Regression Guide – Quantitative Data

<u>Two Variables</u>

Scatterplots, Association, and Correlation

Relationship between two quantitative variables on the same cases (individuals).

1. **1.** : straight, curved, no pattern, other?
2. **2.** : + or – slope?
3. **3.** : how much scatter {how closely points follow the form}
4. **4.** : outliers, clusters, subgroups?

is a deliberately vague term describing the relationship between two variables.

Correlation describes the and of the linear relationship between two quantitative variables, without significant outliers.

C1 Quantitative Variables (units = __ & measure __)
C2 Straight Enough – scatterplot straight.
C3 No Outliers – no isolated points on scatterplot.
"I have two quantitative variables that satisfy the conditions, so correlation is a suitable measure of association."

$$r = \frac{\sum z_x z_y}{n-1}$$

to (= perfect, = no), has no units, and immune to changes of scale or order.

TI: LinReg(a+bx) L1, L2, Y1

r = 0.3 r = -0.8 r = 0 r = 1

Correlation is not a complete description report means and standard deviations as well.

Scatterplots and correlation coefficients never prove (variable)

Linear Regression

Residual
(model underestimates, model overestimates)

Regression line (Line of best fit) - the unique line that the variance of the (sum of the square residuals).
For standardized values:
For actual x and y values:

C1 Quantitative Variables (units = __ & measure __)
C2 Straight Enough – check original scatterplot and residual scatterplot (boring - uniform scatter w/ no direction)
C3 No Outliers – no points on scatterplot with large residuals and/or high leverage.
"I have two quantitative variables that satisfy the conditions, so the relationship can be modeled with a regression line."

1. Find slope,
2. Find y-intercept, b_0:
   plug b_1 and point (x, y) [usually ( )]
   into $\hat{y} = b_0 + b_1x$ and solve for b_0
3. Plug in slope, b_1, and y-intercept, b_0, into $\hat{y} = b_0 + b_1x$

or TI: LinReg(a+bx) L1, L2, Y1 (VARS, Y-VARS, 1:Fn)
(Residuals stored as list: RESID)

the square of the correlation coefficient, r.
The of the regression model:

(differences in x explain XX% of the variability in y)

are dubious predictions of y-values based on x-values outside the range of the original data.

Leverage and residual produce three flavors of outliers:
1)
2)
3)

orders the effects of re-expression
y^2 y $\sqrt{y}$ log y $-1/\sqrt{y}$ $-1/y$

Inference for Regression ( )

One Sample (df =)

between 2 variables

C1 Quantitative Variables (units = __ & measure __)
A2 Linearity and <u>Equal Variance</u> Assumptions
C2 Straight Enough – check and (boring - <u>uniform scatter</u> w/ no direction)
C3 No Outliers – no points on scatterplot with large residuals and/or high leverage.
A4 Independence Assumption
C4 Representative & no trends, clumps in residuals.
A5 Errors around regression line at each x Normal.
C5 OK
"Because the conditions are satisfied, I can model the sampling distribution of the with a model and perform a _____."

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}$$