**Homework 6: Correlation + Regression**

Due: Nov 5th

**Correlation and Regression:**

Questions 1- 6

**SPSS Assignment**:

**Bivariate Correlation and Bivariate Regression**

**(**Use data set from Problem 5 (DO NOT ENTER THE MEANS IN THE LAST ROW).

**A.** Use **Graphs → Legacy Dialogs → Simple Scatter** to create a scatterplot of the relationship between the predictor and criterion variables. Put your predictor on the X axis. Remember to add the regression line in your graph.

**B.** Use **Analyze → Correlate → Bivariate** to run a bivariate correlation in SPSS.
   a. Label the correlation coefficient and significance level.
   b. Explain what the numbers mean in terms of the variables of the study.

**C.** Use **Analyze → Regression → Linear.** Once you are in the LINEAR dialog box, indicate that your independent variable (i.e., predictor variable) is therapist empathy.
   a. On the regression output, label the values for $b_0$, $b_1$, $\beta$, $SS_R$, $SS_T$, $r$, and $R^2$ (see textbook for good examples)
   b. Explain what these numbers mean in terms of the variables in the study.
   c. Use $SS_R$ and $SS_T$ to calculate the model fit $SS_M$ and compare it to $R^2$.
   d. Provide the bivariate regression equation, $\hat{Y} = (b_0 + b_1 X) + \varepsilon$
   e. Compare your hand calculations with the computer output, and identify where your answers are the same or different.

**D.** Repeat these procedures, but this time, predict therapist empathy from patient satisfaction.
   a. Label the values for $b_0$, $b_1$, $\beta$, $SS_R$, $SS_T$, $r$, and $R^2$ (see textbook for good examples).
   b. Provide the bivariate regression equation, $\hat{Y} = (b_0 + b_1 X) + \varepsilon$
   c. Compare these numbers to those from the original output for the problem above, and identify where your answers are the same or different.

**Multiple Regression**

**E.** Add patient self-esteem as a new predictor variable for the original Problem 5: "Patient self-esteem; 50,60,70,70"

**F.** Re-run the regression procedure from the original Problem 5, including the new predictor variable in the equation. (Note: You should have 2 independent variables – empathy and self-esteem – indicated in the dialog box).
   a. On the output, label $b_0$, $b_1$, $b_2$, $\beta_1$, $\beta_2$, $p$ value for 1, $p$ value for 1, R, and $R^2$.
   b. Explain what the numbers for these mean in terms of the variables in the study. (Note: In the multiple regression output, R is the *multiple* correlation coefficient and $R^2$ is the entire model $R^2$).
   c. Give the multiple regression equation.
   d. Did adding a new variable help in predicting test scores? Explain how you know.

**Correlation**

**1.** What are *r*, $r^2$, R, and $R^2$?

**2.** Write the formula for the correlation coefficient, and explain what each part of the formula represents (See class notes).

**Regression**

**3.** Identify the assumptions of the linear model.

**4.** A sports psychologist working with hockey players has found that players' knowledge of physiology predicts the number of injuries received over the subsequent year. The regression constant in the linear prediction rule for predicting injuries from knowledge of physiology is 10.30 and the regression coefficient is -.70.

a. Indicate the predictor variable (IV).
b. Indicate the criterion variable (DV).
c. Provide the regression model for this example, $\hat{Y} = (b_0 + b_1 X) + \varepsilon$, and determine the *predicted* number of injuries for athletes whose scores on the physiology test are:
   i. 0
   ii. 1
   iii. 2
   iv. 5
   v. 6

**5.** A pilot study with four patient-therapist pairs were determined the relationship between psychotherapists' degree of empathy and their patients' satisfaction with therapy. The results are presented here, including the means:

| Pair Number | Therapist Empathy | Patient Satisfaction |
|---|---|---|
| 1 | 70 | 4 |
| 2 | 94 | 5 |
| 3 | 36 | 2 |
| 4 | 48 | 1 |
| *Means* | *62* | *3* |

a. Determine the prediction model, $\hat{Y} = (b_0 + b_1 X) + \varepsilon$, for predicting satisfaction from empathy.
b. Based on that model, draw a regression line. Hint: Centroid and Y-intercept. Predict from your regression line and use the model to predict patient satisfaction; make sure to round to 3 decimal points. Compare your model prediction to that based on your regression line for those whose therapist's empathy was:
   i. 50
   ii. 64
   iii. 80

c. Determine the standardized regression coefficient.
d. Find the proportionate reduction in error (using $SS_R$ and $SS_T$).

**6.** For the following, determine the *multiple linear prediction rule* for predicting criterion variable $Y$ from predictor variables $X_1$, $X_2$, and $X_3$;

$$\hat{Y} = b_0 + (b_1)(X_1) + (b_2)(X_2) + (b_3)(X_3) + \varepsilon$$

| **Regression Constant** | **Regression Coefficient** | **Regression Coefficient** | **Regression Coefficient** |
|---|---|---|---|
| $b_0$ | $b_1$ | $b_2$ | $b_3$ |
| 1.5 | 0.8 | -0.3 | 9.99 |

Use the model to predict a score on the criterion variable for a person with scores on the predictor variables of $X_1 = 2$, $X_2 = 5$, and $X_3 = 9$.

**Homework 6: Correlation + Linear Regression**
**Answers**

## Correlation

**1.** What are *r*, $r^2$, R, and $R^2$?

a. *Correlation Coefficient* = *r*: the numerical measure of association between two quantitative variables. The sign of *r* tells you the direction of the correlation, the magnitude of r tells you about the strength of the correlation. Value of *r* can range from -1 to +1.
b. *Coefficient of Determination* = $r^2$: the percentage of variance in a variable that is accounted for/explained by the variance in another variable.
c. *Multiple Correlation Coefficient* = R: the numeric measure of the correlation between a criterion variable and its predictors.
d. *Coefficient of Determination in linear regression* = $R^2$: In bivariate regression, $R^2$ represents the same concept as $r^2$ (the total variance explained by the predictor in the model). In multivariate regression, $R^2$ represents the amount of variability in the dependent variable explained by using multiple predictor variables (the total variance explained by the predictors in the model).

**2.** Write the formula for Pearson's product-moment correlation coefficient, and explain what each part of the formula represents.

$$r_{xy} = \frac{\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)}{\sqrt{[\Sigma\left(X-\bar{X}\right)^2][\Sigma\left(Y-\bar{Y}\right)^2]}}$$

$\sum\left(X-\bar{X}\right)\left(Y-\bar{Y}\right)$ = sum of cross products $\left(SP_{XY}\right)$. Provides a measure of the nature and magnitude of covariance between X and Y.

$\left[\sum\left(X-\bar{X}\right)^2\right]\left[\sum\left(Y-\bar{Y}\right)^2\right]$ = sum of squared deviations in X $\left(SS_X\right)$ and sum of squared deviations in Y $\left(SS_Y\right)$. Provides a measure of independent variability for each variable.

**Regression**

**3.** Assumptions of the Linear Model

a. **Linearity**: The outcome variable (dependent variable) should be linearly related to any predictors. There should be a linear relationship between variables.

b. **Homoscedasticity:** The variance around the regression line (the residuals) is the same for all values of the predictor variable (e.g., X). This is an equal-variance assumption.

c. **Normally Distributed Errors:** Residuals in the model are random, normally distributed, and have a mean of 0. This assumption does not imply that predictors have to be normally distributed. Small sample sizes will limit the distribution.

d. **Predictors are Uncorrelated with External Variables:** Similar to the "third variable problem addressed in earlier homework. If external variables correlate with predictors, then the conclusions we draw from the model become unreliable. If you know that Variable Z is correlated strongly with a variable of interest, plan early and collect Variable Z too.

e. **Variable Types:** All predictor variables must be either quantitative or categorical with two categories. The outcomes variable must be quantitative, continuous, and unbounded.

f. **No Perfect *Multi*collinearity:** No perfect collinearity between two or more of the predictors; predictor variables should not correlate too highly. When predictors correlate highly, problems occur with estimating regression coefficients because minor fluctuations in the sample (e.g., measurement errors, sampling error) will have a major impact on the beta weights. Collinearity means that within the set of predictors, some of them are (nearly) totally predicted by another predictor.

g. **Non-Zero Variance:** The predictors should have some variation in value.

If these assumptions are not met, we need to correct for them if possible or we cannot use this model.

**4. Sports Question**

a. Predictor Variable (IV) = Score on knowledge of physiology.
b. Criterion Variable (DV) = Number of injuries
c. Linear prediction rule formula: $\hat{Y} = (b_0 + b_1 X) + \varepsilon$

e.g., predicted number of injuries = 10.30 + (-.70)(knowledge of physiology)

i. $X = 0; \hat{Y} = 10.30 + (-.70)(0)$
ii. $X = 1; \hat{Y} = 9.60$
iii. $X = 2; \hat{Y} = 8.90$
iv. $X = 5; \hat{Y} = 6.80$
v. $X = 6; \hat{Y} = 6.10$

**5. Patient/Therapist Predictions**

| Therapist Empathy (X) | $(X - \bar{X})$ | $(X - \bar{X})^2$ | Patient Satisfaction (Y) | $(Y - \bar{Y})$ | $(Y - \bar{Y})^2$ | $(X - \bar{X})(Y - \bar{Y})$ |
|---|---|---|---|---|---|---|
| 70 | 8 | 64 | 4 | 1 | 1 | 8 |
| 94 | 32 | 1024 | 5 | 2 | 4 | 64 |
| 36 | -26 | 676 | 2 | -1 | 1 | 26 |
| 48 | -14 | 196 | 1 | -2 | 4 | 28 |
| $\bar{X} = 62$ | | $SS_X = 1960$ | $\bar{Y} = 3$ | | $SS_Y = 10$ | $SP = 126$ |

a. Determine the linear prediction model.

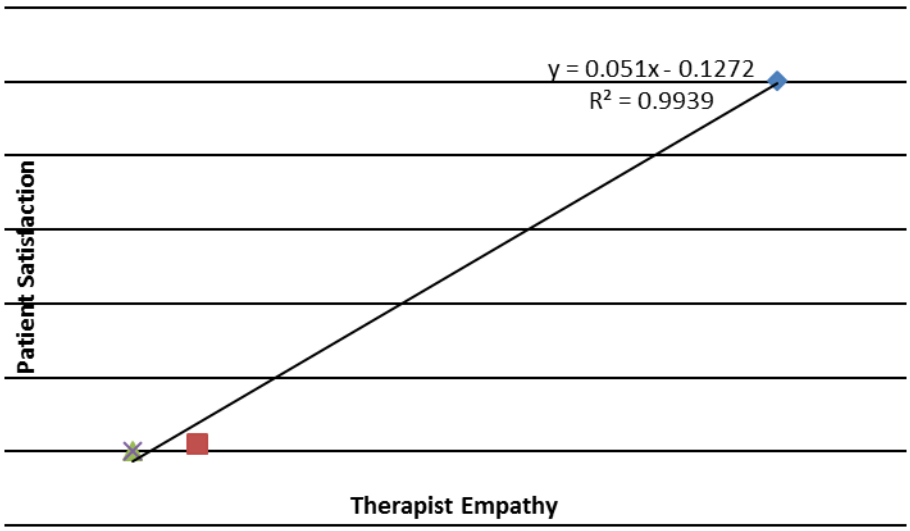
$$\hat{Y}_i = (b_0 + b_1 X_i) + \varepsilon$$

$$b_{1Y \bullet X} = \frac{SP_{XY}}{SS_X} = \frac{126}{1960} = .064$$

$$b_0 = \bar{Y} - (b_1)(\bar{X})$$
$$= 3 - (.064)(62)$$
$$= 3 - 3.968$$
$$= -.968$$

$$\hat{Y} = -.968 + (.064)(X)$$

b. Draw the regression line (using the centroid and the intercept would be best).




i. $X = 50; \hat{Y} = -.968 + (.064)(50) = 2.232$

ii. $X = 64; \hat{Y} = 3.128$

iii. $X = 80; \hat{Y} = 4.152$

c. $\beta = (b_1)\frac{\sqrt{SS_X}}{\sqrt{SS_Y}} = (.064)\frac{\sqrt{1960}}{\sqrt{10}} = (.064)\frac{44.27}{3.16} = (.064)(14) = .896$ or .90

Remember, $\beta$ is standardized, so we can explain this easily in terms of standard deviations. So, for a case where the *empathy score* is 1 **standard deviation** larger, the *satisfaction score* is .84 standard deviations larger. Note: If $\beta$ was negative (because of the slope, $b_1$), we could say that when an *empathy score* is 1 **standard deviation** larger, the *satisfaction score* is .84 standard deviations lower. As a hint for interpretation, consider the slope of the regression line. When the slope is positive (the correlation is positive), increases on X on your graph correspond to increase in Y, whereas when the slope is negative (the correlation is negative), increases on X on your graph correspond to decreases in Y. Doing this also provide you with another opportunity to check that your calculated slope makes sense because your slope corresponds with your correlation.

d. Model fit: $\frac{SS_{Total} - SS_{Error}}{SS_{Total}} = \frac{SS_M}{SS_T}$

| Linear equation | Actual Satisfaction (Y) | $\hat{Y}$ | "Residual" $\left(Y - \hat{Y}\right)$ | $\left(Y - \hat{Y}\right)^2$ |
|---|---|---|---|---|
| -.968+(.064)(70) | 4 | 3.512 | 4-3.512 = .488 | 0.228 |
| -.968+(.064)(94) | 5 | 5.048 | 5-5.048 = -.048 | 0.002 |
| -.968+(.064)(36) | 2 | 1.336 | 2-1.336 = .664 | 0.441 |
| -.968+(.064)(48) | 1 | 2.104 | 1-2.104 = -1.104 | 1.219 |
| | | | $SS_R$ = | 1.900 |

| Actual Satisfaction (Y) | $\overline{Y}$ | $(Y - \overline{Y})$ | $(Y - \overline{Y})^2$ |
|---|---|---|---|
| 4 | 3 | 1 | 1 |
| 5 | 3 | 2 | 4 |
| 2 | 3 | -1 | 1 |
| 1 | 3 | -2 | 4 |
| | | $SS_{Total}$ = | 10 |

$$\frac{SS_{Total} - SS_R}{SS_{Total}} = \frac{SS_M}{SS_T} = \frac{10 - 1.90}{10} = 0.81$$

Thus, 81% of the total error from the mean-based model has been reduced by using therapist empathy to predict patient satisfaction. However, your model does not explain 19% of the variance in patient satisfaction.

**6. Multiple Linear Regression**

Find the multiple prediction model:

$b_0 = 1.5; \quad b_1 = 0.8; \quad b_2 = -0.3; \quad b_3 = 9.99$

$\hat{Y} = 1.5 + (0.8)(X_1) + (-0.3)(X_2) + (9.99)(X_3)$

Find the predicted score given the parameters:

$X_1 = 2; \quad X_2 = 5; \quad X_3 = 9$

$\hat{Y} = 1.5 + (0.8)(2) + (-0.3)(5) + (9.99)(9)$

$= 1.5 + 1.6 - 1.5 + 89.91$

$= 91.51$

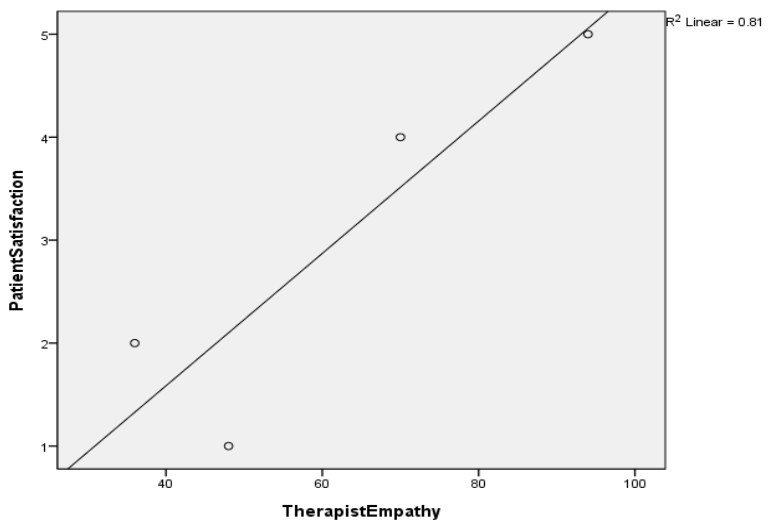
**SPSS 1**

**Bivariate Correlation and Bivariate Regression**

**Syntax:**

```
GRAPH
  /SCATTERPLOT(BIVAR)=TherapistEmpathy WITH PatientSatisfaction
  /MISSING=LISTWISE.

CORRELATIONS
  /VARIABLES=TherapistEmpathy PatientSatisfaction
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT PatientSatisfaction
  /METHOD=ENTER TherapistEmpathy.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT TherapistEmpathy
  /METHOD=ENTER PatientSatisfaction.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT PatientSatisfaction
  /METHOD=ENTER TherapistEmpathy SelfEsteem.
```

**A.**




(See HW 5 for more details about the scatterplot)

B.

**Correlations**

| | | Therapist Empathy | Patient Satisfaction |
|---|---|---|---|
| Therapist Empathy | Pearson Correlation | 1 | .900 |
| | Sig. (2-tailed) | | .100 |
| | N | 4 | 4 |
| Patient Satisfaction | Pearson Correlation | .900 | 1 |
| | Sig. (2-tailed) | .100 | |
| | N | 4 | 4 |

The correlation between therapist empathy and patient satisfaction is .9, indicating a strong positive relationship. However, we fail to reach statistical significance and cannot conclude that the correlation is not .0 in the population. Note that the test of statistical significance is 2-tailed. What would the *p*-value be for a one-tailed test?

**C.**

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .900[a] | .810 | .715 | .975 |

a. Predictors: (Constant), Therapist Empathy

*r* = *R* because this is a bivariate regression, not multiple, regression

$R^2 = .90^2 = .81$

$= (.81*100) = 81\%$

81% of the variance in patient satisfaction can be accounted for by the therapist's empathy.

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8.100 | 1 | 8.100 | 8.526 | .100[a] |
| | Residual | 1.900 | 2 | .950 | | |
| | Total | 10.000 | 3 | | | |

a. Predictors: (Constant), Therapist Empathy

b. Dependent Variable: Patient Satisfaction

$SS_{Total} = \sum(Y - \bar{Y})^2$ = The sum of squared deviations of each patient satisfaction score from the mean patient satisfaction value.

$SS_{error} = \sum\left(Y - \hat{Y}\right)^2$

The sum of squared deviations for how far the patient satisfaction of each pair deviates from its predicted value (the value that falls on the regression line).

*Proportionate Reduction in Error* = $\frac{SS_{Total} - SS_{Error}}{SS_{Total}} = \frac{10 - 1.900}{10} = .81 \equiv R^2$

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -.986 | 1.449 | | -.680 | .567 |
| | Therapist Empathy | .064 | .022 | .900 | 2.920 | .100 |

a. Dependent Variable: Patient Satisfaction

*a*: The value of patient satisfaction when therapist empathy is zero

*b*: The predicted change in patient satisfaction for an increase of 1 in therapist empathy.

*β*: For a case where therapist empathy is 1 standard deviation larger, patient satisfaction is .9 standard deviations larger. Note that this is same as *r*.

*p* >.05. We cannot reject the H0 that the coefficient is zero in the population.

Bivariate regression equation = $\hat{Y} = -.986 + (.064)(X)$

The hand calculations from problem 6 do not completely correspond due to rounding error. The unstandardized regression coefficient =

$$= \frac{\Sigma[(X - \bar{X})(Y - \bar{Y})]}{SS_X} = \frac{126}{1960} = .0642857143$$

In our hand calculations we opted to round to .06, hence producing a different constant (a) and beta weight ($\beta$).

**D.**

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .900[a] | .810 | .715 | 13.646 |

a. Predictors: (Constant), Patient Satisfaction

*r* : Unaffected by the change in predictor.

$R_2$ : Unaffected by the change in predictor.

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1587.600 | 1 | 1587.600 | 8.526 | .100[a] |
| | Residual | 372.400 | 2 | 186.200 | | |
| | Total | 1960.000 | 3 | | | |

a. Predictors: (Constant), Patient Satisfaction

b. Dependent Variable: Therapist Empathy

SSError = This is now the error term for therapist empathy.

$SS_{Total}$ : This is now SS for therapist empathy.

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 24.200 | 14.633 | | 1.654 | .240 |
| | PatientSatisfaction | 12.600 | 4.315 | .900 | 2.920 | .100 |

a. Dependent Variable: TherapistEmpathy

*p* >.05. We cannot reject the H0 that the coefficient is zero in the population.

a: The value of therapist empathy when patient satisfaction is zero.

b: The predicted change in therapist empathy for an increase of 1 in patient satisfaction.

β: The beta is the same as before due to it being standardized. The unstandardized coefficients are different whereby the scale changes when we changed predictor.

**SPSS 2 - Multiple Regression**

**E, F**.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .929[a] | .863 | .589 | 1.170 |

a. Predictors: (Constant), Self Esteem, Therapist Empathy

R = The correlation between the criterion variable and the two predictors.

$R_2 = .929^2 = .86 = 86\%$ of the variance in patient satisfaction can be accounted for by therapist empathy *and* self esteem.

**ANOVA[b]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8.631 | 2 | 4.315 | 3.152 | .370[a] |
| | Residual | 1.369 | 1 | 1.369 | | |
| | Total | 10.000 | 3 | | | |

a. Predictors: (Constant), Self Esteem, Therapist Empathy

b. Dependent Variable: Patient Satisfaction

$p_1$ & $p_2$ are both >.05. We cannot reject the H0:s that either coefficient are zero in the population.

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | S |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | |
| 1 | (Constant) | 3.525 | 7.450 | | .473 | .719 |
| | Therapist Empathy | .050 | .035 | .701 | 1.433 | .388 |
| | Self Esteem | -.058 | .093 | -.305 | -.623 | .645 |

a: Patient Satisfaction

a: The value of patient satisfaction when both predictors are zero.

$\beta_1$ & $\beta_2$: Standardized regression weights; the predicted standard deviation increase in patient satisfaction for 1 standard deviation larger in the corresponding predictor.

$b_1$: The predicted change in patient satisfaction for an increase of 1 in therapist empathy, holding self-esteem constant.

$b_2$: The predicted change in patient satisfaction for an increase of 1 in self-esteem, holding therapist empathy constant.

Adding self-esteem as a predictor increased the $R^2$ from .81 to .86. Though, we should note that neither of the two predictors reached statistical significance.