равносильно следующей системе уравнений

$$X^T X\beta = X^T Y,$$

где

$$X = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix},\ Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \text{ и } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

(b) Пусть $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ есть решение задачи из пункта (a).

В случае если матрица $X^T X$ не вырождена, убедитесь в справедливости формулы $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Указание. При решении пункта (a) используйте, что

$$X^T X = \begin{pmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{pmatrix} \text{ и } X^T Y = \begin{pmatrix} \sum_{i=1}^n x_{i1}Y_i \\ \sum_{i=1}^n x_{i2}Y_i \end{pmatrix}.$$

Задача 9. Покажите, что для моделей $Y_i = \alpha + \beta x_i + \varepsilon_i$, $Z_i = \gamma + \delta x_i + \nu_i$ и $Y_i + Z_i = \mu + \lambda x_i + \xi_i$ МНК-оценки связаны соотношениями $\hat{\mu} = \hat{\alpha} + \hat{\gamma}$ и $\hat{\lambda} = \hat{\beta} + \hat{\delta}$.

Задача 10. Найдите МНК-оценки параметров α и β в модели $Y_i = \alpha + \beta Y_i + \varepsilon_i$.

Ответ: $\hat{\alpha} = 0$, $\hat{\beta} = 1$.

Задача 11. Рассмотрите модели $Y_i = \alpha + \beta(Y_i + Z_i) + \varepsilon_i$, $Z_i = \gamma + \delta(Y_i + Z_i) + \varepsilon_i$ и покажите, что $\hat{\alpha} + \hat{\gamma} = 0$ и $\hat{\beta} + \hat{\delta} = 1$.

Указание. Рассмотрите регрессию

$$Y_i + Z_i = \mu + \lambda(Y_i + Z_i) + \varepsilon_i$$

и воспользуйтесь результатами задач 9 и 10.

Задача 12. Как связаны МНК-оценки параметров α, β и γ, δ в моделях $Y_i = \alpha + \beta x_i + \varepsilon_i$ и $Z_i = \gamma + \delta x_i + \nu_i$, если $Z_i = 2Y_i$.

Ответ: $\hat{\gamma} = 2\hat{\alpha}$, $\hat{\delta} = 2\hat{\beta}$.

Задача 13. Для модели линейной регрессии известны $Y^T = [1 \quad 2 \quad 3 \quad 4 \quad 5]$ и $\hat{Y}^T = [2 \quad 2 \quad 2 \quad 4 \quad 5]$. Найдите R^2.

Ответ: $R^2 = \widetilde{\text{corr}}^2(Y, \hat{Y}) = 0.8$, где через $\widetilde{\text{corr}}(Y, \hat{Y})$ обозначен выборочный коэффициент корреляции Y и $\hat{Y}$.

Задача 14*. Пусть для модели $Y_i = \alpha_1 + \varepsilon_i$, $i = 1, \ldots, m$, $RSS_1 = q_1$, а для модели $Y_i = \alpha_2 + \varepsilon_i$, $i = m+1, \ldots, n$, $RSS_2 = q_2$. Чему равен RSS в модели $Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, \ldots, n$, где

$$x_{i1} = \begin{cases} 1 & \text{при } i \in \{1, \ldots, m\} \\ 0 & \text{при } i \in \{m+1, \ldots, n\} \end{cases} \text{ и } x_{i2} = \begin{cases} 0 & \text{при } i \in \{1, \ldots, m\} \\ 1 & \text{при } i \in \{m+1, \ldots, n\} \end{cases}?$$

(e) $\hat{Y} = X\hat{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 4 \\ 5 \end{bmatrix}$,

$$\hat{\varepsilon} = Y - \hat{Y} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \\ 2 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\varepsilon}_5 = 0.$$

(f) $RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = (-1)^2 + 0^2 + 1^2 + 0^2 + 0^2 = 2.$

(g) $R^2 = 1 - \dfrac{RSS}{TSS} = 1 - \dfrac{2}{10} = 0.8$. Качество регрессии — хорошее. □

Задача 17. Пусть регрессионная модель

$$Y_i = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \varepsilon_i, \; i = 1, \ldots, n$$

задана в матричном виде при помощи уравнения $Y = X\beta + \varepsilon$, где $\beta = \begin{bmatrix} \alpha & \beta_1 & \beta_2 \end{bmatrix}^T$. ~~Известно, что $\mathbb{E}(\varepsilon) = \mathbf{0}$ и $\mathrm{V}(\varepsilon) = \sigma^2 \cdot I$.~~

Имеются следующие наблюдения:

$$Y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}; \; X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}.$$

Для удобства расчетов ниже приведены матрицы

$$X^T X = \begin{bmatrix} 5 & 3 & 1 \\ 3 & 3 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ и } \left(X^T X\right)^{-1} = \begin{bmatrix} 0.5 & -0.5 & 0.0 \\ -0.5 & 1.0 & -0.5 \\ 0.0 & -0.5 & 1.5 \end{bmatrix}.$$

(a) Укажите число наблюдений.

(b) Укажите число регрессоров в модели (с учетом свободного члена).

(c) Рассчитайте $TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$.

(d) Рассчитайте при помощи метода наименьших квадратов оценку для вектора неизвестных коэффициентов.

(e) Чему равен $\hat{\varepsilon}_5$ — МНК-остаток регрессии, который соответствует 5-му наблюдению?

(f) Найдите $RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$.

(g) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оцененного уравнения регрессии.

Ответы:

(a) $n = 5$;

(b) $k + 1 = 3$;

(c) $TSS = 10$;

(d) $\hat{\beta} = \begin{bmatrix} 3/2 & 2 & 3/2 \end{bmatrix}^T$;

(e) $\hat{\varepsilon}_5 = 0$;

(f) $RSS = 1$;

(g) $R^2 = 0.9$. Качество регрессии — хорошее.

Задача 18. Пусть регрессионная модель

$$Y_i = \alpha + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \varepsilon_i,\ i = 1,\ldots,n$$

задана в матричном виде при помощи уравнения $Y = X\beta + \varepsilon$, где $\beta = [\alpha \quad \beta_1 \quad \beta_2]^T$. Известно, что $\mathbb{E}(\varepsilon) = \mathbf{0}$ и $\mathrm{V}(\varepsilon) = \sigma^2 \cdot I$.

Имеются следующие наблюдения:

$$Y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix};\ X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}.$$

Для удобства расчетов ниже приведены матрицы

$$X^T X = \begin{bmatrix} 5 & 3 & 1 \\ 3 & 3 & 1 \\ 1 & 1 & 1 \end{bmatrix} \text{ и } \left(X^T X\right)^{-1} = \begin{bmatrix} 0.5 & -0.5 & 0.0 \\ -0.5 & 1.0 & -0.5 \\ 0.0 & -0.5 & 1.5 \end{bmatrix}.$$

(a) Укажите число наблюдений.

(b) Укажите число регрессоров в модели (с учетом свободного члена).

(c) Рассчитайте $TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$.

(d) Рассчитайте при помощи метода наименьших квадратов оценку для вектора неизвестных коэффициентов.

(e) Чему равен $\hat{\varepsilon}_5$ — МНК-остаток регрессии, который соответствует 5-му наблюдению?

(f) Найдите $RSS = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$.

Указание: $RSS/\sigma^2 \sim \chi^2(n-k-1)$.

Ответы: (a) $\mathbb{E}[RSS] = 20\sigma^2$, (b) $\mathrm{D}(RSS) = 40\sigma^4$,

(c) $\mathbb{P}\{RSS > 10\sigma^2\} \approx 0.9682$,

(d) $\mathbb{P}\{10\sigma^2 < RSS < 30\sigma^2\} \approx 0.8983$.

Задача 30. Рассматривается модель регрессии

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

в которой ошибки $\varepsilon_1, \ldots, \varepsilon_n$ независимы и имеют нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2. Для $n = 13$ найдите

(a) $\mathbb{E}[RSS]$,

(b) $\mathrm{D}(RSS)$,

(c) $\mathbb{P}\{RSS > 10\sigma^2\}$,

(d) $\mathbb{P}\{5\sigma^2 < RSS < 10\sigma^2\}$.

Указание: $RSS/\sigma^2 \sim \chi^2(n-k-1)$.

Ответы: (a) $\mathbb{E}[RSS] = 10\sigma^2$, (b) $\mathrm{D}(RSS) = 20\sigma^4$,

(c) $\mathbb{P}\{RSS > 10\sigma^2\} \approx 0.4404$,

(d) $\mathbb{P}\{5\sigma^2 < RSS < 10\sigma^2\} \approx 0.4507$.

Задача 31. Рассматривается модель регрессии

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

в которой ошибки $\varepsilon_1, \ldots, \varepsilon_n$ независимы и имеют нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2. Для $n = 12$ найдите

(a) $\mathbb{P}\{\hat\alpha > \alpha\}$,

(b) $\mathbb{P}\{\alpha > 0\}$,

(c) $\mathbb{P}\left\{|\hat\alpha - \alpha| < \sqrt{\hat{\mathrm{D}}(\hat\alpha)}\right\}$,

(d) $\mathbb{P}\left\{\hat\beta > \beta + \sqrt{\hat{\mathrm{D}}(\hat\beta)}\right\}$,

(e) $\mathbb{P}\left\{\hat\beta < \beta - \sqrt{\hat{\mathrm{D}}(\hat\beta)}\right\}$,

(f) $\mathbb{E}\left[\dfrac{\hat\alpha - \alpha}{\sqrt{\hat{\mathrm{D}}(\hat\alpha)}}\right]$,

(g) $\mathbb{E}\left[\dfrac{\hat\alpha + \hat\beta - (\alpha+\beta)}{\sqrt{\hat{\mathrm{D}}(\hat\alpha+\hat\beta)}}\right]$,

(h) $\mathrm{D}\left[\dfrac{\hat\alpha - \alpha}{\sqrt{\hat{\mathrm{D}}(\hat\alpha)}}\right]$,

(i) $\mathrm{D}\left[\dfrac{\hat\alpha + \hat\beta - (\alpha+\beta)}{\sqrt{\hat{\mathrm{D}}(\hat\alpha+\hat\beta)}}\right]$,

(j) $\mathbb{P}\{\hat\sigma > \sigma\}$,

(k) $\mathbb{P}\{\hat\sigma < \sigma\}$.

1.4. Теорема Гаусса—Маркова

Пусть модель линейной регрессии задана в матричной форме

$$Y = X\beta + \varepsilon, \tag{1}$$

где

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Определение. Оценка $\tilde\beta = \varphi(X, Y)$ называется несмещенной оценкой для неизвестного вектора параметров β, если $\mathbb{E}\tilde\beta = \beta$ при всех $\beta \in \mathbb{R}^{k\times 1}$.

Определение. Оценка $\tilde{\beta}=\varphi(X,Y)$ называется линейной по переменной Y, если функция φ линейно-однородна по переменной Y, т. е. если

1. $\varphi(X,\lambda\cdot Y)=\lambda\cdot\varphi(X,Y)$, $\forall\lambda\in\mathbb{R}$ (свойство однородности);
2. $\varphi(X,Y+Z)=\varphi(X,Y)+\varphi(X,Z)$, $\forall Y,Z\in\mathbb{R}^{n\times 1}$ (свойство аддитивности по второму аргументу).

Задача 32. Докажите, что МНК-оценки $\hat{\beta}=(X^TX)^{-1}X^TY$ являются несмещенными и линейными по переменной Y.

Решение. Докажем несмещенность МНК-оценок.

$$\mathbb{E}\hat{\beta}=\mathbb{E}\left((X^TX)^{-1}X^TY\right)=(X^TX)^{-1}X^T\mathbb{E}(Y)=$$

$$=(X^TX)^{-1}X^T\mathbb{E}(X\beta+\varepsilon)=(X^TX)^{-1}X^TX\beta=\beta.$$

Обозначим $\varphi(X,Y)=(X^TX)^{-1}X^TY$. Тогда $\hat{\beta}=\varphi(X,Y)$. Покажем, что функция φ линейна по переменной Y.

1. $\varphi(X,\lambda\cdot Y)=(X^TX)^{-1}X^T(\lambda\cdot Y)=$
$$=\lambda\cdot(X^TX)^{-1}X^TY=\lambda\cdot\varphi(X,Y).$$
2. $\varphi(X,Y+Z)=(X^TX)^{-1}X^T(Y+Z)=$
$$=(X^TX)^{-1}X^TY+(X^TX)^{-1}X^TZ=\varphi(X,Y)+\varphi(X,Z).$$

Что и требовалось доказать. □

Задача 33. Являются ли МНК-оценки линейными по переменной X?

Решение. Нет, так как для функции

$$\varphi(X,Y)=(X^TX)^{-1}X^TY$$

не выполнено, например, свойство однородности по переменной X. Действительно,

$$\varphi(\lambda\cdot X,Y)=\left((\lambda\cdot X)^T(\lambda\cdot X)\right)^{-1}(\lambda\cdot X)^TY=$$

$$=\frac{1}{\lambda}\cdot(X^TX)^{-1}X^TY=\frac{1}{\lambda}\cdot\varphi(X,Y).\ \square$$

Задача 34. Приведите пример несмещенной и линейной по переменной Y оценки, отличной от МНК.

Решение. $\tilde{\beta}=(X^TCX)^{-1}X^TCY$, где

$$C=\begin{bmatrix}1&0&0&\cdots&0\\0&2&0&\cdots&0\\0&0&3&\ddots&\vdots\\\vdots&\vdots&\ddots&\ddots&0\\0&0&\cdots&0&n\end{bmatrix}.\ \square$$

Через $\mathcal{K}$ обозначим класс всех несмещенных и линейных по переменной y оценок для вектора β.

Теорема (Гаусс—Марков). Условие теоремы:

1) модель (1) специфицирована правильно, т. е. оцениваемая модель совпадает с моделью, которая порождает данные;

$$X^TX=\begin{bmatrix}5&2&1\\2&2&1\\1&1&1\end{bmatrix} \text{ и } \left(X^TX\right)^{-1}=\begin{bmatrix}1/3&-1/3&0\\-1/3&4/3&-1\\0&-1&2\end{bmatrix}.$$

(a) Укажите число наблюдений.

Ответ: $n=5$.

(b) Укажите число регрессоров в модели (с учетом свободного члена).

Ответ: $k+1=3$.

(c) Рассчитайте $TSS=\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$.

Ответ: $TSS=10$.

(d) Рассчитайте при помощи метода наименьших квадратов оценку для вектора неизвестных коэффициентов.

Решение: $\widehat{\beta}=\begin{bmatrix}\widehat{\alpha}\\\widehat{\beta}_1\\\widehat{\beta}_2\end{bmatrix}=\left(X^TX\right)^{-1}X^TY=\begin{bmatrix}2\\2\\1\end{bmatrix}$.

(e) Найдите $RSS=\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2$.

Ответ: $RSS=2$.

(f) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оцененного уравнения регрессии.

Ответ: $R^2=1-\frac{RSS}{TSS}=0.8$. R^2 — высокий; построенная эконометрическая модель «хорошо» описывает данные.

(g) Сформулируйте основную и альтернативную гипотезу, которые соответствуют тесту на значимость переменной x_1 в уравнении регрессии.

Ответ: основная гипотеза $H_0:\beta_1=0$, альтернативная гипотеза $H_1:\beta_1\neq 0$.

(h) Протестируйте на значимость переменную x_1 в уравнении регрессии на уровне значимости 10%.

- Приведите формулу для тестовой статистики.

Ответ: $T=\frac{\widehat{\beta}_1-\beta_1^{\circ}}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}}=\frac{\widehat{\beta}_1-\beta_1^{\circ}}{\sqrt{\frac{RSS}{n-k-1}\cdot\left[\left(X^TX\right)^{-1}\right]_{22}}}$;

$n=5;\ k=2$.

- Укажите распределение тестовой статистики.

Ответ: $T\overset{H_0}{\sim}t(n-k-1);\ n=5;\ k=2$.

- Вычислите наблюдаемое значение тестовой статистики.

Решение:

$$T_{\text{набл}} = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - 0}{\sqrt{\frac{RSS}{n-k-1}\cdot\left[\left(X^T X\right)^{-1}\right]_{22}}} = \frac{2-0}{\sqrt{\frac{2}{5-2-1}\cdot\frac{4}{3}}} = 1.7321.$$

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $[-T_{\text{кр}}; T_{\text{кр}}] = [-2.920; 2.920]$.

- Сделайте статистический вывод о значимости переменной x_1.

Ответ: поскольку $T_{\text{набл}} = 1.7321 \in [-2.920; 2.920]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10 %.

(i) Найдите *p-value*, соответствующее наблюдаемому тестовой статистики ($T_{\text{набл}}$) из предыдущего пункта. На основе полученного значения *p-value* сделайте вывод о значимости переменной x_1.

Решение.

$$p\text{-}value(T_{\text{набл}}) := \mathbb{P}\left(\{|T| > |T_{\text{набл}}|\}\right) = 2 \cdot F_T(-|T_{\text{набл}}|),$$

где $F_T(-|T_{\text{набл}}|)$ — функция распределения t-распределения с $n-k-1=5-2-1=2$ степенями свободы в точке $(-|T_{\text{набл}}|)$. В программе MATLAB:

$$p\text{-}value(T_{\text{набл}}) = 2 * tcdf(-|T_{\text{набл}}|, n-k-1) = 2 * tcdf(-1.7321, 2) = 0.2253.$$

Поскольку значение *p-value* превосходит уровень значимости 10%, то основная гипотеза $H_0: \beta_1 = 0$ не может быть отвергнута.

(j) Проверьте гипотезу $H_0: \beta_1 = 1$ против альтернативной гипотезы $H_1: \beta_1 \neq 1$. Уровень значимости 10%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \frac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\frac{RSS}{n-k-1}\cdot\left[\left(X^T X\right)^{-1}\right]_{22}}}$;

$n = 5;\ k = 2$.

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} t(n-k-1);\ n = 5;\ k = 2$.

- Вычислите наблюдаемое значение тестовой статистики.

Решение:

$$T_{\text{набл}} = \frac{\widehat{\beta}_1 - 1}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - 1}{\sqrt{\frac{RSS}{n-k-1}\cdot\left[\left(X^T X\right)^{-1}\right]_{22}}} = \frac{2-1}{\sqrt{\frac{2}{5-2-1}\cdot\frac{4}{3}}} = 0.8660.$$

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $[-T_{кр}; T_{кр}] = [-2.920; 2.920]$.

- Сделайте статистический вывод.

Ответ: поскольку $T_{набл} = 0.8660 \in [-2.920; 2.920]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10 %.

(k) Проверьте гипотезу $H_0 : \beta_1 = 1$ против альтернативной гипотезы $H_1 : \beta_1 > 1$. Уровень значимости 10%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\mathrm{D}}(\widehat{\beta}_1)}} = \dfrac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\dfrac{RSS}{n-k-1} \cdot \left[\left(X^T X \right)^{-1} \right]_{22}}}$;

$n = 5;\ k = 2$.

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} t(n-k-1)$; $n = 5;\ k = 2$.

- Вычислите наблюдаемое значение тестовой статистики.

Решение:

$$T_{набл} = \frac{\widehat{\beta}_1 - 1}{\sqrt{\widehat{\mathrm{D}}(\widehat{\beta}_1)}} = \frac{\widehat{\beta}_1 - 1}{\sqrt{\dfrac{RSS}{n-k-1} \cdot \left[\left(X^T X \right)^{-1} \right]_{22}}} = \frac{2-1}{\sqrt{\dfrac{2}{5-2-1} \cdot \dfrac{4}{3}}} = 0.8660.$$

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $(-\infty; T_{кр}] = (-\infty; 1.8856]$.

- Сделайте статистический вывод.

Ответ: поскольку $T_{набл} = 0.8660 \in (-\infty; 1.8856]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10 %.

(l) Проверьте гипотезу $H_0 : \beta_1 = 1$ против альтернативной гипотезы $H_1 : \beta_1 < 1$. Уровень значимости 10%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\mathrm{D}}(\widehat{\beta}_1)}} = \dfrac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\dfrac{RSS}{n-k-1} \cdot \left[\left(X^T X \right)^{-1} \right]_{22}}}$;

$n = 5;\ k = 2$.

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} t(n-k-1)$; $n = 5;\ k = 2$.

- Вычислите наблюдаемое значение тестовой статистики.

Решение:

$$T_{\text{набл}} = \frac{\widehat{\beta}_1 - 1}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}} = \frac{\widehat{\beta}_1 - 1}{\sqrt{\frac{RSS}{n-k-1} \cdot \left[\left(X^T X\right)^{-1}\right]_{22}}} = \frac{2-1}{\sqrt{\frac{2}{5-2-1} \cdot \frac{4}{3}}} = 0.8660.$$

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $[-T_{\text{кр}}; +\infty) = [-1.8856; +\infty)$,

- Сделайте статистический вывод.

Ответ: поскольку $T_{\text{набл}} = 0.8660 \in [-1.8856; +\infty)$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10 %.

(m) Сформулируйте основную гипотезу, которая соответствуют тесту на значимость регрессии «в целом».

Ответ: основная гипотеза $H_0: \begin{cases} \beta_1 = 0, \\ \beta_2 = 0, \end{cases}$

альтернативная гипотеза $H_1: \begin{bmatrix} \beta_1 \neq 0, \\ \beta_2 \neq 0. \end{bmatrix}$

(n) Протестируйте на значимость регрессию «в целом» на уровне значимости 5%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$; $n = 5$; $k = 2$.

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} F(k, n-k-1)$; $n = 5$; $k = 2$.

- Вычислите наблюдаемое значение тестовой статистики.

Решение: $T_{\text{набл}} = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} = \frac{0.8}{1-0.8} \cdot \frac{5-2-1}{2} = 4$.

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $\left[0; T_{\text{кр}}\right] = [0; 19]$

- Сделайте статистический вывод о значимости регрессии «в целом».

Ответ: поскольку $T_{\text{набл}} = 4 \in [0; 19]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5 %. Следовательно, регрессия «в целом» незначима.

(о) Найдите *p-value*, соответствующее наблюдаемому тестовой статистики ($T_{\text{набл}}$) из предыдущего пункта. На основе полученного значения *p*-value сделайте вывод о значимости регрессии «в целом».

Решение.

$$p\text{-}value(T_{\text{набл}}) := \mathbb{P}\left(\{T > T_{\text{набл}}\}\right) = 1 - F_T(|T_{\text{набл}}|),$$

где $F_T(|T_{\text{набл}}|)$ — функция распределения F-распределения с $k=2$ и $n-k-1=5-2-1=2$ степенями свободы в точке $T_{\text{набл}}$. В программе MATLAB:

$$p\text{-}value(T_{\text{набл}}) = 1 - fcdf(-|T_{\text{набл}}|, n-k-1) = 1 - fcdf(4,2,2) = 0.2.$$

Поскольку значение *p-value* превосходит уровень значимости 5%, то основная гипотеза $H_0: \beta_1 = \beta_2 = 0$ не может быть отвергнута. Стало быть, регрессия «в целом» незначима.

(р) Проверьте гипотезу $H_0: \beta_1 + \beta_2 = 2$ против альтернативной гипотезы $H_1: \beta_1 + \beta_2 \neq 2$. Уровень значимости 5%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{\widehat{\beta}_1 + \widehat{\beta}_2 - (\beta_1 + \beta_2)^{0}}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right)}}$, где

$$\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right) = \widehat{\mathrm{D}}\left(\widehat{\beta}_1\right) + 2\widehat{\mathrm{cov}}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) + \widehat{\mathrm{D}}\left(\widehat{\beta}_2\right) =$$
$$= \widehat{\sigma^2} \cdot \left[\left(X^T X\right)^{-1}\right]_{22} + 2 \cdot \widehat{\sigma^2} \cdot \left[\left(X^T X\right)^{-1}\right]_{23} + \widehat{\sigma^2} \cdot \left[\left(X^T X\right)^{-1}\right]_{33} =$$
$$= \frac{RSS}{n-k-1} \cdot \left(\left[\left(X^T X\right)^{-1}\right]_{22} + 2 \cdot \left[\left(X^T X\right)^{-1}\right]_{23} + \left[\left(X^T X\right)^{-1}\right]_{33}\right).$$

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} t(n-k-1);\ n=5;\ k=2.$

- Вычислите наблюдаемое значение тестовой статистики.

$$\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right) = \frac{RSS}{n-k-1} \cdot \left(\left[\left(X^T X\right)^{-1}\right]_{22} + 2 \cdot \left[\left(X^T X\right)^{-1}\right]_{23} + \left[\left(X^T X\right)^{-1}\right]_{33}\right) =$$
$$= \frac{2}{5-2-1} \cdot \left(\frac{4}{3} + 2 \cdot (-1) + 2\right) = \frac{4}{3}.$$

Ответ: $T_{\text{набл}} = \dfrac{\widehat{\beta}_1 + \widehat{\beta}_2 - 2}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right)}} = \dfrac{2+1-2}{\sqrt{\frac{4}{3}}} = 0.8660.$

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $[-T_{\text{кр}}; T_{\text{кр}}] = [-4.3027; 4.3027].$

- Сделайте статистический вывод.

Ответ: поскольку $T_{\text{набл}} = 0.8660 \in [-4.3027; 4.3027]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5 %.

(q) Проверьте гипотезу $H_0 : \beta_1 + \beta_2 = 2$ против альтернативной гипотезы $H_1 : \beta_1 + \beta_2 > 2$. Уровень значимости 5%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{\widehat{\beta}_1 + \widehat{\beta}_2 - (\beta_1 + \beta_2)}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right)}}$, где

$$\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right) = \widehat{\mathrm{D}}\left(\widehat{\beta}_1\right) + 2\widehat{\mathrm{cov}}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) + \widehat{\mathrm{D}}\left(\widehat{\beta}_2\right) =$$
$$= \widehat{\sigma^2} \cdot \left[\left(X^T X\right)^{-1}\right]_{22} + 2 \cdot \widehat{\sigma^2} \cdot \left[\left(X^T X\right)^{-1}\right]_{23} + \widehat{\sigma^2} \cdot \left[\left(X^T X\right)^{-1}\right]_{33} =$$
$$= \frac{RSS}{n-k-1} \cdot \left(\left[\left(X^T X\right)^{-1}\right]_{22} + 2 \cdot \left[\left(X^T X\right)^{-1}\right]_{23} + \left[\left(X^T X\right)^{-1}\right]_{33}\right).$$

- Укажите распределение тестовой статистики.

Ответ: $T \sim t(n-k-1)$; $n = 5$; $k = 2$.

- Вычислите наблюдаемое значение тестовой статистики.

$$\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right) =$$
$$= \frac{RSS}{n-k-1} \cdot \left(\left[\left(X^T X\right)^{-1}\right]_{22} + 2 \cdot \left[\left(X^T X\right)^{-1}\right]_{23} + \left[\left(X^T X\right)^{-1}\right]_{33}\right) =$$
$$= \frac{2}{5-2-1} \cdot \left(\frac{4}{3} + 2 \cdot (-1) + 2\right) = \frac{4}{3}.$$

Ответ: $T_{\text{набл}} = \dfrac{\widehat{\beta}_1 + \widehat{\beta}_2 - 2}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right)}} = \dfrac{2+1-2}{\sqrt{\frac{4}{3}}} = 0.8660.$

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $(-\infty; T_{\text{кр}}] = (-\infty; 2.920]$.

- Сделайте статистический вывод.

Ответ: поскольку $T_{\text{набл}} = 0.8660 \in (-\infty; 2.920]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5 %.

(r) Проверьте гипотезу $H_0 : \beta_1 + \beta_2 = 2$ против альтернативной гипотезы $H_1 : \beta_1 + \beta_2 < 2$. Уровень значимости 5%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{\widehat{\beta}_1 + \widehat{\beta}_2 - (\beta_1 + \beta_2)}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1 + \widehat{\beta}_2\right)}}$, где

$$\widehat{\mathrm{D}}\left(\widehat{\beta}_1+\widehat{\beta}_2\right)=\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)+2\widehat{\mathrm{cov}}\left(\widehat{\beta}_1,\widehat{\beta}_2\right)+\widehat{\mathrm{D}}\left(\widehat{\beta}_2\right)=$$

$$=\widehat{\sigma^2}\cdot\left[\left(X^TX\right)^{-1}\right]_{22}+2\cdot\widehat{\sigma^2}\cdot\left[\left(X^TX\right)^{-1}\right]_{23}+\widehat{\sigma^2}\cdot\left[\left(X^TX\right)^{-1}\right]_{33}=$$

$$=\frac{RSS}{n-k-1}\cdot\left(\left[\left(X^TX\right)^{-1}\right]_{22}+2\cdot\left[\left(X^TX\right)^{-1}\right]_{23}+\left[\left(X^TX\right)^{-1}\right]_{33}\right).$$

- Укажите распределение тестовой статистики.

Ответ: $T\sim t(n-k-1)$; $n=5$; $k=2$.

- Вычислите наблюдаемое значение тестовой статистики.

$$\widehat{\mathrm{D}}\left(\widehat{\beta}_1+\widehat{\beta}_2\right)=$$

$$=\frac{RSS}{n-k-1}\cdot\left(\left[\left(X^TX\right)^{-1}\right]_{22}+2\cdot\left[\left(X^TX\right)^{-1}\right]_{23}+\left[\left(X^TX\right)^{-1}\right]_{33}\right)=$$

$$=\frac{2}{5-2-1}\cdot\left(\frac{4}{3}+2\cdot(-1)+2\right)=\frac{4}{3}.$$

Ответ: $T_{\text{набл}}=\dfrac{\widehat{\beta}_1+\widehat{\beta}_2-2}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1+\widehat{\beta}_2\right)}}=\dfrac{2+1-2}{\sqrt{\frac{4}{3}}}=0.8660$.

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $[-T_{\text{кр}};+\infty)=[-2.920;+\infty)$.

- Сделайте статистический вывод.

Ответ: поскольку $T_{\text{набл}}=0.8660\in[-2.920;+\infty)$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5 %.

Задача 2. Рассмотрим регрессионную модель

$$Y_i=\alpha+\beta_1\cdot x_{i1}+\beta_2\cdot x_{i2}+\beta_3\cdot x_{i3}+\beta_4\cdot x_{i4}+\varepsilon_i\quad i=1,\ldots,n.$$

Известно, что $\varepsilon_1,\ldots,\varepsilon_n$ — независимые нормальные случайные величины с математическим ожиданием ноль и дисперсией σ^2 (параметр σ^2 неизвестен). Ниже (стр. 82) приведены результаты оценки уравнения регрессии.

(a) Сформулируйте основную и альтернативную гипотезу, которые соответствуют тесту на значимость переменной x_1 в уравнении регрессии.

Ответ: $H_0:\beta_1=0\,(\Leftrightarrow x_1-\text{незначимая переменная})$,

$H_1:\beta_1\neq 0\,(\Leftrightarrow x_1-\text{значимая переменная})$.

(b) Протестируйте на значимость переменную x_1 в уравнении регрессии на уровне значимости 10%.

- Приведите формулу для тестовой статистики.

Ответ: $T=\dfrac{\widehat{\beta}_1-\beta_1}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}}$,

где $\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)=\widehat{\sigma^2}\left[\left(X^TX\right)^{-1}\right]_{22}=\dfrac{RSS}{n-k-1}\left[\left(X^TX\right)^{-1}\right]_{22}$.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,991899
R-квадрат	0,983863
Нормированный R-квадрат	0,970954
Стандартная ошибка	1,748775
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	4	932,3089	233,0772	76,21348	0,000114
Остаток	5	15,29108	3,058215		
Итого	9	947,6			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95 %*	*Верхние 95 %*
Y-пересечение	8,378016	3,06332	2,734947	0,041038	0,503502	16,25253
X1	0,743348	0,138405	5,370825	0,003013	0,387567	1,099128
X2	2,269541	0,184886	12,27536	6,35E-05	1,794277	2,744805
X3	0,173608	0,235658	0,736695	0,4944	-0,43217	0,779386
X4	0,464104	0,166876	2,781132	0,03885	0,035136	0,893072

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} t(n-k-1)$; $n=10$, $k=4$.

- Вычислите наблюдаемое значение тестовой статистики.

Ответ: $T_{\text{набл}} = \dfrac{0.743348-0}{0.138405} = 5.370825$.

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $\left[-T_{\text{кр}};T_{\text{кр}}\right]=\left[-2.015;2.015\right]$.

- Сделайте статистический вывод о значимости переменной x_1.

Ответ: Поскольку $T_{\text{набл}} \notin \left[-T_{\text{кр}};T_{\text{кр}}\right]$, то мы вынуждены отвергнуть основную гипотезу $H_0: \beta_1 = 0$ в пользу альтернативной гипотезы $H_1: \beta_1 \neq 0$ на уровне значимости 10%. Следовательно, x_1 — значимая переменная.

(c) Проверьте гипотезу $H_0: \beta_1 = 1$ против альтернативной гипотезы $H_1: \beta_1 > 1$. Уровень значимости 5%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{\widehat{\beta}_1 - \beta_1^0}{\sqrt{\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right)}}$,

где $\widehat{\mathrm{D}}\left(\widehat{\beta}_1\right) = \widehat{\sigma^2}\left[\left(X^T X\right)^{-1}\right]_{22} = \dfrac{RSS}{n-k-1}\left[\left(X^T X\right)^{-1}\right]_{22}$.

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} t(n-k-1)$; $n=10$, $k=4$.

- Вычислите наблюдаемое значение тестовой статистики.

Ответ: $T_{\text{набл}} = \dfrac{0.743348-1}{0.138405} = -1.854354$.

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $\left(-\infty; T_{\text{кр}}\right] = \left(-\infty; 2.015\right]$.

- Сделайте статистический вывод.

Ответ: Поскольку $T_{\text{набл}} \in \left(-\infty; T_{\text{кр}}\right]$, то на основании имеющихся данных мы не можем отвергнуть основную гипотезу $H_0 : \beta_1 = 1$ в пользу альтернативной гипотезы $H_1 : \beta_1 > 1$ на уровне значимости 5%.

(d) Сформулируйте основную гипотезу, которая соответствуют тесту на значимость регрессии «в целом».

Ответ:

$$H_0 : \begin{cases} \beta_1 = 0, \\ \beta_2 = 0, \\ \beta_3 = 0, \\ \beta_4 = 0. \end{cases} \quad (\Leftrightarrow \text{регрессия «в целом» незначима})$$

$H_1 : |\beta_1| + \ldots + |\beta_4| > 0 \quad (\Leftrightarrow \text{регрессия «в целом» значима})$.

(e) Протестируйте на значимость регрессию «в целом» на уровне значимости 1%.

- Приведите формулу для тестовой статистики.

Ответ: $T = \dfrac{R^2}{1-R^2} \cdot \dfrac{n-k-1}{k}$; $n=10$, $k=4$.

- Укажите распределение тестовой статистики.

Ответ: $T \overset{H_0}{\sim} F(k, n-k-1)$; $n=10$, $k=4$.

- Вычислите наблюдаемое значение тестовой статистики.

Ответ: $T_{\text{набл}} = \dfrac{0.983863}{1-0.983863} \cdot \dfrac{10-4-1}{4} = 76.21$.

- Укажите распределение тестовой статистики.

Ответ:

$$T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)} \overset{H_0}{\sim} F(q, n-k-1).$$

- Вычислите наблюдаемое значение тестовой статистики.

Ответ: $T_{\text{набл}} = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)} = \frac{(41.52-15.29)/2}{15.29/(10-4-1)} = 4.28$.

- Укажите область, в которой основная гипотеза не отвергается.

Ответ: $[0; T_{\text{кр}}] = [0; 5.79]$.

- Сделайте статистический вывод.

Ответ: поскольку $T_{\text{набл}} = 4.28 \in [0; 5.79]$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5 %.

Задача 4. Известно, что *p-value* для коэффициента регрессии равно 0.087, а уровень значимости 0.1. Является ли значимым данный коэффициент в регрессии?

Ответ: да.

Задача 5. Известно, что *p-value* для коэффициента регрессии равно 0.078, а уровень значимости 0.05. Является ли значимым данный коэффициент в регрессии?

Ответ: нет.

Задача 6. Известно, что *p-value* для коэффициента регрессии равно 0.09. На каком уровне значимости данный коэффициент в регрессии будет признан значимым?

Ответ: на уровне значимости, большем, чем 0.09.

Задача 7. Известно, что *p-value* для коэффициента регрессии равно 0.07. На каком уровне значимости данный коэффициент в регрессии будет признан незначимым?

Ответ: на уровне значимости, меньшем, чем 0.07.

Задача 8. В таблице ниже (стр. 90) приведены результаты оценивания уравнения линейной регрессии. Перечислите, какие из переменных $x_1, \ldots, x_5$ в регрессии являются значимыми на уровне значимости 5 %.

Ответ: x_1, x_3, x_4.

(a) Спецификация уравнения регрессии с учетом образования родителей:

$$\ln W_i = \alpha + \beta_{Edu} \cdot Edu_i + \beta_{Age} \cdot Age + \\ + \beta_{Age^2} \cdot Age^2 + \beta_{Fedu} \cdot Fedu + \beta_{Medu} \cdot Medu + \varepsilon_i.$$

(b) $H_0: \begin{cases} \beta_{Fedu} = 0, \\ \beta_{Medu} = 0. \end{cases}$ $H_1: \begin{bmatrix} \beta_{Fedu} \neq 0, \\ \beta_{Medu} \neq 0. \end{bmatrix}$

(c) $T = \dfrac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)}$, где $q = 2$ — число линейно независимых уравнений в основной гипотезе H_0; $n = 25$ — число наблюдений; $k = 5$ — число коэффициентов в модели без ограничения (без учета свободного члена).

(d) $T \overset{H_0}{\sim} F(q, n-k-1)$.

(e) $T_{\text{набл}} = \dfrac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)} = \dfrac{(60.4-40.4)/2}{40.4/(25-5-1)} = 4.70$.

(f) $\left[0; T_{\text{кр}}\right] = [0; 3.52]$.

(g) Поскольку $T_{\text{набл}} \notin \left[0; T_{\text{кр}}\right]$, то мы вынуждены отвергнуть основную гипотезу $H_0: \begin{cases} \beta_{Fedu} = 0, \\ \beta_{Medu} = 0 \end{cases}$ в пользу альтернативной гипотезы $H_1: \begin{bmatrix} \beta_{Fedu} \neq 0, \\ \beta_{Medu} \neq 0 \end{bmatrix}$ на уровне значимости 5%.

Иными словами, на основании имеющихся данных мы заключаем, что образование родителей существенно влияет на зарплату.

Задача 24. Пусть задана линейная регрессионная модель

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot X_{i3} + \beta_4 \cdot X_{i4} + \varepsilon_i,\ i = 1, \ldots, 20.$$

По имеющимся данным оценены следующие регрессии:

$$\underset{(s.e.)}{\hat{Y}} = \underset{(0.15)}{10.01} + \underset{(0.06)}{1.05} \cdot X_1 + \underset{(0.04)}{2.06} \cdot X_2 + \underset{(0.06)}{0.49} \cdot X_3 - \underset{(0.06)}{1.31} \cdot X_4,\ RSS = 6.85;$$

$$\underset{(s.e.)}{\widehat{Y - X_1 - 2 \cdot X_2}} = \underset{(0.15)}{10.00} + \underset{(0.07)}{0.50} \cdot X_3 - \underset{(0.06)}{1.32} \cdot X_4,\ RSS = 8.31;$$

$$\underset{(s.e.)}{\widehat{Y + X_1 + 2 \cdot X_2}} = \underset{(3.62)}{9.93} + \underset{(1.48)}{0.56} \cdot X_3 - \underset{(1.42)}{1.50} \cdot X_4,\ RSS = 4310.62;$$

$$\underset{(s.e.)}{\widehat{Y - X_1 + 2 \cdot X_2}} = \underset{(3.26)}{10.71} + \underset{(1.33)}{0.09} \cdot X_3 - \underset{(1.28)}{1.28} \cdot X_4,\ RSS = 3496.85;$$

$$\underset{(s.e.)}{\widehat{Y + X_1 - 2 \cdot X_2}} = \underset{(1.25)}{9.22} + \underset{(0.51)}{0.97} \cdot X_3 - \underset{(0.49)}{1.54} \cdot X_4,\ RSS = 516.23.$$

Проверьте гипотезу $H_0: \begin{cases} \beta_1 = 1, \\ \beta_2 = 2 \end{cases}$ против альтернативной гипотезы $H_1: \begin{bmatrix} \beta_1 \neq 1, \\ \beta_2 \neq 2 \end{bmatrix}$. Уровень значимости 5%.

(a) Приведите формулу тестовой статистики.

(b) Укажите распределение тестовой статистики.

(c) Рассчитайте наблюдаемое значение тестовой статистики.

(d) Укажите область, в которой основная гипотеза не отвергается.

(e) Сделайте статистический вывод.

RSS_2 — это сумма квадратов остатков в модели

$$\ln(W_i) = \alpha^{(2)} + \beta_1^{(2)} \cdot Edu_i + \beta_2^{(2)} \cdot Exp_i + \\ + \beta_3^{(2)} \cdot Exp_i^2 + \beta_4^{(2)} \cdot Fedu_i + \beta_5^{(2)} \cdot Medu_i + \varepsilon_i,\ i = 36,...,58.$$

1. Тестовая статистика:

$$T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-m)},$$

где RSS_R — это сумма квадратов остатков в модели с ограничениями;

RSS_{UR} — это сумма квадратов остатков в модели без ограничений;

q — число линейно независимых уравнений в основной гипотезе H_0;

n — общее число наблюдений;

m — число коэффициентов в модели без ограничений.

2. Распределение тестовой статистики:

$$T \overset{H_0}{\sim} F(q; n-m).$$

3. Наблюдаемое значение тестовой статистики:

$$T_{набл} = \frac{(70.3 - (34.4 + 23.4))/6}{(34.4 + 23.4)/(58 - 12)} = 1.66.$$

4. Область, в которой H_0 не отвергается:

$$[0; T_{кр}] = [0; 2.30].$$

5. Статистический вывод:

Поскольку $T_{набл} \in [0; T_{кр}]$, то на основе имеющихся данных мы не можем отвергнуть гипотезу H_0 в пользу альтернативной гипотезы H_1. Следовательно, имеющиеся данные не противоречат гипотезе об отсутствии дискриминации на рынке труда между мужчинами и женщинами.

Задача 2. По 52 наблюдениям была оценена следующая зависимость цены квадратного метра квартиры $Price$ (в долларах) от площади кухни K (в м^2), времени в пути пешком до ближайшего метро M (в минутах), расстояния до центра города C (в км) и наличия рядом с домом лесопарковой зоны P (1 — есть, 0 — нет)

$$\underset{(s.e)}{\widehat{Price}} = \underset{(3.73)}{16.12} + \underset{(0.14)}{1.7}\,K - \underset{(0.03)}{0.35}\,M - \underset{(0.12)}{0.46}\,C + \underset{(0.98)}{2.22}\,P,$$

$$R^2 = 0.78,\ \sum_{i=1}^{52}\left(Price_i - \overline{Price}\right)^2 = 278.$$

Предположим, что все квартиры в выборке можно отнести к двум категориям: квартиры на севере города (28 наблюдений) и квартиры на юге города (24 наблюдения). Модель регрессии была оценена отдельно только по квартирам на севере и только по квартирам на юге. Ниже приведены результаты оценивания.

Для квартир на севере:

$$\underset{(s.e)}{\widehat{Price}} = \underset{(3.3)}{14} + \underset{(0.23)}{1.6}\,K - \underset{(0.04)}{0.33}\,M - \underset{(0.22)}{0.40}\,C + \underset{(0.78)}{2.1}\,P$$

$$RSS = 21.8.$$

Для квартир на юге:

$$\underset{(s.e)}{\widehat{Price}} = \underset{(3.9)}{16.8} + \underset{(0.4)}{1.62}\,K - \underset{(0.12)}{0.29}\,M - \underset{(0.23)}{0.51}\,C + \underset{(1.28)}{1.98}\,P$$

$$RSS = 19.2.$$

Протестируйте гипотезу о различии в ценообразовании квартир на севере и на юге. Уровень значимости 5%.

$$\sum_{i=1}^{m}(Y_i-(\hat\mu+\hat\gamma)-(\hat\nu+\hat\delta)x_i)^2+\sum_{i=m+1}^{n}(Y_i-\hat\mu-\hat\nu x_i)^2\le$$

$$\le\sum_{i=1}^{m}(Y_i-(\mu+\gamma)-(\nu+\delta)x_i)^2+\sum_{i=m+1}^{n}(Y_i-\mu-\nu x_i)^2. \quad (2)$$

Учитывая, что неравенство (2) справедливо для всех μ, ν, γ и δ, то оно останется верным для $\mu=\hat\mu$, $\nu=\hat\nu$ и произвольных γ и δ. Имеем

$$\sum_{i=1}^{m}(Y_i-(\hat\mu+\hat\gamma)-(\hat\nu+\hat\delta)x_i)^2+\sum_{i=m+1}^{n}(Y_i-\hat\mu-\hat\nu x_i)^2\le$$

$$\le\sum_{i=1}^{m}(Y_i-(\hat\mu+\gamma)-(\hat\nu+\delta)x_i)^2+\sum_{i=m+1}^{n}(Y_i-\hat\mu-\hat\nu x_i)^2.$$

Следовательно,

$$\sum_{i=1}^{m}(Y_i-(\hat\mu+\hat\gamma)-(\hat\nu+\hat\delta)x_i)^2\le\sum_{i=1}^{m}(Y_i-(\hat\mu+\gamma)-(\hat\nu+\delta)x_i)^2. \quad (3)$$

Обозначим $\tilde\alpha:=\hat\mu+\gamma$ и $\tilde\beta:=\hat\nu+\delta$. В силу произвольности γ и δ коэффициенты $\tilde\alpha$ и $\tilde\beta$ также произвольны. Тогда для любых $\tilde\alpha$ и $\tilde\beta$ выполнено неравенство

$$\sum_{i=1}^{m}(Y_i-(\hat\mu+\hat\gamma)-(\hat\nu+\hat\delta)x_i)^2\le\sum_{i=1}^{m}(Y_i-\tilde\alpha-\tilde\beta x_i)^2,$$

которое означает, что $\hat\mu+\hat\gamma$ и $\hat\nu+\hat\delta$ являются МНК-оценками коэффициентов α и β в регрессии $Y_i=\alpha+\beta x_i+\varepsilon_i$, оцененной по наблюдениям $i=1,\dots,m$, т. е. $\hat\alpha=\hat\mu+\hat\gamma$ и $\hat\beta=\hat\nu+\hat\delta$. □

Задача 6*. Выборка содержит 30 наблюдений зависимой переменной y и независимой переменной x. Ниже приведены результаты оценивания уравнения регрессии $Y_i=\alpha+\beta\cdot x_i+\varepsilon_i$

по первым 20-ти и по последним 10-ти наблюдениям соответственно

$$\hat Y=4.0039+2.6632\cdot x,$$

$$\hat Y=1.3780+5.2587\cdot x.$$

По имеющимся данным найдите оценки коэффициентов в модели рассчитанной по 30 наблюдениям

$$Y_i=\alpha+\beta\cdot x_i+\Delta\alpha\cdot d_i+\Delta\beta\cdot x_i\cdot d_i+\varepsilon_i,$$

где фиктивная переменная d определяется следующим образом

$$d_i=\begin{cases}1 & \text{при } i\in\{1,\dots,20\},\\ 0 & \text{при } i\in\{21,\dots,30\}.\end{cases}$$

Задача 7*. Рассмотрите две модели

$$Y_I=X_I\beta_I+\varepsilon_I \text{ и } Y_{II}=X_{II}\beta_{II}+\varepsilon_{II},$$

где матрицы $X_I=[i\ \ x_I]$ и $X_{II}=[i\ \ x_{II}]$ имеют размеры 50×2, кроме того известно, что

$$X_I^TX_I=\begin{bmatrix}50 & 300\\ 300 & 2100\end{bmatrix},\ Y_I^TX_I=[300\ \ 2000],\ Y_I^TY_I=2100,$$

$$X_{II}^TX_{II}=\begin{bmatrix}50 & 300\\ 300 & 2100\end{bmatrix},\ Y_{II}^TX_{II}=[300\ \ 2200],\ Y_{II}^TY_{II}=2500$$

и $[\varepsilon_I^T\ \ \varepsilon_{II}^T]^T\sim N(0,\sigma^2I)$.

На уровне значимости 5 % протестируйте гипотезу $H_0:\beta_I=\beta_{II}$.

Задача 8. В системе MATLAB напишите программу (файл-функцию), которая по заданным матрицам $X1 \in \mathbb{R}^{n_1 \times K}$, $Y1 \in \mathbb{R}^{n_1 \times 1}$, соответствующим данным по первой подвыборке, матрицам $X2 \in \mathbb{R}^{n_2 \times K}$, $Y2 \in \mathbb{R}^{n_2 \times 1}$, соответствующим данным по второй подвыборке, и уровню значимости $SL \in (0;1)$ возвращает T_obs — наблюдаемое значение тестовой статистики в тесте Чоу, T_cr — критическую точку, соответствующую уровню значимости SL, а также текстовую переменную H, которая равна «H0 не отвергается» и «H1 отвергается в пользу H1» в случаях, когда $T\_obs < T\_cr$ и $T\_obs \geq T\_cr$ соответственно. Синтаксис программы должен быть следующим: [H, T_obs, T_cr] = test_Chow(X1, Y1, X2, Y2, SL).

Решение.

```
function [H, T_obs, T_cr] = test_Chow(X1, Y1, X2, Y2, SL)
X = [X1; X2];
Y = [Y1; Y2];
[n, K] = size(X);
RSS_R = get_RSS(X, Y);
RSS_1 = get_RSS(X1, Y1);
RSS_2 = get_RSS(X2, Y2);
RSS_UR = RSS_1 + RSS_2;
T_obs = ((RSS_R - RSS_UR) / K) / (RSS_UR / (n - 2 * K));
T_cr = finv(1 - SL, K, n - 2 * K);
if (T_obs < T_cr)
    H = 'H0 не отвергается';
else
    H = 'H0 отвергается в пользу H1';
end

function [RSS] = get_RSS(X, Y)
b_hat = (X' * X)^(-1) * (X' * Y);
Y_hat = X * b_hat;
e_hat = Y - Y_hat;
RSS = sum(e_hat .^ 2);
```

Глава 5

Гетероскедастичность

Задача 1. Известно, что ошибки $\{\varepsilon_i\}_{i=1}^n$ в регрессионной модели $Y_i = \alpha + \beta x_i + \varepsilon_i$ — независимые ~~одинаково распределенные~~ случайные величины с математическим ожиданием ноль и дисперсией $D(\varepsilon_i) = \sigma^2 \cdot x_i^2$, $i = 1,...,n$. На какие величины надо разделить каждое уравнение регрессии, чтобы устранить гетероскедастичность ошибок?

Решение. Имеем:

$$D(\varepsilon_i) = \sigma^2 \cdot x_i^2 \Rightarrow \frac{1}{x_i^2} D(\varepsilon_i) = \sigma^2 \Rightarrow D\left(\frac{\varepsilon_i}{x_i}\right) = \sigma^2.$$

Стало быть, если разделить каждое уравнение $Y_i = \alpha + \beta x_i + \varepsilon_i$ на x_i, то случайные ошибки $\frac{\varepsilon_i}{x_i}$ в новой регрессионной модели $\frac{Y_i}{x_i} = \alpha \frac{1}{x_i} + \beta + \frac{\varepsilon_i}{x_i}$ являются гомоскедастичными. □

Задача 2. Известно, что ошибки $\{\varepsilon_i\}_{i=1}^n$ в регрессионной модели $Y_i = \alpha + \beta x_i + \varepsilon_i$ — независимые одинаково распределенные случайные величины с математическим ожиданием ноль и дисперсией $\mathrm{D}(\varepsilon_i) = \lambda \cdot |x_i|$, $\lambda > 0$, $i = 1,...,n$. На какие величины надо разделить каждое уравнение регрессии, чтобы устранить гетероскедастичность ошибок?

Ответ: $\sqrt{|x_i|}$.

Задача 3. Известно, что после деления каждого уравнения регрессии $Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,...,n$, на x_i^2 гетероскедастичность ошибок была устранена. Напишите, каким при этом должно быть уравнение для дисперсии ошибок $\mathrm{D}(\varepsilon_i)$.

Решение. Известно, что $\mathrm{D}\left(\frac{\varepsilon_i}{x_i^2}\right) = \sigma^2$ при всех $i = 1,...,n$. Следовательно, $\frac{1}{x_i^4}\mathrm{D}(\varepsilon_i) = \sigma^2$, а значит, $\mathrm{D}(\varepsilon_i) = \sigma^2 x_i^4$. □

Задача 4. Известно, что после деления каждого уравнения регрессии $Y_i = \alpha + \beta x_i + \varepsilon_i$, $i = 1,...,n$, на $\sqrt{x_i}$ гетероскедастичность ошибок была устранена. Напишите, каким при этом должно быть уравнение для дисперсии ошибок $\mathrm{D}(\varepsilon_i)$.

Ответ: $\mathrm{D}(\varepsilon_i) = \sigma^2 x_i$.

Задача 5. Для линейной регрессии $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1,...,30$ была выполнена сортировка наблюдений по возрастанию переменной x_1. Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Используя данные, приведенные ниже, протестируйте ошибки на гетероскедастичность на уровне значимости 5 %.

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 1,...,30$:

$\hat{\alpha} = 1.21$ $\hat{\beta}_1 = 1.89$ $\hat{\beta}_2 = 2.74$ $RSS = 48.69$.

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 1,...,11$:

$\hat{\alpha} = 1.39$ $\hat{\beta}_1 = 2.27$ $\hat{\beta}_2 = 2.36$ $RSS_1 = 10.28$.

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 12,...,19$:

$\hat{\alpha} = 0.75$ $\hat{\beta}_1 = 2.23$ $\hat{\beta}_2 = 3.19$ $RSS_2 = 5.31$.

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 20,...,30$:

$\hat{\alpha} = 1.56$ $\hat{\beta}_1 = 1.06$ $\hat{\beta}_2 = 2.29$ $RSS_3 = 14.51$.

Решение. Протестируем гетероскедастичность ошибок при помощи теста Голдфельда—Квандта.

$H_0 : \mathrm{D}\varepsilon_1 = ... = \mathrm{D}\varepsilon_n$ $H_1 : \mathrm{D}\varepsilon_i = \sigma^2 x_{i1}^2$, $i = 1,...,30$.

1. Тестовая статистика:

$$T = \frac{RSS_3/(n_3-k-1)}{RSS_1/(n_1-k-1)},$$

где n_1 — число наблюдений в первой подгруппе $(n_1=11)$, n_3 — число наблюдений в последней подгруппе $(n_3=11)$, k — число факторов в модели $(k=2)$.

2. Распределение тестовой статистики:

$$T \overset{H_0}{\sim} F(n_3-k-1, n_1-k-1).$$

3. Наблюдаемое значение тестовой статистики:

$$T_{набл} = \frac{14.51/(11-2-1)}{10.28/(11-2-1)} = 1.41.$$

4. Область, в которой H_0 не отвергается:

$$\left[\underline{T}_{кр}; \overline{T}_{кр}\right] = [\text{finv}(0.025,8,8); \text{finv}(0.975,8,8)] = [0.23; 4.43].$$

5. Статистический вывод: поскольку $T_{набл} \in \left[\underline{T}_{кр}; \overline{T}_{кр}\right]$, то на основании имеющихся наблюдений на уровне значимости 5 % основная гипотеза H_0 не может быть отвергнута. Таким образом, тест Голдфельда—Квандта не выявил гетероскедастичность. □

Задача 6. Для линейной регрессии $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1,...,50$ была выполнена сортировка наблюдений по возрастанию переменной x_2. Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Используя данные, приведенные ниже, протестируйте ошибки на гетероскедастичность на уровне значимости 1 %.

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 1,...,50$:

$$\hat{\alpha} = 1.16 \qquad \hat{\beta}_1 = 1.99 \qquad \hat{\beta}_2 = 2.97 \qquad RSS = 174.69.$$

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 1,...,21$:

$$\hat{\alpha} = 0.76 \qquad \hat{\beta}_1 = 2.25 \qquad \hat{\beta}_2 = 3.18 \qquad RSS_1 = 20.41.$$

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 22,...,29$:

$$\hat{\alpha} = 0.85 \qquad \hat{\beta}_1 = 1.81 \qquad \hat{\beta}_2 = 3.32 \qquad RSS_2 = 3.95.$$

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ по наблюдениям $i = 30,...,50$:

$$\hat{\alpha} = 1.72 \qquad \hat{\beta}_1 = 1.41 \qquad \hat{\beta}_2 = 2.49 \qquad RSS_3 = 130.74.$$

Решение.

$$H_0: \mathrm{D}\varepsilon_1 = ... = \mathrm{D}\varepsilon_n \quad H_1: \mathrm{D}\varepsilon_i = \sigma^2 x_{i2}^2, \ i = 1,...,50.$$

1. Тестовая статистика:

$$T = \frac{RSS_3/(n_3-k-1)}{RSS_1/(n_1-k-1)},$$

где n_1 — число наблюдений в первой подгруппе $(n_1=21)$, n_3 — число наблюдений в последней подгруппе $(n_3=21)$, k — число факторов в модели $(k=2)$.

2. Распределение тестовой статистики:

$$T \overset{H_0}{\sim} F(n_3 - k - 1, n_1 - k - 1).$$

3. Наблюдаемое значение тестовой статистики:

$$T_{набл} = \frac{130.74/(21-2-1)}{20.41/(21-2-1)} = 6.40.$$

4. Область, в которой H_0 не отвергается:

$$\left[\underline{T}_{кр}; \overline{T}_{кр}\right] = [\text{finv}(0.005,18,18); \text{finv}(0.995,18,18)] = [0.28; 3.56].$$

5. Статистический вывод: поскольку $T_{набл} \notin \left[0; T_{кр}\right]$, то на основании имеющихся наблюдений на уровне значимости 5 % мы вынуждены отвергнуть основную гипотезу H_0 в пользу альтернативной гипотезы H_1. Следовательно, тест Голдфельда—Квандта выявил гетероскедастичность. □

Задача 7. Рассматривается линейная регрессия

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,\ i = 1,...,50.$$

Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Используя данные, приведенные ниже, протестируйте ошибки на гетероскедастичность на уровне значимости 5 %.

Оценена регрессия $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,\ i = 1,...,50$.

Результаты оценивания приводятся:

$$\hat{\alpha} = 1.21 \qquad \hat{\beta}_1 = 1.11 \qquad \hat{\beta}_2 = 3.15 \qquad R^2 = 0.72.$$

Оценена также вспомогательная регрессия

$\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \delta_3 x_{i1}^2 + \delta_4 x_{i2}^2 + \delta_5 x_{i1} x_{i2} + u_i,\ i = 1,...,50.$

Результаты оценивания следующие:

$$\hat{\delta}_0 = 1{,}50 \quad \hat{\delta}_1 = -2{,}18 \quad \hat{\delta}_2 = 0{,}23 \quad \hat{\delta}_3 = 1{,}87$$
$$\hat{\delta}_4 = -0{,}56 \quad \hat{\delta}_5 = -0{,}09 \quad R^2_{всп} = 0.36.$$

Решение. Протестируем гетероскедастичность ошибок при помощи теста Уайта.

$$H_0 : \mathrm{D}\varepsilon_1 = ... = \mathrm{D}\varepsilon_n$$

$$H_1 : \mathrm{D}\varepsilon_i = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \delta_3 x_{i1}^2 + \delta_4 x_{i2}^2 + \delta_5 x_{i1} x_{i2},\ i = 1,...,30.$$

1. Тестовая статистика:

$$T = n \cdot R^2_{всп},$$

где n — число наблюдений, $R^2_{всп}$ — коэффициент детерминации для вспомогательной регрессии.

2. Распределение тестовой статистики:

$$T \overset{H_0}{\sim} \chi^2(k_{всп}),$$

где $k_{всп}$ — число факторов во вспомогательной регрессии $(k_{всп} = 5)$.

3. Наблюдаемое значение тестовой статистики:

$$T_{набл} = 50 \cdot 0.36 = 18.$$

4. Область, в которой H_0 не отвергается:

$$\left[0; T_{кр}\right] = [0; \text{chi2inv}(0.95, 5)] = [0; 11.07].$$

5. Статистический вывод: поскольку $T_{набл} \notin \left[0; T_{кр}\right]$, то на основании имеющихся наблюдений на уровне значи-

6.2. Задачи

Задача 1. Докажите, что в модели (1) для всех $t = 1, \ldots, T$

(a) $\mathbb{E}\varepsilon_t = 0$,

(b) $\mathrm{D}\,\varepsilon_t = \sigma^2 / (1-\rho^2)$,

(c) $\mathrm{cov}(\varepsilon_{t+h}, \varepsilon_t) = \rho^h \sigma^2 / (1-\rho^2)$ при $h \geq 1$,

(d) $\mathrm{corr}(\varepsilon_{t+h}, \varepsilon_t) = \rho^h$ при $h \geq 1$.

Решение.

(a) $\mathbb{E}[\varepsilon_1] = \mathbb{E}[\rho\varepsilon_0 + u_1] = \rho \underbrace{\mathbb{E}[\varepsilon_0]}_{=0} + \underbrace{\mathbb{E}[u_1]}_{=0} = 0$,

$\mathbb{E}[\varepsilon_2] = \mathbb{E}[\rho\varepsilon_1 + u_2] = \rho \underbrace{\mathbb{E}[\varepsilon_1]}_{=0} + \underbrace{\mathbb{E}[u_2]}_{=0} = 0$, и т. д.

(b) $\mathrm{D}(\varepsilon_1) = \mathrm{D}(\rho\varepsilon_0 + u_1) = \rho^2 \mathrm{D}(\varepsilon_0) + \mathrm{D}(u_1) +$

$$+ 2\rho \underbrace{\mathrm{cov}(\varepsilon_0, u_1)}_{=0} = \rho^2 \frac{\sigma^2}{1-\rho^2} + \sigma^2 = \frac{\sigma^2}{1-\rho^2},$$

$\mathrm{D}(\varepsilon_2) = \mathrm{D}(\rho\varepsilon_1 + u_2) = \rho^2 \mathrm{D}(\varepsilon_1) + \mathrm{D}(u_2) +$

$$+ 2\rho \underbrace{\mathrm{cov}(\varepsilon_1, u_2)}_{=0} = \rho^2 \frac{\sigma^2}{1-\rho^2} + \sigma^2 = \frac{\sigma^2}{1-\rho^2}, \text{ и т. д.}$$

(c) $\mathrm{cov}(\varepsilon_{t+1}, \varepsilon_t) = \mathrm{cov}(\rho\varepsilon_t + u_{t+1}, \varepsilon_t) =$

$$= \rho \underbrace{\mathrm{cov}(\varepsilon_t, \varepsilon_t)}_{=\mathrm{D}(\varepsilon_t)} + \underbrace{\mathrm{cov}(u_{t+1}, \varepsilon_t)}_{=0} = \rho \frac{\sigma^2}{1-\rho^2},$$

$\mathrm{cov}(\varepsilon_{t+2}, \varepsilon_t) = \mathrm{cov}(\rho\varepsilon_{t+1} + u_{t+2}, \varepsilon_t) =$

$$= \rho \underbrace{\mathrm{cov}(\varepsilon_{t+1}, \varepsilon_t)}_{=\rho\frac{\sigma^2}{1-\rho^2}} + \underbrace{\mathrm{cov}(u_{t+2}, \varepsilon_t)}_{=0} = \rho^2 \frac{\sigma^2}{1-\rho^2}, \text{ и т. д.}$$

(d) $\mathrm{corr}(\varepsilon_{t+h}, \varepsilon_t) = \dfrac{\mathrm{cov}(\varepsilon_{t+h}, \varepsilon_t)}{\sqrt{\mathrm{D}(\varepsilon_{t+h})}\sqrt{\mathrm{D}(\varepsilon_t)}} =$

$$= \frac{\rho^h \dfrac{\sigma^2}{1-\rho^2}}{\sqrt{\dfrac{\sigma^2}{1-\rho^2}}\sqrt{\dfrac{\sigma^2}{1-\rho^2}}} = \rho^h. \quad \square$$

Задача 2. Пусть ошибки в модели

$$Y_t = \alpha + \beta x_t + \varepsilon_t, \; t = 1, \ldots, T,$$

удовлетворяют условиям (2) и $\rho \neq 0$. Введем два типа преобразования исходной модели:

1) преобразование $\mathcal{A}$: для каждого $t \in \{2, \ldots, T\}$ уравнение $Y_t = \alpha + \beta x_t + \varepsilon_t$ преобразуем к виду

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(x_t - \rho x_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1},$$

2) преобразование $\mathcal{H}$: первое уравнение $Y_1 = \alpha + \beta x_1 + \varepsilon_1$ преобразуем к виду

$$(1-\rho^2)^{1/2} Y_1 = \alpha(1-\rho^2)^{1/2} + \beta(1-\rho^2)^{1/2} x_1 + (1-\rho^2)^{1/2} \varepsilon_1.$$

Объясните, в чем состоит смысл преобразований $\mathcal{A}$ и $\mathcal{H}$.

Задача 3. По 132 наблюдениям была оценена регрессионная зависимость объема производства Y_t от объема инвестиций I_t: $\widehat{\ln Y_t} = 44.3 + 0.03 \ln I_{t-1} + 0.07 \ln I_{t-2} + 0.5 \ln I_{t-12}$, $R^2 = 0.56$, $DW = 1.95$. При помощи теста Дарбина—Уотсона протестируйте автокорреляцию первого порядка ошибок регрессии на уровне значимости 1 %.

(a) укажите основную и альтернативную гипотезы,
(b) приведите формулу тестовой статистики,
(c) укажите распределение тестовой статистики,
(d) рассчитайте наблюдаемое значение тестовой статистики,
(e) укажите область, в которой основная гипотеза не отвергается,
(f) сделайте статистический вывод.

Решение. (a) $H_0 : \rho = 0$, $H_1 : \rho \neq 0$.

(b) Тестовая статистика (формула):

$$DW = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}.$$

(c) Распределение тестовой статистики: случайная величина DW имеет распределение Дарбина—Уотсона.

(d) Наблюдаемое значение тестовой статистики:

$DW_{\text{набл}} = 1.95$, $n = 132$, $k = 3$.

(e) Область, в которой H_0 не отвергается:

$[d_U; 4 - d_U] = [1.604; 2.396]$.

(f) Статистический вывод: поскольку

$DW_{\text{набл}} = 1.95 \in [d_U; 4 - d_U]$,

гипотеза H_0 не может быть отвергнута в пользу гипотезы H_1 о наличии автокорреляции, т. е. тест Дарбина—Уотсона не выявил автокорреляцию. □

Задача 4. По 132 наблюдениям была оценена регрессионная зависимость объема производства Y_t от объема инвестиций I_t: $\widehat{\ln Y_t} = 44.3 + 0.03 \ln I_{t-1} + 0.5 \ln I_{t-12}$, $R^2 = 0.37$, $DW = 0.33$.

При помощи соответствующего теста протестируйте автокорреляцию первого порядка ошибок регрессии на уровне значимости 5 %.

(a) укажите основную и альтернативную гипотезы,
(b) приведите формулу тестовой статистики,
(c) укажите распределение тестовой статистики,
(d) рассчитайте наблюдаемое значение тестовой статистики,
(e) укажите область, в которой основная гипотеза не отвергается,
(f) сделайте статистический вывод.

Решение.

(a) $H_0 : \rho = 0$, $H_1 : \rho \neq 0$.

(b) Тестовая статистика (формула):

$$DW = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}.$$

(c) Распределение тестовой статистики: случайная величина DW имеет распределение Дарбина—Уотсона.

(d) Наблюдаемое значение тестовой статистики:

$$DW_{\text{набл}} = 0.33,\ n = 100,\ k = 2.$$

(e) Область, в которой H_0 не отвергается:

$$[d_U; 4 - d_U] = [1.715; 2.285].$$

(f) Статистический вывод: поскольку

$$DW_{\text{набл}} = 0.33 < d_L = 1.634,$$

гипотеза H_0 отвергается в пользу гипотезы H_1 о наличии автокорреляции (с положительным коэффициентом ρ). □

Задача 5. По 100 наблюдениям была оценена модель линейной регрессии

$$Y_t = \alpha + \beta x_t + \varepsilon_t,\ RSS = 120,$$

$$\hat{\varepsilon}_1 = -1,\ \hat{\varepsilon}_{100} = 2,\ \sum_{t=2}^{100} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1} = -50.$$

Найдите

(a) DW, (b) $\hat{\rho} = 1 - DW/2$.

Ответ: (a) $DW = 2.7917$, (b) $\hat{\rho} = -0.3958$.

Задача 6. По 21 наблюдению была оценена модель линейной регрессии

$$\underset{\text{(s.e)}}{\hat{Y}_t} = \underset{(0.3)}{1.2} + \underset{(0.18)}{0.9}\, Y_{t-1} + \underset{(0.01)}{0.1}\, t,\ R^2 = 0.6,\ DW = 1.21.$$

Протестируйте гипотезу об отсутствии автокорреляции ошибок на уровне значимости 5 %.

Задача 7. Можно ли при помощи стандартных таблиц Дарбина—Уотсона проверять автокорреляцию в следующих моделях?

(a) $Y_t = \beta x_t + \varepsilon_t$,

(b) $Y_t = \alpha + \beta Y_{t-1} + \varepsilon_t$,

(c) $Y_t = \alpha + \beta t + \gamma Y_{t-1} + \varepsilon_t$,

(d) $Y_t = \beta t + \gamma x_t + \varepsilon_t$,

(e) $Y_t = \alpha + \beta t + \gamma x_t + \delta x_{t-1} + \varepsilon_t$?

Задача 8. Рассматривается модель линейной регрессии $Y = X\beta + \varepsilon$, в которой $\mathrm{V}(\varepsilon) = \sigma^2 I$. Докажите, что $\hat{\beta}_{GLS} = \hat{\beta}_{OLS}$, где $\hat{\beta}_{GLS} = (X^T \mathrm{V}(\varepsilon)^{-1} X)^{-1} X^T \mathrm{V}(\varepsilon)^{-1} Y$ и $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$.

Задача 9. Докажите, что в условиях гетероскедастичности и автокорреляции МНК-оценки остаются несмещенными.

Задача 10. Пусть $u_1, \ldots, u_n$ — независимые случайные величины с $\mathbb{E}[u_t] = 0$ и $\mathrm{D}[u_t] = \sigma^2$. Известно, что ошибки $\{\varepsilon_t\}_{t=1}^n$ в регрессионной модели $Y_t = \beta x_t + \varepsilon_t$ удовлетворяют соотношениям $\varepsilon_1 = u_1$, $\varepsilon_t = u_t - u_{t-1}$ при $t = 2, \ldots, n$.

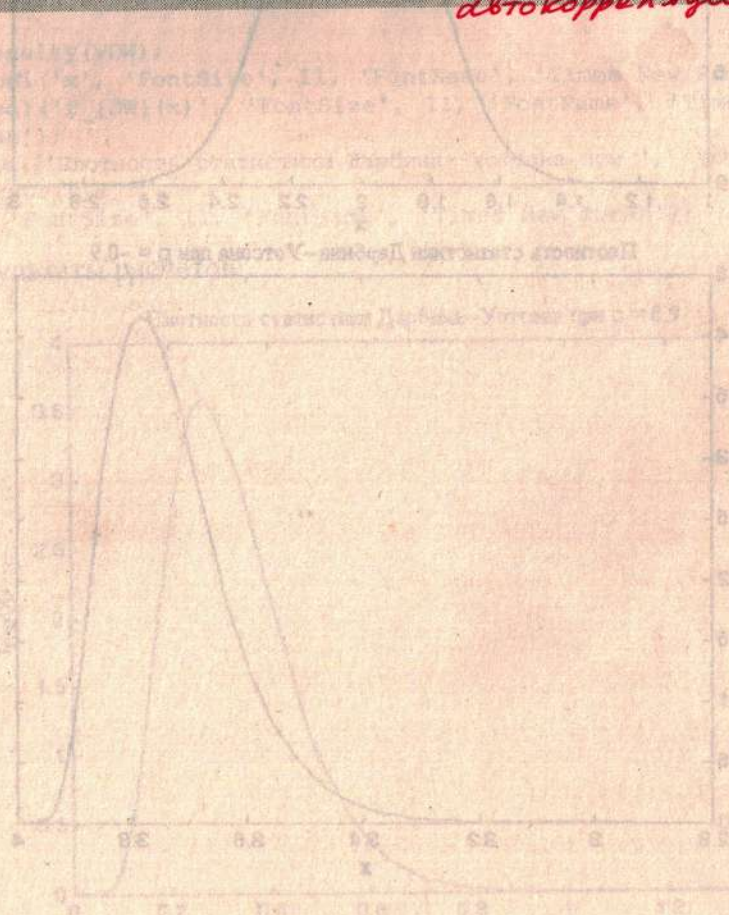
(a) Найдите $\mathrm{D}[\varepsilon_1]$ и $\mathrm{D}[\varepsilon_2]$.

(b) Являются ли ошибки $\{\varepsilon_t\}_{t=1}^n$ гетероскедастичными?

Задача 14. Объясните, с какой целью используются стандартные ошибки в форме Невье—Веста. Приведите развернутый ответ. Верно ли, что стандартные ошибки в форме Невье—Веста позволяют

(a) устранить ~~гетероскедастичность~~ автокорреляцию?

(b) корректно тестировать гипотезы относительно коэффициентов регрессии в условиях ~~гетероскедастичности~~ автокорреляции?



Глава 7

Тесты на правильную спецификацию модели: тест Рамсея и тест Бокса—Кокса

Задача 1. По 25 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, для которой $RSS = 73$. При помощи вспомогательной регрессии $\hat{\hat{Y}} = \hat{\gamma} + \hat{\delta}_1 x_1 + \hat{\delta}_2 x_2 + \hat{\delta}_3 \hat{Y}^2$, для которой $RSS = 70$, выполните тест Рамсея на уровне значимости 5 %.

Решение.

$H_0 : \delta_3 = 0 \Leftrightarrow$ (пропущенных переменных нет),

$H_1 : \delta_3 \neq 0 \Leftrightarrow$ (пропущенные переменные есть).

1) Тестовая статистика:

$$T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)},$$

где RSS_R — сумма квадратов остатков в модели с ограничениями $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, RSS_{UR} — сумма квадратов остатков в модели без ограничений

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_3 \hat{Y}_i^2 + \varepsilon_i,$$

q — число уравнений в гипотезе H_0, n — число наблюдений, k — число переменных в модели без ограничений. В нашем случае $q=1$, $n=25$, $k=3$.

2) Распределение тестовой статистики:

$$T \overset{H_0}{\sim} F(q, n-k-1) = F(1, 25-3-1) = F(1, 21).$$

3) Наблюдаемое значение тестовой статистики

$$T_{\text{набл}} = \frac{(73-70)/1}{70/(25-3-1)} = 0.9.$$

4) Область, в которой H_0 не отвергается:

$$[0; T_{\text{кр}}] = [0; \text{finv}(0.95, 1, 21)] = [0; 4.32].$$

5) Статистический вывод: поскольку $T_{\text{набл}} \in [0; T_{\text{кр}}]$, гипотеза H_0 не может быть отвергнута. Стало быть, пропущенных переменных нет. □

Задача 2. По 20 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, для которой $R^2 = 0.7$. При помощи вспомогательной регрессии $\hat{\hat{Y}} = \hat{\gamma} + \hat{\delta}_1 x_1 + \hat{\delta}_2 x_2 + \hat{\delta}_3 \hat{Y}^2$, для которой $R^2 = 0.75$, выполните тест Рамсея на уровне значимости 5 %.

Указание: тестовая статистика $T = \dfrac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)}$ после деления на TSS числителя и знаменателя дроби может быть приведена к виду $T = \dfrac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(n-k-1)}$. В остальном задача 2 решается так же, как и задача 1.

Задача 3. По 30 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, для которой $RSS = 150$. При помощи вспомогательной регрессии $\hat{\hat{Y}} = \hat{\gamma} + \hat{\delta}_1 x_1 + \hat{\delta}_2 x_2 + \hat{\delta}_3 \hat{Y}^2 + \hat{\delta}_4 \hat{Y}^3$, для которой $RSS = 120$, выполните тест Рамсея на уровне значимости 5 %.

Решение.

$$H_0: \begin{cases} \delta_3 = 0, \\ \delta_4 = 0 \end{cases} \Leftrightarrow \text{(пропущенных переменных нет)},$$

$$H_1: \begin{cases} \delta_3 \neq 0, \\ \delta_4 \neq 0 \end{cases} \Leftrightarrow \text{(пропущенные переменные есть)}.$$

квадратная скобка

1) Тестовая статистика:

$$T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)},$$

где RSS_R — сумма квадратов остатков в модели с ограничениями $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, RSS_{UR} — сумма квадратов остатков в модели без ограничений

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_3 \hat{Y}_i^2 + \delta_4 \hat{Y}_i^3 + \varepsilon_i,$$

q — число уравнений в гипотезе H_0, n — число наблюдений, k — число переменных в модели без ограничений. В нашем случае $q=2$, $n=30$, $k=4$.

2) Распределение тестовой статистики:

$$T \overset{H_0}{\sim} F(q, n-k-1) = F(2, 30-4-1) = F(2, 25).$$

3) Наблюдаемое значение тестовой статистики

$$T_{\text{набл}} = \frac{(150-120)/2}{120/(30-4-1)} = 3.125$$

120

Решение. Функция правдоподобия имеет вид

$$\mathcal{L}(x_1,\ldots,x_n;\lambda)=\prod_{i=1}^{n} f_{X_i}(x_i;\lambda)=\prod_{i=1}^{n}\lambda e^{-\lambda x_i}=\lambda^n e^{-\lambda(x_1+\ldots+x_n)}.$$

Логарифмическая функция правдоподобия

$$l(x_1,\ldots,x_n;\lambda):=\ln\mathcal{L}(x_1,\ldots,x_n;\lambda)=n\ln\lambda-\lambda(x_1+\ldots+x_n).$$

Решая уравнение правдоподобия

$$\frac{\partial l}{\partial\lambda}=\frac{n}{\lambda}-(x_1+\ldots+x_n)=0$$

получаем $\hat{\lambda}_{ML}=\dfrac{n}{x_1+\ldots+x_n}=\dfrac{100}{25}=4$. □

Задача 3. Пусть p — неизвестная вероятность выпадения орла при бросании монеты. Из 100 испытаний 42 раза выпал «орел» и 58 — «решка». При помощи метода максимального правдоподобия найдите оценку неизвестного параметра p.

Решение. Функция правдоподобия имеет вид

$$\mathcal{L}(x_1,\ldots,x_n;p)=\prod_{i=1}^{n}\mathbb{P}_p\left(\{X_i=x_i\}\right)=$$

$$=\prod_{i=1}^{n}p^{x_i}\cdot(1-p)^{1-x_i}=p^{\sum_{i=1}^{n}x_i}\cdot(1-p)^{n-\sum_{i=1}^{n}x_i}.$$

Логарифмическая функция правдоподобия

$$l(x_1,\ldots,x_n;p):=\ln\mathcal{L}(x_1,\ldots,x_n;p)=$$

$$=\left(\sum_{i=1}^{n}x_i\right)\cdot\ln p+\left(n-\sum_{i=1}^{n}x_i\right)\cdot\ln(1-p).$$

Решая уравнение правдоподобия

$$\frac{\partial l}{\partial p}=\frac{\sum_{i=1}^{n}x_i}{p}-\frac{n-\sum_{i=1}^{n}x_i}{1-p}=0,$$

получаем $\hat{p}_{ML}=\overline{x}=0.42$. □

Задача 4. Пусть $x=(x_1,\ldots,x_n)$ — реализация случайной выборки из распределения Пуассона с неизвестным параметром $\lambda>0$. Известно, что выборочное среднее $\overline{x}$ по 80 наблюдениям равно 1.7. При помощи метода максимального правдоподобия найдите оценку неизвестного параметра λ.

Решение. Функция правдоподобия имеет вид

$$\mathcal{L}(x_1,\ldots,x_n;\lambda)=\prod_{i=1}^{n}\mathbb{P}_\lambda\left(\{X_i=x_i\}\right)=\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!}e^{-\lambda}=\frac{\lambda^{\sum_{i=1}^{n}x_i}}{x_1!\ldots x_n!}e^{-\lambda n}.$$

Логарифмическая функция правдоподобия

$$l(x_1,\ldots,x_n;\lambda):=\ln\mathcal{L}(x_1,\ldots,x_n;\lambda)=\left(\sum_{i=1}^{n}x_i\right)\cdot\ln\lambda-\sum_{i=1}^{n}\ln(x_i!)-\lambda n.$$

Решая уравнение правдоподобия

$$\frac{\partial l}{\partial\lambda}=\frac{\sum_{i=1}^{n}x_i}{\lambda}-n=0,$$

получаем $\hat{\lambda}_{ML}=\overline{x}=1.7$. □

Задача 5. Рассматривается модель линейной регрессии $Y_i=\alpha+\varepsilon_i$, $i=1,\ldots,n$, где случайные ошибки $\varepsilon_1,\ldots,\varepsilon_n$ являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием и дисперсией σ^2. При помощи метода максимального правдоподобия найдите оценки неизвестных параметров α и σ^2.

Решение. Поскольку $\varepsilon_i\sim N(0,\sigma^2)$, то

$$Y_i=\alpha+\varepsilon_i\sim N(\alpha,\sigma^2),$$

а значит, логарифмическая функция правдоподобия

$$l(y_1,\ldots,y_n;\beta,\sigma^2) = \\ = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{(y_1-\beta x_1)^2+\ldots+(y_n-\beta x_n)^2}{2\sigma^2}.$$

Найдем точку максимума логарифмической функции правдоподобия:

$$\begin{cases} \dfrac{\partial l}{\partial \beta} = 0, \\ \dfrac{\partial l}{\partial \sigma^2} = 0; \end{cases} \Leftrightarrow \begin{cases} \dfrac{\partial l}{\partial \beta} = -\dfrac{-2x_1(y_1-\beta x_1)-\ldots-2x_n(y_n-\beta x_n)}{2\sigma^2} = 0, \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{(y_1-\beta x_1)^2+\ldots+(y_n-\beta x_n)^2}{2\sigma^4} = 0. \end{cases}$$

Решением первого уравнения системы является

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Выражая из второго уравнения системы параметр σ^2 и подставляя в полученную формулу вместо параметра β найденную выше оценку $\widehat{\beta}$, приходим к выражению для оценки параметра σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{\beta}x_i)^2.$$

Таким образом, оценки параметров β и σ^2 по методу максимального правдоподобия имеют вид:

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \text{ и } \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{\beta}x_i)^2 = \frac{RSS}{n}. \square$$

Задача 7. Рассматривается модель линейной регрессии $Y_i = \beta / x_i + \varepsilon_i$, $i = 1,\ldots,n$, где случайные ошибки $\varepsilon_1,\ldots,\varepsilon_n$ являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием и дисперсией σ^2. При помощи метода максимального правдоподобия найдите оценки неизвестных параметров β и σ^2.

Ответ: $\widehat{\beta} = \dfrac{\sum_{i=1}^n \frac{Y_i}{x_i}}{\sum_{i=1}^n \frac{1}{x_i^2}}$, $\widehat{\sigma^2} = \dfrac{1}{n}\sum_{i=1}^n \left(Y_i - \dfrac{\widehat{\beta}}{x_i}\right)^2 = \dfrac{RSS}{n}$.

Задача 8. Рассматривается модель линейной регрессии $Y_i = \theta x_{i1} + (1-\theta)x_{i2} + \varepsilon_i$, $i = 1,\ldots,n$, где случайные ошибки $\varepsilon_1,\ldots,\varepsilon_n$ являются независимыми нормально распределенными случайными величинами с нулевым математическим ожиданием и дисперсией σ^2. При помощи метода максимального правдоподобия найдите оценки неизвестных параметров θ и σ^2.

Решение. Поскольку $\varepsilon_i \sim N(0,\sigma^2)$, то

$$Y_i = \theta x_{i1} + (1-\theta)x_{i2} + \varepsilon_i \sim N(\theta x_{i1} + (1-\theta)x_{i2}, \sigma^2),$$

а значит,

$$f_{Y_i}(y_i;\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\theta x_{i1}-(1-\theta)x_{i2})^2}{2\sigma^2}}, \ i = 1,\ldots,n.$$

Задача 15. Винни-Пух знает, что мёд бывает правильный, $honey_i = 1$, и неправильный, $honey_i = 0$. Пчелы также бывают правильные, $bee_i = 1$, и неправильные, $bee_i = 0$. По 100 своим попыткам добыть мёд Вини-Пух составил таблицу сопряженности:

	$honey_i = 1$	$honey_i = 0$
$bee_i = 1$	12	36
$bee_i = 0$	32	20

Используя метод максимального правдоподобия Винни-Пух хочет оценить логит-модель для прогнозирования правильности мёда с помощью правильности пчёл:

$$\ln\left(\frac{\mathbb{P}\{honey_i = 1\}}{\mathbb{P}\{honey_i = 0\}}\right) = \beta_1 + \beta_2 bee_i.$$

(a) Выпишите функцию правдоподобия для оценки параметров β_1 и β_2.

(b) Оцените неизвестные параметры.

(c) С помощью теста отношения правдоподобия проверьте гипотезу о том, что правильность мёда не связана с правильностью пчёл на уровне значимости 5 %.

(d) Держась в небе за воздушный шарик, Винни-Пух неожиданно понял, что перед ним неправильные пчёлы. Помогите ему оценить вероятность того, что они делают неправильный мёд.

(e) Найдите информационную матрицу Фишера $I_n(\theta)$ при помощи формулы

$$I_n(\theta) = \mathbb{E}\left[\sum_{i=1}^{n}\left[\frac{\partial l_i}{\partial \theta}\right]\left[\frac{\partial l_i}{\partial \theta}\right]^T\right],$$

где $\theta = (\beta_1, \beta_2)^T$ и $l_i(\theta) := \ln \mathbb{P}\left(\{Y_i = y_i\}\right)$.

(f) Найдите $I_n(\hat{\theta}_{UR})$.

(g) Найдите $I_n(\hat{\theta}_R)$.

(h) Найдите $\hat{\mathrm{V}}(\hat{\theta}) = I_n^{-1}(\hat{\theta}_{UR})$ — оценку ковариационной матрицы для $\hat{\theta}$.

(i) Рассчитайте значение статистики Вальда для проверки гипотезы о том, что правильность мёда не связана с правильностью пчёл.

(j) Рассчитайте значение статистики множителей Лагранжа для проверки гипотезы о том, что правильность мёда не связана с правильностью пчёл.

Решение. Для краткости введем следующие обозначения: $y_i = honey_i$, $d_i = bee_i$.

(a) Функция правдоподобия имеет следующий вид:

$$\mathcal{L}(y_1,\ldots,y_n;\beta_1,\beta_2) =$$

$$= \prod_{i=1}^{n} \mathbb{P}_{\beta_1,\beta_2}\left(\{Y_i = y_i\}\right) = \prod_{i:y_i=1} \mathbb{P}_{\beta_1,\beta_2}\left(\{Y_i = 1\}\right) \cdot \prod_{i:y_i=0} \mathbb{P}_{\beta_1,\beta_2}\left(\{Y_i = 0\}\right) =$$

$$= \prod_{i:y_i=1} \Lambda(\beta_1 + \beta_2 d_i) \cdot \prod_{i:y_i=0} \left[1 - \Lambda(\beta_1 + \beta_2 d_i)\right] =$$

$$= \prod_{i:y_i=1,d_i=1} \Lambda(\beta_1 + \beta_2) \cdot \prod_{i:y_i=1,d_i=0} \Lambda(\beta_1) \cdot$$

$$\cdot \prod_{i:y_i=0,d_i=1} \left[1 - \Lambda(\beta_1 + \beta_2)\right] \cdot \prod_{i:y_i=0,d_i=0} \left[1 - \Lambda(\beta_1)\right] =$$

$$= \Lambda(\beta_1 + \beta_2)^{\#\{i:y_i=1,d_i=1\}} \cdot \Lambda(\beta_1)^{\#\{i:y_i=1,d_i=0\}} \cdot$$

$$\cdot \left[1 - \Lambda(\beta_1 + \beta_2)\right]^{\#\{i:y_i=0,d_i=1\}} \cdot \left[1 - \Lambda(\beta_1)\right]^{\#\{i:y_i=0,d_i=0\}},$$

$$\cdot(0-\Lambda(\hat\beta_{1,R}+\hat\beta_{2,R}\cdot 1))\begin{bmatrix}1\\1\end{bmatrix}+\#\{i: d_i=1, y_i=1\}\cdot$$

$$\cdot(1-\Lambda(\hat\beta_{1,R}+\hat\beta_{2,R}\cdot 1))\begin{bmatrix}1\\1\end{bmatrix}=20\cdot(0-\Lambda(-0.24))\begin{bmatrix}1\\0\end{bmatrix}+$$

$$+32\cdot(1-\Lambda(-0.24))\begin{bmatrix}1\\0\end{bmatrix}+36\cdot(0-\Lambda(-0.24))\begin{bmatrix}1\\1\end{bmatrix}+$$

$$+12\cdot(1-\Lambda(-0.24))\begin{bmatrix}1\\1\end{bmatrix}=\begin{bmatrix}-0.03\\-9.13\end{bmatrix}.$$

Таким образом,

$$LM=\left[\left.\frac{\partial l_{UR}}{\partial\theta}\right|_{\theta=\hat\theta_R}\right]^T\cdot I^{-1}(\hat\theta_R)\cdot\left[\left.\frac{\partial l_{UR}}{\partial\theta}\right|_{\theta=\hat\theta_R}\right]=$$

$$=\begin{bmatrix}-0.03 & -9.13\end{bmatrix}\cdot\begin{bmatrix}24.64 & 11.83\\11.83 & 11.83\end{bmatrix}^{-1}\cdot\begin{bmatrix}-0.03\\-9.13\end{bmatrix}=13.52.\ \square$$

Задача 16. Винни-Пух знает, что мёд бывает правильный, $honey_i=1$, и неправильный, $honey_i=0$. Пчелы также бывают правильные, $bee_i=1$, и неправильные, $bee_i=0$. По 100 своим попыткам добыть мёд Вини-Пух составил таблицу сопряженности:

	$honey_i=1$	$honey_i=0$
$bee_i=1$	40	10
$bee_i=0$	20	30

(a) При помощи метода максимального правдоподобия найдите оценки β_1 и β_2 в logit-модели

$$\ln\left(\frac{\mathbb{P}\{honey_i=1\}}{\mathbb{P}\{honey_i=0\}}\right)=\beta_1+\beta_2 bee_i.$$

(b) При помощи теста отношения правдоподобия протестируйте гипотезу

$$H_0:\begin{cases}\beta_1=0,\\\beta_2=0.\end{cases}$$

на уровне значимости 5 %.

Ответы:

(a) $\hat\beta_{1,UR}\approx -0.41$, $\hat\beta_{2,UR}\approx 1.80$,

(b) $LR_{\text{набл}}\approx 21.29$, $LR_{\text{кр}}\approx 5.99$, следовательно, гипотеза H_0 должна быть отвергнута.

Задача 17. Рассматривается модель

$$Y_t=\mu+\varepsilon_t, \qquad (4)$$

$t=1,\dots,T$, где $\varepsilon_t=\rho\varepsilon_{t-1}+u_t$, случайные величины $\varepsilon_0,u_1,\dots,u_T$ независимы, причем $\varepsilon_0\sim N(0,\sigma^2/(1-\rho^2))$, $u_t\sim N(0,\sigma^2)$.

(a) Выпишите функцию правдоподобия

$$\mathcal{L}(\mu,\rho,\sigma^2)=f_{Y_1}(y_1)\prod_{t=2}^{T}f_{Y_t|Y_{t-1}}(y_t\,|\,y_{t-1}).$$

(b) Для вектора наблюдений $y=(1,2,0,0,1)$ при помощи условной функции правдоподобия

$$\mathcal{L}(\mu,\rho,\sigma^2\,|\,Y_1=y_1)=\prod_{t=2}^{T}f_{Y_t|Y_{t-1}}(y_t\,|\,y_{t-1}),$$

найдите оценки неизвестных параметров модели.

(c) Напишите программу, которая по заданному числу наблюдений T и вектору параметров $\theta=(\mu,\rho,\sigma^2)$ генерирует вектор y длины T согласно модели (4).

Задача 19. Рассматривается модель

$$y_t = \mu + \beta_1 x_{t1} + \dots \beta_k x_{tk} + \varepsilon_t, \qquad (5)$$

$t = 1,\dots,T$, где $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, случайные величины $\varepsilon_0, u_1, \dots, u_T$ независимы, причем $\varepsilon_0 \sim N(0, \sigma^2/(1-\rho^2))$, $u_t \sim N(0,\sigma^2)$.

(a) Выпишите условную логарифмическую функцию правдоподобия

$$l(\mu,\rho,\sigma^2,\beta \mid Y_1 = y_1) = \sum_{t=2}^{T} \ln f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1}).$$

(b) Напишите программу, которая по заданной матрице $X \in \mathbb{R}^{T\times k}$ и вектору параметров

$$\theta = (\mu,\rho,\sigma^2,\beta^T)^T \in \mathbb{R}^{(k+3)\times 1}$$

генерирует вектор y длины T согласно модели (5).

(c) Напишите программу, которая по заданной матрице X, вектору параметров θ и вектору наблюдений y вычисляет значение условной функции правдоподобия.

(d) Напишите программу, которая по заданной матрице X и вектору наблюдений y вычисляет оценку вектора параметров θ по методу условного максимального правдоподобия и находит значение условной функции правдоподобия в точке максимума.

Рассматривается частный случай модели (5):

$$y_t = \mu + \beta_1 t + \varepsilon_t,\ t = 1,\dots,10.$$

Дан вектор наблюдений $y = (3,6,7,9,10,12,15,18,20,22)$.

(e) Найдите $\hat{\mu}_{ML}$, $\hat{\rho}_{ML}$, $\hat{\sigma}^2_{ML}$, $\hat{\beta}_1$.

При помощи теста отношения правдоподобия проверьте гипотезы:

(f) $H_0: \begin{cases} \mu = 0, \\ \rho = 0, \\ \sigma = 1, \\ \beta_1 = 0, \end{cases}$ (g) $H_0: \beta_1 = 0$, (h) $H_0: \rho = 0$.

Решение. (a) Аналогично тому, как это сделано в задаче 17, получаем условную логарифмическую функцию правдоподобия:

$$l(\mu,\rho,\sigma^2,\beta \mid Y_1 = y_1) = \sum_{t=2}^{T} \ln f_{Y_t|Y_{t-1}}(y_t \mid y_{t-1}) =$$

$$= -\frac{T-1}{2}\ln(2\pi) - \frac{T-1}{2}\ln\sigma^2 -$$

$$-\frac{1}{2\sigma^2}\sum_{t=2}^{T}(y_t - \rho y_{t-1} - (1-\rho)(\mu + \beta_1 x_{t1} + \dots + \beta_k x_{tk}))^2.$$

(b) Код программы в среде MATLAB:

```
function [y] = get_yX(X, theta)
[~,m] = size(theta);
if (m ~= 1)
    error('вектор theta должен быть столбцом');
end
[T, k] = size(X);
X = [zeros(1,k); X];
y = zeros(T+1,1);
e = zeros(T+1,1);
e(1) = sqrt(theta(3) / (1 - theta(2)^2)) * randn;
u = sqrt(theta(3)) * randn(T+1,1);
for t = 2:(T+1)
    e(t) = theta(2) * e(t-1) + u(t);
    y(t) = theta(1) + X(t,:) * theta(4:end,1) + e(t);
end
y = y(2:end,1);
```

(c) Код программы в среде MATLAB:

```
function [l] = get_LLF_X(theta, y, X)
[~,m] = size(theta);
if (m ~= 1)
    error('вектор theta должен быть столбцом');
end
T = length(y);
l = 0;
for t = 2:T
    l = l + (y(t) - theta(2) * y(t-1) - (1 - theta(2)) *
(theta(1) + X(t,:) * theta(4:end,1)))^2;
end
l = -0.5 * (T - 1) * log(2 * pi) - 0.5 * (T - 1) *
log(theta(3)) - 0.5 * l / theta(3);
```

(d) Код программы в среде MATLAB:

```
function [theta_ML, LLF] = get_theta_ML_X(y, X)
[~, k] = size(X);
fun = @(theta) get_NegLLF_X(theta, y, X);
x0 = [0, 0, 1, zeros(1, k)]';
A = [];
b = [];
Aeq = [];
beq = [];
lb = [-10000, -0.9999, 0.0001,   -10000 * ones(1,k)]';
ub = [10000,   0.9999,    10000,  10000 * ones(1,k)]';
nonlcon = [];
options = optimset('Algorithm', 'interior-point',
'Display', 'off');
[theta_ML, NegLLF] =
fmincon(fun,x0,A,b,Aeq,beq,lb,ub,nonlcon,options);
LLF = -NegLLF;

% Вспомогательная функция: отрицательная логарифмическая
функция правдоподобия
function [l] = get_NegLLF_X(theta, y, X)
[~,m] = size(theta);
if (m ~= 1)
    error('вектор theta должен быть столбцом');
end
T = length(y);
l = 0;
for t = 2:T
```



```
    l = l + (y(t) - theta(2) * y(t-1) - (1 - theta(2)) *
(theta(1) + X(t,:) * theta(4:end,1)))^2;
end
l = -0.5 * (T - 1) * log(2 * pi) - 0.5 * (T - 1) *
log(theta(3)) - 0.5 * l / theta(3);
l = -l;
```

(e) **Ответ:**

$$\hat{\mu}_{ML} = 2.6068,\ \hat{\rho}_{ML} = 0.5228,\ \hat{\sigma}^2_{ML} = 0.4066,\ \hat{\beta}_1 = 2.1547.$$

Задача 20. Рассматривается модель

$$y_t = \mu + \varepsilon_t, \qquad (6)$$

$t = 1,\dots,T$, где $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + u_t$, случайные величины $\varepsilon_{-1}, \varepsilon_0, u_1,\dots,u_T$ независимы,

$\varepsilon_{-1} \sim N(0, \sigma^2/(1-\rho_1^2-\rho_2^2))$, $\varepsilon_0 \sim N(0, \sigma^2/(1-\rho_1^2-\rho_2^2))$, $u_t \sim N(0,\sigma^2)$, и на параметры ρ_1 и ρ_2 накладываются следующие ограничения:

$$\begin{cases} \rho_2 < 1-\rho_1, \\ \rho_2 < 1+\rho_1, \\ -1 < \rho_2 < 1. \end{cases}$$

(a) Выпишите условную логарифмическую функцию правдоподобия

$$l(\mu,\rho_1,\rho_2,\sigma^2 \mid Y_1 = y_1, Y_2 = y_2) = \sum_{t=3}^{T} \ln f_{Y_t|Y_{t-1},Y_{t-2}}(y_t \mid y_{t-1}, y_{t-2}).$$

(b) Напишите программу, которая по заданному числу наблюдений T и вектору параметров $\theta = (\mu,\rho_1,\rho_2,\sigma^2)$ генерирует вектор y длины T согласно модели (6).

(c) Напишите программу, которая по заданному вектору параметров θ и вектору наблюдений y вычисляет

$$\begin{cases} \dfrac{\partial l}{\partial \alpha} = 0, \\ \dfrac{\partial l}{\partial \sigma^2} = 0; \end{cases} \Leftrightarrow \begin{cases} \dfrac{\partial l}{\partial \beta} = -\dfrac{-2x_1(y_1-\beta x_1)-\ldots-2x_n(y_n-\beta x_n)}{2\sigma^2} = 0, \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{(y_1-\beta x_1)^2+\ldots+(y_n-\beta x_n)^2}{2\sigma^4} = 0. \end{cases}$$

Решением первого уравнения системы является

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Выражая из второго уравнения системы параметр σ^2 и подставляя в полученную формулу вместо параметра β найденную выше оценку $\hat{\beta}$, приходим к выражению для оценки параметра σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{\beta}x_i)^2.$$

Таким образом, оценки параметров β и σ^2 по методу максимального правдоподобия имеют вид:

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2} \text{ и } \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}x_i)^2 = \frac{RSS}{n}.$$

Следовательно, $\hat{\beta} = 1.6428$, $\widehat{\sigma^2} = 0.0714$.

Глава 10

Бутстрап

Задача 1. Рассматривается модель линейной регрессии $Y_i = \beta x_i + \varepsilon_i$. Имеются следующие наблюдения

$$x_1 = 1,\ x_2 = 2,\ y_1 = 3,\ y_2 = 4.$$

(a) Постройте таблицу распределения бутстраповской оценки $\hat{\beta}_*$.

(b) Найдите математическое ожидание бутстраповской оценки $\hat{\beta}_*$.

(c) Постройте функцию распределения бутстраповской оценки $\hat{\beta}_*$.

(d) Для найденной в предыдущем пункте функции распределения $F_{\hat{\beta}_*}(x)$ найдите квантили уровней: 0.025, 0.1, 0.3, 0.8, 0.975.

(e) Для неизвестного параметра β постройте 95 %-й бутстраповский доверительный интервал.

(f) При помощи доверительного интервала, полученного в предыдущем пункте, протестируйте гипотезу о значимости коэффициента β на уровне значимости 5 %.

x	y
1	2
3	4
5	6
7	8

С помощью компьютера выполните следующие задания;

(a) Найдите приближенно математическое ожидание бутстраповской оценки $\hat{\beta}_*$.

(b) Для неизвестного параметра β постройте приближенный 95 %-й бутстраповский доверительный интервал.

(c) При помощи доверительного интервала, полученного в предыдущем пункте, протестируйте гипотезу о значимости коэффициента β на уровне значимости 5 %.

Решение. Код программы в среде MATLAB:

```
clear; clc;

X = [1 3 5 7]';
Y = [2 4 6 8]';

Z = [Y, X];

SL = 0.05;
[n, K] = size(X);

S = 1000000;
b_BOOT = zeros(K, S);
for s = 1:S
    Z_boot = Z(unidrnd(n, 1, n), :);
    Y_boot = Z_boot(:,1);
    X_boot = Z_boot(:, 2:end);
    b_BOOT(:,s) = (X_boot' * X_boot)^(-1) * (X_boot' * Y_boot);
end

mean_b_boot = mean(b_BOOT);

CI_b_boot = zeros(K,2);
for j = 1:K
    CI_b_boot(j,1) = quantile(b_BOOT(j,:),SL/2);
    CI_b_boot(j,2) = quantile(b_BOOT(j,:),1 - SL/2);
end
```


Ответы:

(a) $\mathbb{E}[\hat{\beta}_*] \approx 1.22$; (b) $\left[Q_{\hat{\beta}_*}(0.025);\ Q_{\hat{\beta}_*}(0.975)\right] \approx [1.15; 1.40]$;

(c) $0 \notin \left[Q_{\hat{\beta}_*}(0.025);\ Q_{\hat{\beta}_*}(0.975)\right]$, следовательно, коэффициент β значим.

Задача 12. Рассматривается модель линейной регрессии $Y_i = \beta x_i + \varepsilon_i$. В следующей таблице приведены наблюдения

x	y
1	2
3	4
5	6
7	8
9	10

С помощью компьютера выполните следующие задания;

(a) Найдите приближенно математическое ожидание бутстраповской оценки $\hat{\beta}_*$.

(b) Для неизвестного параметра β постройте приближенный 95 %-й бутстраповский доверительный интервал.

(c) При помощи доверительного интервала, полученного в предыдущем пункте, протестируйте гипотезу о значимости коэффициента β на уровне значимости 5 %.

Ответы:

(a) $\mathbb{E}[\hat{\beta}_*] \approx 1.165$; (b) $\left[Q_{\hat{\beta}_*}(0.025);\ Q_{\hat{\beta}_*}(0.975)\right] \approx [1.12; 1.29]$;

(c) $0 \notin \left[Q_{\hat{\beta}_*}(0.025);\ Q_{\hat{\beta}_*}(0.975)\right]$, следовательно, коэффициент β значим.