Лекция 12.

Построение и анализ уравнения множественной линейной регрессии

Постановка задачи: пусть выявлена зависимость результата X от фактора $Z_1, Z_2, \ldots, Z_k$
При $n > k$ получены результаты экспериментов $\vec{X} = (X_1, \ldots, X_n)$ при соответствующих значениях факторов $\vec{Z}^{(j)} = (Z_1^{(j)}, Z_2^{(j)}, \ldots, Z_k^{(j)})$
Предполагаем, что зависимость всех факторов от X линейная. Требуется по данным построить линейную модель, наилучшим образом объясняющую предсказывающую поведение X

Мультиколлинеарность

Def. Мультиколлинеарность - наличие линейной связи между всеми или несколькими факторами. Неприятные последствия:

- Оценки параметров становятся ненадежными, имеют большие стандартные ошибки и малую значимость. Небольшое изменение данных приводит к заметному изменению оценок
- Трудно определить изолированное влияние конкретного фактора на результат и выявить смысл данного влияния

Ex. Исследовалась зависимость веса от роста и размера обуви. В одной группе студентов модель оказалась такая: X - вес, Z_1 - рост, Z_2 - размер обуви,
В первой группе уравнение получилось таким: $X - \overline{x} = 0.9(Z_1 - \overline{Z_1}) + 0.1(Z_2 - \overline{Z_2})$
А во второй группе таким: $X - \overline{x} = 0.2(Z_1 - \overline{Z_1}) + 0.8(Z_2 - \overline{Z_2})$

Отбор факторов в уравнении регрессии

Чтобы избавиться от неприятных последствий, нужно отобрать факторы, которые мы в дальнейшем будем учитывать. Для этого строим корреляционную матрицу, состоящую из коэффициентов корреляции между результатом, факторами и факторов между собой

$$r = \begin{pmatrix} 1 & r_{X,Z_1} & r_{X,Z_2} & \ldots & r_{X,Z_k} \\ r_{X,Z_1} & 1 & r_{Z_1,Z_2} & \ldots & r_{Z_1,Z_k} \\ r_{X,Z_2} & r_{Z_2,Z_1} & 1 & \ldots & r_{X,Z_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{X,Z_k} & r_{Z_k,Z_1} & r_{Z_k,Z_2} & \ldots & 1 \end{pmatrix}$$

Алгоритм следующий:

1. Первым берем фактор, имеющий наибольшую корреляцию с результатом

2. Затем добавляем факторы, которые с одной стороны имеют наибольшую корреляцию с результатом, а с другой стороны наименьшую корреляцию с уже имеющимися факторами

Ex. Для данной корреляционной матрицы лучше всего брать Z_2 (фактор имеет лучшую корреляцию с X), а затем Z_3 (этот фактор меньше всего коррелирует с Z_2)

	X	Z_1	Z_2	Z_3
X	1	-	-	-
Z_1	0.81	1	-	-
Z_2	0.85	0.93	1	-
Z_3	-0.65	-0.38	-0.28	1

Анализ уравнения линейной регрессии

Пусть $X = \beta_0 + \beta_1 Z_1 + \cdots + \beta_k Z_k + \varepsilon$ - теоретическая модель линейной регрессии, $\varepsilon \in N(0, \sigma^2)$ и независимы

По методу наименьших квадратов была построена модель $\hat{X} = b_0 + b_1 Z_1 + \cdots + b_k Z_k$. Тогда $\hat{\varepsilon}_i = X_i - \hat{X}_i$ - экспериментальная (или эмпирическая) ошибка

Согласно следствию из теоремы $S^2 = \dfrac{1}{n-k-1} \sum\limits_{i=1}^{n} \hat{\varepsilon}_i^2$ - несмещенная оценка σ^2

Def. Величину S называют стандартной ошибкой регрессии

Из прошлых лекций знаем, что $Db_i = \sigma^2 (A^{-1})_{ii}$, где $A = ZZ^T$, $Z = (Z : (i)_j)$ - матрица плана

Из этого $\hat{D}b_i = S^2 (A^{-1})_{ii}$ - оценка дисперсии

Def. $S_{b_i} = S\sqrt{(A^{-1})_{ii}}$ - стандартная ошибка коэффициента регрессии b_i

Уравнение регрессии стандартных масштабов

Nota. При обычном уравнении линейной регрессии трудно сравнить влияние различных факторов, так как они имеют разную природу и разные единицы измерения. Поэтому делают стандартизацию данных.

Пусть есть выборка $\vec{X} = (X_1, \ldots, X_n)$ объема n. Тогда стандартизованный вектор $\vec{t}_x$ состоит из значений $\dfrac{X_i - \overline{x}}{\hat{\sigma}_X}$

Получаем новые данные, которые можем считать новой выборкой стандартизированной случайной величины $\dfrac{X - EX}{\sigma_X}$ и которая не имеет единиц измерения

Свойства стандартизированных данных:

1. $\overline{t_x} = 0$
2. $D^*_{t_x} = 1$

3. $r_{x,y} = r_{t_x,t_y} = \overline{t_x t_y}$

$$r_{x,y} = \frac{\widehat{\operatorname{cov}}(x,y)}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{1}{n}\sum_{i=1}^{n}\frac{x_i - \overline{x}}{\hat{\sigma}_x}\frac{y_i - \overline{y}}{\hat{\sigma}_y} = \overline{t_x t_y}$$

$r_{t_x,t_y} = \overline{t_x t_y}$, так как $\overline{t_x} = \overline{t_y} = 0$, $D^* t_x = D^* t_y = 1$

В уравнении регрессии данные результата и факторов заменяем на стандартизированные:
$t_j = \left(\frac{Z_i^{(j)} - \overline{Z^{(j)}}}{\sigma_Z^{(j)}}\right)$ - стандартизация j-ого фактора, $t_x = \left(\frac{X_i - \overline{X}}{\sigma_X}\right)$ - стандартизация результата
Так как стандартизация - линейная операция, то при соответствующей замене в уравнении получаем линейной уравнение, называемое уравнением регрессии стандартных масштабов:
$t_x = \gamma_1 t_1 + \gamma_2 t_2 + \cdots + \gamma_k t_k$

Nota. Заметим, что γ_0 (свободный член) равен 0: $\overline{t_j} = 0$, поэтому линия уравнения пройдет через начало координат. При этом система нормальных уравнений приобретает более простой и наглядный вид:

$$\begin{cases} \gamma_1 + r_{Z_1,Z_2}\gamma_2 + \cdots + r_{Z_1,Z_k}\gamma_k = r_{Z_1,x} \\ r_{Z_1,Z_2}\gamma_1 + \gamma_2 + \cdots + r_{Z_2,Z_k}\gamma_k = r_{Z_2,x} \\ \cdots \\ r_{Z_k,Z_1}\gamma_1 + r_{Z_k,Z_2}\gamma_2 + \cdots + \gamma_k = r_{Z_k,x} \end{cases}$$

Или в матричной форме: $r\Gamma = r_x$, где r - корреляционная матрица, Γ - столбец коэффициентов регрессии, r_x - столбец из коэффициентов корреляции факторов с результатом
Нормальное уравнение регрессии: $A\vec{B} = Z\vec{X}$, где $A = ZZ^T$, Z - матрица плана
Допустим, что все данные стандартизированные, тогда для i-ого элемента

$$\begin{pmatrix} Z_1^{(i)} & \dots & Z_n^{(i)} \end{pmatrix}\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \sum_{j=1}^{n} Z_j^{(i)} X_j = n\overline{Z^{(i)}X} = n r_{Z_i,X}$$

$$a_{ij} = \begin{pmatrix} Z_1^{(i)} & \dots & Z_n^{(i)} \end{pmatrix}\begin{pmatrix} Z_1^{(j)} \\ \vdots \\ Z_n^{(j)} \end{pmatrix} = \begin{cases} n\overline{Z_i Z_j} = \gamma_i r_{Z_i,Z_j}, & \text{при } i \neq j \\ n\overline{Z_i^2} = nDZ_i = n, & \text{при } i = j \end{cases}$$

Перевод коэффициентов можно выполнять по формулам:
$b_i = \gamma_i \frac{\sigma_X}{\sigma_{Z_i}}; \quad \gamma_i = b\frac{\sigma_{Z_i}}{\sigma_X}$, $b_0 = \overline{x} - \sum_{i=1}^{n} b_i \overline{z_i}$

Ex. Уравнение парной линейной регрессии: $\frac{x - \overline{x}}{\hat{\sigma}_x} = \hat{r}\frac{z - \overline{z}}{\hat{\sigma}_z}$
В стандартных масштабах получаем $t_x = \hat{r} t_z$
Тогда $\gamma_1 = \hat{r}$

Смысл стандартизированных коэффициентов

Значение γ_i можно трактовать как величину прямого влияния i-ого фактора на результат. Остальные слагаемые в i-ом уравнении можно трактовать как величины косвенного влияния остальных факторов

$r_{Z_iZ_1}\gamma_1 + \cdots + \gamma_i + \cdots + r_{Z_iZ_k}\gamma_k = r_{Z_i,X}$

Величину $r_{Z_i,X}$ можно рассматривать как сумму прямого и косвенного влияний

Nota. Для измерения тесноты отдельной связи между отдельным фактором и результатом при очищении влиянии других факторов есть понятие коэффициента частной корреляции. При $k=2$: $r_{X,Z_1/Z_2} = \dfrac{r_{X,Z_1} - r_{X,Z_2}r_{Z_1,Z_2}}{\sqrt{(1-r^2_{X,Z_2})(1-r^2_{Z_1,Z_2})}}$

Коэффициенты детерминации и множественной корреляции

Допустим, что как и в случае парной линейной регрессии дисперсию результата X можно разложить на объясненную и необъясненную $D(X) = D(\hat{X}) + D(\hat{\varepsilon})$

Nota. В зарубежной литературе используется $\text{RSS} = D(\hat{\varepsilon}) = \sum \hat{\varepsilon}_i^2$

Def. Коэффициентом детерминации R^2 называется величина $R^2 = 1 - \dfrac{D(\hat{\varepsilon})}{D(X)}$

Его можно трактовать как долю объясненной дисперсии, а $1-R^2$ - как долю необъясненной
Свойства:

1. $0 \le R^2 \le 1$
2. Если $R^2 = 1$, то $D(\hat{\varepsilon}) = 0$, то есть все данные лежат в гиперплоскости построенного уравнения регрессии
3. Если $R^2 = 0$, то $D(\hat{X}) = 0$, то есть все $\hat{X}_i = \overline{x} \Longrightarrow b_1 = b_2 = \cdots = b_k = 0$, то есть модель ничего не объясняет
4. Чем больше R^2, тем лучше
5. В случае линейного уравнения регрессии $R^2 = \sum\limits_{i=1}^{k} \gamma_i r_{Z_i,X}$

Def. Величина R называется коэффициентом множественной корреляции, который показывает силу линейной связи

Скорректированный коэффициент детерминации

При добавлении в модель новых факторов R^2 как правило увеличивается, хотя не всегда есть смысл добавлять новые факторы. Для выяснения того, следует ли это делать, исполь-

зуется скорректированный коэффициент детерминации $\overline{R^2} = 1 - \frac{k-1}{n-k-1}\frac{D(\hat{\varepsilon})}{D(X)}$, где n - число экспериментов, а k - число факторов

Проверка гипотез по значимости уравнения регрессии

а) F-тест: проверка гипотезы о значимости уравнения регрессии в целом $H_0 : R^2_{\text{теор}} = 0$ (уравнение регрессии статистически незначимо) против $H_1 : R^2_{\text{теор}} \neq 0$

> **Th.** Если $H_0 : R^2_{\text{теор}} = 0$ верна, то $F = \frac{R^2}{1-R^2}\frac{n-k-1}{k} \in F(k, n-k-1)$

Получаем критерий, называемый F-тестом: t_α - квантиль $F(k, n-k-1)$ уровня значимости α

$$\begin{cases} H_0 : R^2_{\text{теор}} = 0, & \text{если } F < t_\alpha \\ H_0 : R^2_{\text{теор}} \neq 0, & \text{если } F \geq t_\alpha \end{cases}$$

б) T-тест: проверка гипотезы о значимости отдельного коэффициента регрессии $H_0 : \beta_i = 0$ (b_i статистически незначим) против $H_1 : \beta_i \neq 0$

> **Th.** Если $H_0 : \beta_i = 0$ верна, то $T_i = \frac{b_i}{S_{b_i}} \in T_{n-k-1}$

Получаем критерий, называемый T-тестом: t_α - квантиль двухстороннего распределения Стьюдента $|T_{n-k-1}|$ уровня значимости α

$$\begin{cases} H_0 : \beta_i = 0, & \text{если } |T_i| < t_\alpha \\ H_0 : \beta_i \neq 0, & \text{если } |T_i| \geq t_\alpha \end{cases}$$

Nota. T-тест служит для отсева несущественных факторов из модели при условии, что все другие факторы включены в модель

Nota. При мультиколлинеарности возможно, что уравнение имеет высокую значимость, а большинство коэффициентов не проходит T-тест

Nota. При применении T-теста убираем только один фактор, далее строим новую модель и для нее опять проводим T-тест. Удаление 2 факторов может привести к неопределенным результатам