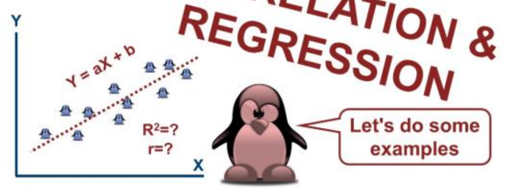
SE

CORRELATION & REGRESSION

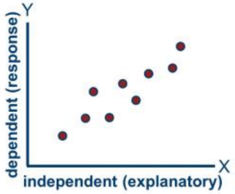
STEP-BY-STEP EXAMPLES



REGRESSION/CORRELATION

Goal: identify any non-random relationship between X and Y.

- Check assumptions.
- Estimate parameters. $Y = \alpha + \beta X,\ r,\ R^2$
- Test for significance. ANOVA, t-test
- Calculate CI.
- Interpret results.



ASSUMPTIONS

- Independence of data points.
- Treatment (i.e., independent) variables being fixed/variable.
- Linearity of the pattern.
- Normal distribution of residuals around best fit line.
- Equal variance of Y values along the X range.

If assumptions are violated:

- Transform data values into ones that meet assumptions.
- Use a different method (e.g., nonlinear regression).

INTERPRETING RESULTS

- Significant slopes in a *regression* analysis *can* imply causation, but the causation may be indirect.
- Significant slopes in a *correlation* analysis *may* imply causation, but other factors may be driving the pattern.
- Correlation does not prove causation, but it's still useful.
  - Best-fit line can be used to make predictions.
  - Significant correlations are a first step to finding causation.
  - Lack of a correlation is strong evidence against causation.
- A significant slope implies a non-random relationship, but is it relevant or trivial?
- The results only hold for the range of X values studied.

PROCEDURE FOR CALCULATIONS

- Calculate SS_X, SS_Y, and SP_{XY}.
- Use these to estimate slope and Y-intercept for best-fit line.
- Use best-fit line to calculate SS_{reg} and SS_{error}
- Use these to perform ANOVA and t-tests of slope significance.
- Calculate r and R^2.
- (optional) Perform significance test for r.
- (optional) Calculate confidence/inclusion intervals for line/slope.

EXAMPLE 1: 1st and 2nd exam scores — t-test analysis

H_0: $\beta = \beta_0 = 0$ $df = n - 2 = 8 - 2 = 6$

H_A: $\beta \neq 0$

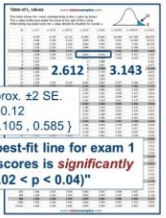
$$t_{calc} = \frac{b - \beta_0}{SE_b} = \frac{0.345 - 0}{0.1203} = 2.867$$

$$SE_b = \sqrt{\frac{\left(\frac{SS_Y - b^2 SS_X}{n-2}\right)}{SS_X}}$$

$$SE_b = \sqrt{\frac{\left(\frac{102 - (0.345)^2 496}{8-2}\right)}{496}} = 0.1203$$

Recall: 95% CI is approx. ±2 SE.
Slope is 0.345, SE is 0.12
95% CI is approx. { 0.105 , 0.585 }

"The slope of the best-fit line for exam 1 scores vs exam 2 scores is *significantly larger* than zero (0.02 < p < 0.04)"



EXAMPLE 1: 1st and 2nd exam scores — ANOVA analysis

H_0: $\beta = \beta_0$

H_A: $\beta \neq \beta_0$

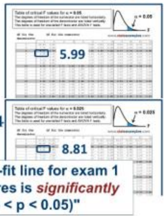
$$F_{calc} = \frac{MS_{reg}}{MS_{error}}$$

$$MS_{reg} = \frac{SS_{reg}}{df_{reg}} = \frac{SS_{reg}}{1} = \frac{58.954}{1} = 58.954$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{SS_{error}}{n-2} = \frac{43.046}{8-2} = 7.174$$

$$F_{calc} = \frac{58.954}{7.174} = 8.217$$

"The slope of the best-fit line for exam 1 scores vs exam 2 scores is *significantly larger* than zero (0.025 < p < 0.05)"

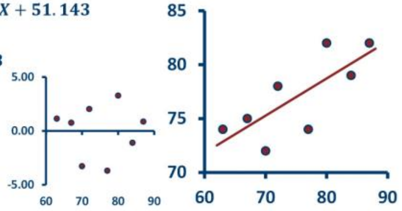


EXAMPLE 1: 1st and 2nd exam scores — Summary & graphs

$Y = 0.345\,X + 51.143$

$r = 0.760$

$R^2 = 0.578$



EXAMPLE 1: 1st and 2nd exam scores — Calculations

$$SS_X = \sum (X_i - \bar{X})^2 = 496 \qquad SS_Y = \sum (Y_i - \bar{Y})^2 = 102 = SS_{total}$$

$$SP_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 171 \qquad b = \frac{SP_{XY}}{SS_X} = \frac{171}{496} = 0.345$$

$$\bar{Y} = 51.143 + (0.345)\bar{X}$$

$$SS_{reg} = \sum (\hat{Y}_i - \bar{Y})^2 = 58.954 \qquad SS_{error} = \sum (Y_i - \hat{Y}_i)^2 = 43.046$$

$$r = \frac{SP_{XY}}{\sqrt{SS_X SS_Y}} = \frac{171}{\sqrt{(496)(102)}} = 0.760$$

$$R^2 = \frac{SS_{reg}}{SS_Y} = \frac{58.954}{102} = 0.578 = 0.760^2 = r^2$$

EXAMPLE 1: 1st and 2nd exam scores

1^{st}	2^{nd}	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	SP_{xy}	$\hat{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$	$y_i-\hat{y}$	$(y_i-\hat{y})^2$
63	74	-12	-3	144	9	36	72.86	-4.14	17.12	1.14	1.29
67	75	-8	-2	64	4	16	74.24	-2.76	7.61	0.76	0.57
70	72	-5	-5	25	25	25	75.28	-1.72	2.97	-3.28	10.73
72	78	-3	1	9	1	-3	75.97	-1.03	1.07	2.03	4.14
77	74	2	-3	4	9	-6	77.69	0.69	0.48	-3.69	13.61
80	82	5	5	25	25	25	78.72	1.72	2.97	3.28	10.73
84	79	9	2	81	4	18	80.10	3.10	9.63	-1.10	1.22
87	82	12	5	144	25	60	81.14	4.14	17.12	0.86	0.74
600	616	0	0	496	102	171		0	58.954	0	43.046
75	77										
				SS_X	SS_Y SS_{total}	SP_{XY}			SS_{reg}		SS_{error}

EXAMPLE 1: 1st and 2nd exam scores — t-test of r value

H_0: $\rho = \rho_0$

H_A: $\rho \neq \rho_0$

Same t_{calc} as we got for the slope

$$t_{calc} = r\sqrt{\frac{n-2}{1-r^2}}$$

$$= (0.760)\sqrt{\frac{8-2}{1-(0.760)^2}} = 2.867$$

"The correlation coefficient for the relationship between exam 1 scores and exam 2 scores is *significantly larger* than zero (0.02 < p < 0.04)"

EXAMPLE 2: ID number and exam scores

1^{st}	2^{nd}	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	SP_{xy}	$\hat{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$	$y_i-\hat{y}$	$(y_i-\hat{y})^2$
51	74	4	-2	16	4	-8	76.07	0.07	0.01	-2.07	4.30
67	72	20	-4	400	16	-80	76.37	0.37	0.14	-4.37	19.12
27	75	-20	-1	400	1	20	75.63	-0.37	0.14	-0.63	0.39
27	80	-20	4	400	16	-80	75.63	-0.37	0.14	4.37	19.12
5	74	-42	-2	1764	4	84	75.22	-0.78	0.61	-1.22	1.48
67	78	20	2	400	4	40	76.37	0.37	0.14	1.63	2.65
85	79	38	3	1444	9	114	76.71	0.71	0.50	2.29	5.25
329	532	0	0	4824	54	90		0	1.679	0	52.321
47	76										
				SS_X	SS_Y SS_{total}	SP_{XY}			SS_{reg}		SS_{error}

EXAMPLE 2: ID number and exam scores — Calculations

$$SS_X = \sum (X_i - \bar{X})^2 = 4824 \qquad SS_Y = \sum (Y_i - \bar{Y})^2 = 54 = SS_{total}$$

$$SP_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = 90 \qquad b = \frac{SP_{XY}}{SS_X} = \frac{90}{4824} = 0.019$$

$$\bar{Y} = 75.123 + (0.019)\bar{X}$$

$$SS_{reg} = \sum (\hat{Y}_i - \bar{Y})^2 = 1.679 \qquad SS_{error} = \sum (Y_i - \hat{Y}_i)^2 = 52.321$$

$$r = \frac{SP_{XY}}{\sqrt{SS_X SS_Y}} = \frac{90}{\sqrt{(4824)(54)}} = 0.176$$

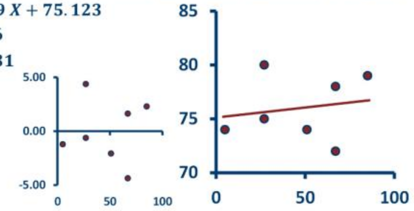
$$R^2 = \frac{SS_{reg}}{SS_Y} = \frac{1.679}{54} = 0.031 = 0.176^2 = r^2$$

EXAMPLE 2: ID number and exam scores — Summary & graphs

$Y = 0.019\,X + 75.123$

$r = 0.176$

$R^2 = 0.031$



EXAMPLE 2: ID number and exam scores — ANOVA analysis

H_0: $\beta = 0$

H_A: $\beta \neq 0$

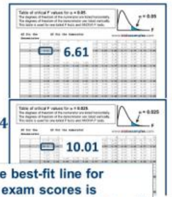
$$F_{calc} = \frac{MS_{reg}}{MS_{error}}$$

$$MS_{reg} = \frac{SS_{reg}}{df_{reg}} = \frac{SS_{reg}}{1} = \frac{1.679}{1} = 1.679$$

$$MS_{error} = \frac{SS_{error}}{df_{error}} = \frac{SS_{error}}{n-2} = \frac{52.231}{7-2} = 10.464$$

$$F_{calc} = \frac{1.679}{10.464} = 0.160$$

"The slope of the best-fit line for ID numbers vs exam scores is *not significantly different* from zero (0.05 < p)"



INTERPRETING RESULTS

	Correlation	Regression
A significant (i.e., non-random) relationship exists	Causation hinted	Causation confirmed
	p < 0.05 for tests Slope 95% CI excludes zero Can predict each variable from the other Real, but might be unimportant	
A nonsignificant (i.e., random) relationship exists	Causation claims weakened	Causation claims severely weakened
	p > 0.05 for tests Slope 95% CI includes zero Can't predict either variable from the other Not real or important	

EXAMPLE 2: ID number and exam scores — Real data

$Y = -0.041\,X + 66.628$

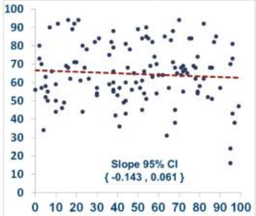
$r = -0.072$

$R^2 = 0.005$

$F_{calc} = 0.660\ (p \approx 0.42)$

$t_{calc} = -0.809\ (p \approx 0.42)$

"The slope of the best-fit line for ID numbers vs exam 1 scores is *not significantly different* from zero (0.4 < p)"



EXAMPLE 1: 1st and 2nd exam scores — Real data

$Y = 0.641\,X + 23.397$

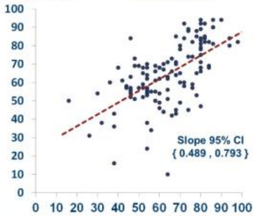
$r = 0.604$

$R^2 = 0.365$

$F_{calc} = 71.7\ (p < 10^{-13})$

$t_{calc} = 8.43\ (p < 10^{-13})$

"The slope of the best-fit line for exam 1 scores vs exam 2 scores is *significantly larger* than zero ($p < 10^{-13}$)"



EXAMPLE 2: ID number and exam scores — t-test of r value

H_0: $\rho = \rho_0$

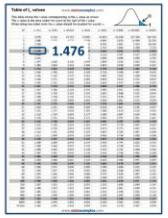
H_A: $\rho \neq \rho_0$

Same t_{calc} as we got for the slope

$$t_{calc} = r\sqrt{\frac{n-2}{1-r^2}}$$

$$= (0.176)\sqrt{\frac{7-2}{1-(0.176)^2}} = 0.401$$

"The correlation coefficient for the relationship between ID numbers and exam 1 scores is *not significantly different* from zero (0.2 < p)"



EXAMPLE 2: ID number and exam scores — t-test analysis

H_0: $\beta = \beta_0 = 0$ $df = n - 2 = 7 - 2 = 5$

H_A: $\beta \neq 0$

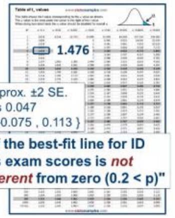
$$t_{calc} = \frac{b - \beta_0}{SE_b} = \frac{0.019 - 0}{0.0466} = 0.401$$

$$SE_b = \sqrt{\frac{\left(\frac{SS_Y - b^2 SS_X}{n-2}\right)}{SS_X}}$$

$$SE_b = \sqrt{\frac{\left(\frac{54 - (0.019)^2 4824}{7-2}\right)}{4824}} = 0.0466$$

Recall: 95% CI is approx. ±2 SE.
Slope is 0.019, SE is 0.047
95% CI is approx. { -0.075 , 0.113 }

"The slope of the best-fit line for ID numbers vs exam scores is *not significantly different* from zero (0.2 < p)"



StatsExamples.com