Логистическая регрессия (Logit)

Преподаватель: Герард Костин

Цели регрессионного анализа

1. Предсказание значения зависимой переменной с помощью независимых переменных.

$$y_i = x_i'\beta + \varepsilon_i$$

2. Определение вклада отдельных независимых переменных в вариацию зависимой переменной.

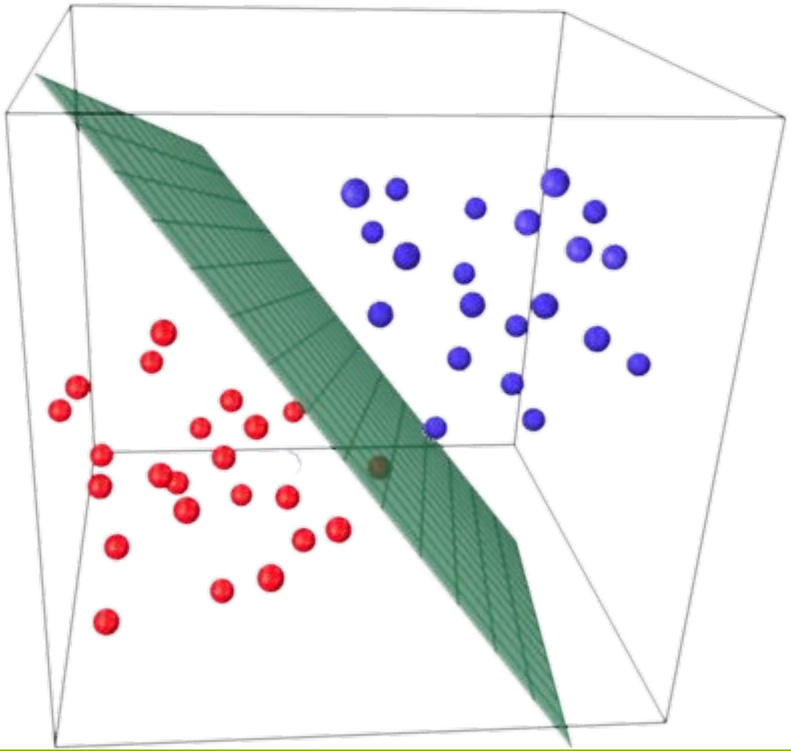
3. Регрессионный анализ нельзя использовать для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для применения этого вида анализа.

assumptions регрессионного анализа

1. Переменные модели должны иметь распределение, близкое к нормальному.
2. Для построения линейных регрессий, зависима и независимые переменные должны иметь линейную связь.
3. Отсутствие мультиколлинеарности – независимость между собой переменных-предикторов.
4. Независимость наблюдений
5. Гомоскедастичность - дисперсия остатков одинакова для каждого значения.

Логистическая (дискретная) модель

Дискретная модель - – модель регрессии, в которой зависимая переменная является дискретной.



Модель бинарного выбора – частный случай модели дискретного выбора, при котором зависимая переменная может принимать только два значения (1 или 0)

Логистическая регрессия

Уравнение логистической регрессии

$$Z = B_0 + B_1X_1 + \ ... + B_p$$

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

отношение шансов может быть записано в следующем виде:

P / 1 − P = $e^{B_0 + B_1X^1 + B_2X^2 + \ldots + B_pX^p} = e^{B_0} e^{B_1X^1}$

Отсюда получается, что, изменение x k на единицу вызывает изменение отношения шансов в e^ bk раз

- Зависимая переменная – дихотомическая.
- Цель – построение модели прогноза вероятности события {Y=1} в зависимости от независимых переменных X1,…,Xp путём подгонки данных к логистической кривой.
- Отношение вероятности того, что событие произойдет к вероятности того, что оно не произойдет P / 1−P называется отношением шансов.

Проблемы МНК

Подход

Подбор функции, область значений которой описывается отрезком [0;1], неубывающей на этом отрезке и обладающей свойством непрерывности.

Наиболее распространенные функции – стандартного нормального и логистического распределения

Проблемы

- Биномиальное распределение остатков
- Гетероскедастичность и смещенность оценок
- Расчетные значений зависимой переменной могут выходить за пределы интервала [0 ; 1]

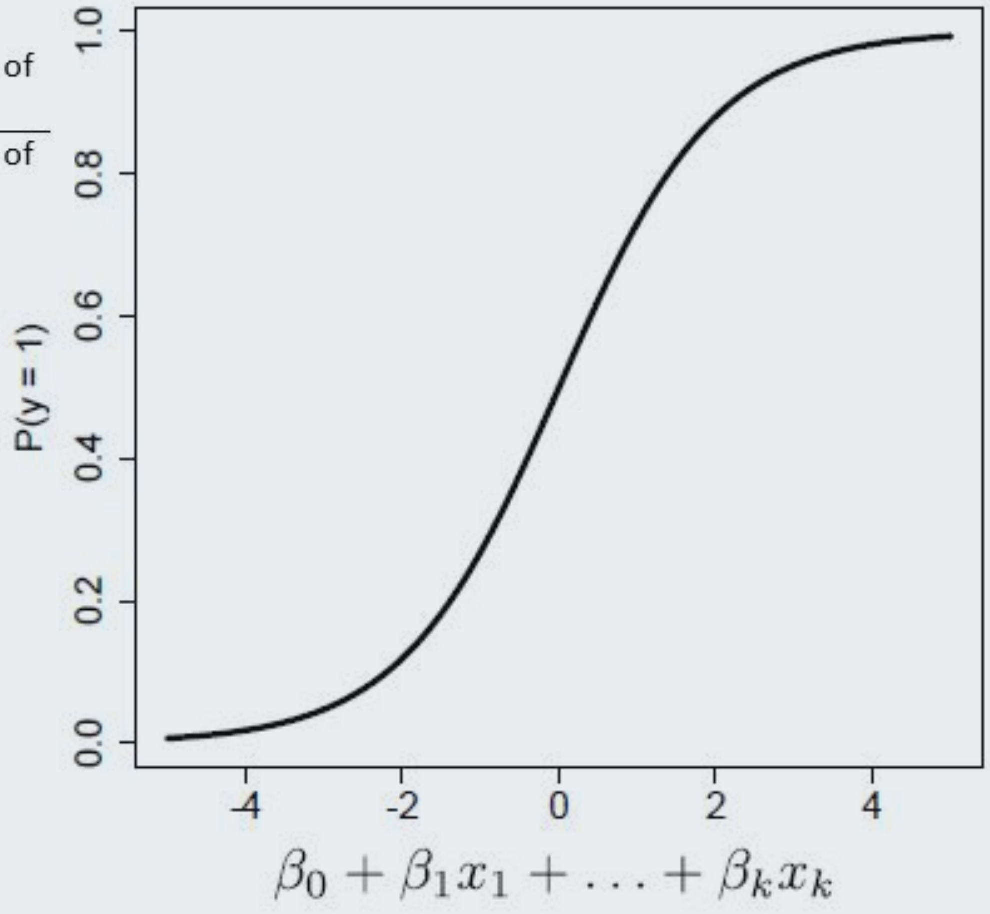
Логарифм шансов (функция Logit)

$$Odds = \frac{P(y=1)}{P(y=0)}$$

$$\frac{\text{The Odds will be > 1 when there is a higher probability of predicting y = 1}}{\text{The Odds will be < 1 when there is a higher probability of predicting y = 0}}$$

$$Odds = e^{\beta_0+\beta_1 x_1+\cdots+\beta_n x_n} \xrightarrow{applying\ log} \log(Odds) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

- *Коэффициенты бета 0, бета 1, бета К выбраны, чтобы максимизировать вероятность для наблюдений, принадлежащих к классу 1,*
- *И прогнозировать малую вероятности для наблюдений, фактически принадлежащих к классу 0.*



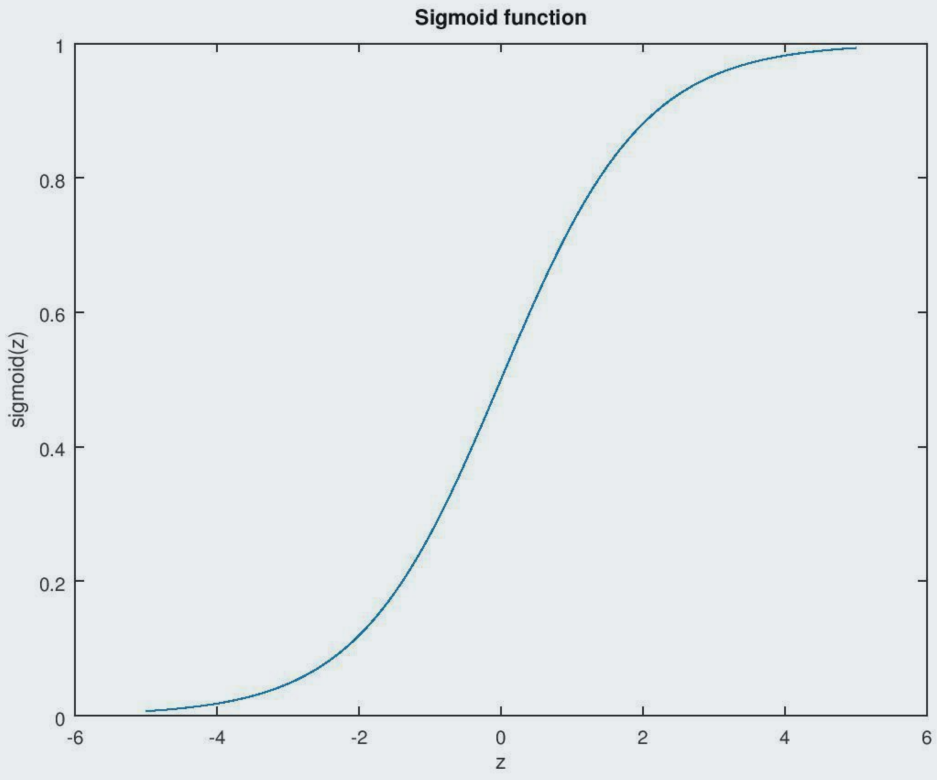
Оценка параметров модели LOGIT

Метод максимального правдоподобия для:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_1 + \beta_2 X_i$$

Интерпретация расчетных значений результата

Вероятность того, что зависимая переменная примет значение 1 при заданном значении объясняющих переменных



Полиномиальная (мультиклассовая) регрессия

Модель множественного выбора –модель дискретного выбора, при котором зависимая переменная может принимать более двух значений

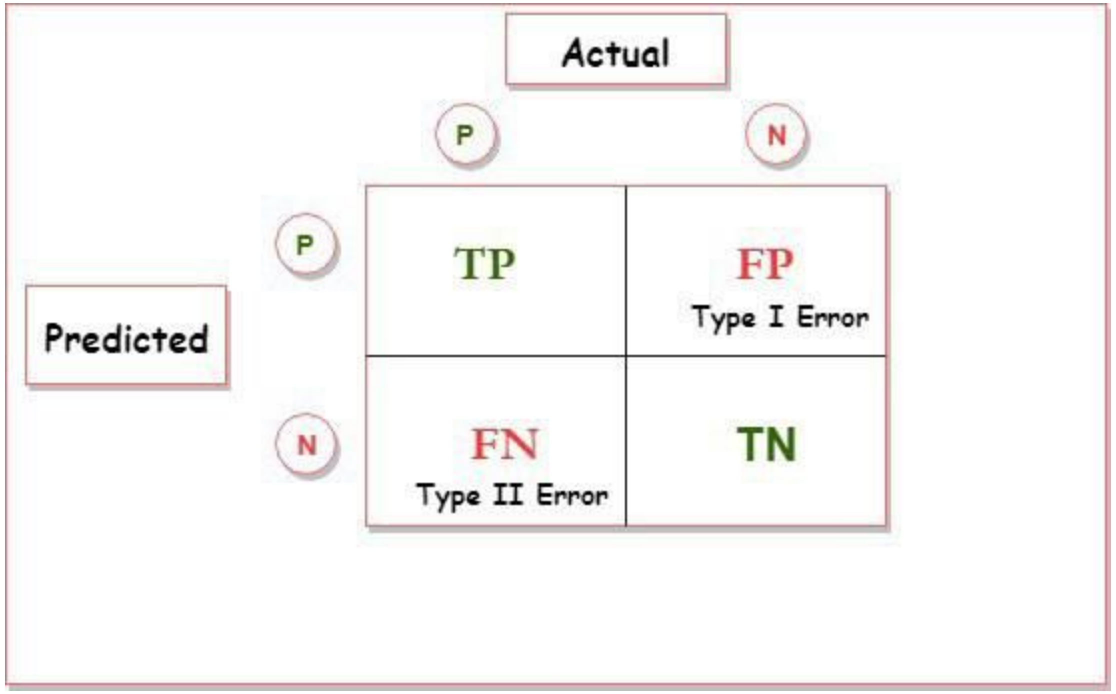
Log-Loss/Кросс-энтропия

$$P(\vec{y} \mid \sigma(\vec{w}^T X)) = \prod_{i=1}^{n} \sigma(\vec{w}^T \vec{x}_i)^{[y_i=+1]} \; (1 - \sigma(\vec{w}^T \vec{x}_i)^{[y_i=-1]}) \rightarrow max$$

$$CE = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

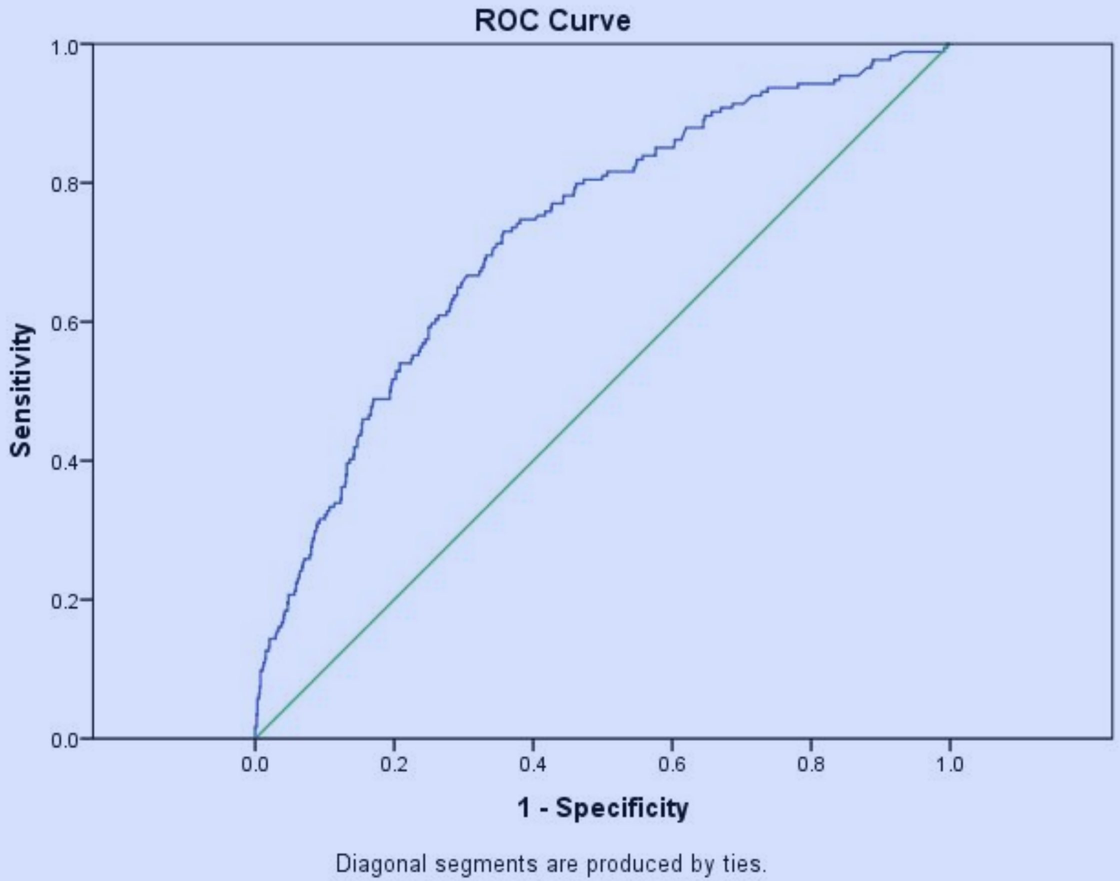
Кросс-энтропия (или логарифмическая функция потерь – log loss): Кросс-энтропия измеряет расхождение между двумя вероятностными распределениями. Если кросс-энтропия велика, это означает, что разница между двумя распределениями велика, а если кросс-энтропия мала, то распределения похожи друг на друга.

Confusion matrix



Матрица ошибок, является предиктором производительности модели для задачи классификации. Количество правильных и неправильных прогнозов суммируется со значениями количества и разбивается по каждому классу.

ROC-AUC



Receiver operating characteristic (ROC)или кривая

ROC - это графический график, который иллюстрирует производительность системы двоичного классификатора при изменении порога дискриминации. Кривая создается путем нанесения истинного положительного показателя (чувствительности) на уровень ложного положительного результата (1 - специфичность) при различных настройках пороговых значений.

Ridge Regression

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2$$

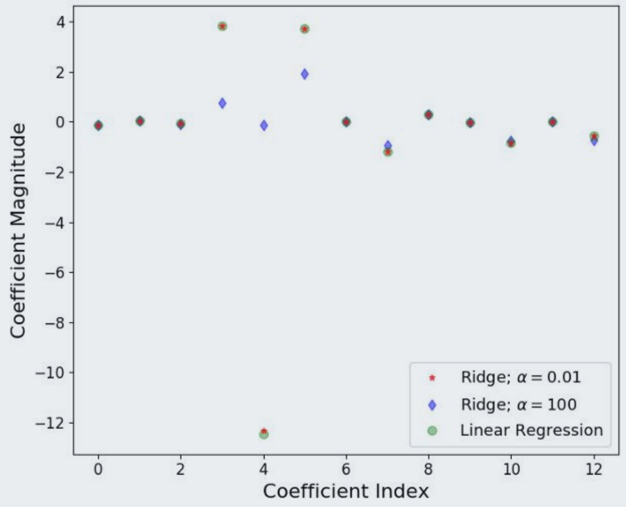
Простая линейная регрессия

$$\sum_{i=1}^{M}(y_i-\hat{y}_i)^2=\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}w_j^2$$

Ridge регрессия

For some $c>0,\ \sum_{j=0}^{p}w_j^2< c$

Ограничение



Lasso Regression

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2$$

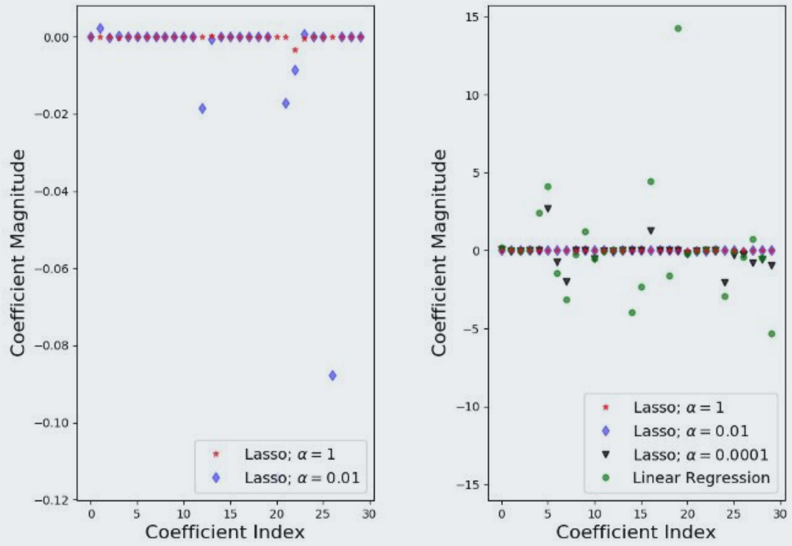
Простая линейная регрессия

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Ridge регрессия

$$\text{For some } t > 0, \ \textstyle\sum_{j=0}^{p} |w_j| < \ t$$

Ограничение



Lasso VS Ridge