REGRESSION

regression is used to find the relationships among different variables. This can be used to just find the relationship or to predict values.

Definitions:

Heteroscedastic: the variance of a residual term varies widely

Homoscedastic: the variance is nearly constant

Intrinsic Scatter: the deviations from the fit even if all measurements were perfect

Latent variable: a random variable that is not measured

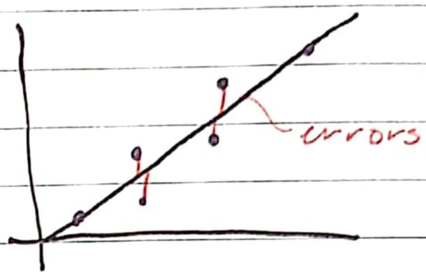
Systemic Error: when measurements are made with some sort of bias.

Methods

Ordinary Least Squares

The goal is to find the slope & intercept that best matches the data

$$y_n = \sum_{i=0}^{K} \beta_i X_{ni} + \beta_n$$



to find the β coef. it minimizes

$$m\left(\sum_{i=1}^{n} Y_i - \beta_0 - \beta_1 X_i\right)^2$$ → aka the sum of the squared distances from point to line

Algorithm's to estimate this:
Orthogonal distance & bivariate correlated errors & intrinsic scatter

M-estimation

This is a form of robust regression which is useful if the data has a small number of point & large residuals

β value to minimize $\sum_{i=1}^{n} \rho(y_i - x_i^T \beta)$

ρ is function (like $\rho(x) = x^2$ which is least squares)

One choice is Huber's Method

$$\rho(x) = \begin{cases} -c & x < -c \\ x & |x| < c \\ c & x > c \end{cases}$$ → changes large residuals to a set value

Or another option is Tukey's Bisquare:

$$\rho(x) = \begin{cases} x(c^2 - x^2)^2 & |x| < c \\ 0 & \text{otherwise} \end{cases}$$

eliminates large outliers, and weights mid-size outliers less

So, with M-estimation you change the way large residuals influence the regression line to hopefully have a better more accurate line

Thiel-Sen Median Slope:

another option for robust regression:

$$\beta_{ij} = \frac{Y_i - Y_j}{X_i - X_j}$$

find all the slopes between all pairs of the data points

the fitted slope is the median of all of them.

Quantile Regression:

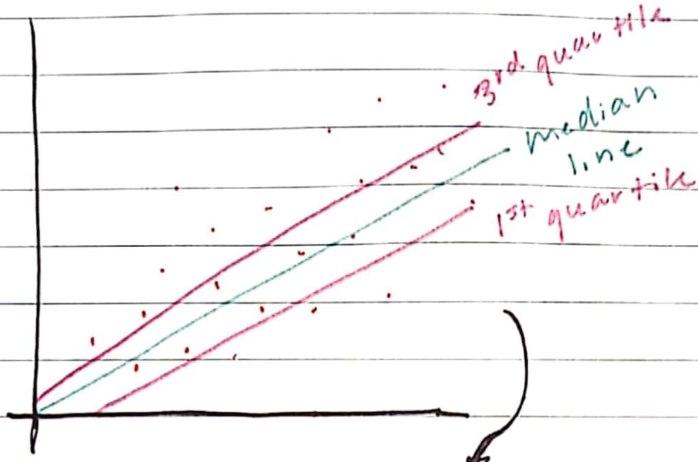
quantiles are the inverse of the cdf

↳ 1st quartile, 3rd quartile, median

quantile regression aims to relate the quantiles while minimizes quantile loss

↓

These have to be solved by iterative process



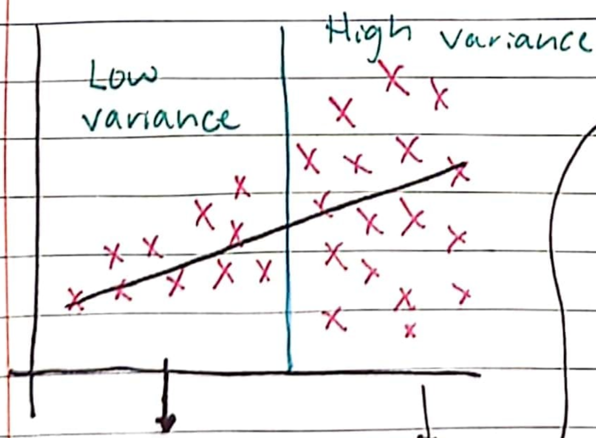
for normal dist. they will be equally spaced, but for non-normal they won't be.

The slopes can be different too

Quantile regression is more robust in that its not as sensitive to outliers (compared to linear)

Weighted Least Squares

In WLS, there isn't the assumption that the variance is always σ^2, instead it may depend on the X value.



But in OLS it treats them the same. So, the low variance should have a higher weight

If an extra point is here, it probably should change the line more

extra points here should affect the line less

WLS minimizes

$$S_{r,wt} = \sum \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2_{Y,i}}$$

the variance can change

Poisson Regression

If the Y variable takes on positive integer values poisson regression may work.

$$P_{\mu_i}(Y=y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

So, when measuring counted responses that follow poisson dist.

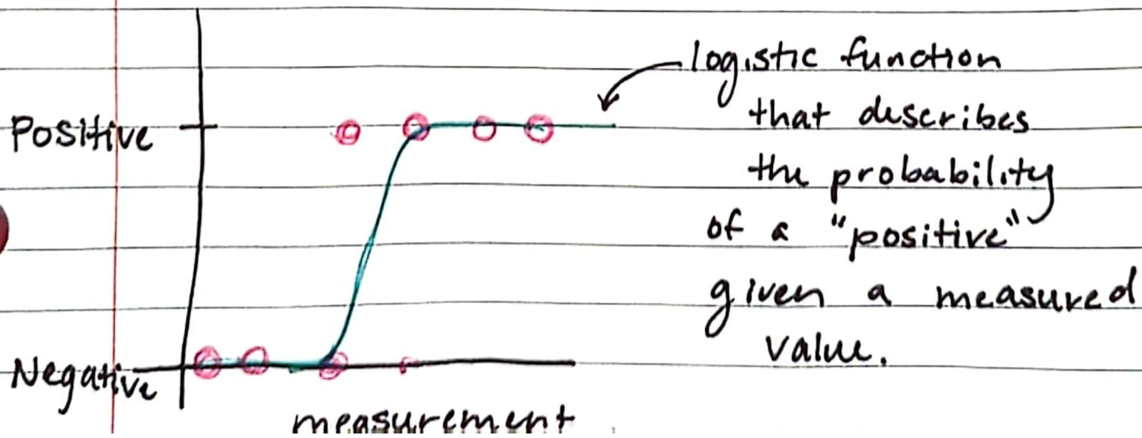
$$E(Y|x) = e^{\beta x}$$

the model is: $\log(\mu) = \alpha + \beta x$

- if $\beta = 0$, Y & x are not related
- if $\beta > 0$, then the count $\mu = E(Y)$ is e^{β} times larger than at $x=0$
- if $\beta < 0$, the $\mu = E(Y)$ is e^{β} times smaller than at $x=0$

Logistic Regression

Used when there are 2 specific outcomes, it doesn't fit a line, but a "logistic function".

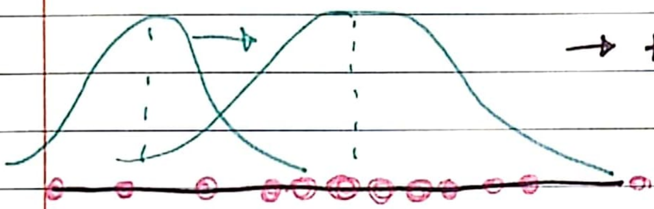


the measurement can be discrete or continuous,
see if different measurements are better predictors

Instead of minimizes squared errors, it uses maximum likelihood estimators, find the line w/ the highest likelihood

Maximum Likelihood

Want to find the distribution that describes the data the best, by finding what is most likely



→ try different curves
for each calculate how likely the data is → the probability of observing the data given the curve

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\left[\frac{-1}{2\sigma}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_i)^2\right]}$$

→ likelihood of the data given parameters

$$\beta_{1,MLE} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1(MLE)\bar{X}, \quad \hat{\sigma}^2_{MLE} = \frac{RSS}{n}$$

Coefficient of Determination (r^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

} want this to be close to 1 (R^2)

ratio of the errors of squares to total sum of squares another option is adjusting R^2

$$R_a^2 = 1 - \frac{n-1}{n-p}(1-R^2)$$

p is the # of parameters & you are panalized for having too many

It is also a good idea to plot the residuals to look for patterns