1 Контрольная работа 2

На выполнение заданий отводится 50 минут. Разрешается пользоваться калькуляторами, компьютерами, любыми материалами. Не разрешается совместное (коллективное) решение задач. Для зачтения контрольной нужно получить не менее 6 баллов из 10.

1.1 Разбор пробного варианта

Задача 1 (1б). 110 точек вида (y, x_1, x_2) описано зависимостью вида $y = ax_1 + bx_2 + c$. Рассчитайте число степеней свободы. Найдите коэффициенты f_1 и f_2 в $F_{\text{crit}}(p, f_1, f_2)$.

Решение. Число коэффициентов модели $k = 3$, число точек $n = 110$. Число степеней свободы $f = n - k = 107$.

Выражение для F_{emp} выглядит следующим образом:

$$F_{\text{emp}} = \frac{ESS/f_1}{RSS/f_2} < F_{\text{crit}}(p, f_1, f_2) \tag{1}$$

В нашем случае $f_1 = k - 1 = 2$ и $f_2 = n - k = 107$.

Задача 2 (1б). Что такое мультиколлинеарность и как она влияет на результаты многомерной линейной регрессии?

Решение. Мультиколлинеарность — наличие линейной зависимости между объясняющими переменными (факторами) регрессионной модели. Выраженная мультиколлинеарность может приводить к плохой обусловленности системы линейных уравнений в МНК и вести к большим погрешностям в коэффициентах модели и их неустойчивости по отношению к небольшим изменениям в исходных данных.

Задача 3 (2б). Пусть в нелинейной регрессии функция имеет вид $y = \beta_1 + \beta_2 \ln(\beta_3 x)$. Запишите якобиан $J_{ij} = \frac{\partial f(x_i)}{\partial \beta_j}$.

Решение. Найдём производные функции $y(\beta)$ по переменным β_i:

$$\begin{cases} \frac{\partial y}{\partial \beta_1} = 1 \\ \frac{\partial y}{\partial \beta_2} = \ln(\beta_3 x) \\ \frac{\partial y}{\partial \beta_3} = \beta_2 \left(\frac{1}{x\beta_3} \cdot x\right) = \frac{\beta_2}{\beta_3} \end{cases} \tag{2}$$

Запишем матрицу Якоби

$$J = \begin{pmatrix} 1 & \ln(\beta_3 x_1) & \frac{\beta_2}{\beta_3} \\ 1 & \ln(\beta_3 x_2) & \frac{\beta_2}{\beta_3} \\ \vdots & \ddots & \vdots \\ 1 & \ln(\beta_3 x_n) & \frac{\beta_2}{\beta_3} \end{pmatrix} \tag{3}$$

Задача 4 (1б). Линеаризовать функцию $y = a\exp(bx)$. Указать независимую и зависимую переменную после преобразования.

Решение. Прологарифмируем функцию:

$$\ln y = \ln a + bx \tag{4}$$

Новые переменные: $\ln y$ (зависимая) и x (независимая).

Задача 5 (1б). Линеаризовать функцию $C_p(T) = AT + B/T + C$. Выписать матрицы X и y из системы уравнений $X\hat{\beta} + e = y$.

Решение. Перейдём от одномерной нелинейной регрессии к многомерной линейной, используя следующие подстановки:

$$\begin{cases} y = C_p \\ x_1 = T \\ x_2 = T^{-1} \end{cases} \tag{5}$$

Матрицы X, y и $\hat{\beta}$ из переопределённой системы уравнений $X\hat{\beta}+e = y$ (столбец из единиц соответствует константе C):

$$X = \begin{pmatrix} T_1 & T_2^{-1} & 1 \\ T_2 & T_2^{-1} & 1 \\ \vdots & \ddots & \vdots \\ T_n & T_n^{-1} & 1 \end{pmatrix};\ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix};\ \hat{\beta} = \begin{pmatrix} \hat{A} \\ \hat{B} \\ \hat{C} \end{pmatrix} \tag{6}$$

Задача 6 (4б). Описать данные методом наименьших квадратов, используя зависимость $y = ax$. Записать сумму квадратов отклонений RSS в явном виде и решением уравнения $\frac{\partial RSS}{\partial a} = 0$ получить выражения и численные значения для a (значения коэффициента регрессии) и s_a (его стандартного отклонения).

```
X = [-2, -1, 0, 1, 2]
Y = [-3.93, -1.55, -0.60, 3.59, 4.43]
```

Решение. Записать сумму квадратов отклонений RSS в явном виде и решением уравнения $\frac{\partial RSS}{\partial a} = 0$ получить выражение для a.

Найдём производную суммы квадратов отклонений по параметру модели:

$$\frac{\partial RSS}{\partial a} = \frac{\partial \sum_i (\hat{y}_i - y_i)^2}{\partial a} = \frac{\partial \sum_i (ax_i - y_i)^2}{\partial a} = \sum_i 2\,(ax_i - y_i)\,x_i = 0 \Leftrightarrow$$
$$a \sum_i x_i^2 - \sum_i x_i y_i = 0 \Leftrightarrow a = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{\overline{xy}}{\overline{x^2}} \tag{7}$$

Рассчитаем значения $\overline{x^2}$, $\overline{xy}$ и коэффициента a

$$\overline{x^2} = \frac{1}{5}\left((-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\right) = \frac{2\cdot 4 + 2\cdot 1}{5} = \frac{10}{5} = 2 \tag{8}$$

$$\overline{xy} = \frac{1}{5}\left(-2\cdot(-3.93) - 1\cdot(-1.55) + 0\cdot(-0.60) + 1\cdot(3.59) + 2\cdot(4.43)\right) =$$
$$\frac{2\cdot 3.93 + 1.55 + 3.59 + 2\cdot 4.43}{5} = 4.372 \tag{9}$$

$$a = \frac{\overline{xy}}{\overline{x^2}} = \frac{4.372}{2} = 2.186 \tag{10}$$

Найдём стандартное отклонение s_a:

$$s_a^2 = \hat{\sigma}^2 \left(X^\top X\right)^{-1} = \frac{\sum_i \left(\hat{y}_i - y_i\right)^2}{n-1} \cdot \left(\begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \right)^{-1} =$$

$$\frac{\sum_i \left(\hat{y}_i - y_i\right)^2}{(n-1)\sum_i x_i^2} = \frac{(0.442^2 + 0.636^2 + 0.60^2 + 1.404^2 + 0.058^2)}{4 \cdot 10} = \frac{2.934}{40} = 0.07335 \Rightarrow s_a = 0.271 \tag{11}$$

1.2 Полный список теоретических вопросов (задача 2)

- Что такое мультиколлинеарность и как она влияет на результаты многомерной линейной регрессии?
- Означает ли нулевой коэффициент корреляции отсутствие зависимости между величинами? Ответ проиллюстрируйте графиком.
- Как проверить, является ли регрессионная зависимость статистически значимой?
- Как связаны коэффициент корреляции и коэффициент детерминации?
- Как проверить, является ли один из коэффициентов регрессии статистически значимым?
- Докажите, что формула для ковариационной матрицы коэффициентов регрессии $C = \operatorname{cov}\left(\hat{\beta}, \hat{\beta}\right) = \hat{\sigma}^2 \left(J^\top J\right)^{-1}$ (нелинейная регрессия) переходит в $C = \hat{\sigma}^2 \left(X^\top X\right)^{-1}$ в случае линейной регрессии.
- В чём преимущество метода Левенберга-Марквардта по сравнению с методом Гаусса-Ньютона?
- В чём разница между доверительным (confidence interval) и предсказательным (prediction interval) интервалом значения $\hat{y}$? Ответ проиллюстрируйте графиком.