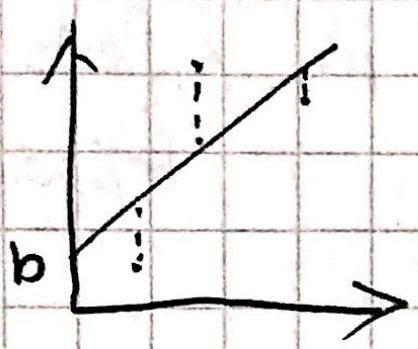
线性回归中的平方误差及公式推导

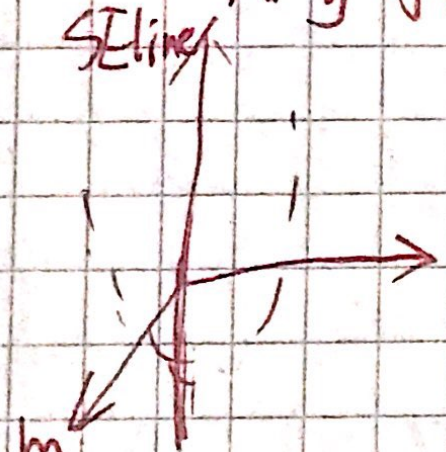


$$SE_{line} = (y_1 - (mx_1+b))^2 + \cdots + (y_n - (mx_n+b))^2$$

$$= \sum y_n^2 - 2m\sum x_n y_n - 2b\sum y_n$$
$$+ m^2\sum x_n^2 + 2mb\sum x_n + nb^2$$

$$= n\overline{y^2} + 2nm\overline{xy} - 2bn\overline{y_n} + m^2 n\overline{x^2}$$
$$+ 2mnb\overline{x} + nb^2$$

$\overline{x}$、$\overline{y}$ 可视作已知，仅有 m、b 未知



求 $\overline{y}$

$$m = \frac{\overline{x}\,\overline{y} - \overline{xy}}{(\overline{x})^2 - \overline{x^2}}$$

决定系数 R^2

y的波动多大程度可被x描述的波动

$$R^2 = 1 - \frac{SE_{line}}{SE_{\bar{y}}}$$

SE_{line}：y离回归线的波动

$SE_{\bar{y}}$：y总波动

$$SE_{\bar{y}} = \sum (y_n - \bar{y})^2$$

$$SE_{line} = \sum (y_n - (mx_n + b))^2$$

Date

协方差 & 回归线

covariance

$$Cov(X,Y) = E[(X-E[X])(Y-E[Y])]$$

与最小二乘系数 $m = \frac{cov(X,Y)}{var(X)}$

$$E[(X-E[X])(Y-E[Y])$$

$$= E[XY - XE[Y] - E[X]Y + E[X]E[Y]]$$

把 $E[X] \cdot E[Y]$ 看作常数

$$E[XY] - E[X]E[Y] - E[X]E[Y] - E[X]E[Y]$$

$$= \underbrace{E[XY]}_{\approx \overline{XY}} - \underbrace{E[Y]E[X]}_{\overline{Y}\,\overline{X}}$$

$$Cov(X,Y) = E[XY] - E[X][Y]$$

$$(= \overline{XY} - \overline{X}\,\overline{Y})$$ 回归的分子

χ^2 分布

$X \sim N(0,1)$

$Q = X^2$

↳ ~~X^2~~ ~~$Q \sim \chi^2_{df}$~~

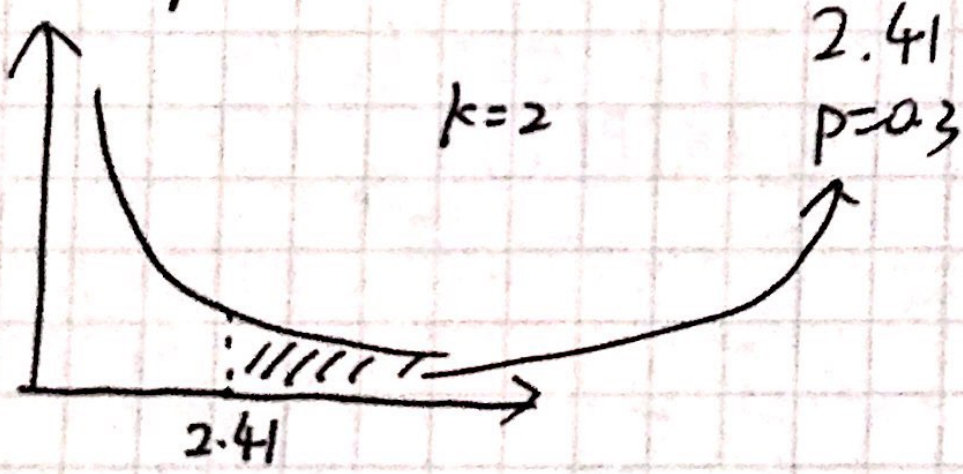
χ^2 分布

$X \sim N(0,1)$

$Q = x_1^2 + \cdots + x_k^2$

↓

$Q \sim \chi^2_k$



皮尔逊 χ^2 检验

$$\chi^2 = \sum \frac{(f - f_e)^2}{f_e}$$

$$df = n - 1$$

列联表 χ^2 检验

	A	B	C		
Sick				80	21%
Not Sick	D	E	F	300	79%
				380	

$$\chi^2 = \sum \frac{(f - f_e)^2}{f_e}$$

$$df = (\text{行} - 1)(\text{列} - 1)$$