Example Problems: Correlation and Regression

A researcher has heard that the more telephone poles a city has, the more murders there are per year in that city. Intrigued, he drives around 8 cities and counts how many telephone poles there are in each one. Then, he goes to the police station in each city and records how many people had been murdered in the last year. The data follow:

City	Telephone Poles (X)	Murders (Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})*(Y-\bar{Y})$
Booger Holler, AR	42	34	3	3.25	9	10.5625	9.75
Frog Suck, WY	50	37	11	6.25	121	39.0625	68.75
Hornytown, NC	35	12	-4	-18.75	16	351.5625	75
Medicine Hat, Alberta	37	19	-2	-11.75	4	138.0625	23.5
Monkey's Eyebrow, KY	52	56	13	25.25	169	637.5625	328.25
Moosejaw, Saskatchewan	12	8	-27	-22.75	729	517.5625	614.25
Rudeville, NJ	47	55	8	24.25	64	588.0625	194
Sandwich, IL	37	25	-2	-5.75	4	33.0625	11.5
Mean	**39**	**30.75**		**Sum**	**1116**	**2315.5**	**1325**
StDev	**12.627**	**18.188**					

Part 1: Correlation

A. Calculate the correlation between amount of telephone poles (X) and the number of murders per year (Y) and test for significance using all seven steps.

1. **State Null Hypothesis:** H_0: ρ __ 0 (there is no relationship between X and Y)
2. **Alternative Hypothesis:** H_1: ρ __ 0 (there is a relationship between X and Y)
3. **Decide on α (usually .05):** α = _____
4. **Decide on type of test (distribution; *z*, *t*, F, r, etc.).**
   Questions to ask:
   a. Are you comparing group means in any way?
      i. If yes then you might be using the z, t or F distributions
      ii. If no then you are not using the z, t or F distributions (continue)
   b. Are we looking for a bidirectional relationship between the variables?
   c. Are we treating both of the variables as continuous?
      If yes then continue with the calculation of a pearson correlation coefficient.
      If no, then continue with the calculation but you may have some other kind of correlation (e.g. spearman, phi-coefficient, point biserial)
5. **Find critical value & state decision rule**
   a. The degrees of freedom for a correlation is DF = N-2 = ____ - 2 = _____
   b. Using table D.3 and DF = _____, the r_{crit} = _____
   c. If $r_{observed} > r_{crit}$, reject H_0 or if $r_{observed} >$ _____, reject H_0
6. **Calculate test**
   a. Covariance Method:
      i. Calculate $\text{cov}_{xy} = \sum \frac{(X-\bar{X})(Y-\bar{Y})}{N-1} = \frac{\_\_\_\_\_}{\_\_} =$ _____
      ii. Calculate $r = \frac{\text{cov}_{xy}}{s_y * s_x} = \frac{\_\_\_\_\_}{(\_\_\_\_\_ * \_\_\_\_\_)} =$ _____

b. Z score method

City	Telephone Poles (X)	Murders (Y)	Z_X	Z_Y	Z_X*Z_Y
Booger Holler, AR	42	34	0.238	0.179	0.042
Frog Suck, WY	50	37	0.871	0.344	0.299
Hornytown, NC	35	12	-0.317	-1.031	0.327
Medicine Hat, Alberta	37	19	-0.158	-0.646	0.102
Monkey's Eyebrow, KY	52	56	1.030	1.388	1.429
Moosejaw, Saskatchewan	12	8	-2.138	-1.251	2.675
Rudeville, NJ	47	55	0.634	1.333	0.845
Sandwich, IL	37	25	-0.158	-0.316	0.050
Mean	**39**	**30.75**			
StDev	**12.627**	**18.188**			
Sum					**5.770**

i. Calculate the sum of the products of Z-scores by completing the table above.

ii. Calculate $r = \frac{Z_X Z_Y}{N-1} = \frac{\_\_\_\_\_}{\_\_-1} = \frac{\_\_\_\_\_}{\_\_} = \_\_\_\_\_$

7. Since _____ (> < =) ._____, reject H_0. The correlation between Telephone Poles and Murder Rate is/is not significant.

B. What does it mean for a correlation to be significant?

C. What important point about correlations does this example demonstrate (Hint: think about the variables)?

Part 2: Regression

A. Compute the regression equation using number of telephone poles to predict murders.

1. Compute b:

a. $b = \frac{\text{cov}_{xy}}{s_X^2} = \frac{\_\_\_\_\_}{(\_\_\_\_\_)^2} = \frac{\_\_\_\_\_}{\_\_\_\_\_} = \_\_\_\_\_$ **OR**

b. $b = r * \left(\frac{s_y}{s_x}\right) = \_\_\_\_\_ * \left(\frac{\_\_\_\_\_}{\_\_\_\_\_}\right) = \_\_\_\_\_$

2. Compute a

- $a = \bar{Y} - (b * \bar{X}) = \_\_\_\_\_ - (1.187 * \_\_\_\_\_) = \_\_\_\_\_$

D. Write out equation: $\hat{Y} = (\_\_\_\_\_ * X) - \_\_\_\_\_$

E. If a city has 300 telephone poles, how many murders would you expect to take place in that city? Why might it be problematic to make this prediction?

$\hat{Y} = (\_\_\_\_ * 300) + \_\_\_\_ = \_\_\_\_$

F. Compute the coefficient of determination and the coefficient of alienation. What do these values tell you?

1. Coefficient of Determination : r^2 = _____
2. Coefficient of Alienation: $(1-r^2)$ = _____

G. Compute the Standard Error of Estimate. What does this value tell you?

$$S_{Y-\hat{Y}} = S_Y \sqrt{(1-r^2)*\left(\frac{N-1}{N-2}\right)} = \_\_\_\_\_ * \sqrt{\_\_\_\_ * \left(\frac{\_\_}{\_\_}\right)} = \_\_\_\_\_ * \sqrt{\_\_\_\_ * (\_\_\_\_)} = \_\_\_\_\_$$

H. Test the regression for significance.

1. With a single predictor you can do this through ANOVA:
   a. Compute SS_{Total}, $SS_{regression}$, and $SS_{residual}$. Hint: ($SS_{Total} = SS_Y$)
      i. $SS_{Total} = \sum (Y - \bar{Y})^2 = \_\_\_\_\_$
      ii. $SS_{\text{Regression}} = SS_{Total} * r^2 = \_\_\_\_\_ * \_\_\_\_ = \_\_\_\_\_$
      iii. $SS_{\text{Residual}} = SS_{Total} - SS_{\text{Regression}} = \_\_\_\_\_ - \_\_\_\_\_ = \_\_\_\_\_$
   b. Compute the Degrees of Freedom
      i. $df_{Total} = N - 1 = \_\_\_ - 1 = \_\_\_$
      ii. $df_{regression} = \#\ predictors = \_\_\_$
      iii. $df_{redidual} = df_{Total} - df_{regression} = \_\_\_ - 1 = \_\_\_$
   c. Compute $MS_{\text{Regression}}$ and MS_{Residual} and F.
      i. $MS_{regression} = \frac{SS_{regression}}{df_{regression}} = \frac{\_\_\_\_\_}{\_\_} = \_\_\_\_\_$
      ii. $MS_{residual} = \frac{SS_{residual}}{df_{residual}} = \frac{\_\_\_\_\_}{\_\_} = \_\_\_\_\_$
      iii. $F = \frac{MS_{regression}}{MS_{residual}} = \frac{\_\_\_\_\_}{\_\_\_\_\_} = \_\_\_\_\_$
      iv. F_{crit} (___, ___) = _____
      v. Since _____ (> < =) _____, our regression equation is/is not significant, meaning that number of telephone poles in a city does not/significantly predicts murder rate with number of telephone poles accounting for about 68% of the variability in murder rate.
2. Or you can test test the slope for significance.
   a. Calculate the standard error for the slope.

$$SEb = \frac{S_{Y-\hat{Y}}}{s_x * \sqrt{N-1}} = \frac{\_\_\_\_\_}{\_\_\_\_\_ * \sqrt{\_\_}} = \_\_\_\_\_$$

   b. Divide the slope by its standard error for t_{obs}. $t = \frac{b}{SEb} = \frac{\_\_\_\_\_}{\_\_\_\_\_} = \_\_\_\_\_$
   c. T_{crit} (___)= _____
   d. Since _____ (> < =) _____, the slope is significant.

CAVEAT: This data is completely made up, although a correlation like this does exist. The number of murders in these cities are completely made up, so don't hesitate to visit the wonderful cities of Frog Suck, Booger Holler, or Sandwich. I'm sure they are very nice. And yes, these ARE actual city names. I didn't make THOSE up. ☺