\sum\limits_{i=1}^{k} e_i^2}{n}.$$
4. Оцениваем параметры регрессии
$$\frac{e_i^2}{\hat{\sigma_i}^2} = \gamma_1 + \gamma_2 z_{i2} + ... + \gamma_p z_{ip} + \varepsilon_i$$
и вычисляем для нее ESS.
5. Рассчитываем тестовую статистику:
$$\chi^2 = \frac{ESS}{2} \sim \chi^2_p$$
6. Если P-значение $< \alpha$, гипотеза H_0 о гомоскедастичности отвергается.

7.2 Оценивание параметров множественной линейной регрессии в условиях гетероскедастичности ошибок

- Изменение модели с линейной формы на логарифмическую может привести к тому, что в тестах на гетероскедастичность не будет отвергаться основная гипотеза о гомоскедастичности.
- Если дисперсии всех ошибок заранее известны, то для устранения гетероскедастичности достаточно было бы оценить исходное уравнение, поделенное почленно на стандартные отклонения ошибок
$$\frac{y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{x_{i2}}{\sigma_i} + ... + \beta_k \frac{x_{ik}}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i}, i = 1, ..., n$$
и оценить регрессию с новыми факторами
$$y_i^* = \frac{y_i}{\sigma_i}, 1_i^* = \frac{1}{\sigma_i}, x_{ij}^* = \frac{x_{ij}}{\sigma_i}, \varepsilon_i^* = \frac{\varepsilon_i}{\sigma_i}, j = 1, .., k, i = 1, ..., n$$
В уравнении $y_i^* = \beta_1 1_i^* + \beta_2 x_{i2}^* + ... + \beta_k x_{ik}^* + \varepsilon_i^*$ все ошибки имеют одну и ту же дисперсию, равную единице, поскольку $\mathrm{Var}(\varepsilon_i^*) = \mathrm{Var}(\frac{\varepsilon_i}{\sigma_i}) = \frac{1}{\sigma_i^2}\mathrm{Var}(\varepsilon_i) = 1$. Тогда МНК-оценки коэффициентов в регрессии с преобразованными данными имеют вид
$$\hat{\beta}^* = (X^{*T}X^*)^{-1}(X^{*T}Y^*)$$
Эту оценку можно выразить через исходные данные
$$\hat{\beta}^* = (X^T\Omega^{-1}X)^{-1}(X^T\Omega^{-1}Y)$$
где $\Omega = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ 0 & \dots & 0 \\ 0 & \dots & \sigma_n^2 \end{bmatrix}$

Данная оценка называется оценкой **взвешенного МНК**. Ковариационная матрица оценок имеет вид $V(\hat{\beta}) = (X^T\Omega^{-1}X)^{-1}$

- Наиболее распространенным способом коррекции гетероскедастичности в общем виде является использование оценок Уайта для дисперсий коэффициентов:
$$\hat{V}(\hat{\beta}) = \frac{1}{n}(\frac{1}{n}X^TX)^{-1}(\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i^T)(\frac{1}{n}X^TX)^{-1}$$
где x_i – i-я строка матрицы $X, i = 1, ..., n$
- Для выявления гетероскедастичности рекомендуется проверить выборку на наличие вертикальных выбросов.
- Можно также рассмотреть стандартизированные остатки в двух формах:
$$e_i^* = \frac{e_i}{\hat{\sigma}_\varepsilon\sqrt{1-h_{ii}}}$$
$$e_i^* = \frac{e_i}{\hat{\sigma}_{\varepsilon(-i)}\sqrt{1-h_{ii}}},$$
где $e = (I - H)\varepsilon$,
$H = X(X'X)^{-1}X'$ – матрица-шляпница,
$\hat{Y} = HY$,
$\hat{y}_i = \sum\limits_{i=1}^{n} h_{ii}y_i, h_{ii}$ – диагональный элемент h,
$\mathrm{Var}(e) = (I - H)\sigma_\varepsilon^2$.
Если $|e_i^*| > 2$, то это выброс.
При обнаружении выбросов можно:

1. Оценить без них;
2. Создать фиктивную переменную на выброс;
3. Применить робастное (устойчивое к выбросам) оценивание.

В чем заключается робастное оценивание?
Медианная регрессия: $RSS = \sum\limits_{i=1}^{n} |e_i| \to \min\limits_{\beta}$
R-регрессия: $\sum\limits_{i=1}^{n} \rho\{\frac{e_i}{\sigma}\} \to min$ тождественно равно $\sum\limits_{i=1}^{n} w_i e_i^2 \to min$
Функция потерь Хьюбера:
$$\rho(u) = \begin{cases} \frac{1}{2}u^2, |u| \leq c, \\ c|u| - \frac{1}{2}c^2, |u| > c \end{cases}$$
Веса корректируются на каждом шаге по формуле $u_i = \frac{1}{e_i^2}\rho(\frac{e_i}{\sigma})$. Корректировка заканчивается, когда веса почти не меняются. Берутся последние значения β.

8.0 Метод максимального правдоподобия

Метод максимального правдоподобия, в отличие от МНК, работает при более слабых предположениях, чем выполнение условий ТГМ. Если $x_1, ..., x_n$ – реализации случайных величин $X_1, ..., X_n$, то необходимо максимизировать функцию правдоподобия:
$$L(x_1, ..., x_n|\theta) = P(X_1 = x_1, ..., X_n = x_n|\theta) \to \max_\theta,$$
где θ – вектор оцениваемых параметров.
Однако на практике чаще ищут максимум логарифмической функции правдоподобия:
$$l(x_1, ..., x_n|\theta) = lnL(x_1, ..., x_n|\theta)$$
Для случая непрерывных величин $X_1, ..., X_n$ в качестве функции правдоподобия выбирается плотность их совместного распределения в точке $x_1, ..., x_n$:
$$L(x_1, ..., x_n|\theta) = f(x_1, ..., x_n|\theta)$$

8.1 Свойства ML-оценок

1. Инвариантность
Если $\hat{\theta}_{\mathrm{ML}}$ – ML-оценка параметра θ и $g(\cdot)$ – непрерывная функция, то $g(\hat{\theta}_{\mathrm{ML}})$ является ML-оценкой параметра $g(\theta)$.
2. Состоятельность
3. Асимптотическая нормальность
При $n \to \infty$ оценка вектора параметров имеет нормальное распределение:
$$\hat{\theta}_{\mathrm{ML}} \overset{as}{\sim} N(\theta; I^{-1}(\theta)),$$
где $I(\theta)$ – информационная матрица Фишера.
4. Асимптотическая эффективность

8.2 Проверка гипотез

Следующие три теста позволяют проверить гипотезу об ограничениях:
$$H_0 : \begin{cases} g_1(\theta) = 0 \\ ... \\ g_r(\theta) = 0 \end{cases} \qquad H_a : \begin{cases} g_1(\theta) \neq 0 \\ ... \\ g_r(\theta) \neq 0 \end{cases}$$

- Тест Вальда
$$W = g^T(\hat{\theta}_{ur})[\frac{\partial g}{\partial\theta}(\hat{\theta}_{ur})I^{-1}(\hat{\theta}_{ur})\frac{\partial g^T}{\partial\theta}(\hat{\theta}_{ur})]^{-1}g(\hat{\theta}_{ur}) \sim \chi^2_r$$
- Тест отношения правдоподобия (LR-тест)
$$LR = -2(l(\hat{\theta}_r) - l(\hat{\theta}_{ur})) \sim \chi^2_r$$
- Тест множителей Лагранжа (LM-тест)
$$LM = \frac{\partial l}{\partial\theta}(\hat{\theta}_r)^T I^{-1}(\hat{\theta}_r)\frac{\partial l}{\partial\theta}(\hat{\theta}_r) \sim \chi^2_r$$

Во всех случаях r – количество ограничений.

9.0 Модели бинарного выбора

В случаях, когда объясняемая переменная y – бинарная (принимает значения 0 или 1), также можно использовать линейную модель
$y_i = \beta_1 + \beta_1 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$.

Пусть $p_i = P(y_i = 1), i = 1, ..., n$. Тогда $P(y_i = 0) = 1 - p_i$ и верно, что $E(y_i) = P(y_i = 1) \times 1 + P(y_i = 0) \times 0$. При этом также верно, что $E(y_i) = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}$. Отсюда получается модель линейной вероятности:
$$P(y_i = 1) = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

Недостатки модели линейной вероятности

1. Оцененное значение вероятности может не принадлежать отрезку [0; 1].
2. Ошибки не будут распределены нормально, т.к. ε_i тоже принимает значения 0 или 1. Следовательно, привычные t-статистики для проверки гипотез использовать нельзя.
3. Дисперсии ошибок равны
$$(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik})(1 - (\beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ik})), i = 1, ..., n,$$
т.е. имеет место гетероскедастичность, МНК-оценки не будут эффективными.

9.1 Логит и пробит модели

Для исправления данных недостатков используют модели бинарного выбора:
$$P(y_i = 1) = F(z_i), \quad z_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}, \quad i = 1, ..., n,$$
где F – сигмоидная функция, принимающая значения на интервале [0;1].
Предположим, что существует $y_i^* = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i, i = 1, ..., n$, где ошибки $\varepsilon_i, ..., \varepsilon_n$ независимы и имеют одну и ту же симметричную функцию плотности $f(x) = f(-x)$, функцию распределения $F(\cdot)$, $E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma_\varepsilon^2$.

Величины y_i и y_i^* связаны следующим образом:
$$\begin{cases} y_i = 1, y_i^* \geq 0, \\ y_i = 0, y_i^* < 0. \end{cases}$$

Тогда $P(y_i = 1) = P(y_i^* \geq 0) = P(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \geq 0) =$
$= P(\varepsilon_i \geq -(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}))$.
В силу симметрии функции плотности f относительно нуля
$$P(y_i = 1) = P(\varepsilon_i \leq \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik})$$
Обозначим за $F(\cdot)$ функцию распределения нормированных ошибок $\frac{\varepsilon_i}{\sigma_\varepsilon}$.
Отсюда можно выразить формулу для моделей бинарного выбора:
$$P(y_i = 1) = F(\frac{\beta_1}{\sigma_\varepsilon} + \frac{\beta_2}{\sigma_\varepsilon}x_{i2} + ... + \frac{\beta_k}{\sigma_\varepsilon}x_{ik})$$
Наиболее употребимые в эконометрических моделях функции:

- Логит-модель
$F(z) = \frac{1}{1+e^{-z}}$ – логистическая функция
$f(z) = \frac{e^{-z}}{(1+e^{-z})^2}$
- Пробит-модель
$F(z) = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{z} e^{-t^2/2}dt$
$f(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$

9.2 Оценивание параметров моделей бинарного выбора

Для оценки параметров моделей бинарного выбора используется ММП.
Функция правдоподобия выглядит следующим образом:
$L(\beta) = \prod\limits_{y_i=1} F(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}) \prod\limits_{y_i=0}(1 - F(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik})) =$
$= \prod\limits_{i=1}^{n}(F(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}))^{y_i}(1 - F(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}))^{1-y_i}$

Логарифмическая функция правдоподобия:
$l(\beta) = lnL(\beta) = \sum\limits_{i=1}^{n}[y_i lnF(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik}) + (1 - y_i)ln(1 - F(\beta_1 +$
$+ \beta_2 x_{i2} + ... + \beta_k x_{ik}))]$

Если $F(z) = \Lambda(z) = \frac{1}{1+e^{-z}}$, то для логит-моделей можно вывести следующую формулу:
$$\sum_{i=1}^{n}[y_i - \Lambda(\beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik})]x_{ij} = 0, \quad j = 0, ..., k$$

9.3 Интерпретация результатов оценивания логит- и пробит- моделей

При использовании ММП дисперсии оценок параметров являются соответствующими диагональными элементами матрицы $I^{-1}(\theta)$. Для моделей бинарного выбора можно проверить гипотезу:
$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$
При $\hat{\theta} \overset{as}{\sim} N(\theta; I^{-1}(\theta))$ тестовая статистика рассчитывается как:
$$z = \frac{\hat{\beta}_j}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j)}}$$

Если $|z| \geq z_{\alpha/2}$, где α - выбранный уровень значимости, то гипотеза H_0 о незначимости отвергается. Однако интерпретировать можно только знак оценки коэффициента (если $\hat{\beta}_j > 0$, то при увеличении x_j вероятность того, что $y_i = 1$ увеличивается, и наоборот), а не его абсолютное значение.

9.4 Предельные эффекты

Функция $F(x)$ является нелинейной, поэтому для интерпретации влияния каждого фактора рассчитываются предельные эффекты (частные производные):

$$\frac{\partial \hat{p}_i}{\partial x_j}(x_1, ..., x_k) = f(x_1, ..., x_k)\hat{\beta}_j, \quad i = 1, ..., n, j = 1, ..., k$$

Чаще всего в качестве точки, в которой рассчитывается предельный эффект, выбирается $\bar{x}_1, ..., \bar{x}_k$.

Для фиктивных переменных предельные эффекты рассчитываются по формуле:

$$\frac{\partial \hat{p}_i}{\partial x_j}(x_1, ..., x_k) = \hat{p}(x_1, ..., x_{j-1}, D = 1, x_{j+1}, ..., x_k) -$$
$$-\hat{p}(x_1, ..., x_{j-1}, D = 0, x_{j+1}, ..., x_k)$$

9.5 Показатели качества оценки моделей бинарного выбора

Ниже приведены несколько популярных метрик для оценки моделей бинарного выбора.

- R^2 МакФаддена:
$$R^2_{MF} = 1 - \frac{\hat{l}_{ML}}{l_0},$$
где $\hat{l}$ – значение логарифмической функции правдоподобия в точке максимума, а l_0 – максимум логарифмической функции правдоподобия для модели, в которую включена только константа.
- Псевдо-R^2:
$$R^2_p = \frac{1}{1 + \frac{2}{n}(\hat{l}_{ML} - l_0)}$$

Альтернативным критерием оценки качества модели является *сравнение точности прогнозирования* по оцененной модели бинарного выбора. Прогноз строится следующим образом:

$$\hat{y}_i = \begin{cases} 1, \text{если } F(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik}) > 0.5 \\ 0, \text{если } F(\hat{\beta}_1 + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik}) \leq 0.5 \end{cases}$$

Доля неверных прогнозов вычисляется по формуле:

$$W_1 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Данная доля интересна в сравнении с долей неверных прогнозов в самой простой бинарной модели, в которую включена только константа. Прогнозы для простой модели строятся так: $\hat{y}_i = 1$, если в выборке доля единиц превышает 0.5. Если больше нулей, то $\hat{y}_i = 0$.
Доля неверных прогнозов для простой модели:

$$W_0 = \begin{cases} 1 - \hat{p}, \text{если } \hat{p} > 0.5 \\ \hat{p}, \text{если } \hat{p} \leq 0.5 \end{cases}$$

Показатель качества подгонки модели:

$$R^2_{fit} = 1 - \frac{W_1}{W_0}$$

10.0 Эндогенность

Эндогенность – это коррелированность регрессоров и случайных ошибок:
$$\text{Cov}(X, \varepsilon) \neq 0$$

Утверждение 10.1 Если регрессоры не коррелирует с ошибкой ε, т.е. $\underset{n\to\infty}{\text{plim}}(\frac{1}{n}X^T\varepsilon) = 0$, то ТГМ не нарушается:

$$\underset{n\to\infty}{\text{plim}}\,\hat{\beta} = \beta + \underset{n\to\infty}{\text{plim}}(\frac{1}{n}X^TX)^{-1}\underset{n\to\infty}{\text{plim}}(\frac{1}{n}X^T\varepsilon) = \beta$$

При этом существование предела $\underset{n\to\infty}{\text{plim}}(\frac{1}{n}X^TX)^{-1}$ является дополнительным требованием.

Утверждение 10.2 Если существует такой регрессор x_j, что $\underset{n\to\infty}{\text{plim}}(\frac{1}{n}x_j^T\varepsilon) \neq 0$, то имеет место проблема эндогенности, и оценки МНК не будут состоятельны.

10.1 Причины возникновения эндогенности

1) Пропуск существенной переменной;
2) Ошибки измерения;
3) Одновременность изменения зависимой и эндогенной объясняющей переменной;
4) Проблема самоотбора;

10.2 Способы борьбы с эндогенностью

10.2.1 Инструментальные переменные

Утверждение 10.3 Переменная z_i является инструментальной для проблемного регрессора x_i если:

$$\text{Corr}(x_i, z_i) > 0 - \textit{свойство релевантности}$$

$$\text{Corr}(z_i, \varepsilon_i) = 0 - \textit{свойство валидности}$$

При этом второе условие можно заменить более слабым:

$$\underset{n\to\infty}{\text{plim}}(\frac{1}{n}z_i^T\varepsilon) = 0$$

Метод инструментальных переменных (IV)

Когда используется? При возникновении проблемы эндогенности.

- Для модели парной регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ с эндогенной переменной x и инструментом z:
$$\hat{\beta}_2^{IV} = \frac{\text{sCov}(z, y)}{\text{sCov}(z, x)}$$
- Для модели множественной регрессии $y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$ с рядом эндогенных переменных и инструментами $z_1, ..., z_m$:

Случай 1: **m=k** (количество инструментов совпадает с количеством переменных в исходной модели)

$$\hat{\beta}^{IV} = (Z^TX)^{-1}Z^Ty$$

Эта оценка несмещенная и состоятельная, а также отвечает свойствам релевантности и валидности.

Случай 2: **m>k** (количество инструментов превышает количество переменных в исходной модели)

В этом случае необходимо применить двухшаговый МНК (2SLS):

Шаг 1. Построить регрессию всех регрессоров, коррелированных с ошибками, на инструментальные переменные:

$$X = Z\gamma + \varepsilon$$

Получить $\hat{X} = Z(Z^TZ)^{-1}Z^TX$.
Шаг 2. Оценить исходную модель, подставив полученный $\hat{X}$:

$$y = \hat{X}\beta + \varepsilon$$

Тогда при $\hat{X} = Z(Z^TZ)^{-1}Z^TX$ оценка методом инструментальных переменных будет следующая:

$$\hat{\beta}^{2SLS} = (\hat{X}^T\hat{X})^{-1}\hat{X}^Ty = (X^TZ(Z^TZ)^{-1}Z^TX)^{-1}X^TZ(Z^TZ)^{-1}Z^Ty$$

10.2.2 Обобщенный метод моментов (GMM)

Если имеется m инструментов $z_1, ..., z_m$, то можно определить функцию $g_i(\beta) = z_i(y_i - x_i^T\beta)$, которая называется моментом.

Если инструменты экзогенны, должно быть выполнено моментное тождество $E(g_i(\beta)) = 0$

Выборочный аналог моментного тождества будет иметь вид:

$$\bar{g}(\beta) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i^T\beta)$$

Применение GMM для получения оценок линейной регрессионной модели.

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$
$$(z_1 \ldots z_m) - \text{инструменты}$$

Случай 1: **m=k** (Число моментных тождеств, которое соответствует числу инструментов, совпадает с количеством регрессоров)
Оценка вектора коэффициентов $\hat{\beta}$ будет решением уравнения $\bar{g}(\beta) = 0$:

$$\hat{\beta}^{IV} = \hat{\beta}^{GMM}$$

Случай 2: **m>k**

$$Q(\beta) = (\frac{1}{n}Z^T(y - X\beta))^T W(\frac{1}{n}Z^T(y - X\beta)) \to \min_\beta$$

$$\hat{\beta}^{GMM} = (X^TZWZ^TX)^{-1}X^TZWZ^Ty$$

где W – некая весовая матрица. Наиболее эффективными являются GMM-оценки с весовой матрицей, равной обратной ковариационной матрице моментных функций.

10.3 Проверка качества инструментов

При использовании инструментальных переменных необходимо убедиться в том, что они удовлетворяют двум основным условиям:

1) Условие релевантности (z коррелирует с x)
Это условие проверяется следующим образом:
Шаг 1. Оценивается регрессия

$$x_i^* = z_i^T\beta + \varepsilon$$

где x_i^* - вектор эндогенных переменных.
Шаг 2. Рассчитывается статистика F для проверки гипотезы об адекватности регрессии. Если её значение выше 10, то инструменты приемлемы.

2) Условие валидности (z не коррелирует с ошибками)
При **m=k** проверить валидность невозможно.
Однако при **m>k** возможны два варианта:

Случай 1: При гомоскедастичности.
При гомоскедастичности используется **Тест Саргана**:

$$H_0 : \text{инструменты валидны}$$

$$H_a : \text{инструменты не валидны}$$

В тесте используется статистика вида:

$$\zeta(\hat{\beta}^{GMM}) = \frac{1}{\hat{\sigma}^2_\varepsilon}\hat{\varepsilon}^TZ(Z^TZ)^{-1}Z^T\hat{\varepsilon} \sim \chi^2_{m-k}$$

Если рассчитанное значение превышает критическое при выбранном уровне значимости, то гипотеза H_0 отвергается, инструменты слабые.

Случай 2: При гетероскедастичности.
При гетероскедастичности используется **Тест Хансена**:

$$H_0 : \text{инструменты валидны}$$

$$H_a : \text{инструменты не валидны}$$

В тесте используется статистика вида:

$$\zeta(\hat{\beta}^{GMM}) = \hat{\varepsilon}^TZ(Z^T\hat{\Omega}Z)^{-1}Z^T\hat{\varepsilon} \sim \chi^2_{m-k}$$

Если рассчитанное значение превышает критическое при выбранном уровне значимости, то гипотеза H_0 отвергается, инструменты слабые.

10.4 Тест Хаусмана

Для чего используется? Чтобы определить, имеет ли место проблема эндогенности в модели

$$H_0 : \text{нет эндогенности; } \hat{\beta}_{OLS} \text{ и } \hat{\beta}_{IV} - \text{состоятельные}$$

$$H_a : \text{есть эндогенность; } \hat{\beta}_{IV} - \text{состоятельна, } \hat{\beta}_{OLS} - \text{несостоятельна}$$

Пусть $\hat{q} = \hat{\beta}_{IV} - \hat{\beta}_{OLS}$. Тогда в тесте используется статистика вида:

$$\chi^2 = \hat{q}^T(\hat{\text{Var}}(\hat{\beta}_{IV}) - \hat{\text{Var}}(\hat{\beta}_{OLS}))^{-1}\hat{q} \sim \chi^2_{k+1}$$

Если рассчитанное значение статистики превышает критическое при выбранном уровне значимости, то гипотеза H_0 отвергается, и нужно использовать инструментальные переменные.

Также можно провести процедуру **Дарбина-Ву-Хаусмана**:
Пусть $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$; z_i - инструмент.

$$H_0 : \text{нет эндогенности}$$

$$H_a : \text{есть эндогенность}$$

Проверка гипотезы осуществляется следующим образом:
Шаг 1. Строим вспомогательную регрессию $x_i = \alpha_1 + \alpha_2 z_i + \nu_i$, получаем $\hat{\nu}_i$.
Шаг 2. Строим регрессию $y_i = \beta_1 + \beta_2 x_i + \delta\hat{\nu}_i + \varepsilon_i$.
Если коэффициент δ значим, то гипотеза H_0 отвергается.

11.0 Автокорреляция

Автокорреляция – это нарушение условия теоремы Гаусса-Маркова о некоррелируемости ошибок регрессии:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$$

Автокорреляция 1-го порядка: $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, где $u_t \sim$ i.i.d.$(0, \sigma_u^2)$
При этом переменная ρ – коэффициент автокорреляции, $|\rho| < 1$.
Ошибка текущего момента может зависеть не только от ошибки предыдущего периода, но и от ошибок предшествовавших периодов, поэтому могут существовать автокорреляции более высоких порядков:

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \ldots + \rho_p\varepsilon_{t-p} + u_t$$

Утверждение 11.1 Если $\text{Cov}(u_i, u_j) = 0, \ i \neq j$, то:

$$\sigma^2_\varepsilon = \text{Var}(\varepsilon_t) = \sigma^2_u(1 + \rho^2 + \rho^4 + \ldots) = \frac{\sigma^2_u}{1 - \rho^2}, \ |\rho| < 1$$

$$\text{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \rho\sigma^2_u(1 + \rho^2 + \rho^4 + \ldots) = \rho\frac{\sigma^2_u}{1 - \rho^2} = \rho\sigma^2_\varepsilon$$
$$\text{Cov}(\varepsilon_t, \varepsilon_{t-s}) = \rho^s\sigma^2_\varepsilon$$

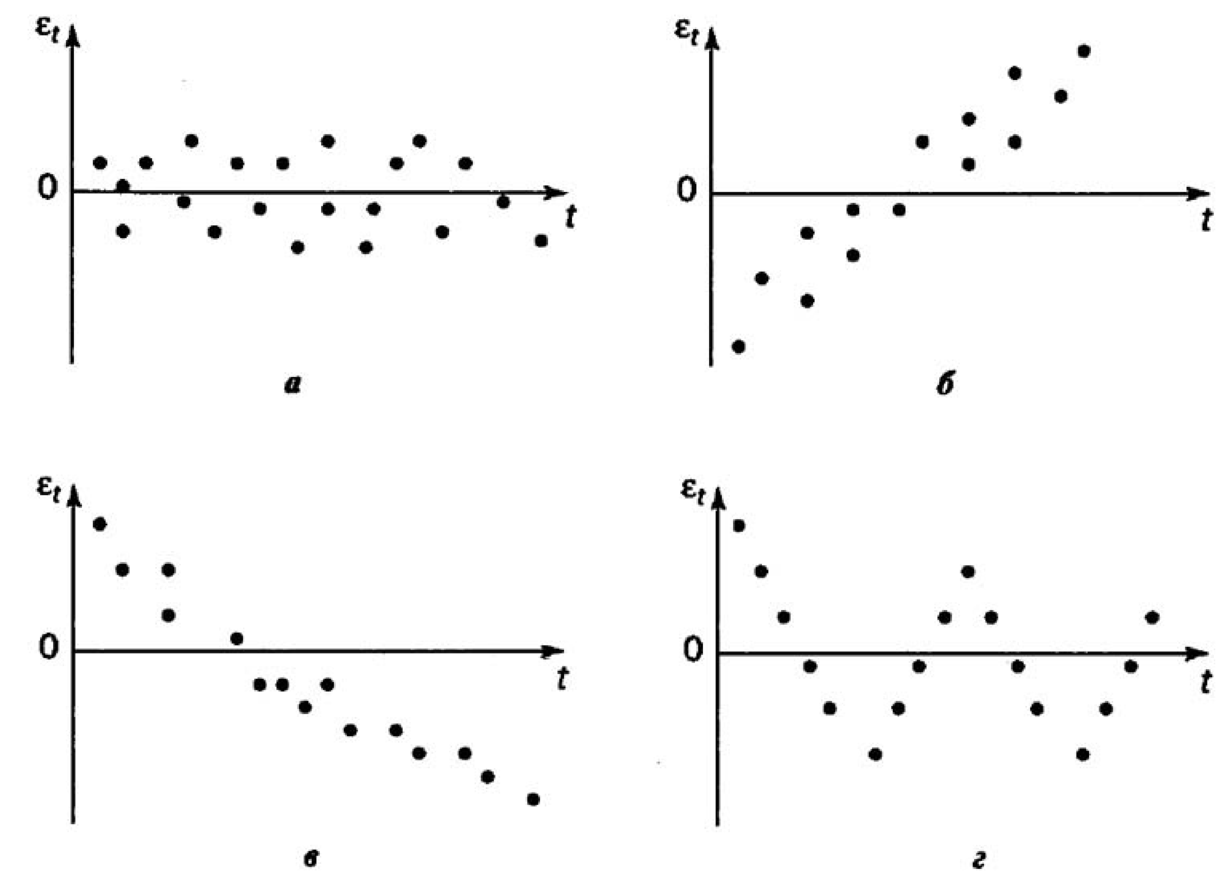
При автокорреляции оценки МНК будут несмещенными, но неэффективными. Стандартные ошибки коэффициентов будут недооценены, t-статистики будут слишком большие.

11.1 Причины возникновения автокорреляции

1) Пропуск существенной переменной;
2) Инертность макроэкономических данных;
3) Сглаживание первоначальных данных;
4) Неправильная форма модели;

11.2 Способы обнаружения автокорреляции

Первый способ: Графический.



На картинке (а) изображен график нормальных остатков по времени. На всех остальных изображены графики остатков с различными видами автокорреляции.

Второй способ: Проведение тестов на выявление автокорреляции.

• **Тест серий**.

Для проведения теста выписываются знаки остатков по очереди, при этом *серия* – набор остатков одного знака.
Дано: T – общее число наблюдений, N_1 – число положительных остатков, N_2 – число отрицательных остатков, K – число серий.

$$H_0: \text{нет автокорреляции}$$

$$H_a: \text{положительная автокорреляция/отрицательная автокорреляция}$$

Если $K \leq K_{min}$ – автокорреляция положительная;
Если $K \geq K_{max}$ – автокорреляция отрицательная.
При этом $K_{min} = \min\{N_1; N_2\} - 1$, $K_{max} = \max\{N_1, N_2\} + 1$.
Если $T > 40$, справедливо следующее:

$$K \overset{as}{\sim} N\left(\frac{2N_1N_2}{N_1+N_2} + 1; \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1+N_2)^2(N_1+N_2-1)}\right)$$

• **Тест Дарбина - Уотсона**.

В этом тесте используется особая статистика, которая так и называется – статистика Дарбина - Уотсона (DW-статистика). Рассчитывается она следующим образом:

$$DW = \frac{\sum\limits_{t=2}^{T}(e_t - e_{t-1})^2}{\sum\limits_{t=1}^{T} e_t^2} \approx 2 - 2\rho,\ |\rho| < 1 \Rightarrow 0 < 2 - 2\rho < 4$$

Тестирование на автокорреляцию с помощью DW-статистики проводится по следующей гипотезе:

$$H_0: \rho = 0, \text{ т.е. отсутствие автокорреляции}$$

$$H_a: \rho > 0 \text{ ЛИБО } \rho < 0$$

Для проведения расчетов используются две вспомогательные переменные d_u и d_L, значения которых заданы таблично. При этом: $0 < d_L < d_u < 2$.

При $H_a: \rho > 0$ (положительная автокорреляция):
1)$DW < d_L \Rightarrow$ Присутствует положительная автокорреляция.
2)$d_L \leq DW \leq d_U \Rightarrow$ Определить наличие автокорреляции невозможно.
3)$DW > d_U \Rightarrow$ Положительная автокорреляция отсутствует.

При $H_a: \rho < 0$ (отрицательная автокорреляция:
1)$DW > (4 - d_L) \Rightarrow$ Присутствует отрицательная автокорреляция.
2)$(4 - d_U) \leq DW \leq (4 - d_L) \Rightarrow$ Определить наличие автокорреляции невозможно.
3)$DW < (4 - d_U) \Rightarrow$ Отрицательная автокорреляция отсутствует.

• **Тест Бройша - Годфри**.

В отличие от предыдущих тестов, этот тест используется для выявления автокорреляции более 1-го порядка.

Пусть $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$, при этом $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \ldots + \rho_p\varepsilon_{t-p} + u_t$
Гипотеза выглядит следующим образом:

$$H_0: \rho_1 = \ldots = \rho_p = 0 \text{ (автокорреляции в остатках нет)}$$

$$H_a: \exists \rho_j \neq 0,\ j = 1\ldots T \text{ (автокорреляция есть)}$$

Шаг 1. Оценивается начальная модель и рассчитывается ряд остатков.
Шаг 2. Оценивается дополнительная модель:

$$e_t = \beta_1 + \beta_2 x_t + r_1 e_{t-1} + r_2 e_{t-2} + \ldots + r_p e_{t-p} + u_t$$

Шаг 3. Для дополнительной модели рассчитывается R_e^2, а затем статистика вида:

$$\chi^2 = T \times R_e^2 \sim \chi_p^2$$

Если значение статистики превышает критическое при выбранном уровне значимости, основная гипотеза об отсутствии автокорреляции отвергается.

11.3 Что делать, если обнаружилась автокорреляция?

• **Переход к взвешенным разностям.**
Пусть модель выглядит как $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$.
Шаг 1. Необходимо получить оценку параметра ρ. Например, можно оценить его из статистики Дарбина-Уотсона по формуле $\hat{\rho} = 1 - \frac{d}{2}$.
Шаг 2. Далее исходное уравнение сдвигается на шаг назад и домножается на ρ:

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{t-1} + \varepsilon_{t-1}$$

Шаг 3. Вычесть из исходного уравнения преобразованное. Получаем:

$$y_t - \rho y_{t-1} = \beta_1(1-\rho) + \beta_2(x_t - \rho x_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1}$$

Теперь автокорреляция отсутствует, но теряется одно наблюдение.

• **Поправка Прайса - Уинстона.**
Заменяем первые наблюдения:

$$y_1^* = y_1\sqrt{1-\rho^2}, \quad x_1^* = x_1\sqrt{1-\rho^2}$$

12.0 Временные ряды

Временной ряд – ряд некоторых числовых величин, измеренных в последовательные моменты времени.

$$\{y_t\}_{-\infty}^{+\infty},\ t = 1\ldots T$$

Ключевым отличием временных рядов от пространственных выборок является то, что наблюдения строго упорядочены по времени, невозможно произвольно менять их местами, а также выкидывать из выборки.

12.1 Стационарные процессы

Утверждение 12.1 $\{y_t\}$ называется стационарным в узком смысле, если распределение $y_{t1}, y_{t2}, y_{t3}, \ldots, y_{tk}$ совпадает с распределением $y_{t1+s}, y_{t2+s}, y_{t3+s}, \ldots, y_{tk+s}$ $\forall t_1 \ldots t_{k+s},\ s > 0, k > 0$

Утверждение 12.2 $\{y_t\}$ называется стационарным в широком смысле, если:

$$\mathrm{E}(y_t) = \mu$$

$$\mathrm{Var}(y_t) = \sigma^2$$

$$\mathrm{Cov}(y_t, y_{t-k}) = \mathrm{Cov}(y_{t+s}, y_{t-k+s}) = \gamma(k)$$

Стационарный временной ряд – это такой ряд, который не меняется со временем с каким-либо трендом, то есть не растет и не снижается. Формально это означает, что математическое ожидание стационарного временного ряда – постоянно, равно как и его дисперсия, а его значения на равном расстоянии друг от друга связаны одинаково.

12.2 Процессы AR, MA, ARMA и ARIMA

Определение 12.0 Белый шум $y_t \sim WN(0, \sigma^2)$ – слабо стационарный процесс, обладающий следующими свойствами:

$$E(y_t) = E(y_s) = 0$$

$$\mathrm{Var}(y_t) = Var(y_s) = \sigma^2$$

$$\mathrm{Cov}(y_t, y_s) = 0$$

Определение 12.1 $AR(p)$ – процесс, зависящий только от своих предыдущих значений и ошибки в текущий момент времени:

$$y_t = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t = c + \sum_{i=1}^{p}\alpha_i y_{t-i} + \varepsilon_t$$

$$t = 1, \ldots, T; \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Определение 12.2 $MA(q)$ – процесс, зависящий только от лагов случайной ошибки:

$$y_t = c + \varepsilon_t + \beta_1\varepsilon_{t-1} + \ldots + \beta_q\varepsilon_{t-q} = c + \sum_{i=1}^{q}\beta_i\varepsilon_{t-i} + \varepsilon_t$$

$$t = 1, \ldots, T; \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Определение 12.3 $ARMA(p, q)$ – процесс, зависящий и от своих предыдущих значений, и от текущей и предыдущих ошибок:

$$y_t = c + \sum_{i=1}^{p}\alpha_i y_{t-i} + \sum_{j=1}^{q}\beta_j\varepsilon_{t-j} + \varepsilon_t$$

$$t = 1, \ldots, T; \quad \varepsilon_t \sim WN(0, \sigma^2)$$

Как выбрать оптимальное число лагов? Можно использовать информационные критерии Акаике (AIC) и Шварца (BIC):

$$AIC(p, q) = ln\hat{\sigma}^2 + 2\frac{p+q+1}{T} \longrightarrow \min_{p,q}$$

$$BIC(p, q) = ln\hat{\sigma}^2 + \frac{p+q+1}{T}lnT \longrightarrow \min_{p,q}$$

Также можно провести тесты Льюнг - Бокса и Бокса - Пирса.

Оба проверяют следующую гипотезу:

$$H_0: \text{процесс - } WN(0, \sigma^2)$$

$$H_a: \text{процесс - } ARMA(p, q)$$

1. Q-тест Льюнг - Бокса. Статистика рассчитывается как:

$$Q_k = T(T+2)\sum_{k=1}^{K}\frac{1}{T-K}r_k^2 \sim \chi^2_{K-p-q-1}$$

2. Q-статистика Бокса - Пирса. Статистика рассчитывается как:

$$Q_k = T\sum_{k=1}^{K} r_k^2 \sim \chi^2_{K-p-q-1}$$

В обоих случаях, если при выбранном уровне значимости величина статистики превышает критическое значение, основная гипотеза отвергается.

Определение 12.4 Если для некоторого ряда y_t необходимо взять d разностей, чтобы привести его к стационарному виду, а ряд Δy_t описывается моделью $ARMA(p, q)$, то ряд y_t описывается моделью $ARIMA(p, d, q)$.

12.3 Тест Дики - Фуллера на стационарность ряда

Для начала необходимо привести исходный ряд к виду
$\Delta y_t = \delta + \gamma t + (\theta - 1)y_{t-1} + \varepsilon_t$, где δ – константа, а γt отвечает за тренд.
Гипотеза формулируется как:

$$H_0: \theta - 1 = 0 \text{ - ряд нестационарный}$$

$$H_a: \theta - 1 < 0 \text{ - ряд стационарный}$$

При этом предполагается, что $\varepsilon_i \sim N(0, \sigma^2)$.
Тестовая статистика (DF-статистика) рассчитывается по формуле, аналогичной с формулой t-статистики на значимость коэффициента:

$$DF = \frac{\hat{\theta} - 1}{\sqrt{\hat{\mathrm{Var}}(\hat{\theta} - 1)}}$$

У статистики Дики-Фуллера своя таблица критических значений и свое распределение, если тестовая статистика находится правее критического значения, то нулевая гипотеза отвергается. При этом критическое значение рассчитывается исходя из спецификации теста:
1. Без константы;
2. С константой;
3. С константой и трендом.

12.4 Ложная корреляция и коинтеграция

Ложная корреляция – эффект, при котором между двумя переменными наблюдается значимая корреляция, при этом отсутствует качественная причинно-следственная связь. Может возникать из-за того, что обе переменные являются нестационарными, и в них обеих наблюдается стохастический тренд.

Коинтеграция – свойство нескольких нестационарных временных рядов, заключающееся в существовании некоторой их стационарной линейной комбинации.

Говорят, что нестационарный ряд является интегрированным порядка d, если необходимо взять d разностей, чтобы его к стационарному виду. Такие ряды означаются как $I(d)$.

Рассмотрим два нестационарных ряда: $y_t \sim I(d)$ и $x_t \sim I(d)$,где $I(d)$. Если существует такой вектор (α, β) : $\alpha \neq 0,\ \beta \neq 0$, что $\alpha y_t + \beta x_t \sim I(d-b), b > 0$, то ряды называются *коинтегрированными* порядка b.