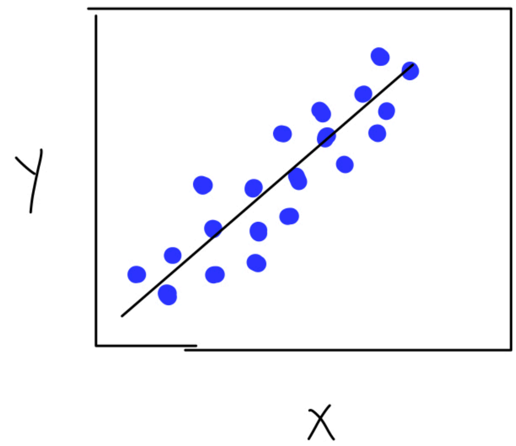
Regression

- Continous x
- continuous y



Mathematical

$$Y = mx + b$$

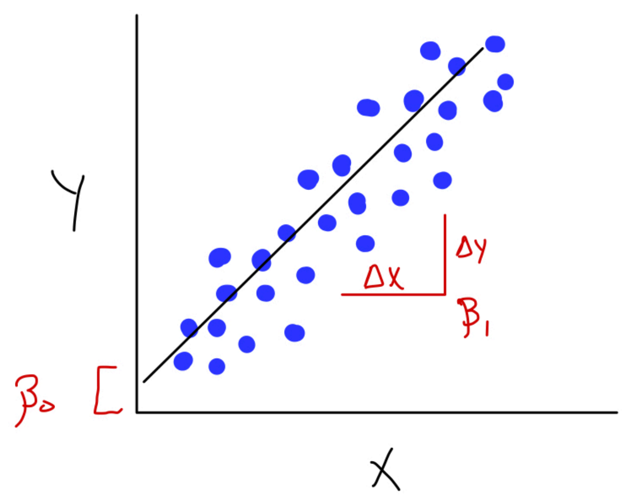
slope — intercept

$$\frac{\text{Rise}}{\text{Run}} = \frac{\Delta y}{\Delta x}$$

Statistical

$$\hat{Y} = \beta_0 + \beta_1 x$$

"hat" = equation — intercept — slope



* NOTE: What we are trying to do?

associate points to each other.

New equation

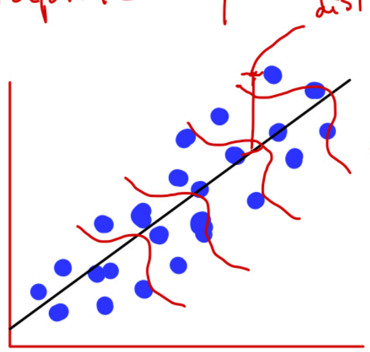
- change $\hat{y}$ to y_i and x to x_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon \sim N(0, \sigma)$$

error with a normal distribution of mean 0 and Standard deviation σ

In English
- we have the line + some error.

graphically



we have a bell curve around the line.

- mean of residual error is 0 and standard deviation equal to sigma.

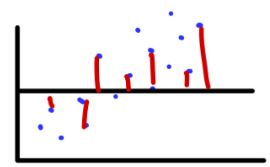
* goal is to estimate slope, intercept, Standard deviation

- Small standard deviation - close to line
- large standard deviation - far from line.

How? - we want to fit the best fit line.

- minimize error

  ↳ $\bar{\varepsilon} = 0$? NO



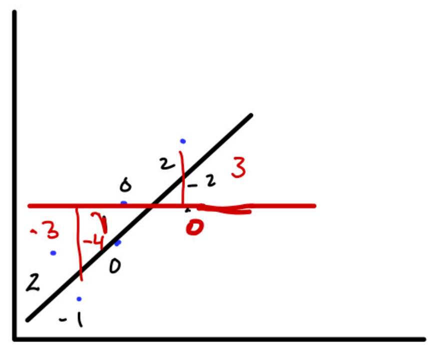
$\bar{\varepsilon} = 0$

- minimize the sum of squares (SSE)

$$\sum_{i=1}^{n} (y_i - \hat{y})^2$$

y_i: each point

$\hat{y}$: Best fit line.



$2 + \cancel{-1} + \cancel{0} + \cancel{1} + \cancel{2 + -2} = 2^2 = 4 = SSE$

$-\cancel{3} + -4 + -1 + 0 + 0 + \cancel{3} = -5^2 = 25 = SSR$

$4 + 25 = 29 = TSS$

We also have

- Total sum of Squares (TSS) $(y_i - \bar{y})^2$

  y_i: each point

  $\bar{y}$: y bar, mean of everything.

- Sum of squares due to regression. (SSR)

$$\sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$

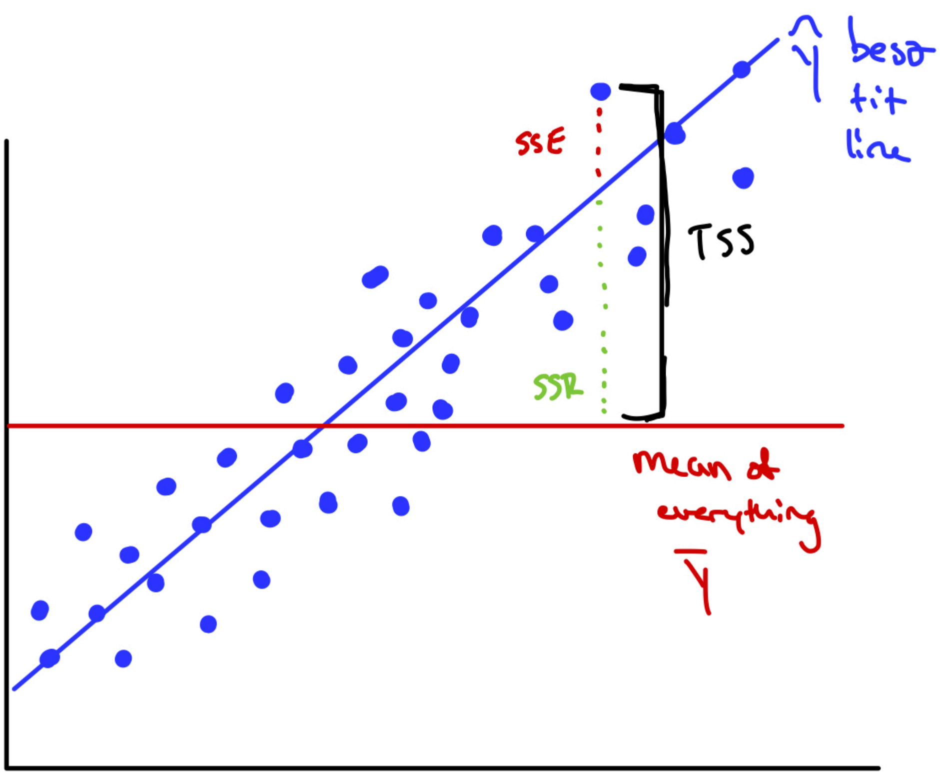
Is it statistically sig?

- what is the chance of getting the result by chance?

H_0: No slope or $\beta_1 = 0$

- slope ↑ p↓
- Sample size ↑ p↓
- noise (Error) ↓ P↓
  SSE

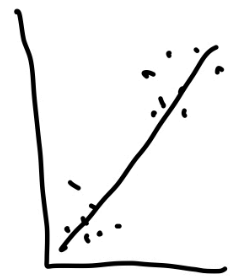
- Test p-value by partitioning the TSS.



Assumptions of Regressions

1. Continuous Y

2. normally distributed error term.
   ↳ not Y term



3. Relationship is linear.
   ↳ only one that truely matters at end. because can mess up slope.

Bimodal

4. homoskedasticity of sigma.
   - No effect on slope
   - no effect on pvalue

5. Independent samples.
   ↳ no autocorrelation.
   ↳ previous measurment is correlated with the previous.
   ↳ space and time.

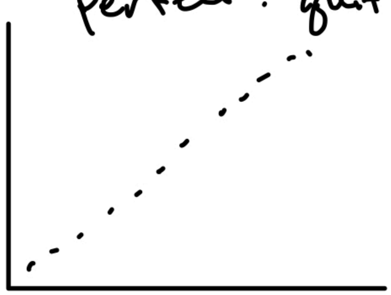
In reallity assumptions are silly because regressions are robust to moderate departures of the assumptions.

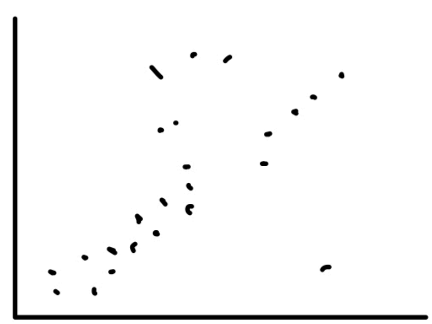
- what really matters? - linearity.
- There are formal tests - KS-test
  Shapero-wilks
  really sensitive to large sample size.

What to do?

look at your data!

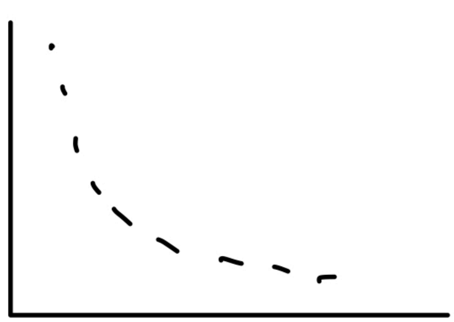
perfect! quit science





not normally distributed residuals

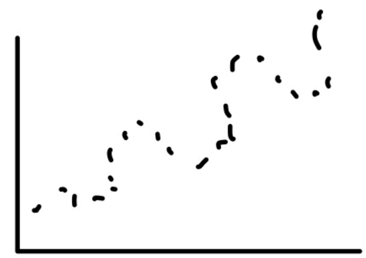
- slope? no
-



not linear - will alter slope.

heteroskedastic

- not alter slope
- not p-value



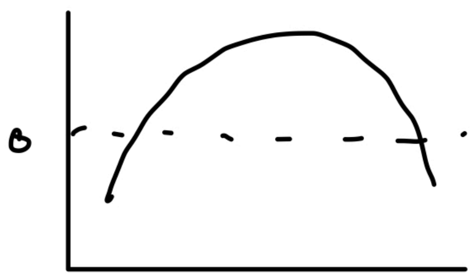
auto correlated

- no slope
- no p-value.

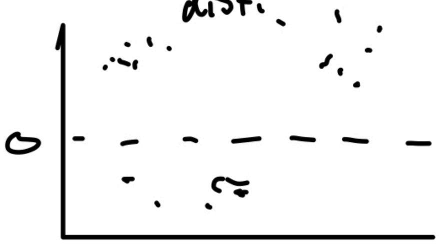
Residual plots

perfect

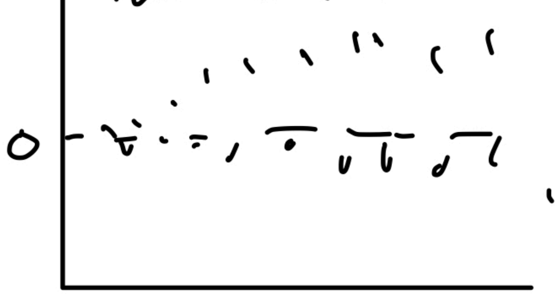
non-linear.

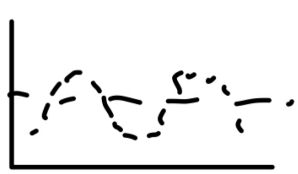


not normally dist.



heteroskedastic = Fan





auto correlation.

T-test

- Continuous Y
- Categorical X

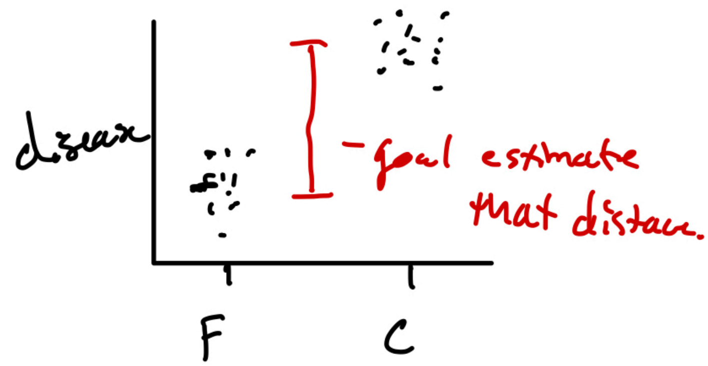
$X_1 = F$

$X_2 = C$

Y = disease

H_0: no difference

$$\mu_F = \mu_C$$



$$\text{disease} = \beta_0 + \beta_1(\text{Fungicide}) + \varepsilon \sim N(0, \sigma)$$

F 0

1 C

$$\text{disease}(F) = \beta_0 + \cancel{\beta_1(0)} + \varepsilon \sim N(0, \sigma)$$

intercept.

$$\text{disease}(C) = \beta_0 + \beta_1(1) + \varepsilon \sim N(0, \sigma)$$

Holy moly its a regression!