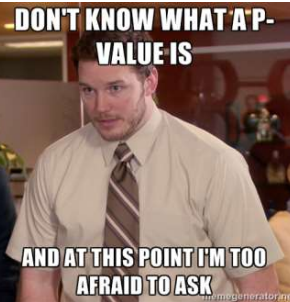


Hopefully everyone feels like this: ... not this:



LETS LEARN

REGRESSION



DON'T KNOW WHAT A P-VALUE IS

AND AT THIS POINT I'M TOO AFRAID TO ASK

(the Risk that you would be wrong if you reject the null)
But PLEASE stop me if you have ANY question(s)!

Interval Covariation

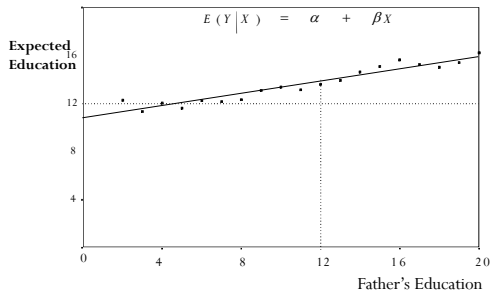
- Remember the definition of a “relationship”:
 - Values of the DV change across values of the IV
- If both variables are interval, 3 options:
 - ANOVA**
 - compares IV categories with 2 kinds of DV variance
 - Uses the F test
 - Correlation** (“r” value, and associated sig. level)
 - summary measure closely related to regression analysis
 - usually relating only two variables
 - Regression** (model, equation, and “r-squared”)
 - omnibus test of entire model, including multiple variables
 - summary equation: coefficients (beta weights), alpha, eta
 - Summary statistic: r^2 , the Holy Grail of statistics

6 SOC424 @ CSUN - Ellis Godard

Regression

SOC424 – Statistics in Soft Pants
w/ Dr. Ellis Godard

Let's talk about that 3rd line from last time

$$E(Y|X) = \alpha + \beta X$$


7 SOC424 @ CSUN - Ellis Godard

Outline for Today...

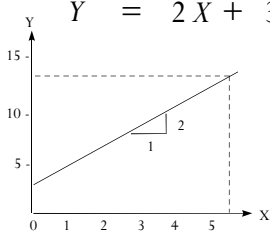
- Regression Line**
 - Errors (Residuals)
 - Line of Best Fit
 - Hypothesis Test
- Examples**
- Goals**
- Lab**

5 SOC424 @ CSUN - Ellis Godard

Remember Algebra?

- The formula for a *slope* was “ $Y = m \cdot x + b$ ”

X	Y
0	3
1	5
2	7
3	9
4	11
5	13

$$Y = 2X + 3$$


8 SOC424 @ CSUN - Ellis Godard

Line Errors Fit Test Example Goals Lab

Same basic idea, but...

- May be that *no* points are *on* the line
- The line is “predicted” values
 - For any given X, the value of Y that – when all the {x,y}’s are together, gives a straight line that’s closest to all the points
- The vertical distance between the line, at any given x, and each actual *observed* value of y for that given x, is a residual

9 SOC424 w/ Dr. Ellis Godard -- Slide

Line Errors Fit Test Example Goals Lab

Basis for Calculations

- Sum of Squared Errors (SSE) is the sum of squared differences between the actual Y value (of the dependent variable) and the expected value of Y (or the predicted value, given knowledge of X):

$$SSE = \sum (Y_i - \hat{Y})^2$$
- “Line of best fit” is the line that minimizes SSE, i.e. that involves the least total squared deviations from the actual data.
 - Called “ordinary least squares”, or OLS Regression.

12 SOC424 @ CSUN - Ellis Godard

Line Errors Fit Test Example Goals Lab

Residuals...

The vertical distance between the line (which represents predicted values) and each actual observed data point is the “residual” (a regression “error”, though different from the implied “error term”)

10 SOC424 @ CSUN - Ellis Godard

Line Errors Fit Test Example Goals Lab

Core Calculations (a & b)

- For the regression equation $Y = \alpha + Bx$,
 - B (“beta”) is the population parameter slope (expected increase in Y for an increase of 1 in X) which we estimate w/ “b”, where:

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$
 - α (“alpha”) is the population parameter Y-intercept (predicted Y when X=0) which we estimate w/ “a”:

$$a = \bar{Y} - b * \bar{X}$$

13 SOC424 @ CSUN - Ellis Godard

Line Errors Fit Test Example Goals Lab

Implied Error Term

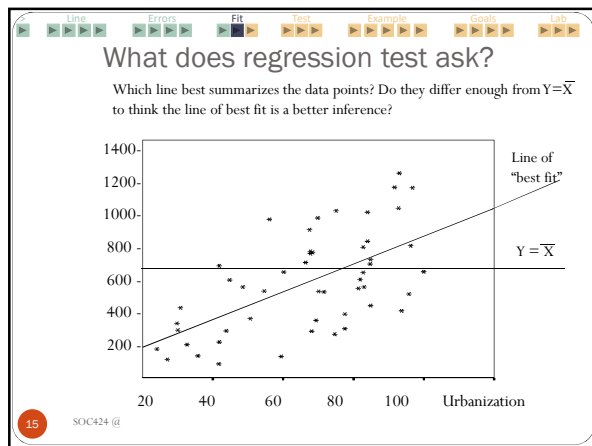
• Every regression equation has one...

11 SOC424 @ CSUN - Ellis Godard

Line Errors Fit Test Example Goals Lab

The “Line of Best Fit”

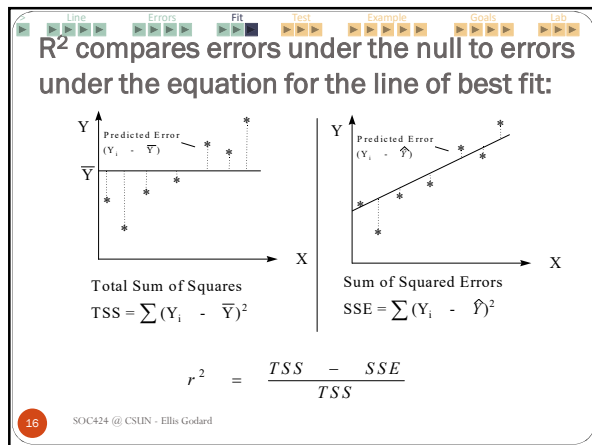
14 SOC424 @ CSUN - Ellis Godard



Assumptions for Regression

- Random Sample
- Interval Data
- Only Linear Association
 - Not curvilinear
- Conditional $Y =$ normal distribution
- Homoskedasticity
 - Y 's stdev is the same for each X , which is of course almost never true

18 SOC424 @ CSUN - Ellis Godard



P-Value Interpretation (one-tailed test)

If the null hypothesis is true, if X and Y are independent (the slope equals zero), then the probability of obtaining a sample slope equal to or greater than b (i.e. the probability of being wrong in rejecting the null) is equal to p .

If urbanization and violent crime rates are independent, then the probability of obtaining a sample slope of or greater is equal to ____.

19 SOC424 @ CSUN - Ellis Godard

Hypothesis Testing

- $H_0: b = 0$ (X and Y are independent)
- $H_A: b$ is not equal to 0 (X and Y are dependent)
 - or $b > 0$ (the relationship between X and Y is positive)
 - or $b < 0$ (the relationship between X and Y is negative)
- For the test statistic, $t = b / \sigma_b$
 - where we estimate σ_b (est. standard error of the slope) with $\sigma_b = \sigma / \sqrt{\sum (X_i - \bar{X})^2}$
 - and we estimate σ (standard deviation of the conditional distribution of y) as $\sigma_{y|x} = \sqrt{SSE / (n-2)}$

17 SOC424 @ CSUN - Ellis Godard

Example Hypotheses: Crime and Urbanization

For $Y = Bx + a \dots$

$H_0: \beta = 0$ There is no association between of violent crime rates and urbanization.

$H_a: \beta > 0$ Rates of violent crime are positively associated with levels of urbanization.

Looked at scatterplot for this in last lecture

20 SOC424 @ CSUN - Ellis Godard

Example SPSS Output:
Crime and Urbanization

	(coefficients)	(std error)	(r)	(t-stat)	(p-value)
Variable	B	SE B	Beta	T	Sig T
URBAN	8.162241	1.546742	.605929	5.277	.0000
(Constant)	17.606701	108.430912		.162	.8717

Those are the "parameters" of the equation...

Note: $T = B / (SE\ B) = 8.162 / 1.546 = 5.277$

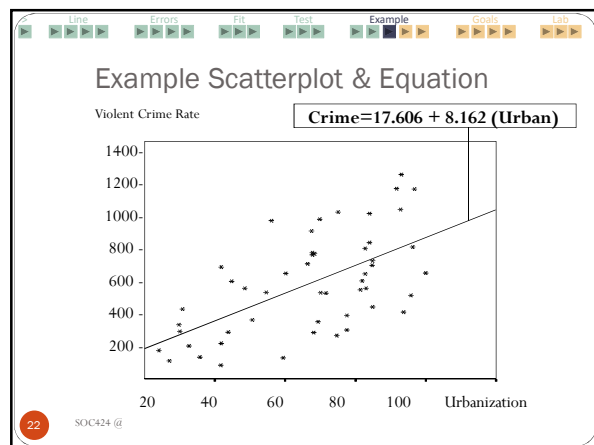
21 SOC424 @ CSUN - Ellis Godard

Example's Conclusion:
Crime and Urbanization

The given significance level for the test statistic for the slope is practically 0 (significant to four decimals, at 0.0000).

There is thus practically no risk of being wrong in rejecting the null hypothesized slope of 0 –that crime rates and levels of urbanization are independent – and we are able to find support for our claim that more urbanized states have higher violent crime rates than less urbanized states.

24 SOC424 @ CSUN - Ellis Godard



Summary statistics in SPSS

VARIABLE	B	SE B	Beta	T	Sig T
IV name (constant)	"b"	se of "b"		test-stat for b	p-value
	"a"	se of "a"		test-stat for a	p-value

These are the hypothesized values (t and p).

These coefficients are standardized in units of standard deviation
The standardized Beta = $(\alpha_x / \alpha_y) * b$
These are *not* the coefficients you want this week.

These are the coefficients you want for the equation $Y = a + bx$ (estimating $Y = \alpha + Bx$)
 b = change, on average, in Y for each increase of 1 in X.
 a = predicted value of Y when $X=0$ (but don't infer beyond observed range of X)

25 SOC424 @ CSUN - Ellis Godard

Stating Your Conclusion

- If the p-value is relatively small (less than 0.05, for example), then we can reject the null hypothesis that the population slope is zero and find support for the claim (H_a) that the slope is not zero, or (if we have a one-tailed test) greater than or less than zero.
- Conversely, if our p-value is relatively large, then there is too great a risk that our results are due to sampling error. We are thus unable to reject the null hypothesis, and cannot claim to have found support for our research (alternate) hypothesis.

23 SOC424 @ CSUN - Ellis Godard

Four+ Things R^2 (& its p) Tells us...

- Tells whether there is a relationship (hypothesis test)
- Assesses the strength
- Describes the strength of the relationship (% of variance in Y explained by the variance in X)
- Tells the direction of the relationship (positive if same direction, e.g. up and up, or down and down; negative if opposite directions)
- Describes the form of the relationship (assume linear for now, but possible to analyze error terms in equations)
- Allows you to predict the mean of the dependent variable, given any value of the independent variable
- Provides a PRE interpretation (proportional reduction in error) – how much we reduce our error in predicting Y by knowing values of X

26 SOC424 @ CSUN - Ellis Godard

When is Correlation Better?

- Difficult to compare regression coefficients from sample to sample, esp. with single IV regression
 - College GPA = $2.3 + 0.0023 * \text{SAT}$
 - College GPA = $2.56 + 0.12 * \text{ACT}$
- You can't tell whether ACT or SAT scores better predict college GPA
 - conditional means of Y for ACT were higher so the entire line is higher
 - those who take the ACT have higher GPAs, esp those with high ACTs

27

SOC424 @ CSUN - Ellis Godard

DEMO EXERCISE (2 of 2; also *not* the lab)

- regdemo-grades.sav, which uses three components from your performance:
 - Exam1 grade, average on 1st 3 HWs, and whether or not you've been absent more than once
 - Several outliers have been removed
- ANALYZE – REGRESSION – LINEAR
 - Exam1 is the dependent variable (the “effect”)
 - HWAVG and ABSENT are the IVs (the “cause”)
- Write and plot the equation

30

SOC424 @ CSUN - Ellis Godard

Not all are linear – e.g. logorhythmic

Happiness & Income: A Global Lens

World Happiness Index 2023 Score Vs. GDP Per Capita [Index/10, USD]



28

Source: World Happiness Report

11/26/23/

LAB EXERCISE

- Your research question is whether job satisfaction (on a scale of 1-100) varies with the amount of pay.
- Use SPSS and the reg-lab.sav dataset to find the parameters for an equation which predicts a respondent's job satisfaction using their hourly wage.
- Write the regression equation, using the intercept (“B” for the “Constant”) and slope (“B” for HOURWAGE) as given by SPSS.
- Using that equation, what level of job satisfaction do we predict for someone who is paid nothing?
- What about for someone who earns \$10/hour?
- What percent of the variance in jobsat is explained by the variance in hourwage?

31

SOC424 @ CSUN - Ellis Godard

SPSS Demonstration (this is *not* the lab!)

- regdemo1.sav
- ANALYZE – REGRESSION – LINEAR
 - X is the independent (the “cause”)
 - Y is the dependent (the “effect”)
- Write and plot the equation
- R-squared interpretations (at least 1st 2)
 - % of the variance in Y explained by the % of the variance in X
 - Strength (<.3 weak, >.7 strong)
 - Proportional reduction in error

29

SOC424 @ CSUN - Ellis Godard