Mathematical Model

Equation, formula

$E = mc^2$ $V = IR$ $y = x^{2/3}\cos^2(\log\theta)$

Mathematical Model

$PV = NRT$

P : pressure V : Volume T : Temperature

n : number of moles R : universal gas constant

Assumptions : ideal gas, static and close environment

Mathematical Model

Ideal gas law : $PV = NRT$

Q1 : Is this relationship true?
Q2 : What is the value of the constant R?

Answer these questions by a set of measurements :

$(P_i, V_i, T_i, N_i) \rightarrow R_i = \frac{P_i V_i}{N_i T_i}$

Errors due to unknown outside factors exists.

Statistical Model

Observed data $p = P + \delta_p$ $v = V + \delta_v$ $t = T + \delta_t$ $n = N + \delta_n$

δ —— Unobserved measurement errors (random)

Ideal gas law :

$$pv = nRt + (v\delta_p + p\delta_v - \delta_p\delta_v - Rt\delta_n - Rn\delta_t + R\delta_n\delta_t)$$

Statistical Model — [Systematic component + Random errors] — Data

Model parameter —— Unknown parameter in systematic component e.g. universal gas constant R

Analysis of Variance Model (ANOVA)

One-way ANOVA —— Compare multiple populations

$N(\mu_1, \sigma^2) \rightarrow Y_{11}, Y_{12}, ..., Y_{1n_1}$

$N(\mu_2, \sigma^2) \rightarrow Y_{21}, Y_{22}, ..., Y_{2n_2}$

……..

$N(\mu_a, \sigma^2) \rightarrow Y_{a1}, Y_{a2}, ..., Y_{an_a}$

Assumptions

1. Normal
2. Equal Variances
3. Independence

One-way ANOVA

Total sample size $N = \sum_{i=1}^{a} n_i$

Overall population mean (grand mean) $\mu = \frac{1}{N}\sum_{i=1}^{a} n_i\mu_i$

i^{th} treatment effect $\alpha_i = \mu_i - \mu$ $\left(\sum_{i=1}^{a} n_i\alpha_i = 0\right)$

Random errors $\varepsilon_{ij} = Y_{ij} - \mu_i = Y_{ij} - \mu - \alpha_i$

ANOVA model

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $j = 1,2,...,n_i$ $i = 1,2,...,a$

$\sum_{i=1}^{a} n_i\alpha_i = 0$ $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

One-way ANOVA

ANOVA model

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $j = 1,2,...,n_i$ $i = 1,2,...,a$

$\sum_{i=1}^{a} n_i\alpha_i = 0$ $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

Between group Within group

μ → $\mu + \alpha_1$, $\mu + \alpha_2$, ……, $\mu + \alpha_a$

$\mu + \alpha_2 + \varepsilon_{21} = Y_{21}$

$\mu + \alpha_2 + \varepsilon_{22} = Y_{22}$

……

$\mu + \alpha_2 + \varepsilon_{2n_2} = Y_{2n_2}$

Test for Treatment Effects

$H_0: \alpha_1 = \alpha_2 = \cdots \alpha_a = 0$ vs $H_1:$ some $\alpha_i \neq 0$ (Treatment effects.)

i^{th} sample mean $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$

overall sample mean $\bar{\bar{Y}} = \frac{1}{N}\sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N}\sum_{i=1}^{a} n_i \bar{Y}_i$

Total sum of squares $SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{\bar{Y}}\right)^2$

Between Group Variation $SS_A = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 = \sum_{i=1}^{a} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$

Within Group Variation $SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$

Test for Treatment Effects

Break down of sum of squares

$$\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{\bar{Y}}\right)^2 \quad SS_T = SS_A + SS_E \quad \sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$$

Treatment mean squares $MS_A = \frac{SS_A}{a-1} = \frac{1}{a-1}\sum_{i=1}^{a} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$

Error mean squares $MS_E = \frac{SS_E}{N-a} = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$

H_1 true → μ_i not all the same → MS_A tends to be large

MS_E is unaffected by the population means.

Test for Treatment Effects

Treatment mean squares $MS_A = \frac{SS_A}{a-1} = \frac{1}{a-1}\sum_{i=1}^{a} n_i\left(\bar{Y}_i - \bar{\bar{Y}}\right)^2$

Error mean squares $MS_E = \frac{SS_E}{N-a} = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$

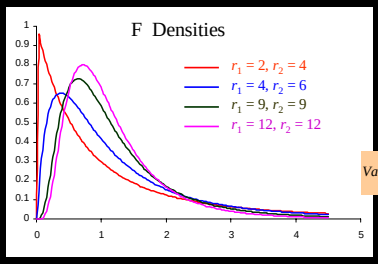
Test statistic $F = \frac{MS_A}{MS_E}$

Reject H_0 if $F_{obs} > F(a\text{-}1, N\text{-}a, \alpha)$.

$F(a-1, N-a, \alpha)$ —— Obtained from F distribution table

F Distribution

$$f(x) = \frac{\Gamma\left(\frac{r_1 + r_2}{2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)}\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} x^{\frac{r_1}{2}-1}\left(1 + \frac{r_1}{r_2}x\right)^{-\frac{r_1+r_2}{2}}, \quad x > 0$$

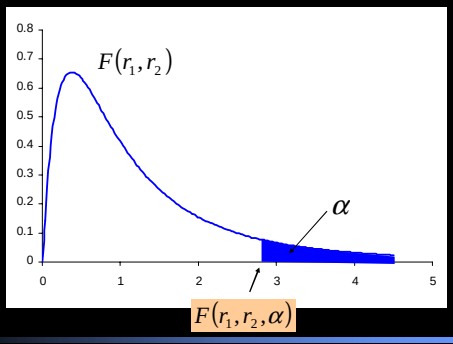


$X \sim F(r_1, r_2)$

$E(X) = \frac{r_2}{r_2 - 2}$

$Var(X) = \frac{2r_2^2(r_1 + r_2 - 2)}{r_1(r_2 - 2)^2(r_2 - 4)}$

F Distribution



F Distribution Table

α	$P(F \le f)$	Den. d.f. r_2	Numerator Degrees of Freedom, r_1						
			1	2	3	4	5	6	7
0.05	0.95	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8
0.025	0.975		647.79	799.50	864.16	899.58	921.85	937.11	948.22
0.01	0.99		4052	4999.5	5403	5625	5764	5859	5928
0.05	0.95	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
0.025	0.975		38.51	39.00	39.17	39.25	39.30	39.33	39.36
0.01	0.99		98.50	99.00	99.17	99.25	99.30	99.33	99.36
0.05	0.95	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
0.025	0.975		17.44	16.04	15.44	15.10	14.88	14.73	14.62
0.01	0.99		34.12	30.82	29.46	28.71	28.24	27.91	27.67
0.05	0.95	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
0.025	0.975		12.22	10.65	9.98	9.60	9.36	9.20	9.07
0.01	0.99		21.20	18.00	16.69	15.98	15.52	15.21	14.98
0.05	0.95	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
0.025	0.975		10.01	8.43	7.76	7.39	7.15	6.98	6.85
0.01	0.99		16.26	13.27	12.06	11.39	10.97	10.67	10.46
0.05	0.95	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
0.025	0.975		8.81	7.26	6.60	6.23	5.99	5.82	5.70
0.01	0.99		13.75	10.92	9.78	9.15	8.75	8.47	8.26
0.05	0.95	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
0.025	0.975		8.07	6.54	5.89	5.52	5.29	5.12	4.99
0.01	0.99		12.25	9.55	8.45	7.85	7.46	7.19	6.99

$F(3,4,0.05) = 6.59$ $F(4,6,0.01) = 9.15$

ANOVA Table

$H_0 : \alpha_1 = \alpha_2 = \cdots \alpha_a = 0$ vs $H_1 : \text{some } \alpha_i \neq 0$

Test statistic $F = \dfrac{MS_A}{MS_E}$ Reject H_0 if $F_{obs} > F(a\text{-}1, N\text{-}a, \alpha)$.

Source	SS	d.f.	MS	F-ratio
Treatment	SS_A	a - 1	$SS_A/(a$ - 1)	MS_A / MS_E
Error	SS_E	N - a	$SS_E/(N$ - a)	
Total	SS_T	N - 1		

Computational Formulae

i^{th} total $T_i = \sum_{j=1}^{n_i} Y_{ij}$ overall total $T_{..} = \sum_{i=1}^{a} T_i$

$$SS_A = \sum_{i=1}^{a} n_i \left(\bar{Y}_i - \bar{\bar{Y}}\right)^2 = \sum_{i=1}^{a} \frac{T_i^2}{n_i} - \frac{T_{..}^2}{N}$$

$$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{a} \frac{T_i^2}{n_i}$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{\bar{Y}}\right)^2 = \sum_{i=1}^{a}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{T_{..}^2}{N}$$

One-way ANOVA

Example : Color brightness of films

$a = 3$ $n_1 = n_2 = n_3 = 15$ $N = 45$

$T_1 = 452$ $T_2 = 578$ $T_3 = 378$ $T_{..} = 1408$ $\sum_{i=1}^{3}\sum_{j=1}^{15} Y_{ij}^2 = 46040$

$H_0 : \alpha_1 = \alpha_2 = \alpha_3$ vs $H_1 : \text{not } H_0$ $\alpha = 0.05$

$SS_A = 1363.38 \dfrac{578^2}{15} + \dfrac{378^2}{15} - \dfrac{1408^2}{45}$ $SS_E = 621.86 - 1363.38$

$SS_T = 1985.24 \dfrac{1408^2}{45}$

One-way ANOVA

Source	SS	d.f.	MS	F-ratio
Treatment	1363.38	2	681.69	46.03
Error	621.86	42	14.81	
Total	1985.24	44		

From F distribution table $F(2,42,0.05) \approx F(2,40,0.05) = 3.23$

$F - ratio = 46.03 > 3.23$

Reject H_0 at $\alpha = 0.05$.

The color brightness of the three brands of films are significantly different.

Estimation

Treatment effect : α_i

Point $\bar{Y}_i - \bar{\bar{Y}}$ Interval $\left(\bar{Y}_i - \bar{\bar{Y}}\right) \pm t_{N-a,\alpha/2} \sqrt{MS_E \left(\dfrac{1}{n_i} - \dfrac{1}{N}\right)}$

Difference in treatment effects : α_i - α_j

Point $\bar{Y}_i - \bar{Y}_j$ Interval $\left(\bar{Y}_i - \bar{Y}_j\right) \pm t_{N-a,\alpha/2} \sqrt{MS_E \left(\dfrac{1}{n_i} + \dfrac{1}{n_j}\right)}$

Estimation

Example : Color brightness of films

$\bar{Y}_1 = \dfrac{452}{15} = 30.13$ $\bar{Y}_2 = \dfrac{578}{15} = 38.53$ $\bar{Y}_3 = \dfrac{378}{15} = 25.2$ $\bar{\bar{Y}} = \dfrac{1408}{45} = 31.29$

95% C.I. For α_1 : $[-2.80, 0.48]) \pm (2.021) \sqrt{(14.81)\left(\dfrac{1}{15} - \dfrac{1}{45}\right)}$

95% C.I. For α_2 - α_3 : $[10.49, 16.17] \pm (2.0$ $\alpha_2 > \alpha_3$ $31)\left(\dfrac{1}{15} + \dfrac{1}{15}\right)$

95% C.I. For α_1 - α_2 : $[-11.24, -5.56]$ $\alpha_1 < \alpha_2$

95% C.I. For α_1 - α_3 : $[2.09, 7.77]$ $\alpha_1 > \alpha_3$

$\alpha_2 > \alpha_1 > \alpha_3$ Overall confidence < 95%

Two way ANOVA

Example : Brightness of synthetic fabric

Time (cycles)	Temperature 350°F	375°F	400°F
40	38, 32, 30	37, 35, 40	36, 39, 43
50	40, 45, 36	39, 42, 46	39, 48, 47

Two-way factorial ANOVA model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \quad k = 1,2,3 \quad j = 1,2,3 \quad i = 1,2$$

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \qquad \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$$

Two way ANOVA

Example : Brightness of synthetic fabric

```
MTB > ANOVA 'Bright' = Time Temp Time*Temp.
```

Analysis of Variance (Balanced Designs)

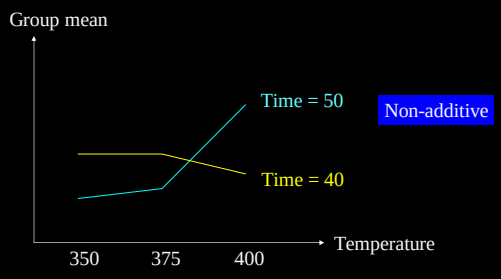
```
Factor      Type Levels Values
Time       fixed      2    40    50
Temp       fixed      3   350   375   400

Analysis of Variance for Bright

Source       DF         SS         MS       F      P
Time          1     150.22     150.22    9.69  0.009
Temp          2      80.78      40.39    2.61  0.115
Time*Temp     2       3.44       1.72    0.11  0.896
Error        12     186.00      15.50
Total        17     420.44
```

significant

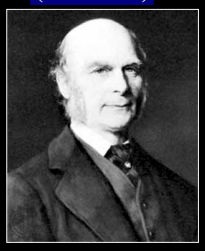
Interaction

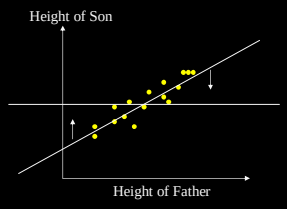


Non-additive

Regression

Sir Francis Galton
(1822 – 1911)





Height of the sons of fathers regressed towards the mean height of the population

Regression

Simple Linear Regression —— Linear

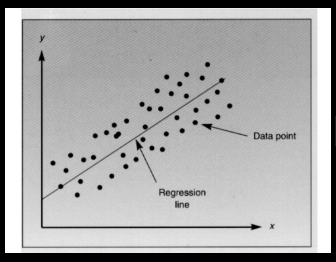
Model the relationship between dependent variable and one independent variable

Examples

Dependent variable (Y)	Independent variable (X)
Job performance	Extent of training
Return of a stock	Risk of the stock
Overall CGA	A-Level Score
Tree age (by C_{14})	Tree age (by tree rings)

Simple Linear Regression

Scatterplot



Regression line

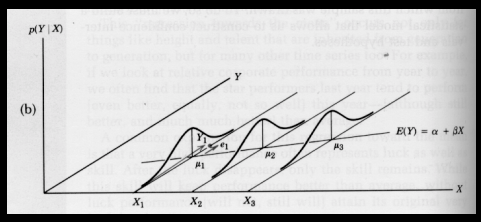
A line well fit the data

Simple Linear Regression Model

Data : $\{(X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n)\}$

$Y_i = \alpha + \beta X_i + \varepsilon_i \quad , i = 1,2,...,n$

$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ ← assumptions



Simple Linear Regression Model

Example : Y = Height of son (in cm)
X = Height of father (in cm)

More reasonable relationship : $E(Y) = 0.9X + 15$

X	Y	0.9X + 15	ε (Random Error)	Y
170	169.3	168	1.3	169.3
175	171.7	172.5		
180	174.6	177		
185	182.2	181.5	0.7	182.2
	↑ Observed	↑ Unobserved	↑ Unobserved	↑ Observed

Fit a regression line to the data

Estimation of Model Parameters

Sample statistics

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ $\quad S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$

$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$ $\quad S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$

$\hat{\beta} = b = \frac{S_{xy}}{S_{xx}}$ $\quad \hat{\alpha} = a = \bar{Y} - b\bar{X}$

Fitted regression line : $\hat{Y} = a + bX$

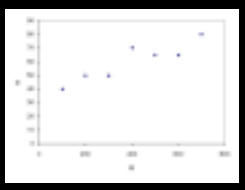
≠ True regression line : $E(Y) = \alpha + \beta X$

Fitting Regression Line

Example : Study of how wheat yield depends on fertilizer.

X = Fertilizer (in lb/acre) Y = Yield (in bu/acre)

X	100	200	300	400	500	600	700
Y	40	50	50	70	65	65	80



$\bar{X} = 400$ $\quad \bar{Y} = 60$

$\sum_{i=1}^{7} X_i^2 = 1400000$ $\quad \sum_{i=1}^{7} Y_i^2 = 26350$

$\sum_{i=1}^{7} X_i Y_i = 184500$

Fitting Regression Line

$\bar{X} = 400$ $\quad \bar{Y} = 60$

$\sum_{i=1}^{7} X_i^2 = 1400000$ $\quad \sum_{i=1}^{7} Y_i^2 = 26350$

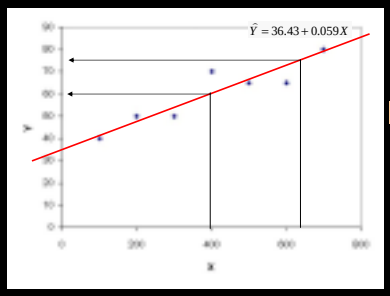
$\sum_{i=1}^{7} X_i Y_i = 184500$

$S_{xx} = 280000 - (7)(400)^2$ $\quad S_{yy} = 1150\,0 - (7)(60)^2$ $\quad S_{xy} = 16500 - (7)(400)(60)$

$b = \frac{16500}{280000} = 0.059$ $\quad a = 60 - (0.059)(400) = 36.43$

Fitted regression line : $\hat{Y} = 36.43 + 0.059X$

Fitting Regression Line



Prediction

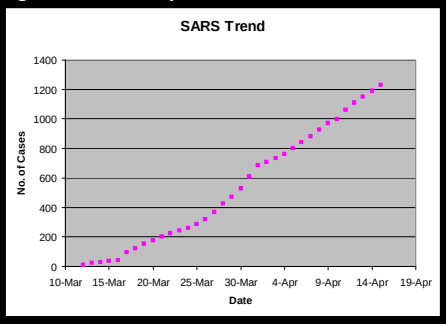
$X_0 = 650$

$\hat{Y}_0 = 3(\hat{Y}_0 = 74.78\ 400)$

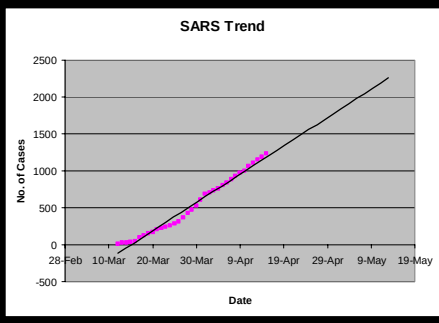
$X_0 = 0$

$\hat{Y}_0 = 36.43$?

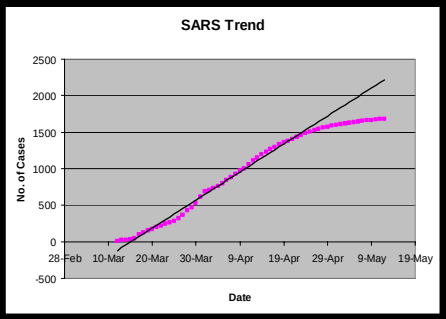
Danger of Extrapolation



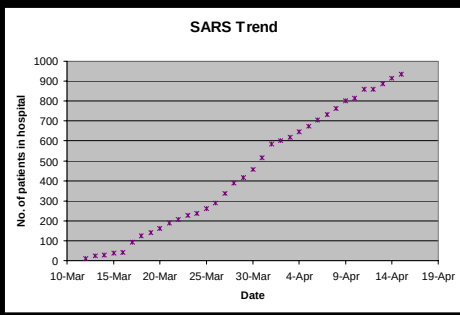
Danger of Extrapolation



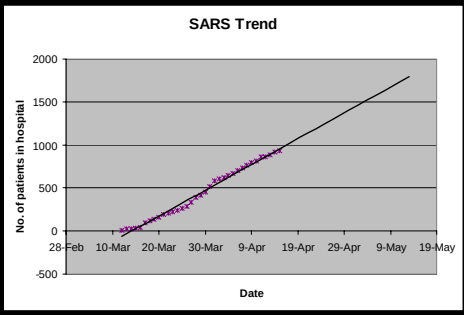
Danger of Extrapolation



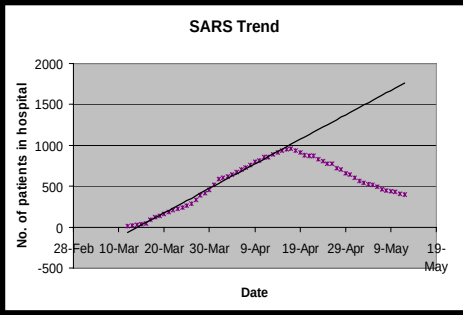
Danger of Extrapolation



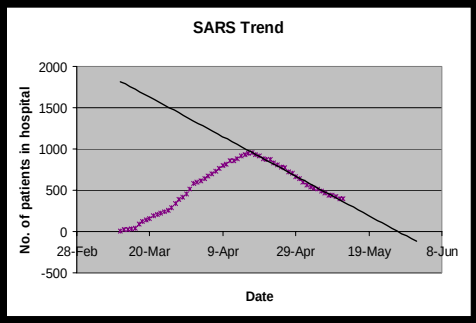
Danger of Extrapolation



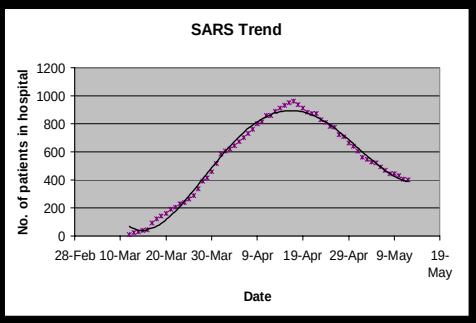
Danger of Extrapolation



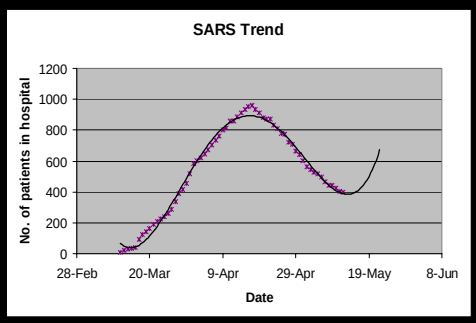
Danger of Extrapolation



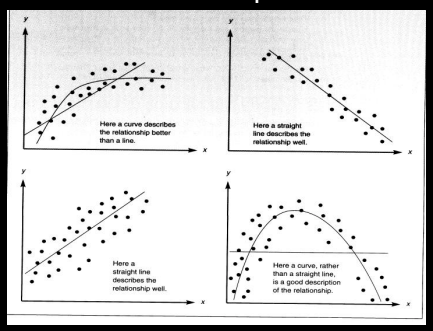
Danger of Extrapolation



Danger of Extrapolation



Nonlinear Relationships



Association ≠ Causation

Example : Price and Demand for gas

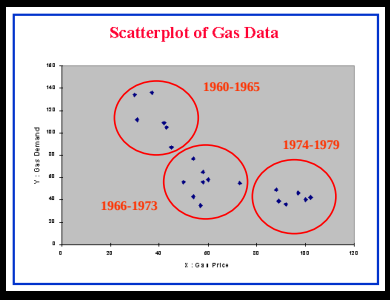
Year	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
Price	30	31	37	42	43	45	50	54	54	57
Demand	134	112	136	109	105	87	56	43	77	35

Year	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
Price	58	58	60	73	88	89	92	97	100	102
Demand	65	56	58	55	49	39	36	46	40	42

Fitted regression line : Demand = 139.24 – 1.11 Price

? ~~Low demand is due to high price.~~ ?

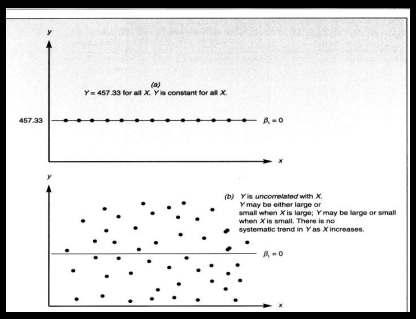
Simpson's Paradox



Year

Price Demand

Test For Regression Effect



Test For Regression Effect

Test $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

Fitted values $\hat{Y}_i = a + bX_i$

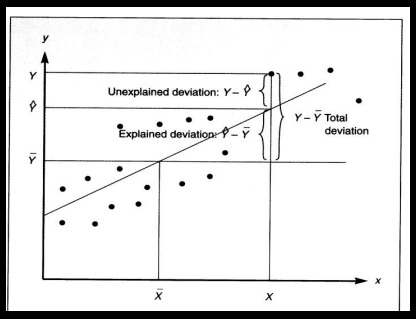
Residuals $r_i = Y_i - \hat{Y}_i$ $\neq$ Random Error $\varepsilon_i = Y_i - \alpha - \beta X_i$

Decomposition of Variation

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Variation of Y | Explained variation | Unexplained variation

Test For Regression Effect



Test For Regression Effect

Break down of sum of squares

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

$$SS_T = SS_R + SS_E$$

Total sum of squares $SS_T = S_{yy}$

Regression sum of squares $SS_R = b^2 S_{xx} = \frac{S_{xy}^2}{S_{xx}}$

Error sum of squares $SS_E = S_{yy} - b^2 S_{xx} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

Test For Regression Effect

$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

$MS_R = \frac{SS_R}{1} = SS_R$ $\quad$ $MS_E = \frac{SS_E}{n-2}$

Test statistic $F = \frac{MS_R}{MS_E}$ Reject H_0 if $F_{obs} > F(1, n-2, \alpha)$.

ANOVA table

Source	SS	d.f.	MS	F-ratio
Regression	SS_R	1	SS_R	MS_R / MS_E
Error	SS_E	n - 2	$SS_E/(n-2)$	
Total	SS_T	n - 1		

Test For Regression Effect

Example : Wheat yield example

Regression line $\hat{Y} = 36.43 + 0.059X$

Source	SS	d.f.	MS	F-ratio
Regression	974.68	1	974.68	27.805
Error	175.32	5	35.064	
Total	1150	6		

$F(1,5,0.05) = 6.61 < 27.805$

Reject H_0 at $\alpha = 0.05$.

Coefficient of Determination

Strong relationship → High prediction power

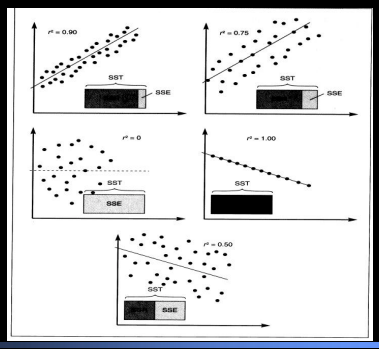
$R^2 = \frac{SS_R}{SS_T}$ — Explained variation / Total variation

$0 \le R^2 \le 1$

No linear relationship — Perfect linear relationship

Example : $R^2 = \frac{974.68}{1150} = 84.8\%$

Coefficient of Determination



C.I. For Regression Parameters

100(1 - α)% C.I. for β : $b \pm t_{n-2,\alpha/2}\sqrt{\frac{MS_E}{S_{xx}}}$

100(1 - α)% C.I. for α : $a \pm t_{n-2,\alpha/2}\sqrt{MS_E\left(\frac{1}{n}+\frac{\bar{X}}{S_{xx}}\right)}$

Large S_{xx} → More accurate estimates

C.I. For Regression Parameters

Example : Wheat yield example

Regression line $\hat{Y} = 36.43 + 0.059X$

Source	SS	d.f.	MS	F-ratio
Regression	974.68	1	974.68	27.805
Error	175.32	5	35.064	
Total	1150	6		

95% C.I. for β : $[0.0302, 0.0878] \sqrt{\frac{35.064}{280000}}$

95% C.I. for α : $[32.892, 37.236] \sqrt{35.64\left(\frac{1}{7}+\frac{(400)^2}{280000}\right)}$

Prediction

Predict the value of Y_0 at a fixed value of $X = X_0$

Point prediction : $\hat{Y}_0 = a + bX_0$

100(1 - α)% prediction interval (P.I.)

$$\hat{Y}_0 \pm t_{n-2,\alpha/2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{S_{xx}}\right)}$$

Prediction

Example : Wheat yield example

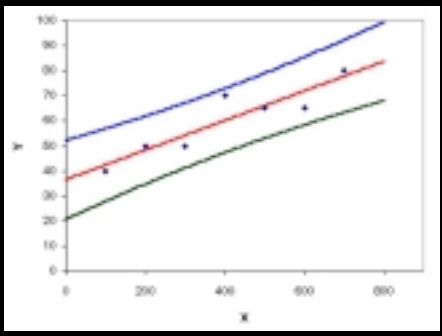
Regression line $\hat{Y} = 36.43 + 0.059X$

Source	SS	d.f.	MS	F-ratio
Regression	974.68	1	974.68	27.805
Error	175.32	5	35.064	
Total	1150	6		

At $X_0 = 450$, $\hat{Y}_0 = 62.98 + (0.059)(450)$

90% prediction interval $[50.143, 75.817] \sqrt{4\left(1+\frac{1}{7}+\frac{(450-400)^2}{280000}\right)}$

Prediction



Multiple Linear Regression

Example : Fuel consumption data

Data Display

Row	State	POP	TAX	NLIC	INC	ROAD	FUELC	DLIC
1	ME	1029	9.00	540	3.571	1.976	557	52.4781
2	NH	771	9.00	441	4.092	1.250	404	57.1984
3								087
4								771
5	RI	968	8.00	527	4.399	0.431	397	54.4422
6	CN	3082	10.00	1760	5.342	1.333	1408	57.1058
7	NY	18366	8.00	8278	5.319	11.868	6312	45.0724
8	NJ	7367	8.00	4074	5.126	2.138	3439	55.3007
9	PA	11926	8.00	6312	4.447	8.577	5528	52.9264
10	OH	10783	7.00	5948	4.512	8.507	5375	55.1609
11	IN	5291	8.00	2804	4.391	5.939	3068	52.9957
12	IL	11251	7.50	5903	5.126	14.186	5301	52.4664

$$FUEL = \beta_0 + \beta_1 TAX + \beta_2 DLIC + \beta_3 INC + \beta_4 ROAD + \varepsilon$$

Multiple Linear Regression

Example : Fuel consumption data

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	3991.92	997.98	22.68	0.000
Error	43	1892.05	44.00		
Total	47	5883.96			

Unusual Observations

Obs.	TAX	FUEL	Fit	Stdev.Fit	Residual	St.Resid
37	5.0	63.963	64.758	3.723	-0.795	-0.14 X
40	7.0	96.812	73.371	2.102	23.441	3.73R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.