Обобщения регрессионной модели

Обобщения регрессионной модели

1. Нелинейные модели

(используется линеаризация)

- показательная (экспоненциальная) модель:

$$Y = a_0 e^{a_1 X} \varepsilon$$

Обобщения регрессионной модели

1. Нелинейные модели

(используется линеаризация)

- показательная (экспоненциальная) модель:

$$Y = a_0 e^{a_1 X} \varepsilon$$

Метод линеаризации - логарифмирование

$$\ln(Y) = \ln(a_0) + a_1 X + \ln(\varepsilon)$$

Обобщения регрессионной модели

1. Нелинейные модели

(используется линеаризация)

- показательная (экспоненциальная) модель:

$$Y = a_0 e^{a_1 X} \varepsilon$$

Метод линеаризации - логарифмирование

$$\ln(Y) = \ln(a_0) + a_1 X + \ln(\varepsilon)$$

Введение новых переменных и параметров:

$$Y^* = \ln(Y) \quad b_0 = \ln(a_0) \quad b_1 = a_1 \quad \varepsilon^* = \ln(\varepsilon)$$

Получим линейную модель

$$Y^* = b_0 + b_1 X + \varepsilon^*$$

Полиномиальная модель:

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_k x^k + \varepsilon$$

Полиномиальная модель:

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_k x^k + \varepsilon$$

Новые переменные:

$$z_1 = x; \quad z_2 = x^2; \quad z_3 = x^3; \quad ... \quad ; z_k = x^k$$

Полиномиальная модель:

$$Y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_k x^k + \varepsilon$$

Новые переменные:

$$z_1 = x; \quad z_2 = x^2; \quad z_3 = x^3; \quad \ldots \quad ; z_k = x^k$$

Переход к новым переменным →

линейная модель множественной регрессии:

$$Y = a_0 + a_1 z_1 + a_2 z_2 + \ldots + a_k z_k + \varepsilon$$

Гиперболическая модель

$$Y = a_0 + a_1 \frac{1}{X} + \varepsilon$$

Гиперболическая модель

$$Y = a_0 + a_1 \frac{1}{X} + \varepsilon$$

Новая переменная $z = \frac{1}{X}$

Подстановка $\rightarrow$уравнение парной регрессии:

Гиперболическая модель

$$Y = a_0 + a_1 \frac{1}{X} + \varepsilon$$

Новая переменная $z = \dfrac{1}{X}$

Подстановка →уравнение парной регрессии:

$$Y = a_0 + a_1 z + \varepsilon$$

Степенная модель (нелинейная по параметрам)

$$Y = a_0 x_1^{a_1} x_2^{a_2} \cdot \varepsilon$$

Степенная модель (нелинейная по параметрам)

$$Y = a_0 x_1^{a_1} x_2^{a_2} \cdot \varepsilon$$

Метод линеаризации – логарифмирование с последующим введением новых переменных:

$$\log(Y) = \log(a_0) + a_1 \log(x_1) + a_2 \log(x_2) + \log \varepsilon$$

Степенная модель (нелинейная по параметрам)

$$Y = a_0 x_1^{a_1} x_2^{a_2} \cdot \varepsilon$$

Метод линеаризации – логарифмирование с последующим введением новых переменных:

$$\log(Y) = \log(a_0) + a_1 \log(x_1) + a_2 \log(x_2) + \log \varepsilon$$

Вводятся новые переменные и параметры:

$$Y^* = \log(Y) \quad z_1 = \log(x_1) \quad z_2 = \log(x_2) \quad \varepsilon^* = \log \varepsilon \quad b_0 = \log(a_0) \quad b_1 = a_1 \quad b_2 = a_2$$

Степенная модель (нелинейная по параметрам)

$$Y = a_0 x_1^{a_1} x_2^{a_2} \cdot \varepsilon$$

Метод линеаризации – логарифмирование с последующим введением новых переменных:

$$\log(Y) = \log(a_0) + a_1 \log(x_1) + a_2 \log(x_2) + \log\varepsilon$$

Вводятся новые переменные и параметры:

$$Y^* = \log(Y) \quad z_1 = \log(x_1) \quad z_2 = \log(x_2) \quad \varepsilon^* = \log\varepsilon \quad b_0 = \log(a_0) \quad b_1 = a_1 \quad b_2 = a_2$$

В новых переменных исходное уравнение принимает вид уравнения множественной регрессии:

$$Y^* = b_0 + b_1 z_1 + b_2 z_2 + \varepsilon^*$$

Пример степенной модели:

Производственная функция Кобба-Дугласа:

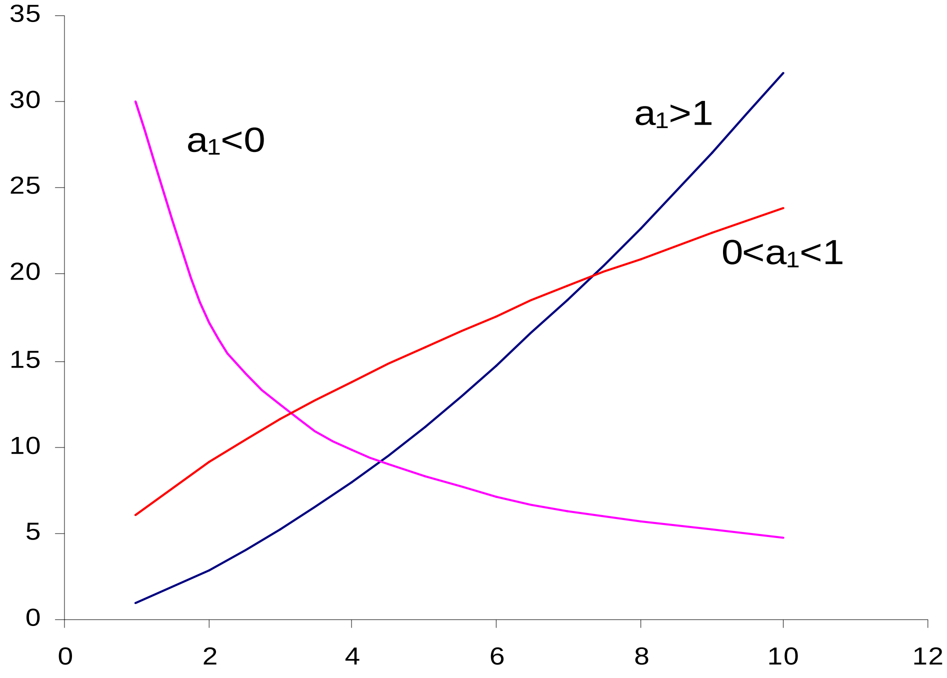
$$Y = a_0 K^{a_1} L^{(1-a_1)}$$

Пример степенной модели:

Производственная функция Кобба-Дугласа:

$$Y = a_0 K^{a_1} L^{(1-a_1)}$$

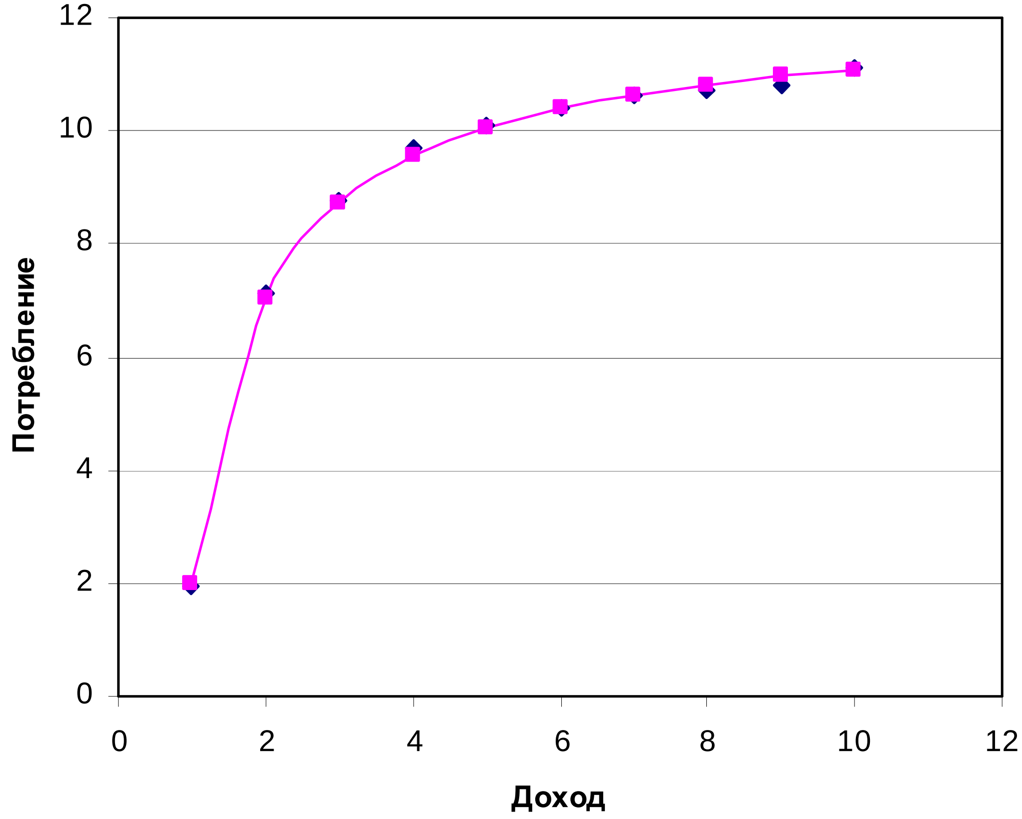
Является нелинейной как по переменным, так и параметру a_1



Логарифмическая модель $Y = a_0 + a_1 \ln\left(x\right) + \varepsilon$

Логарифмическая модель $Y = a_0 + a_1 \ln(x) + \varepsilon$

С помощью модели описывают процессы, обладающие свойством насыщения, например, кривые Энгеля для товаров повседневного спроса.



2. Модели с качественными переменными

2. Модели с качественными переменными

введение **фиктивных** переменных:

Например, Y, X_1, X_2 количественные, $X_3 \in \{a, b\}$ - качественная. **Фиктивная** переменная

$$\tilde{X}_3 = \begin{cases} 1, & \text{если } X_3 = a; \\ 0, & \text{иначе.} \end{cases}$$

2. Модели с качественными переменными

введение **фиктивных** переменных:

Например, Y, X_1, X_2 количественные, $X_3 \in \{a, b\}$ - качественная. **Фиктивная** переменная

$$\tilde{X}_3 = \begin{cases} 1, & \text{если } X_3 = a; \\ 0, & \text{иначе.} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \tilde{X}_3 + \varepsilon$$

2. Модели с качественными переменными

введение **фиктивных** переменных:

Например, Y, X_1, X_2 количественные, $X_3 \in \{a, b\}$ - качественная. **Фиктивная** переменная

$$\tilde{X}_3 = \begin{cases} 1, & если\ X_3 = a; \\ 0, & иначе. \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \tilde{X}_3 + \varepsilon$$

Коэффициент β_3 : ожидаемое изменение Y при $X_3 = a$ по сравнению с $X_3 = b$.

Если $L \geq 2$, то вводятся $L-1$ фиктивных переменных.

Если $L \geq 2$, то вводятся $L-1$ фиктивных переменных.

Например, если $X_4 = \{a, b, c\}$, то вводятся фиктивные переменные

$$\tilde{X}_4^{(1)} = \begin{cases} 1, \textit{если} \quad X_4 = a; \\ 0, \textit{иначе}; \end{cases}$$

$$\tilde{X}_4^{(2)} = \begin{cases} 1, \ \textit{если} \quad X_4 = b; \\ 0, \textit{иначе}. \end{cases}$$

Если $L \geq 2$, то вводятся $L-1$ фиктивных переменных.

Например, если $X_4 = \{a, b, c\}$, то вводятся фиктивные переменные

$$\tilde{X}_4^{(1)} = \begin{cases} 1, \text{ } \textit{если} \quad X_4 = a; \\ 0, \textit{ иначе}; \end{cases}$$

$$\tilde{X}_4^{(2)} = \begin{cases} 1, \text{ } \textit{если} \quad X_4 = b; \\ 0, \textit{ иначе}. \end{cases}$$

Если $\tilde{X}_4^{(1)} = 0$ и $\tilde{X}_4^{(2)} = 0 \Rightarrow X_4 = c$.

Если $L \geq 2$, то вводятся $L-1$ фиктивных переменных.

Например, если $X_4 = \{a, b, c\}$, то вводятся фиктивные переменные

$$\tilde{X}_4^{(1)} = \begin{cases} 1, \textit{если} \;\; X_4 = a; \\ 0, \textit{иначе}; \end{cases}$$

$$\tilde{X}_4^{(2)} = \begin{cases} 1, \; \textit{если} \;\; X_4 = b; \\ 0, \textit{иначе}. \end{cases}$$

Если $\tilde{X}_4^{(1)} = 0$ и $\tilde{X}_4^{(2)} = 0 \Rightarrow X_4 = c$.

Коэффициенты при фиктивных переменных в линейной модели имеют смысл ожидаемого изменения Y по сравнению с базовым уровнем.

Проблема мультиколлинеарности - коррелированность (зависимость) двух или нескольких объясняющих переменных в модели.

Проблема мультиколлинеарности - коррелированность (зависимость) двух или нескольких объясняющих переменных в модели.

Последствия:

оценки коэффициентов регрессии - **ненадежные** (определитель матрицы объясняющих переменных *det* X^TX близок к нулю); **неустойчивые**, т. е. сильно меняются при исключении небольшой части наблюдений; результаты проверки значимости переменных недостоверны.

Проблема мультиколлинеарности - коррелированность (зависимость) двух или нескольких объясняющих переменных в модели.

Последствия:

оценки коэффициентов регрессии - **ненадежные** (определитель матрицы объясняющих переменных $det\ X^T X$ близок к нулю); **неустойчивые**, т. е. сильно меняются при исключении небольшой части наблюдений; результаты проверки значимости переменных недостоверны.

Устранение мультиколлинеарности

- исключение коррелированных переменных;
- пошаговый отбор информативных переменных.

Пошаговая регрессия

- Найти переменную, максимально коррелированную с Y;
- Включить эту переменную в модель;
- Найти следующую максимально коррелированную переменную; включить ее и т.д.

Метод включения-исключения – аналогично (исключается наименее коррелированная переменная).

Гребневая регрессия

Штраф на сумму квадратов коэффициентов (L_2 регуляризация).

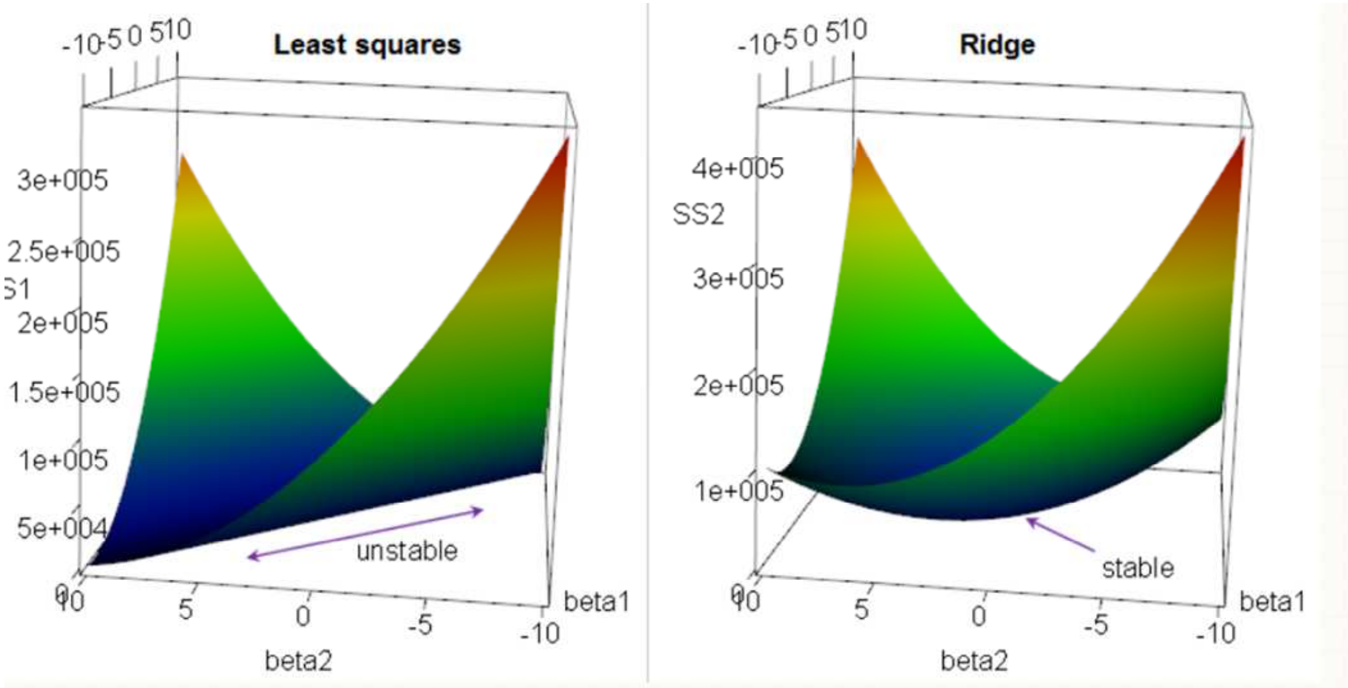
$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_i (y^{(i)} - \sum_{j=0}^{m} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \right\}.$$

Решение:

$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda I_m)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, где I_m - единичная матрица.

Добавление «гребня» увеличивает все собственные значения матрицы $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, не меняя собственных векторов.

Пример: оптимизируемые функционалы



Метод LASSO (Least Absolute Shrinkage and Selection Operator)

Вместо штрафа введем ограничения:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \{ \sum_{i} (y_i - \sum_{j=0}^{m} x_{i,j}\beta_j)^2 \}$$

$$subject \text{ to } \sum_{j=1}^{m} |\beta_j| \le s$$

либо

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \underbrace{\left\{ \sum_{i} \left(y_i - \sum_{j=0}^{m} x_{i,j}\beta_j \right)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\}}_{J(\beta)}$$

(L_1 регуляризация).

Получаем задачу квадратичного программирования.

Решение: метод градиентного спуска.

β^0 - начальное значение вектора параметров;

$$\beta^{i+1} = \beta^i - \tau \cdot \nabla J(\beta^i),$$

τ - длина шага.

Можно показать, что при уменьшении параметра s все больше коэффициентов β_j принимают нулевое значение – происходит отбор информативных переменных.