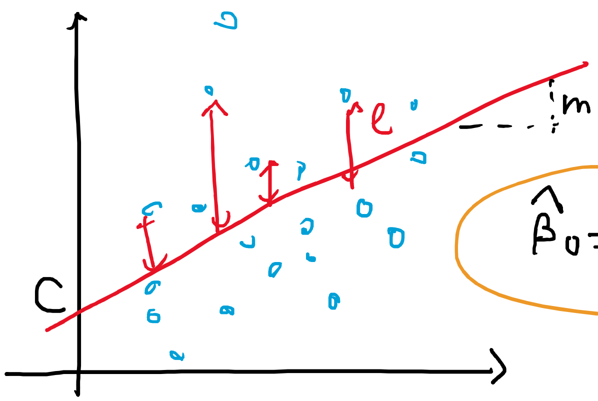
Saturday, August 24, 2024 10:02 PM

Simple Linear Regression

$y = mx + C$



$y = \beta_0 + \beta_1 x + \cancel{\epsilon}$

$\bar{y} = \frac{\sum y}{n_y}$ → num of y

prediction

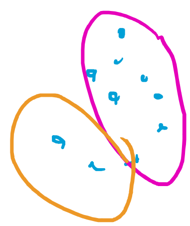
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$= \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$

$= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$

Data (ML)
- Supervised (Label, numeric target)
  - x, y
  - Classification → labelling
  - Regression → Prediction
- Unsupervised



t-test / z

	H_0	H_1
①	$\beta_1 =$ Certain value (not zero) K	$\beta_1 \neq K$
②	$\beta_0 =$ certain value (not zero)	$\beta_0 \neq K$

(2) β_0 = certain value (not zero) | $\beta_0 \neq k$

(3) $\beta_1 = 0$ | $\beta_1 \neq 0$

ANOVA

Analysis of variance.

(1) Test Stat.

$$H_0 : \beta_1 = k \quad \text{vs.} \quad H_1 : \beta_1 \neq k$$

Which test? t-value or z-value?

- small size < 30
- $\hat{\sigma}^2 \longrightarrow MSE$

(Population variance is it known?)

If not, we use t-value with MS_E

If σ^2 is given.

(1) Test β_1

$$t = \frac{\hat{\beta}_1 - k}{} = \frac{\hat{\beta}_1 - k}{}$$

$$t = \frac{\hat{\beta}_1 - k}{Se(\beta_1)} = \frac{\hat{\beta}_1 - k}{\sqrt{MS_E / S_{xx}}}$$

$$Z = \frac{\hat{\beta}_1 - k}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

(2) Test β_0 . No z-value.

$$t = \frac{\hat{\beta}_0 - k}{Se(\beta_0)} = \frac{\hat{\beta}_0 - k}{\sqrt{MS_E\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}}$$

$t > t_{\frac{\alpha}{2}, n-2}$, Reject H_0

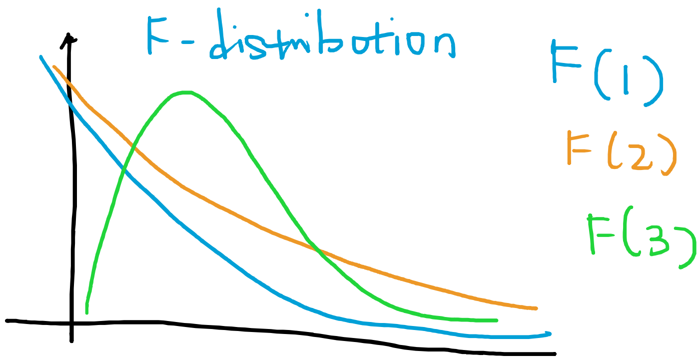
(3) Test $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$

Using F-test, construct ANOVA table

Simple Linear!) $MS_R = SS_R$ only for Simple LM

Source	Degree of freedom	Sum of Squares	Mean Squares	
Regression	1	SS_R	$MS_R = SS_R$	$F^* = MS_R$

Regression	1	SS_R	$MS_R = \frac{SS_R}{1}$	$F^* = \frac{MS_R}{MS_E}$
	+	+		
Error	$n-2$	SS_E	$MS_E = \frac{SS_E}{n-2}$	
	=			
Treatment	$n-1$	SS_T		



$$\text{Mean Square} = \frac{\text{Sum of square}}{\text{deg. of freedom}}$$

```
Call:
lm(formula = Sweet ~ Pectin, data = juice.dat)
Residuals:
     Min       1Q   Median       3Q      Max
-0.54373 -0.11039  0.06089  0.13432  0.34638
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.2520679  0.2366220  26.422   <2e-16 ***
Pectin      -0.0023106  0.0009049  -2.554   0.0181 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.215 on 22 degrees of freedom
Multiple R-squared:  0.2286,    Adjusted R-squared:  0.1936
F-statistic:  6.52 on 1 and 22 DF,  p-value: 0.01811
```

From <http://127.0.0.1:37401/>

β_0 β_1

$S_e(\beta_0)$

R^2

Degree of freedom.

$S_e(\beta_1)$

$H_0: \beta_1 = 0 \quad H_1: \beta_1 \neq 0$

p-value = 0.01811 < 0.05

Do not reject H_0.

```
Analysis of Variance Table
Response: Sweet
          Df Sum Sq  Mean Sq F value  Pr(>F)
Pectin     1 0.3014 0.301402  6.5204 0.01811 *
Residuals 22 1.0169 0.046224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSR

SSE

From <http://127.0.0.1:37401/>

Source	df	SS	MS	
Regression	1	0.3014 (SS_R)	0.3014	$F = \frac{0.3014}{0.046224}$.
Error	22	1.0169 (SS_E)	0.046224	
Treatment	23 ‖ 24 − 1			