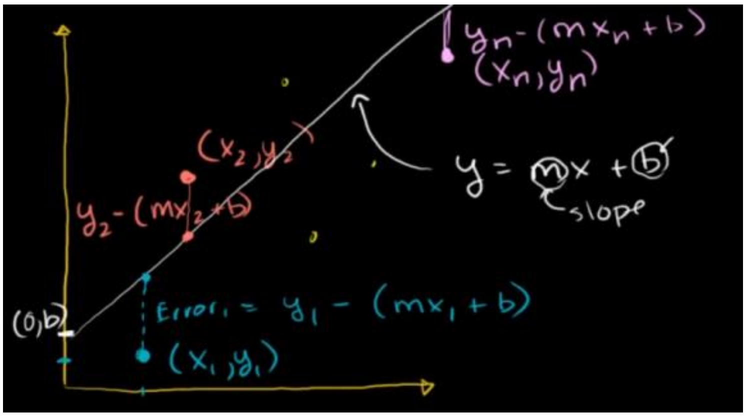
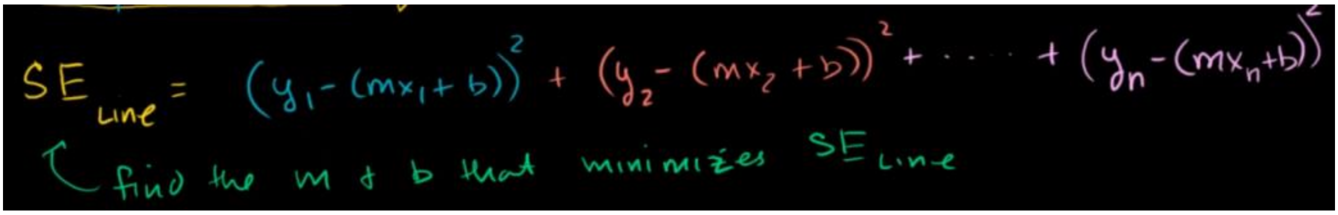
Thursday, May 16, 2019 9:30 PM

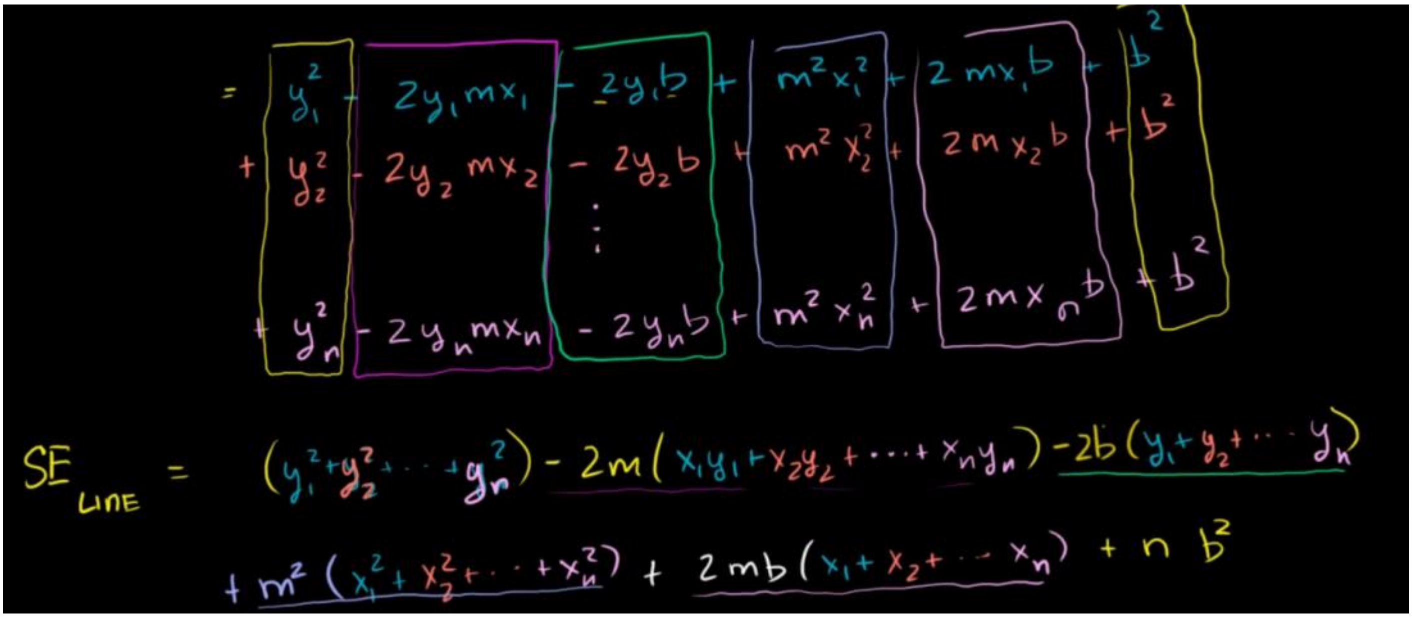
Squared error of regression line:



Minimize the distance

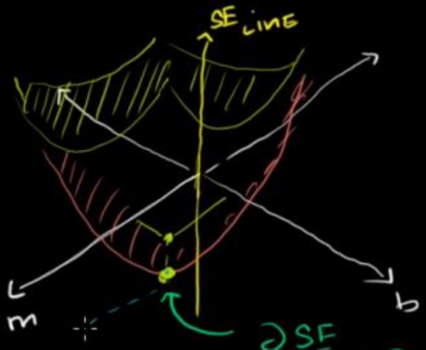


Find the m and n to minimize SE



The variable is m and b, and when the derivative is 0, SE has the minimum value

$$\boxed{SE_{LINE} = n\overline{y^2} - 2mn\overline{xy} - 2bn\bar{y} + m^2 n\overline{x^2} + 2mbn\bar{x} + nb^2}$$



$$\frac{\partial SE}{\partial m} = 0 \quad \& \quad \frac{\partial SE}{\partial b} = 0$$

Dm:

$$-2n\overline{xy} + 2n\overline{x^2}m + 2bn\bar{x}$$

Dn:

$$-2n\bar{y} + 2mn\bar{x} + 2bn = 0$$

$$-\overline{xy} + m\overline{x^2} + b\bar{x} = 0$$

$$-\bar{y} + m\bar{x} + b = 0$$

$$\left\{\begin{array}{l} m\overline{x^2} + b\bar{x} = \overline{xy} \\ m\bar{x} + b = \bar{y} \\ y = mx + b \\ \quad (\bar{x}, \bar{y}) \text{ lies on the line} \end{array}\right.$$

$$\Rightarrow m\frac{\overline{x^2}}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}}$$

$$\Downarrow$$

$(\bar{x}, \bar{y})$, $\left(\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}}\right)$ lies on the line

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

Coefficient of determination:
r^2

Estimating how good the line is fitting to those points:

What percentage of the total variation in y in described by the variation in x, so y is more correlated with x.

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

The total variation in y: sum of the distances of each y

$SE_{line} / SE_{\bar{y}}$

The percentage of total variation not explained by the line (the variation in x), because whatever the value of x, the square error of line is constant

$r^2 = 1 - SE_{line} / SE_{\bar{y}}$

The closer is r^2 to 1, the more correlated x is with y, that is the variation of y in defined by the variation of x to a more extent.

Covariance between two random variables:

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

(expected value)

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

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

$$= E[XY] - E[Y]E[X] - \cancel{E[X]E[Y]} + \cancel{E[X]E[Y]}$$

$$= \boxed{\overline{XY} - \bar{Y}\bar{X}}$$ ← numerator

$$\hat{m} = \frac{\overline{XY} - \bar{Y}\bar{X}}{\overline{X^2} - (\bar{X})^2}$$

Chi-square distribution:

X ~ N(0,1)

$$Q_1 = X^2$$

$$Q \sim \chi^2_1 \rightarrow 1 \text{ degree of freedom}$$

↓

chi² distribution

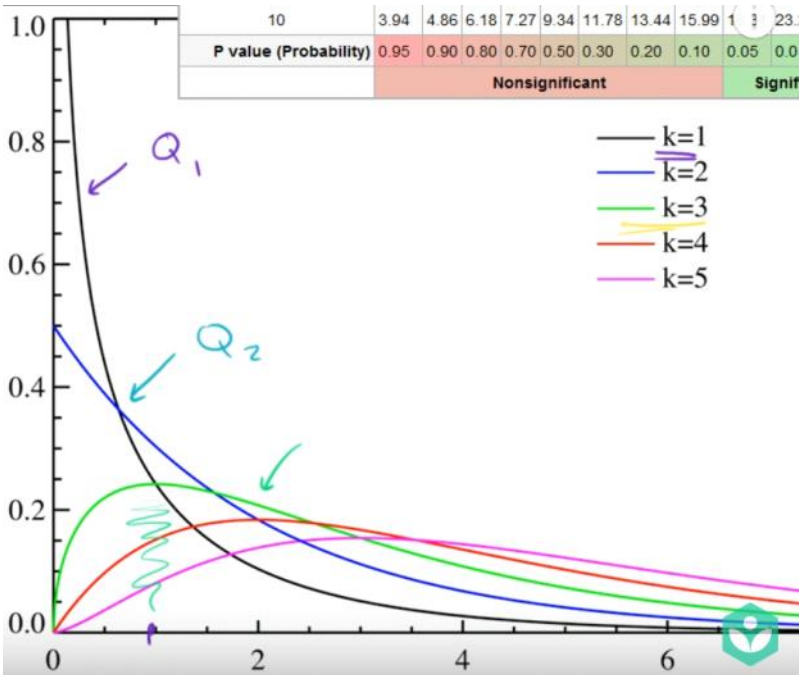
$$Q_2 = X_1^2 + X_2^2$$

$$Q \sim \chi^2_2$$

X1 and X2 are two independent individual from the N(0,1)

As the degree of freedom(k) increase, the peak of chi square distribution moves to right. And you can easily prove it.

(analogy of sample distribution)



Degrees of freedom (df)	χ^2 value [9]									
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01
	Nonsignificant								Signific	

The probability of Q2 be greater than 2.41 is 30%

Pearson's chi square test:

Example:

Expected customers pecentage and observed customers percentage

Day:	M	T	W	T	F	S
Expected %:	10	10	15	20	30	15
Observed:	30	14	34	45	57	20

H0: owners' distribution is correct

H1: not correct

$$\text{chi-square statistic} = \chi^2 = \frac{(30-20)^2}{20} + \frac{(14-20)^2}{20} + \frac{(34-30)^2}{30} + \frac{(45-40)^2}{40} + \frac{(57-60)^2}{60} + \frac{(20-30)^2}{30}$$

$$= \frac{100}{20} + \frac{36}{20} + \frac{16}{30} + \frac{25}{40} + \frac{9}{60} + \frac{100}{30} = \underline{\underline{11.44}}$$

(normalization to make the mean of chi-square distribution is close to 0)

If H0 is right, this chi-square distribution is N(0,)

Degree of freedom: n-1

$$\boxed{11.44}$$

$$11.44 > 11.07$$

This is to say, our chi value is more extreme than the critical value, in a random test, there is less than 5% of chance that H0 is true. So we reject this hypothesis

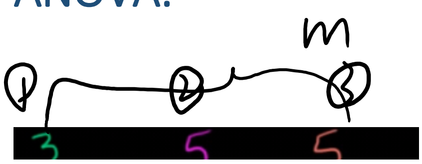
Chi-square of contingency table

	Herb 1	Herb 2	Placebo (sugar pill)	
# sick	20	30	30	80
Expected:	25.3	29.4	25.3	21%
# not sick	100	110	90	300
Expected:	94.7	110.6	94.7	79%
Total	120	140	120	380

Degree of freedom: (number of column -1)*(number of row-1)

K = (3-1)*(2-1)

ANOVA:



1	2	3
3	5	5
2	3	6
1	4	7

n

$\bar{x}_1 = 2 \quad \bar{x}_2 = 4 \quad \bar{x}_3 = 6$

Grand mean (mean of mean):

$$\bar{\bar{X}} = \frac{3+2+1+5+3+4+5+6+7}{9} =$$

How is SST correlated with (Variation within each group, variation between different groups)

SST(total sum of squares):

$$SST = (3-4)^2+(2-4)^2+(1-4)^2+(5-4)^2+(3-4)^2+(\ldots)^2 + (5-4)^2+(6-4)^2+(7-4)^2 = 30$$

For calculating SST, Degree of freedom: m*n-1 = 8

SSW(sum of squares within groups):

$$SSW = (3-2)^2+(2-2)^2+(1-2)^2 + (5-4)^2+(3-4)^2+(4-4)^2 + (5-6)^2+(6-6)^2+(7-6)^2 = 6$$

For calculating SSW, degree of freedom: m*(n-1) =6

SST is 30, SSW is 6, we can say that 6 of the variation 30 is coming from SSW

SSB(sum of squares between groups):

$$SSB = (2-4)^2+(2-4)^2+(2-4)^2 + \ldots + (4-4)^2 + (4-4)^2 = 24$$

$$SSB = (2-4)^2 + (2-4)^2 + (2-4)^2 + (4-4)^2 + (4-4)^2 + (4-4)^2 + (6-4)^2 + (6-4)^2 + (6-4)^2 = 24$$

For calculating SSB, freedom is m-1 = 2

SST = SSW + SSB

Degree of freedom : 8 = 6 + 2

F test:

Example:

Food1	Food2	Food3
3	5	5
2	3	6
1	4	7

We want to figure out does the foods make a difference

That is if $u_1 = u_2 = u_3$

population mean

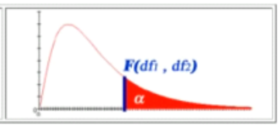
F statistics can be seen as the radio of two chi-square distribution that may or may not have different degrees of freedom

F statistics = $\dfrac{\frac{SSB}{m-1}}{\frac{SSW}{m(n-1)}} = 12$

So if the numerator is relative large, we can conclude that the total variation is mostly due to the variation between groups rather than the individual differences within each group, that is the different foods does make a difference.

What is special about f statistics, every alpha has a table
Df1 is the numerator df, df2 is the denominator df

F Table for $\alpha = 0.10$



\	$df_1=1$	2	3	4	5	6	7	8	9
$df_2=1$	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.857
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.380
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.240
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.935
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.316
6	3.77595	$\boxed{3.46330}$	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.957
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.724
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

12>>3.46
So we reject the null hypothesis

Correlation and causality: