<u>Задача 4</u>

№1

Классическая линейная нормальная регрессионная модель

Если регрессионная модель отвечает данным условиям:

1. $Y_i = \beta_1 + \beta_2 \cdot X_{2,i} + ... + \beta_k \cdot X_{k,i} + \varepsilon_i$ т.е. $y = X\beta + \varepsilon$
2. Все $X_{j,i}$ — детерминированы, нет линейной зависимости между объясняющими переменными. т.е. X − детерм. матрица, $rank(x) = k$
3.
    a. $E(\varepsilon_i) = 0\ \forall i \in [1;\ n]$ (дисперсия ошибки постоянна - гомоскедастичность)
    b. $D(\varepsilon_i) = \sigma^2{}_\varepsilon;\ Cov(\varepsilon_i,\ \varepsilon_j) = 0;\ i \neq j$
    c. $\varepsilon_i \sim N(0,\ \sigma^2{}_\varepsilon)$ и все ε_i независимы

то она называется классической линейной нормальной регрессионной моделью.

№2

Метод наименьших квадратов и теорема Гаусса-Маркова

МНК заключается в нахождении таких коэффициентов регрессии, при которых сумма квадратов ошибок будет наименьшей:

$$\sum_{i=1}^{n}(Y_i - (\hat{\beta}_1 + \hat{\beta}_2 \cdot X_{2,i} + ... + \hat{\beta}_k \cdot X_{k,i}))^2 \to 0$$

Берется частная производная по каждому коэффициенту, приравнивается к нулю. Из таких уравнений составляется и решается система.

Теорема Гаусса Маркова: при выполнении (1 - 3b) $\widehat{\beta} = (X^T X)^{-1} X^T Y$ - BLUE (**B**est **L**inear **U**nbiased **E**stimator) для β . т.е.

- Полученные оценки коэффициентов несмещенные
- Оценки состоятельны
- Оценки эффективны, т.е. имеют наименьшую дисперсию среди всех возможных линейных оценок
- Оценки распределены нормально

№3

Оценка дисперсии случайной составляющей и ковариационной матрицы оценок коэффициентов регрессии

Ответ:

$$\hat{\sigma}^2(\varepsilon) = \frac{RSS}{n-k}$$

k — количество оцениваемых коэффициентов

$$\hat{V}(\beta) = \hat{\sigma}^2(\varepsilon) \cdot (X^T \cdot X)^{-1}$$

№4

Коэффициент детерминации

Ответ: $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

Это доля дисперсии зависимой переменной, объяснённая моделью. Принимает значения от 0 до 1. Чем он выше, тем лучше подобрана модель и больше зависимость объясняемой переменной от объясняющих.

№5

Доверительный интервал для β_j ***с уровнем доверия*** $1 - \alpha$

$$\widehat{\beta}_j - t_{n-k,\frac{\alpha}{2}} \widehat{\sigma}_{\widehat{\beta}_i} < \beta_j < \widehat{\beta}_j + t_{n-k,\frac{\alpha}{2}} \widehat{\sigma}_{\widehat{\beta}_i}$$

№6

Проверка гипотезы о значении коэффициента и значимости регрессии в целом

Критерий значения коэффициента модели регрессии

H_0: $\beta_j = \beta_j{}^0$

Если $\beta_j{}^0 = 0$, то говорится, что проверяется **значимость** коэффициента

$$t = \frac{\hat{\beta}_j - \beta_j{}^0}{\sqrt{\hat{\sigma}^2(\beta_j)}} \sim t(n - \mathrm{k}),$$

k – кол-во коэффициентов (с учётом свободного)

$\beta_j{}^0$ – значение, на равенство которому проверяется выбранный коэффициент. Если равен 0, то тогда

$\hat{\beta}_j$ – оценка выбранного коэффициента

$\hat{\sigma}^2(\beta_j)$ – оценка дисперсии выбранного коэффициента (квадрат стандартной ошибки)

Критическое правило: $|t| > t_{n-k,\frac{\alpha}{2}} \Rightarrow H_0$ отвергается.

Критерий значимости модели регрессии в целом

H^0: $\beta_2 = \ldots = \beta_k = 0$ (т.е. хреновая модель, толку от неё немного)

↑ Начинается с β_2, потому что свободный член в гипотезу не включается!

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

k – кол-во коэффициентов (с учётом свободного)

На всякий случай: $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$

Критическое правило: $F > F_{k-1,\, n-k,\, \alpha} \Rightarrow H_0$ отвергается, регрессия значима в целом.

№7

Проверка гипотезы о линейном ограничении

Критерий линейного ограничения регрессии

(если вы с трудом понимаете это – это норма, я тоже)

H_0: $R \cdot \beta = r$

Например: $\beta_2 = 3\beta_3$, т. е. $\begin{pmatrix} 0 & 1 & -3 \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = 1$

Или: $\begin{cases} \beta_2 = 2 \\ \beta_3 = 4 \\ \beta_4 = \beta_1 \end{cases}$, т. е. $\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}$

Далее считаются RSS_{ur} и RSS_r:

RSS_{ur} – это RSS без условий, просто по МНК (ur – unrestricted, без ограничений)

RSS_r – это RSS с условием, т.е. найдены коэффициенты по МНК, а потом подставлены условия (r – restricted, с ограничениями)

Ну или вместо этого можно найти R^2_{ur} *и* R^2_r.

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

q – ранг матрицы R (количество знаков “=” в ограничении)

k – кол-во коэффициентов регрессии (с учётом свободного члена)

Задача 5

№1

Интерпретация коэффициентов линейной, полулогарифмической и логарифмической моделей регрессии.

Линейная зависимость: $y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$. Интерпретация коэффициентов такова: увеличение x_j на единицу соответствует увеличению y на β_j при прочих равных условиях (то есть при неизменных значениях всех остальных регрессоров и случайной составляющей).

Логарифмическая зависимость: $\ln y = \beta_1 + \beta_2 \ln x_2 + \ldots + \beta_k \ln x_k + \varepsilon$. Увеличение x_j на один процент приблизительно соответствует увеличению y на β_j процентов при прочих равных условиях (точнее, в $1{,}01^{\beta_j}$ раз, но приближение очень хорошее). Иначе говоря, коэффициент β_j есть частная эластичность y по x_j.

Полулогарифмическая зависимость: $\ln y = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$. Увеличение x_j на единицу соответствует при прочих равных условиях увеличению y в e^{β_j} раз, или на $(\mathrm{e}^{\beta_j} - 1) \cdot 100\,\%$. В этой модели интерпретируются потенцированные коэффициенты, но можно пользоваться тем, что $\mathrm{e}^{\beta_j} - 1 \sim \beta_j$ по базе $\beta_j \to 0$. Так что если значение коэффициента невелико, то увеличение x_j на единицу соответствует увеличению y на $\approx \beta_j \cdot 100\,\%$.

№2

Тесты на правильность спецификации: график «остатки-прогнозы», тест Рамсея

Ответ: Читаем статью Фурманова К.К.!