CHAPTER 10:
REGRESSION AND CORRELATION

NOTE THAT UP TO NOW, WE WERE INTERESTED IN A <u>SINGLE</u> ATTRIBUTE (VARIABLE) OF A POPULATION. IN THIS CHAPTER, WE EXTEND THIS TO <u>TWO</u> VARIABLES (USUALLY DENOTED X AND Y), SUCH AS: WEIGHT AND HEIGHT, SALARY AND YEARS OF SERVICE, AGE AND PRICE OF A CAR MODEL, ETC. BOTH OF THEM MUST BE OF <u>NUMERICAL</u> TYPE, PREFERABLY HAVING AN INTERVAL SCALE.

WE ARE THEN INTERESTED IN THEIR (THE TWO VARIABLES') RELATIONSHIP, BEST SEEN BY PLOTTING THEIR **SCATTER DIAGRAM** (SCATTERGRAM).

TO SIMPLIFY MATTERS (YET BE ABLE TO COVER THE VAST MAJORITY OF PRACTICAL SITUATIONS), WE WILL ASSUME THAT THE RELATIONSHIP (BETWEEN X AND Y) IS **LINEAR** (STRAIGHT-LINE).

WE WILL STUDY THIS RELATIONSHIP IN THE CONTEXT OF RANDOM <u>SAMPLES</u> (TRYING TO EXTEND THE SAMPLE RESULTS TO THE WHOLE POPULATION).

FIRST WE LEARN HOW TO **FIT**, THROUGH A GIVEN SAMPLE OF n PAIRS OF (x, y) VALUES, THE <u>BEST</u> **LEAST-SQUARES** STRAIGHT LINE.

WE CALL X THE **EXPLANATORY** (INDEPENDENT) AND Y THE **RESPONSE** (DEPENDENT) VARIABLE, AND WE WANT TO KNOW HOW Y DEPENDS ON (AND CAN BE PREDICTED FROM) THE VALUE OF X.

ASSUMING THAT THE RELATIONSHIP IS LINEAR, I.E. OF THE $Y = \alpha + \beta \cdot X$ TYPE (WHERE ″ IS THE **INTERCEPT** AND \$ THE **SLOPE** - COLLECTIVELY, THESE ARE KNOWN AS THE **REGRESSION COEFFICIENTS**), WE NEED TO FIND GOOD SAMPLE <u>ESTIMATES</u> OF THESE (DENOTED a AND b RESPECTIVELY).

TO DO THIS, WE MUST FIRST COMPUTE THE TWO SAMPLE MEANS $\overline{x}$ AND $\overline{y}$, THE QUANTITY WE USED TO CALL

$$SS_x = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

AND YET ANOTHER, SIMILAR QUANTITY

$$SP_{xy} \equiv \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

(TEXTBOOK'S SS_{xy}).

THEN, $b = SP_{xy} / SS_x$ AND $a = \overline{y} - b \cdot \overline{x}$

FOR ANY SPECIFIC *x,* WE CAN NOW <u>PREDICT</u> THE CORRESPONDING y BY: $y_p \equiv a + b \cdot x$

THE DIFFERENCE BETWEEN EACH OF THE ACTUALLY OBSERVED VALUES OF y AND THE CORRESPONDING y_{p} (WHICH, UNLIKE y ITSELF, LIES <u>ON</u> THE FITTED STRAIGHT LINE) IS CALLED THE **RESIDUAL**.

THE RESIDUALS ARE ASSUMED INDEPENDENT OF EACH OTHER, AND NORMALLY DISTRIBUTED.

THE **RESIDUAL STANDARD DEVIATION** (TEXTBOOK'S S_e) IS COMPUTED BY

$$s_r \equiv \sqrt{\frac{\Sigma(y-y_p)^2}{n-2}}$$

THE NUMERATOR (SUM OF SQUARES OF ALL *n* RESIDUALS) CAN BE MORE EASILY COMPUTED FROM: $SS_y - b \cdot SP_{xy}$ WHERE

$$SS_y = \Sigma y^2 - \frac{\left(\Sigma y\right)^2}{n}$$

WE CAN NOW USE THESE RESULTS TO PREDICT A VALUE OF THE RESPONSE VARIABLE *Y* FOR A NEW OBSERVATION, TAKEN AT A SPECIFIC VALUE OF *X* (THIS, AS WE ALREADY KNOW, IS DONE BY COMPUTING THE CORRESPONDING y_p).

BETTER YET, WE CAN CONSTRUCT A 'CONFIDENCE' (NOW CALLED **PREDICTION**) INTERVAL FOR THE NEW VALUE OF *Y*, THUS:

$$y_p \pm t_c \cdot s_r \cdot \sqrt{1+\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_x}}$$

(USE *n* - 2 DEGREES OF FREEDOM).

EXAMPLE: THE FOLLOWING TABLE REPRESENTS A RANDOM SAMPLE OF EMPLOYEES IN A CERTAIN INDUSTRY, PROVIDING US WITH THEIR:

< YEARS OF EDUCATION BEYOND GRADE 12 (*X*)

< ANNUAL SALARY (IN THOUSANDS) - *Y*:

X	4	2	7	0	4	3	1	2	4	6
Y	45	32	60	27	39	33	30	28	43	52

WE WANT TO FIT THE REGRESSION LINE THROUGH THIS DATA, AND COMPUTE A 95% PREDICTION INTERVAL FOR A SALARY OF AN EMPLOYEE WITH 2 YEARS OF EDUCATION BEYOND HIGH SCHOOL.

FIRST WE COMPUTE:

$$\Sigma x = 33,\ \Sigma y = 389,\ \Sigma x^2 = 151,\ \Sigma y^2 = 16225,\ \Sigma xy = 1489$$

THIS IS NOW CONVERTED TO:

$$\overline{x} = 3.3,\ \overline{y} = 38.9,\ SS_x = 151 - \frac{33^2}{10} = 42.1,$$

$$SS_y = 16225 - \frac{389^2}{10} = 1092.9,\ SP_{xy} = 1489 - \frac{33 \times 389}{10} = 205.3$$

HAVING THESE, WE CAN NOW EASILY COMPLETE THE EXERCISE:

$$b = \frac{205.3}{42.1} = 4.876, \;\; a = 38.9 - 4.876 \times 3.3 = 22.808$$

WHICH MEANS THAT THE BEST (LEAST-SQUARES) STRAIGHT LINE IS:

$$y = 4.876\, x + 22.808$$

NOW, WE WILL ALSO NEED

$$s_r = \sqrt{\frac{1092.9 - 4.876 \times 205.3}{8}} = 3.387$$

IT WOULD BE POSSIBLE - BUT QUITE TEDIOUS - TO VERIFY THAT THE NUMERATOR DOES AGREE WITH THE SUM OF SQUARES OF THE RESIDUALS, WHICH ARE COMPUTED BY: 45 - (4.876×4 +22.808) = 2.688, 32 - (4.876×2 +22.808) = -0.56, ...

THE PREDICTION INTERVAL FOR A SALARY OF AN EMPLOYEES WITH 2 EXTRA YEARS OF EDUCATION IS THUS:

$$4.876 \times 2 + 22.808 \pm 2.306 \times 3.387 \times \sqrt{1 + \frac{1}{10} + \frac{(2-3.3)^2}{42.1}} =$$

32.561 ± 8.339 OR (24.22, 40.90) IN THOUSAND OF DOLLARS.

NOTE THAT THIS TIME, WE WOULD <u>NOT</u> BE ABLE TO REDUCE THE INTERVAL'S WIDTH ARBITRARILY BY EXTRA SAMPLING.

ONE CAN ALSO SHOW THAT

$$\frac{b-\beta}{s_r \big/ \sqrt{SS_x}}$$

HAS THE t DISTRIBUTION WITH n - 2 DEGREES OF FREEDOM.

THIS CAN BE UTILIZED FOR EITHER CONSTRUCTING A CONFIDENCE INTERVAL FOR THE ACTUAL VALUE OF β (WE WILL NOT GO INTO THAT), OR TESTING THE NULL HYPOTHESIS THAT $\beta = 0$ (THIS IS USUALLY DONE AS A ONE-TAIL TEST).

(EXTENSION OF THE PREVIOUS) EXAMPLE: TO TEST H_0: $\beta = 0$ AGAINST H_1: $\beta > 0$ AT 1% LEVEL OF SIGNIFICANCE, WE FIRST COMPUTE THE VALUE OF THE TEST STATISTIC

$$\frac{b}{s_r \big/ \sqrt{SS_x}} = \frac{4.876}{3.387 \big/ \sqrt{42.1}} = 9.341$$

AND COMPARE IT WITH THE CORRESPONDING $t = 2.896$. CLEARLY, WE HAVE A HIGHLY SIGNIFICANT PROOF THAT $\beta > 0$.

CORRELATION

WE WOULD LIKE A GOOD MEASURE OF HOW GOOD (CLOSE) IS THE (LINEAR) RELATIONSHIP BETWEEN X AND Y. ONE MAY FEEL THAT s_r ITSELF PROVIDES THIS INFORMATION, BUT IS THE VALUE $9234 LARGE OR SMALL?

THE QUANTITY ONE USES FOR THIS PURPOSE IS CALLED (SAMPLE) **CORRELATION COEFFICIENT**, AND IS DEFINED AS FOLLOWS:

$$r \equiv \frac{SP_{xy}}{\sqrt{SS_x \cdot SS_y}}$$

ONE CAN SHOW THAT ITS VALUE IS ALWAYS BETWEEN -1 AND 1 (THE SIGN DEPENDING ON THE SLOPE OF THE REGRESSION LINE). FURTHERMORE, r IS ALWAYS **DIMENSIONLESS** (NO UNITS).

THE RELATIONSHIP BETWEEN *X* AND *Y* IS WEAK (OR NONEXISTENT) WHEN r IS CLOSE TO ZERO (HOW CLOSE IS 'CLOSE'? - WE WILL LOOK AT THAT SHORTLY), AND STRONG WHEN $|r|$ APPROACHES 1 (r =1 OR -1 WOULD IMPLY THAT THE RELATIONSHIP IS PERFECT - ALL OUR OBSERVATIONS LIE EXACTLY ON THE REGRESSION LINE).

TO ELIMINATE THE SIGN, WE CAN SIMPLY SQUARE r TO GET THE SO CALLED **COEFFICIENT** OF **DETERMINATION**

$$r^2 = \frac{SP_{xy}^2}{SS_x \cdot SS_y}$$

IT REPRESENTS THE <u>RELATIVE</u> REDUCTION IN SS_y ACHIEVED BY FITTING THE REGRESSION LINE (I.E. AS COMPARED TO THE SUM OF SQUARES OF THE RESULTING RESIDUALS).

CONTINUING THE PREVIOUS EXAMPLE: WE CAN EASILY COMPUTE $r = {}^{205.3}\!\big/\!\sqrt{42.1 \times 1092.9}$ = 0.9571 (GETTING A FAIRLY HIGH, POSITIVE CORRELATION). THE COEFFICIENT OF DETERMINATION IS $r^2 = 91.6\%$.

NOTE THAT r IS COMPUTED BASED ON OUR <u>SAMPLE</u> OF n PAIRS OF (x, y) VALUES, AND IS THUS ONLY AN <u>ESTIMATE</u> OF THE EXACT (ALBEIT UNKNOWN) <u>POPULATION</u> CORRELATION COEFFICIENT D .

WHEN X AND Y ARE UNCORRELATED (I.E. X DOES <u>NOT</u> EFFECT THE VALUE OF Y, AND THE KNOWLEDGE OF X THUS <u>CANNOT</u> HELP US PREDICTING THE VALUE OF Y), THE POPULATION CORRELATION COEFFICIENT D HAS THE VALUE OF ZERO.

YET, THE CORRESPONDING SAMPLE CORRELATION COEFFICIENT r WILL PRACTICALLY ALWAYS BE NON-ZERO (BUT, IN THE D = 0 CASE, IT SHOULD BE ‘SMALL’).

FOR TESTING THE NULL HYPOTHESIS OF D = 0 AGAINST EITHER A ONE- OR A TWO-TAIL ALTERNATIVE, WE USE THE FOLLOWING TEST STATISTIC

$$\frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

WHICH (UNDER H_0) HAS THE t DISTRIBUTION WITH n - 2 DEGREES OF FREEDOM.

THIS TEST (EVEN THOUGH SEEMINGLY DIFFERENT) PROVES TO BE <u>EQUIVALENT</u> TO OUR OLD TEST FOR POPULATION SLOPE (BEING ZERO, OR NOT).

ONE MORE EXAMPLE: (IT IS POSSIBLE THAT EITHER VARIABLE MAY HAVE NEGATIVE VALUES - LET US PRACTICE WITH THESE).

X	-10	-5	0	5	10	15
Y	56	49	36	18	6	-11

E x = 15, E y = 154, E x^2 = 475, E y^2 = 7314, E $x \times y$ = -820

$\bar{x}$ = 15/6 = 2.5 $\bar{y}$ = 154/6 = $25.\overline{6}$ SSx = 475 - 15^2/6 = 437.5

$SS_y = 7314 - 154^2/6 = 3361.\overline{3}$ $SP_{xy} = -820 - 15\times154/6 = -1205$

$b = -1205/437.5 = -2.7543$ $a = 25.667 + 2.7543\times2.5 = 32.55$

BEST (LEAST-SQUARES) REGRESSION LINE IS THUS:

$$y = -2.7543\,x + 32.55$$

THE RESIDUAL STANDARD DEVIATION:

$$s_r = \sqrt{\frac{3361.33 - 1205\times2.7543}{4}} = \sqrt{\frac{42.40}{4}} = 3.256$$

THE 95% PREDICTION INTERVAL FOR A NEW Y OBSERVATION, TAKEN AT $X = 12$ (DON'T EXTRAPOLATE, I.E. USE AN X VALUE <u>OUTSIDE</u> THE ORIGINAL INTERVAL):

$$-2.7543\times12 + 32.55 \pm 2.776\times3.256\times\sqrt{1+\frac{1}{6}+\frac{(12-2.5)^2}{437.5}} =$$

$$-0.50 \pm 10.59 \quad = \quad (-11.09,\ 10.09)$$

TEST $H_0: \beta = 0$ AGAINST $H_1: \beta < 0$

TEST STATISTIC: $\dfrac{b}{s_r}\sqrt{SS_x} = \dfrac{-2.7543}{3.256}\sqrt{437.5} = -17.69$

IS A LOT SMALLER THAN THE CRITICAL VALUE OF -3.747 (USING THE t_4 DISTRIBUTION AND $\alpha = 0.01$) $\Rightarrow$ HIGHLY SIGNIFICANT EVIDENCE TO REJECT H_0 IN FAVOUR OF H_1.

CORRELATION COEFFICIENT: $\dfrac{-1205}{\sqrt{437.5 \times 3361.33}}$ = - 0.99367

COEFFICIENT OF DETERMINATION: $(-0.99367)^2 = 98.73\%$

TESTING H_0: D=0 ....

TEST STATISTIC: $\dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{-0.99367 \times \sqrt{4}}{\sqrt{0.01262}}$ = -17.69 (CHECK)