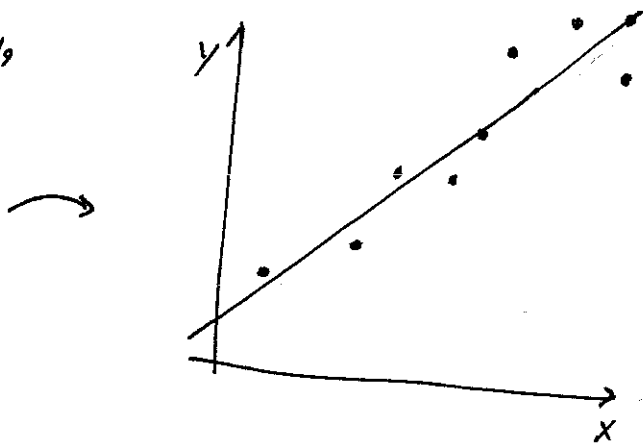
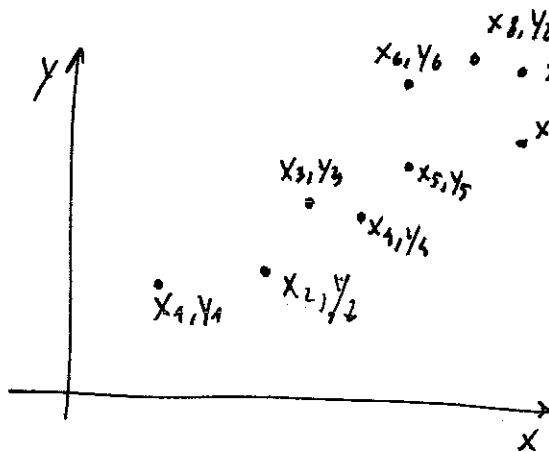


8.0 A

REGRESSION

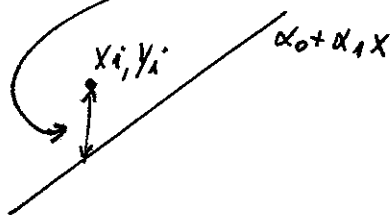


$$Y = \beta_0 + \beta_1 X$$

HOW TO ESTIMATE β_0 AND β_1 ?

LEAST SQUARES

$$F(\alpha_0, \alpha_1) = \sum_{i=1}^n [y_i - (\alpha_0 + \alpha_1 x_i)]^2$$



WE LOOK FOR α_0 AND α_1 WHICH MINIMIZE F

$$\frac{\partial F}{\partial \alpha_0} = 0$$

$$\frac{\partial F}{\partial \alpha_1} = 0$$

$$\begin{cases} 2 \sum_{i=1}^n [y_i - (\alpha_0 + \alpha_1 x_i)] \cdot (-1) = 0 \\ 2 \sum_{i=1}^n [y_i - (\alpha_0 + \alpha_1 x_i)] \cdot (-x_i) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^n y_i - (\alpha_0 + \alpha_1 x_i) = 0 \\ \sum_{i=1}^n x_i y_i - \alpha_0 x_i - \alpha_1 x_i^2 = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^n y_i - n \alpha_0 - \alpha_1 \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i - \alpha_0 \sum_{i=1}^n x_i - \alpha_1 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

$$\begin{cases} n \cdot \bar{y}_n - n \alpha_0 - \alpha_1 n \bar{x}_n = 0 \\ \sum_{i=1}^n x_i y_i - \alpha_0 n \bar{x}_n - \alpha_1 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

$$\begin{cases} \alpha_0 = \bar{y}_n - \alpha_1 \bar{x}_n \\ \sum_{i=1}^n x_i y_i - \bar{y}_n n \bar{x}_n + \alpha_1 \bar{x}_n^2 n - \alpha_1 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

8.0 B

$$\begin{cases} \alpha_0 = \bar{y}_m - \alpha_1 \bar{x}_m \\ \left(\sum_{i=1}^n x_i y_i - n \bar{x}_m \bar{y}_m \right) - \alpha_1 \left(\sum_{i=1}^n x_i^2 - n \bar{x}_m^2 \right) = 0 \end{cases}$$

↓
SS_{xy}

↓
SS_{xx}

$$\alpha_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\alpha_0 = \bar{y}_m - \frac{SS_{xy}}{SS_{xx}} \bar{x}_m$$

↓
 $\hat{\beta}_1$

↓
 $\hat{\beta}_0$

ESTIMATIONS FOR β_1 AND β_0

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \text{ESTIMATED } y_i$$

$$F(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 = \text{AFTER SOME CALCULATION}$$

$$= SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}$$

THIS IS THE SUM OF SQUARES OF ERRORS

S.S.E.

$$d = 1 - \frac{sse}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}}$$

COEFFICIENT OF DETERMINATION
 $d \in [0, 1]$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

PEARSON CORRELATION

$r \in [-1, 1]$

AND CLEARLY $r^2 = d$

d MEANS: THE PROPORTION OF y VARIANCE WHICH IS EXPLAINED, THROUGH THE MODEL, BY THE VARIANCE OF x . $d=1 \Rightarrow$ PERFECT EXPLANATION. $d=0$ NO EXPLANATION.

8.0 c

UNTIL NOW NO HYPOTHESIS WAS USED!

NOW WE SUPPOSE

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{E}_i$$

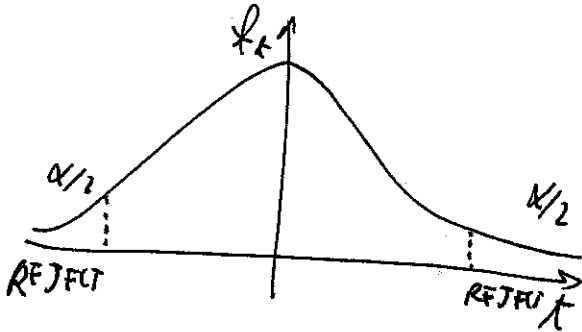
\hat{E}_i ARE THE VALUES COMING FROM

R.V. E_i . WE SUPPOSE E_i INDEPENDENT AND NORMALLY DISTRIBUTED WITH 0 MEAN AND THE SAME VARIANCE σ^2

TEST:
$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

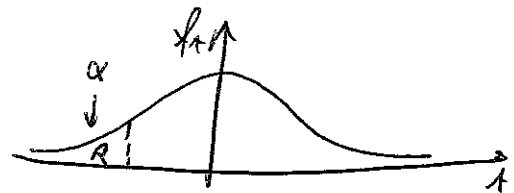
STATISTICS
$$T = \frac{\sqrt{n-2} SS_{xy}}{\sqrt{SS_{yy} SS_{xx} - SS_{xy}^2}}$$

IT IS DISTRIBUTED LIKE A STUDENT-T WITH $n-2$ DEGREES OF FREEDOM



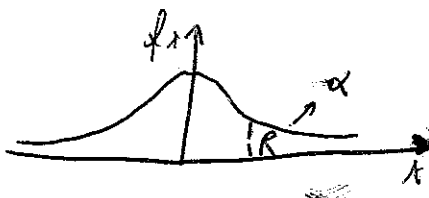
FOR n LARGE ENOUGH (≥ 30), STUDENT-T $\approx N(0,1)$

ONE-TAILED TEST:
$$\begin{cases} H_0: \beta_1 \geq 0 \\ H_1: \beta_1 < 0 \end{cases}$$



SINCE β_1 AND T HAVE THE SAME SIGN, IT IS EASY TO LOCATE THE CORRECT REJECTION REGION

$$\begin{cases} H_0: \beta_1 \leq 0 \\ H_1: \beta_1 > 0 \end{cases}$$



8.0

IN CASE OF SEVERAL X VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + E_i$$

$$F(\alpha_0, \alpha_1, \dots, \alpha_k)$$

CALCULATING $\frac{\partial F}{\partial \alpha_0} = 0$ $\frac{\partial F}{\partial \alpha_1} = 0 \dots \frac{\partial F}{\partial \alpha_k} = 0$ WE FIND THE ESTIMATES FOR $\beta_0, \beta_1, \dots, \beta_k$

$$d = 1 - \frac{F(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)}{SS_{yy}} = 1 - \frac{S.S.E.}{SS_{yy}}$$

IN CASE OF NON LINEAR RELATIONSFOR
EXAMPLE

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 \ln X_i + \beta_4 \exp(X_i) + E_i$$

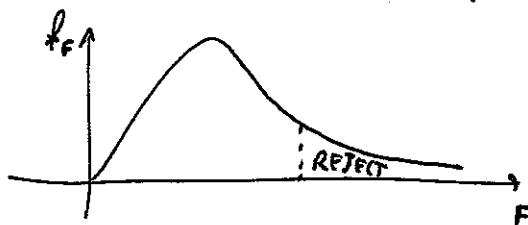
EVERYTHING ELSE IS THE SAME!

$$d = 1 - \frac{S.S.E.}{SS_{yy}}$$

TEST FOR SEVERAL X
OR FOR NONLINEAR

$$\left\{ \begin{array}{l} H_0: \beta_1 = 0 \quad \beta_2 = 0 \dots \beta_k = 0 \\ H_1: \beta_1 \neq 0 \quad \text{OR} \quad \beta_2 \neq 0 \dots \text{OR} \quad \beta_k \neq 0 \end{array} \right.$$

STATISTICS IS $F = \frac{\frac{d}{k}}{\frac{1-d}{n-(k+1)}}$

K IS THE NUMBER
OF β , EXCLUDING β_0 IS DISTRIBUTION LIKE A FISHER F WITH $(k; n - (k+1))$ D.F.

8.0 E

DUMMY VARIABLES

WHEN WE INTRODUCE CATEGORICAL VARIABLES X , WE MUST INTRODUCE $K-1$ VARIABLES WITH VALUES 0 AND 1 FOR K CATEGORIES.

FOR EXAMPLE, DEGREE COURSE (E&M, BS, PPE, MASTER)

$$X_1 = \begin{cases} 0 & \text{NO E\&M} \\ 1 & \text{E\&M} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{NO BS} \\ 1 & \text{BS} \end{cases}$$

$$X_3 = \begin{cases} 0 & \text{NO PPE} \\ 1 & \text{PPE} \end{cases}$$

WE DO NOT NEED A FOURTH VARIABLE FOR MASTER, BECAUSE WHEN $X_1=0$ $X_2=0$ $X_3=0$ IT IS CLEAR THAT THE SUBJECT BELONGS TO CATEGORY MASTER

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{E}_i$$

NOTE THAT TAKING BS STUDENTS ONLY WE HAVE

$$Y_i = \hat{\beta}_0 + \hat{\beta}_2 + \hat{E}_i$$

NOTE THAT, IN ORDER TO PERFORM TESTS, WE NEED E_i INDEPENDENT AND $N(0, \sigma^2)$

WHenever X_i is TIME, IT IS A TYPICAL CASE WHERE ERRORS E_i MAY BE DEPENDENT.

TO CHECK THIS WE WILL USE DURBIN-WATSON TEST

8.1 A

TEST WHETHER THE NUMBER OF EMPLOYEES IN CASINOS CAN PREDICT THE CRIME RATE USING THESE DATA ON TOWN IN MISSISSIPPI

EMPL.	15	18	24	22	25	29	30	32	35	38
CRIMES BY 1000 INHABITANTS	1,35	1,63	2,33	2,41	2,63	2,93	3,41	3,26	3,63	4,15

WE TRY TO ESTIMATE Y_i USING X_i

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

$$\hat{\beta}_1 = \frac{SS_{YX}}{SS_{XX}}$$

$$\bar{X}_{10} = 26,8$$

$$\bar{Y}_{10} = 2,773$$

$$SS_{XX} = \sum_{i=1}^{10} X_i^2 - n \bar{X}_{10}^2 = 15^2 + 18^2 + \dots + 38^2 - 7182,4 =$$

$$= 7668 - 7182,4 = 485,6$$

$$SS_{YX} = \sum_{i=1}^{10} X_i Y_i - 10 \bar{Y}_{10} \bar{X}_{10} = 15 \cdot 1,35 + 18 \cdot 1,63 + \dots + 38 \cdot 4,15 - 10 \cdot 26,8 \cdot 2,773 =$$

$$= 800,62 - 743,164 = 57,456$$

$$\hat{\beta}_1 \approx 0,1183$$

$$\hat{\beta}_0 \approx -0,3974$$

$$\hat{Y}_i = -0,3974 + 0,1183 X_i$$

$$\text{LET'S EVALUATE } R^2 = \frac{SS_{XUY}}{SS_{XX} SS_{YY}} = \frac{57,456^2}{485,6 \cdot 6,9780} = \frac{57,456^2}{3388,52} \approx 0,9742$$

THE MODEL EXPLAINS 97% OF THE VARIANCE OF Y_i USING X_i

8.1 B

LET NOW TEST

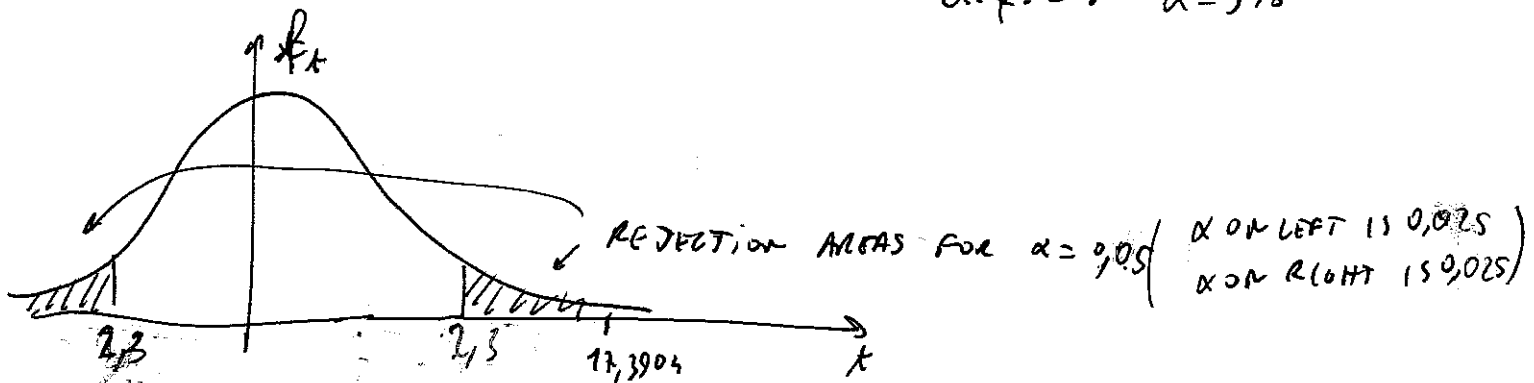
$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

WE HAVE $\hat{\beta}_1 \approx 0,1183$

$$t \approx \frac{\sqrt{n-2} \cdot SS_{yx}}{\sqrt{SS_{yy} SS_{xx} - SS_{yx}^2}} = \frac{\sqrt{8} \cdot 57,456}{\sqrt{6,979 \cdot 475,6 - 57,456^2}}$$

$$= \frac{162,51}{1 \cdot 9,3448} = 17,3906$$

d.f. = 8 $\alpha = 5\%$



WE REJECT $H_0 \Rightarrow$ model is useful

D

8.1 c

FOR PREVIOUS EXERCISE, CALCULATE \hat{E}_j AND SUM OF SQUARES OF ERRORS

\hat{Y}_j ARE THE PREDICTED Y_j AND $\hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$

$$\hat{Y}_1 = -0,3974 + 0,1183 \cdot 15 = 1,377 \quad \hat{Y}_2 = -0,3974 + 0,1183 \cdot 18 = 1,732$$

$$\hat{Y}_3 = -0,3974 + 0,1183 \cdot 24 = 2,442 \quad \hat{Y}_4 = 2,205 \quad \hat{Y}_5 = 2,56$$

$$\hat{Y}_6 = 3,033 \quad \hat{Y}_7 = 3,152 \quad \hat{Y}_8 = 3,388 \quad \hat{Y}_9 = 3,743$$

$$\hat{Y}_{10} = 4,098$$

$$\hat{E}_j = Y_j - \hat{Y}_j \quad \hat{E}_1 = 1,35 - 1,377 = -0,027$$

$$\hat{E}_2 = 1,63 - 1,732 = -0,102 \quad \hat{E}_3 = -0,112 \quad \hat{E}_4 = 0,205$$

$$\hat{E}_5 = 0,070 \quad \hat{E}_6 = -0,103 \quad \hat{E}_7 = 0,258 \quad \hat{E}_8 = -0,128$$

$$\hat{E}_9 = -0,113 \quad \hat{E}_{10} = 0,052$$

$$S.S.E. = \sum_{j=1}^{10} \hat{E}_j^2 = (-0,027)^2 + (-0,102)^2 + \dots + (0,052)^2 = 0,1798$$

$$\text{OR } S.S.E. = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}} = 6,978 - \frac{57,456^2}{485,6} = 0,1798$$

□

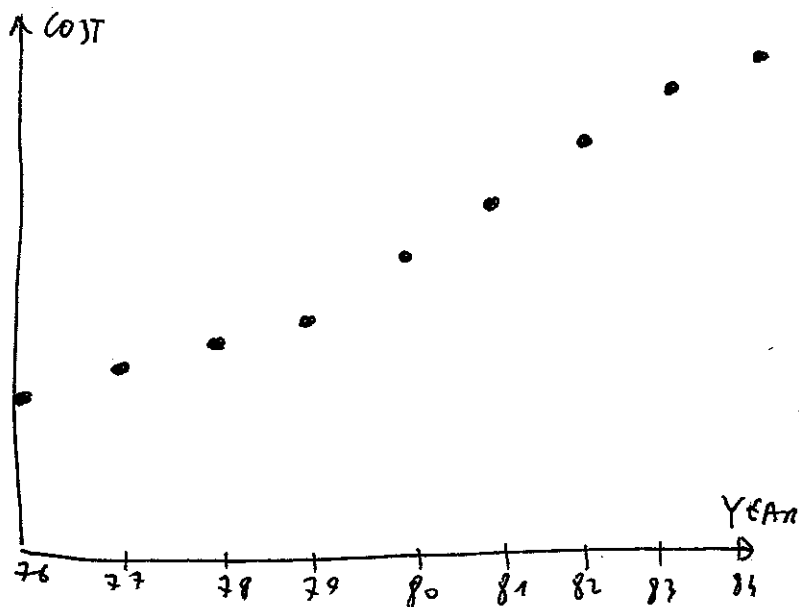
8.2 AVERAGE DISEASE INSURANCE COST IN U.S.
HAS GROWTH BETWEEN 1976-1984.

TEST WHETHER IT IS RELATED TO TIME
OR NOT WITH A LINEAR REGRESSION MODEL

YEAR	1976	1977	1978	1979	1980	1981	1982	1983	1984
AV. COST	184,7	202,4	219,4	239,7	265,9	294,5	328,7	357,3	378,0

ASSUME $\hat{\epsilon}_i$ ARE INDEPENDENT

THE SCATTER PLOT IS



IT SEEMS A LINEAR RELATION

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

WE CAN TAKE $X_1 = 1976 \dots X_9 = 1984$
OR, TO SIMPLIFY CALCULATIONS,

$$X_1 = 0 \quad X_2 = 1 \dots X_9 = 8$$

$$\bar{X}_9 = 4 \quad \bar{Y}_9 = 274,5 \quad SS_{xx} = 60 \quad SS_{xy} = 1511,3 \quad \hat{\beta}_1 = \frac{1511,3}{60} \approx 25,18$$

$$\hat{\beta}_0 = \bar{Y}_9 - \hat{\beta}_1 \bar{X}_9 = 274,5 - 100,75 = 173,75$$

$$\hat{Y}_i = 173,75 + 25,18 X_i \Rightarrow \text{EACH YEAR THE AV. INSR. COST RAISES BY 25,18}$$

$$r^2 = \left(\frac{SS_{xy}}{SS_{xx} SS_{yy}} \right)^2 = \frac{1511,3^2}{60 \cdot 38489,19} = 0,989 \quad \hat{r} = 0,994$$

$r^2 = 0,989 \Rightarrow 98,9\%$ OF Y VARIANCE IS DUE TO TIME

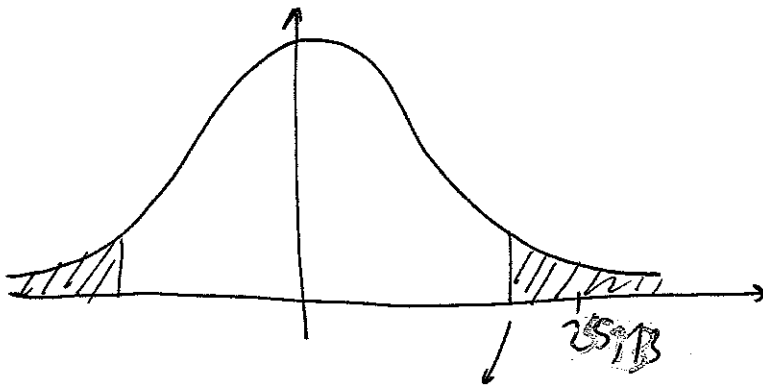
8.2 B

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$$

$$\beta_1 \approx 25,18$$

HYPOTHESES: E_i INDEPENDENT
AND $E_i \sim N(0, \sigma^2)$

$$t = \frac{\sqrt{n-2} \cdot SS_{xy}}{\sqrt{SS_{yy} \cdot SS_{xx} - SS_{xy}^2}} = \frac{\sqrt{7} \cdot 1511,3}{\sqrt{38489,19 \cdot 60 - 1511,3^2}} = \frac{3998,5}{159,13} = 25,13$$



d.f. = 7

THIS VALUE IS 2,36 FOR $\alpha = 0,05$
3,50 FOR $\alpha = 0,01$

WE REJECT H_0 (AT LEAST FOR EVERY $\alpha > 0,01$)

↓

$\beta_1 = 0$ IS REJECTED

NOTE: $\hat{y}_i = 173,75 + 25,18 x_i$ IS AN APPROXIMATION OF y_i USING ONLY FOR x_i VALUES CLOSE TO THE INTERVAL $[0,9]$ (WHICH IS 1926 - 1984). AFTER 1984 OR BEFORE 1926 WE DO NOT HAVE ANY INFORMATION ON THE REAL RELATION BETWEEN INSURANCE AND TIME

□

NOTE: TO DO THIS TEST WE ASSUME THAT ERRORS ARE NOT CORRELATED FROM A YEAR TO YEAR

8.3A

Y	90	72	54	42	30	12	$\bar{Y} = 50$
X ₁	3	5	6	8	12	14	$\bar{X}_1 = 8$
X ₂	16	10	7	4	3	2	$\bar{X}_2 = 7$

TEST WHETHER Y CAN BE PREDICTED BY X₁ AND X₂

$$F = \sum_{i=1}^6 (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

$$0 = \frac{\partial F}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^6 () = 0 \Rightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

$$0 = \frac{\partial F}{\partial \hat{\beta}_1} = -2 \left(\sum_{i=1}^6 Y_i X_{1i} - \hat{\beta}_1 \sum_{i=1}^6 X_{1i}^2 - \hat{\beta}_2 \sum_{i=1}^6 X_{2i} X_{1i} - \hat{\beta}_0 \sum_{i=1}^6 X_{1i} \right) = 0$$

$$0 = \frac{\partial F}{\partial \hat{\beta}_2} = -2 \left(\sum_{i=1}^6 Y_i X_{2i} - \hat{\beta}_1 \sum_{i=1}^6 X_{1i} X_{2i} - \hat{\beta}_2 \sum_{i=1}^6 X_{2i}^2 - \hat{\beta}_0 \sum_{i=1}^6 X_{2i} \right) = 0$$

FROM THE SECOND $\sum_{i=1}^6 Y_i X_{1i} = \left(\sum_{i=1}^6 \hat{\beta}_1 X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^6 X_{2i} X_{1i} \right) + 6 \bar{X}_1 \bar{Y} - 6 \hat{\beta}_1 \bar{X}_1^2 - 6 \hat{\beta}_2 \bar{X}_1 \bar{X}_2$

$$\sum_{i=1}^6 X_{1i} Y_i - 6 \bar{X}_1 \bar{Y} = \hat{\beta}_1 \left(\sum_{i=1}^6 X_{1i}^2 - 6 \bar{X}_1^2 \right) + \hat{\beta}_2 \left(\sum_{i=1}^6 X_{2i} X_{1i} - 6 \bar{X}_1 \bar{X}_2 \right)$$

FROM THE THIRD THE RESULT IS VERY SIMILAR

$$\sum_{i=1}^6 X_{2i} Y_i - 6 \bar{X}_2 \bar{Y} = \hat{\beta}_1 \left(\sum_{i=1}^6 X_{1i} X_{2i} - 6 \bar{X}_1 \bar{X}_2 \right) + \hat{\beta}_2 \left(\sum_{i=1}^6 X_{2i}^2 - 6 \bar{X}_2^2 \right)$$

$$\begin{cases} \hat{\beta}_1 SS_{11} + \hat{\beta}_2 SS_{12} = SS_{1Y} \\ \hat{\beta}_1 SS_{12} + \hat{\beta}_2 SS_{22} = SS_{2Y} \end{cases}$$

IN OUR CASE $SS_{11} = 474 - 384 = 90$
 $SS_{12} = 236 - 336 = -100$ $SS_{1Y} = 1818 - 2400 = -582$
 $SS_{22} = 434 - 294 = 140$ $SS_{2Y} = 2820 - 2100 = 720$

8.3 B

$$\begin{cases} 90 \hat{\beta}_1 - 100 \hat{\beta}_2 = -582 \\ -100 \hat{\beta}_1 + 140 \hat{\beta}_2 = 720 \end{cases}$$

$$\hat{\beta}_0 = 50 - \hat{\beta}_1 \cdot 8 - \hat{\beta}_2 \cdot 7$$

FROM THE SECOND ONE

$$\hat{\beta}_1 = -7,2 + 1,4 \hat{\beta}_2$$

THE FIRST ONE BECOMES

$$-90 \cdot 7,2 + 90 \cdot 1,4 \hat{\beta}_2 - 100 \hat{\beta}_2 = -582$$

$$\downarrow$$

$$-648 + 26 \hat{\beta}_2 = -582$$

$$\hat{\beta}_2 = +\frac{66}{26} = +\frac{33}{13} \approx 2,54$$

$$\hat{\beta}_1 = -7,2 + \frac{46,2}{13} = \frac{-47,4}{13} = -\frac{237}{65} \approx -3,65$$

$$\begin{aligned} \hat{\beta}_0 &= 50 + 8 \cdot \frac{237}{65} - 7 \cdot \frac{33}{13} = 50 + \frac{1896}{65} - \frac{231}{13} = \frac{47250 + 24648 - 15015}{845} \\ &= \frac{51883}{845} = 61,4 \end{aligned}$$

$$\hat{Y} = 61,4 - 3,65 X_1 + 2,54 X_2$$

$$\begin{aligned} SSE &= \sum_{i=1}^6 (Y_i - 61,4 + 3,65 X_{1i} - 2,54 X_{2i})^2 = \\ &= (90 - 91,09)^2 + (72 - 68,55)^2 + (54 - 57,28)^2 + (42 - 42,36)^2 + \\ &+ (30 - 25,22)^2 + (11 - 15,38)^2 = 52,6914 \end{aligned}$$

$$SS_{YY} = 4008$$

$$d = 1 - \frac{52,6914}{4008} = 0,9869$$

$$\Downarrow$$

THE MODEL EXPLAINS A
LOT OF Y VARIATION

8.3 C

LET'S TEST USEFULNESS $\begin{cases} H_0: \beta_1 = 0 \text{ or } \beta_2 = 0 \\ H_1: \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \end{cases}$

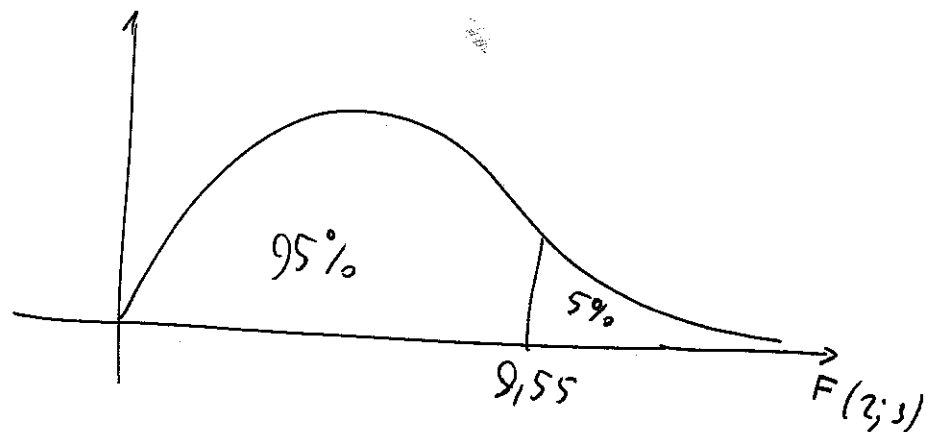
HYPOTHESIS: E_i INDEPENDENT AND $E_i \sim N(0, \sigma^2)$

THE STATISTICS IS $F = \frac{d/k \text{ SSE}}{(1-d)/(n-k-1)}$

$$n=6 \quad k=2$$

$$F_{\text{VALUE}} = \frac{0,49345}{0,00436} = 113,13$$

DEGREE OF FREEDOM 2; 3



H_0 IS REJECTED

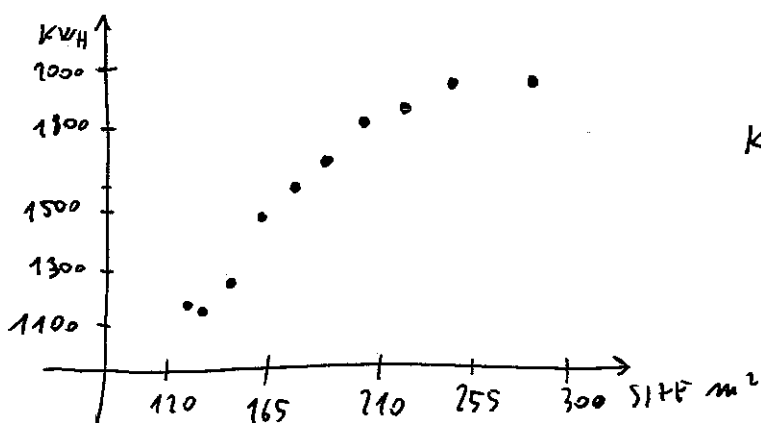
⇓

THE MODEL IS USEFUL

8.4

TEST WHETHER MONTHLY ELECTRICAL POWER USAGE CAN BE PREDICTED BY A PARABOLIC REGRESSION MODEL BASED ON THE SIZE OF HOME

SITE	129	135	147	160	171	184	198	223	240	293
KWH	1182	1172	1264	1493	1571	1711	1904	1840	1956	1954



$$KWH = \beta_0 + \beta_1 \text{SITE} + \beta_2 \text{SITE}^2$$

$$F = \sum_{i=1}^{10} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

$$\frac{\partial F}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{10} y_i - \beta_1 x_i - \beta_2 x_i^2 = 10 \cdot \beta_0$$

$$\frac{\partial F}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{10} y_i x_i - \beta_1 x_i x_i - \beta_2 x_i^2 x_i - \beta_0 x_i = 0$$

$$\frac{\partial F}{\partial \beta_2} = 0 \Rightarrow \sum_{i=1}^{10} y_i x_i^2 - \beta_1 x_i x_i^2 - \beta_2 x_i^2 x_i^2 - \beta_0 x_i^2 = 0$$

$$\hat{\beta}_0 = -1216,1 \quad \hat{\beta}_1 = 23,989 \quad \hat{\beta}_2 = -0,045 \quad \text{S.S.E.} = 15332,8$$

$$SS_{yy} = 846402$$

$$d = 1 - \frac{\text{S.S.E.}}{SS_{yy}} = 1 - \frac{15332,8}{846402} = 0,98188$$

$$\begin{cases} H_0: \beta_1 = 0 \text{ or } \beta_2 = 0 \\ H_1: \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \end{cases}$$

$$F = \frac{d/2}{(1-d)/(10-3)} = 189,9$$

$$F_{0,05;2;7} = 4,74 \Rightarrow \text{WE REJECT } H_0$$

8.5 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$

$n = 20$ $\sum_{i=1}^{20} (Y_i - \hat{Y}_i)^2 = 11,37$ $\sum_{i=1}^{20} (Y_i - \bar{Y})^2 = 23,75$

EVALUATE d . EVALUATE USEFULNESS WITH $\alpha = 0,05$

FROM THEORY $d = 1 - \frac{SSE}{SS_{yy}}$ AND

$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ AND $SS_{yy} = \sum_{i=1}^{20} (Y_i - \bar{Y})^2$

\Downarrow
 $d = 1 - \frac{11,37}{23,75} = \frac{11,38}{23,75} = 0,479$

ONLY LESS THAN 50% OF Y VARIANCE IS EXPLAINED BY X_1 AND X_2 VARIANCES.

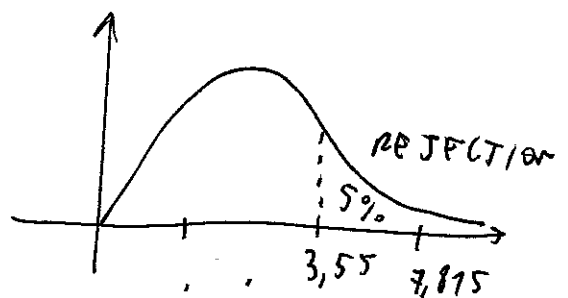
TO TEST ITS USEFULNESS
 HYP0: E_i INDEP. AND $E_i \sim N(0, \sigma^2)$ $\left\{ \begin{array}{l} H_0: \beta_1 = 0 \quad \beta_2 = 0 \\ H_1: \beta_1 \neq 0 \text{ OR } \beta_2 \neq 0 \end{array} \right.$

$F = \frac{\frac{d}{k}}{(1-d)/(n-(k+1))} = \frac{0,479}{2} \cdot \frac{17}{0,521} = \frac{8,143}{1,072} = 7,615$

$F_{2,18}$ FOR $\alpha = 0,05$ IS $= 3,55$

\Downarrow
 WE REJECT H_0 .

THE MODEL IS USEFUL FOR $\alpha = 0,05$



D

8.6] BUILD A MODEL TO PREDICT GRADUATION MARK BASED ON SUBJECT'S SEX

MALE: 100 105 87 90 83 77
 FEMALE: 105 101 93 90 95

$$X = \begin{cases} 1 & \text{FEMALE} \\ 0 & \text{NOT FEMALE} \end{cases}$$

$$Y = \beta_0 + \beta_1 X + E$$

$$SS_{xx} = 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 - 11 \cdot 0,455^2 \approx 2,73$$

$$SS_{yy} = 100^2 + 105^2 + \dots + 95^2 - 11 \cdot 93,3^2 = 96512 - 95687,81 \approx 814,19$$

$$SS_{xy} = 0 \cdot 100 + 0 \cdot 105 + 0 \cdot 87 + 0 \cdot 90 + 0 \cdot 83 + 0 \cdot 77 + 1 \cdot 105 + 1 \cdot 101 + 1 \cdot 93 + 1 \cdot 90 + 1 \cdot 95 - 11 \cdot 0,455 \cdot 93,3 \approx 17,6$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{17,6}{2,73} = 6,45 \rightarrow \text{BEING FEMALE IMPROVES GRADUATION MARK BY 6,45}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 93,28 - 6,45 \cdot 0,455 \approx 90,35 \rightarrow \text{AVERAGE MALE MARK}$$

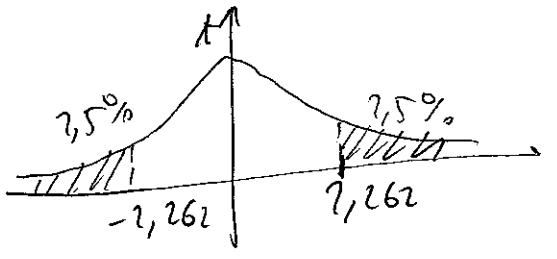
$$d = \frac{SS_{xy}^2}{SS_{xx} \cdot SS_{yy}} = 0,142$$

USEFULNESS TEST

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \quad Y_j = \beta_0 + \beta_1 X_j + E_j \quad E_j \text{ INDEPENDENT AND } N(0, \sigma^2)$$

$$t = \frac{\sqrt{9} \cdot 17,6}{\sqrt{2,73 \cdot 814,19 - 17,6^2}} = \frac{52,8}{43,74} = 1,207$$

D.F. = 9



H_0 IS NOT REJECTED
 MODEL IS NOT USEFUL

8.8

HOW MANY DUMMY VARIABLES ARE NEEDED TO PREDICT A PRICE BASED ON THE MONTH?

HW

$$Y = f(\text{MONTH})$$

$$X_1 = \begin{cases} 1 & \text{JANUARY} \\ 0 & \text{OTHER} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{FEBRUARY} \\ 0 & \text{OTHER} \end{cases}$$

$$\dots X_{11} = \begin{cases} 1 & \text{NOVEMBER} \\ 0 & \text{OTHER} \end{cases}$$

X_{12} IS NOT NEEDED, SINCE $X_1=0 \dots X_{11}=0 \Rightarrow \text{MONTH} = \text{DECEMBER}$

WE DO NOT EITHER NEED $X_i \cdot X_j$ IN THE FUNCTION, SINCE $X_i \cdot X_j = 0$ FOR EVERY $i \neq j$

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{11} X_{11}$$

IN GENERAL, IF A VARIABLE HAS k CATEGORIES WE NEED $k-1$ DUMMY VARIABLES AND WE DO NOT NEED NONLINEAR ELEMENTS USING THESE VARIABLES.

WHILE IF WE HAVE 2 VARIABLES WITH k_1 AND k_2 CATEGORIES, WE NEED k_1-1 AND k_2-1 DUMMY VARIABLES AND WE SHOULD INSERT IN THE MODEL ALL NONLINEAR TERMS $X_i \cdot X_j$ $i: 1 \rightarrow k_1-1$ $j: 1 \rightarrow k_2-1$

8.9 A FAST FOOD CHAIN WANTS TO PREDICT THE SALES BASED ON TRAFFIC AND CITY (1;2;3;4) USING THE MODEL:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

WITH A SAMPLE OF 24 RESTAURANTS

X_1 = TRAFFIC FLOW

X_2 = INDICATOR FOR CITY 2

X_3 = INDICATOR FOR CITY 3

X_4 = INDICATOR FOR CITY 4

$$\hat{\beta}_0 = 1,083388 \quad \hat{\beta}_1 = 0,103673 \quad \hat{\beta}_2 = -1,215762 \quad \hat{\beta}_3 = -0,530757$$

$$\hat{\beta}_4 = -1,076575 \quad d = 0,9791$$

TEST HYPOTHESIS $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ $\alpha = 0,05$

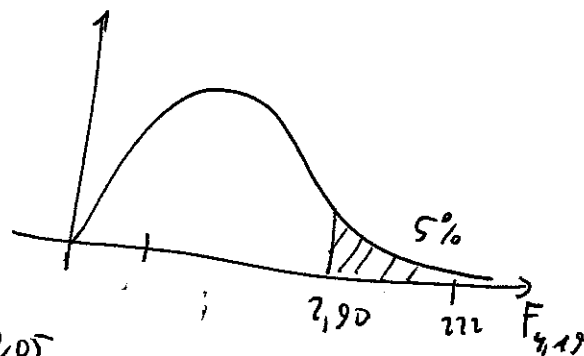
DRAW A GRAPH SALES BY TRAFFIC FOR EACH CITY

$$H_0: \beta_i = 0 \quad \forall i$$

H_1 : AT LEAST ONE $\beta_i \neq 0$

$$f = \frac{\frac{d}{4}}{(1-d)/(24-5)} = \frac{0,9791}{4} \cdot \frac{19}{0,0209} = 222,52$$

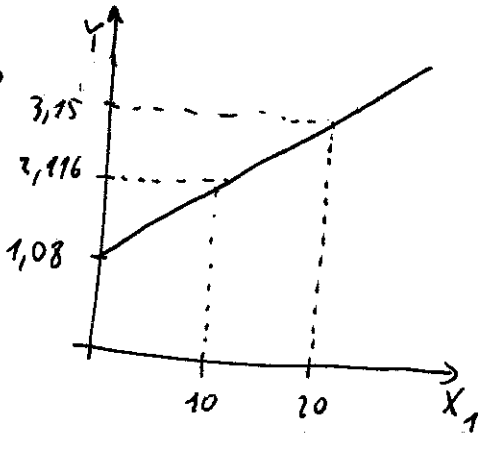
$$F_{4,19} = 2,90$$



H_0 IS REJECTED WITH $\alpha = 0,05$

□

8.9 B

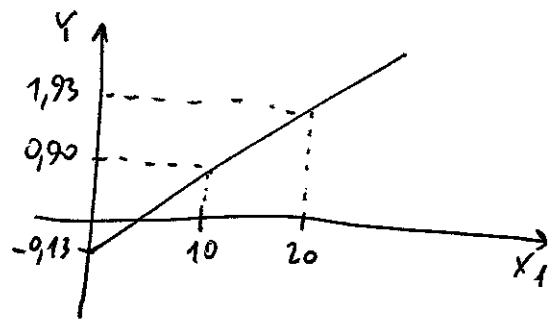


CITY 1 $X_2 = X_3 = X_4 = 0$

$$Y = 1,093388 + 0,103673 X_1$$

CITY 2 $Y = 1,093388 - 1,215762 \cdot 1 + 0,103673 X_1 = -0,132374 + 0,103673 X_1$

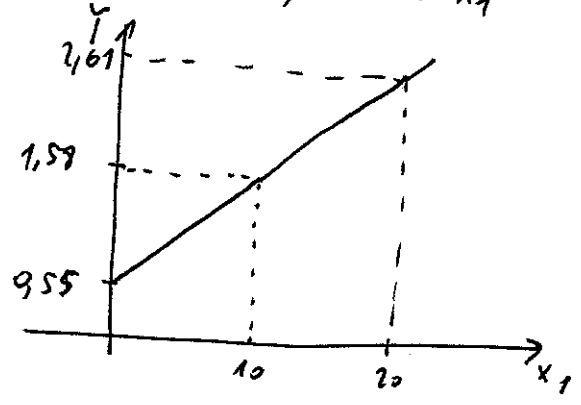
THE LINE HAS THE SAME INCLINATION BUT A DIFFERENT CROSSING WITH Y AXIS



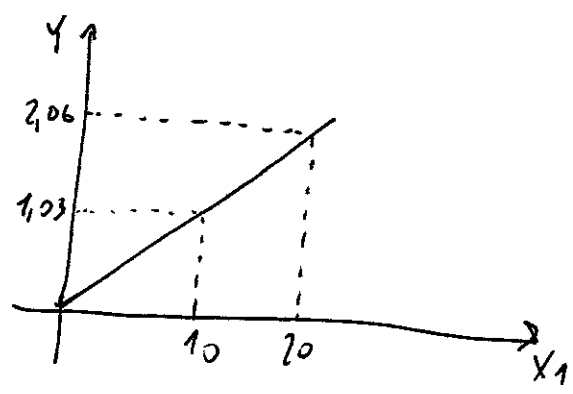
CITY 3: $Y = 1,093388 - 0,530257 \cdot 1 +$

$$+ 0,103673 X_1 =$$

$$= 0,563131 + 0,103673 X_1$$



CITY 4: $Y = 1,093388 - 1,026525 +$
 $+ 0,103673 X_1 = 0,066863 + 0,103673 X_1$



THE LINE IS ALWAYS THE SAME \Rightarrow THE EFFECT OF TRAFFIC IS THE SAME FOR EVERY CITY!

TO HAVE DIFFERENT EFFECTS ACCORDING TO EACH CITY, OTHER COEFFICIENTS NEED TO BE ADDED

$$+ \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \beta_7 X_1 X_4$$

D

8.10

SIMPLE LINEAR REGRESSION MODEL $d = 0,96$

IS IT POSSIBLE THAT PEARSON CORRELATION BE

$r < 0$, $r = 0$, $r = 0,96$?

AND $\hat{\beta}_1 < 0$, $\hat{\beta}_1 = 0$, $\hat{\beta}_1 = 0,96$?

$$d = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}} = (\text{PEARSON})^2$$

⇓

$$\text{PEARSON} = \pm \sqrt{0,96} = \pm 0,9797 \quad \text{BOTH SIGNS ARE POSSIBLE!}$$

THEREFORE PEARSON MAY BE < 0 IF $SS_{xy} < 0$, MAY NOT BE 0, NOT $-0,96$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = 0,96 \cdot \frac{SS_{yy}}{SS_{xx}}$$

WITH $SS_{yy} > 0$ $SS_{xy} > 0$
SINCE $R^2 > 0$

THEREFORE $\hat{\beta}_1$ MAY NOT BE ZERO AND MAY BE < 0 IF $SS_{xy} < 0$
FOR $\hat{\beta}_1 = 0,96$, SS_{yy} MUST BE $= SS_{xy}$

D

8.11

MODEL THE TOTAL WEEKLY SALES OF BEER AS FUNCTION OF BRAND (A, B, C).

- WRITE A MODEL

- USING 30 WEEKS, THE RESEARCHER FOUND THE VALUE OF d . WRITE HYPOTHESIS TEST AND ASSUMPTIONS

HW

- GIVEN $d = 0.92$ WHAT CAN YOU SAY ABOUT USEFULNESS OF THE MODEL? USE $\alpha = 0.01$

$$X_A = \begin{cases} 1 & \text{BEER IS A} \\ 0 & \text{BEER IS NOT A} \end{cases}$$

$$X_B = \begin{cases} 1 & \text{BEER IS B} \\ 0 & \text{BEER IS NOT B} \end{cases}$$

$$S_i = \beta_A X_{A_i} + \beta_B X_{B_i} + \beta_0 + E_i$$

$$n = 30 \quad \begin{cases} H_0: \beta_A = 0 \quad \beta_B = 0 \\ H_1: \beta_A \neq 0 \quad \text{OR} \quad \beta_B \neq 0 \end{cases}$$

$$\text{STATISTICS IS } F = \frac{\frac{d/2}{1-d}}{\frac{27}{30-(3)}} = \frac{d}{2} \cdot \frac{27}{1-d} = \frac{27d}{2(1-d)}$$

ASSUMPTIONS ARE: E_i INDEPENDENT, DISTRIBUTION $N(0, \sigma^2)$

$$d = 0.92 \quad F = \frac{27 \cdot 0.92}{2 \cdot 0.08} = 155.25 \quad F_{2,27} = 5.49$$

H_0 IS REJECTED

↓

THE MODEL IS USEFUL \square