Бинарные модели

$$P(Y_i=1)=F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i}+\varepsilon_i)$$

Как оценивать качество модели?

- R^2 **МакФаддена** — $R^2_{MF}=1-\dfrac{\hat{\ell}(1\dots X_k)}{\hat{\ell}(1)}$
- **Псевдо** $-R^2$ — $R^2_p=\dfrac{1}{1+\frac{2}{n}(\hat{\ell}(1\dots X_k)-\hat{\ell}(1))}$
- Сравнение точности прогнозирования — $R^2_{fit}=1-\dfrac{W_1}{W_0},\quad$ где $W_1=\dfrac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$

Как оценивать модель?

МНК

Посмотрим на распределение ошибок регрессии:

$$\varepsilon_i=Y_i-F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i})$$

Ошибки принимают 2 значения:

С вероятностью $1-p_i=P(Y_i=0):\quad F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i})$

С вероятностью $p_i=P(Y_i=1):\quad 1-F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i})$

Выводы:

1. Распределение ε_i нельзя считать нормальным $\to$ нельзя проводить тесты на значимость.
2. $Var(\varepsilon_i)=F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i})(1-F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i}))\to$ имеет место гетероскедастичность $\to$ оценки мнк — **неэффективны**

ММП

$$\mathcal{L}(Y,X|\beta)=\prod_{Y_i=1}F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i})*\prod_{Y_i=0}(1-F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i}))=$$

$$=\prod_{i=1}^{n}F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i})^{Y_i}(1-F(\beta_0+\beta_1 X_{1_i}+\dots+\beta_k X_{k_i}))^{1-Y_i}=$$

$$\max_{\beta_0\dots\beta_k}\ell(Y,X|\beta)=\sum_{i=1}^{n}(Y_i*log(F(\beta_0+\dots+\beta_k X_{k_i}))+(1-Y_i)log(1-F(\beta_0+\dots+\beta_k X_{k_i}))$$

$$\frac{\partial\ell}{\partial\beta_j}=\sum_{i=1}^{n}X_{j_i}\left(\frac{Y_i f(Z_i)}{F(Z_i)}-\frac{(1-Y_i)f(Z_i)}{1-F(Z_i)}\right)=0\quad\forall j=1\dots k$$

Используемые функции для F

- Логит-модель — $F(Z_i)=\dfrac{1}{1+e^{-Z_i}}\quad f(Z_i)=\dfrac{e^{-Z_i}}{(1+e^{-Z_i})^2}$
- Пробит-модель — $F(Z_i)=\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{Z_i}e^{\frac{-t^2}{2}}dt\quad f(Z_i)=\dfrac{1}{\sqrt{2\pi}}e^{\frac{-Z_i^2}{2}}$

Интерпретация полученных оценок

Дисперсии оценок находятся на диагоналях матрицы $I^{-1}(\hat{\beta})=-E\left(\dfrac{\partial^2\ell}{\partial\beta\partial\beta^T}\right)$

Для проверки коэффициентов на значимость пользуемся свойствами оценок ММП, а именно:

$$\hat{\beta}_{ML}\sim N(\beta,I^{-1}(\beta))$$

$$z=\frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}\sim N(0,1)$$

Имеем следующее уравнение после оценки: $Y_i=F(Z_i)=F(\hat{\beta}_0+\hat{\beta}_1 X_{1_i}+\dots+\hat{\beta}_k X_{k_i})$

Как коэффициенты влияют на предсказание модели?

$$\frac{\partial\hat{Y}_i}{\partial\hat{\beta}_j}=\frac{\partial F}{\partial Z_i}\frac{\partial Z_i}{\partial\hat{\beta}_j}=f(Z_i)X_j$$

Какое влияние непрерывный фактор оказывает на зависимую переменную?

$$\frac{\partial\hat{Y}_i}{\partial X_j}=\frac{\partial F}{\partial Z_i}\frac{\partial Z_i}{\partial X_j}=f(Z_i)\hat{\beta}_j,\quad Z_i\text{ считается в точке }(\bar{X}_1,\bar{X}_2,\dots,\bar{X}_j)$$

Какое влияние оказывает дамми-фактор на предсказание модели?

$$\frac{\partial\hat{Y}_i}{\partial X_{dum}}=\hat{Y}_i(X_1,\dots,X_{dum}=1)-\hat{Y}_i(X_1,\dots,X_{dum}=0)$$

Иными словами: $\dfrac{\partial\hat{Y}_i}{\partial X_{dum}}=P(Y=1|X_{dum}=1)-P(Y=1|X_{dum}=0)$