Advanced Statistics
Homework-7
Due: After April 25, 2025 and on or before May 05, 2025

Question 1: Bivariate data often arises from the use of two different techniques to measure the same quantity. As an example, the accompanying observations on x = hydrogen concentration (ppm) using a gas chromatography method and y = concentration using a new sensor method were read from a graph in the article **"A New Method to Measure the Diffusible Hydrogen Content in Steel Weldments Using a Polymer Electrolyte-Based Hydrogen Sensor"** (*Welding Res., July 1997: 251s–256s*).

x	47	62	65	70	70	78	95	100	114	118
y	38	62	53	67	84	79	93	106	117	116
x	124	127	140	140	140	150	152	164	198	221
y	127	114	134	139	142	170	149	154	200	215

Construct a scatterplot. Does there appear to be a very strong relationship between the two types of concentration measurements? Do the two methods appear to be measuring roughly the same quantity? Explain your reasoning.

Question 2: The accompanying data on y = ammonium concentration (mg/L) and x = transpiration (ml/h) was read from a graph in the article **"Response of Ammonium Removal to Growth and Transpiration of *Juncus effusus* During the Treatment of Artificial Sewage in Laboratory-Scale Wetlands"** (*Water Research, 2013: 4265–4273*). The article's abstract stated "a linear correlation between the ammonium concentration inside the rhizosphere and the transpiration of the plant stocks implies that an influence of plant physiological activity on the efficiency of N-removal exists." (The rhizosphere is the narrow region of soil at the plant root–soil interface, and transpiration is the process of water movement through a plant and its evaporation.) The article reported summary quantities from a simple linear regression analysis. Based on a scatterplot, how would you describe the relationship between the variables, and does simple linear regression appear to be an appropriate modeling strategy?

x	5.8	8.8	11.0	13.6	18.5	21.0	23.7
y	7.8	8.2	6.9	5.3	4.7	4.9	4.3
x	26.0	28.3	31.9	36.5	38.2	40.4	
y	2.7	2.8	1.8	1.9	1.1	0.4	

Question 3: The flow rate y (m^3/min) in a device used for air-quality measurement depends on the pressure drop x (in. of water) across the device's filter. Suppose that for x values between 5 and 20, the two variables are related according to the simple linear regression model with true regression line

$$y = -0.12 + 0.095x.$$

a. What is the expected change in flow rate associated with a 1-in. increase in pressure drop? Explain.

b. What change in flow rate can be expected when pressure drop decreases by 5 in.?

c. What is the expected flow rate for a pressure drop of 10 in.? A drop of 15 in.?

d. Suppose $\sigma = 0.025$ and consider a pressure drop of 10 in. What is the probability that the observed value of flow rate will exceed 0.835? That observed flow rate will exceed 0.840?

e. What is the probability that an observation on flow rate when pressure drop is 10 in. will exceed an observation on flow rate made when pressure drop is 11 in.?

Question 4: Refer back to the data in Exercise 2, in which y = ammonium concentration (mg/L) and x = transpiration (ml/h). Summary quantities include $n = 13$, $\sum x_i = 303.7$, $\sum y_i = 52.8$, $S_{xx} = 1585.230769$, $S_{xy} = -341.959231$, and $S_{yy} = 77.270769$.

a. Obtain the equation of the estimated regression line and use it to calculate a point prediction of ammonium concentration for a future observation made when ammonium concentration is 25 ml/h.

b. What happens if the estimated regression line is used to calculate a point estimate of true average concentration when transpiration is 45 ml/h? Why does it not make sense to calculate this point estimate?

c. Calculate and interpret s.

d. Do you think the simple linear regression model does a good job of explaining observed variation in concentration? Explain.

Question 5: The article **"Characterization of Highway Runoff in Austin, Texas, Area"** (*J. of Envir. Engr., 1998: 131–137*) gave a scatterplot, along with the least squares line, of x = rainfall volume (m^3) and y = runoff volume (m^3) for a particular location. The accompanying values were read from the plot.

x	5	12	14	17	23	30	40	47
y	4	10	13	15	15	25	27	46
x	55	67	72	81	96	112	127	
y	38	46	53	70	82	99	100	

a. Does a scatterplot of the data support the use of the simple linear regression model?

b. Calculate point estimates of the slope and intercept of the population regression line.

c. Calculate a point estimate of the true average runoff volume when rainfall volume is 50.

d. Calculate a point estimate of the standard deviation σ.

e. What proportion of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall?

Question 6: No-fines concrete, made from a uniformly graded coarse aggregate and a cement–water paste, is beneficial in areas prone to excessive rainfall because of its excellent drainage properties. The article **"Pavement Thickness Design for No-Fines Concrete Parking Lots,"** *J. of Trans. Engr., 1995: 476–484* employed a least squares analysis in studying how y = porosity (%) is related to x = unit weight (pcf) in concrete specimens. Consider the following representative data:

x	99.0	101.1	102.7	103.0	105.4	107.0	108.7	110.8
y	28.8	27.9	27.0	25.2	22.8	21.5	20.9	19.6
x	112.1	112.4	113.6	113.8	115.4	115.4	120.0	
y	17.1	18.9	16.0	16.7	13.0	13.6	10.8	

Relevant summary quantities are:
$\sum x_i = 1640.1$, $\sum y_i = 299.8$, $\sum x_i^2 = 179{,}849.73$, $\sum x_i y_i = 32{,}308.59$, $\sum y_i^2 = 6430.06$

a. Obtain the equation of the estimated regression line. Then create a scatterplot of the data and graph the estimated line. Does it appear that the model relationship will explain a great deal of the observed variation in y?

b. Interpret the slope of the least squares line.

c. What happens if the estimated line is used to predict porosity when unit weight is 135? Why is this not a good idea?

d. Calculate the residuals corresponding to the first two observations.

e. Calculate and interpret a point estimate of σ.

f. What proportion of observed variation in porosity can be attributed to the approximate linear relationship between unit weight and porosity?

Question 7: For the past decade, rubber powder has been used in asphalt cement to improve performance. The article **"Experimental Study of Recycled Rubber-Filled High-Strength Concrete"** (*Magazine of Concrete Res., 2009: 549–556*) includes a regression of y = axial strength (MPa) on x = cube strength (MPa) based on the following sample data:

x	112.3	97.0	92.7	86.0	102.0	99.2	95.8	103.5	89.0	86.7
y	75.0	71.0	57.7	48.7	74.3	73.3	68.0	59.3	57.8	48.5

a. Obtain the equation of the least squares line, and interpret its slope.

b. Calculate and interpret the coefficient of determination.

c. Calculate and interpret an estimate of the error standard deviation σ in the simple linear regression model.

Question 8: The bond behavior of reinforcing bars is an important determinant of strength and stability. The article **"Experimental Study on the Bond Behavior of Reinforcing Bars Embedded in Concrete Subjected to Lateral Pressure"** (*J. of Materials in Civil Engr., 2012: 125–133*) reported the results of one experiment in which varying levels of lateral pressure were applied to 21 concrete cube specimens, each with an embedded 16-mm plain steel round bar, and the corresponding bond capacity was determined.

Due to differing concrete cube strengths (f_{cu} in MPa), the applied lateral pressure was equivalent to a fixed proportion of the specimen's f_{cu} (0, $0.1f_{cu}$, ..., $0.6f_{cu}$). Also, since bond strength can be heavily influenced by the specimen's f_{cu}, bond capacity was expressed as the ratio of bond strength (MPa) to $\sqrt{f_{cu}}$.

Pressure	0	0	0	0.1	0.1	0.1	0.2
Ratio	0.123	0.100	0.101	0.172	0.133	0.107	0.217
Pressure	0.2	0.2	0.3	0.3	0.3	0.4	0.4
Ratio	0.172	0.151	0.263	0.227	0.252	0.310	0.365
Pressure	0.4	0.5	0.5	0.5	0.6	0.6	0.6
Ratio	0.239	0.365	0.319	0.312	0.394	0.386	0.320

a. Does a scatterplot of the data support the use of the simple linear regression model?

b. Use the accompanying Minitab output to give point estimates of the slope and intercept of the population regression line.

c. Calculate a point estimate of the true average bond capacity when lateral pressure is $0.45f_{cu}$.

d. What is a point estimate of the error standard deviation σ, and how would you interpret it?

e. What is the value of total variation, and what proportion of it can be explained by the model relationship?

Question 9: During oil drilling operations, components of the drilling assembly may suffer from sulfide stress cracking. The article "Composition Optimization of High-Strength Steels for Sulfide Cracking Resistance Improvement" (Corrosion Science, 2009: 2878–2884) reported on a study in which the composition of a standard grade of steel was analyzed. The following data on y = threshold stress (% SMYS) and x = yield strength (MPa) was read from a graph in the article (which also included the equation of the least squares line).

x	635	644	711	708	836	820	810	870	856	923	878	937	948
y	100	93	88	84	77	75	74	63	57	55	47	43	38

Also given:

$$\sum x_i = 10,576, \quad \sum y_i = 894, \quad \sum x_i^2 = 8,741,264, \quad \sum y_i^2 = 66,224, \quad \sum x_i y_i = 703,192$$

(a) What proportion of observed variation in stress can be attributed to the approximate linear relationship between the two variables?

(b) Compute the estimated standard deviation $s_{\hat{\beta}_1}$.

(c) Calculate a confidence interval using confidence level 95% for the expected change in stress associated with a 1 MPa increase in strength. Does it appear that this true average change has been precisely estimated?

Question 10: Electromagnetic technologies offer effective nondestructive sensing techniques for determining characteristics of pavement. The propagation of electromagnetic waves through the material depends on its dielectric properties. The following data, kindly provided by the authors of the article "Dielectric Modeling of Asphalt Mixtures and Relationship with Density" (J. of Transp. Engr., 2011: 104–111), was used to relate y = dielectric constant to x = air void (%) for 18 samples having 5% asphalt content:

y	4.55	4.49	4.50	4.47	4.47	4.45	4.40	4.34	4.43
x	4.35	4.79	5.57	5.20	5.07	5.79	5.36	6.40	5.66

y	4.43	4.42	4.40	4.33	4.44	4.40	4.26	4.32	4.34
x	5.90	6.49	5.70	6.49	6.37	6.51	7.88	6.74	7.08

The R output below is from a simple linear regression of y on x:

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  4.858691    0.059768   81.283     <2e-16
AirVoid     -0.074676    0.009923   -7.526   1.21e-06

Residual standard error: 0.03551 on 16 DF Multiple
R-squared:   0.7797,   Adjusted   R-squared:   0.766
F-statistic: 56.63 on 1 and 16 DF, p-value: 1.214e-06

Analysis of Variance Table

Response: Dielectric
           Df    Sum Sq   Mean Sq  F value     Pr(>F)
Airvoid     1  0.071422  0.071422   56.635  1.214e-06
Residuals  16   0.20178  0.001261
```

(a) Obtain the equation of the least squares line and interpret its slope.

(b) What proportion of observed variation in dielectric constant can be attributed to the approximate linear relationship between dielectric constant and air void?

(c) Does there appear to be a useful linear relationship between dielectric constant and air void? State and test the appropriate hypotheses.

(d) Suppose it had previously been believed that when air void increased by 1 percent, the associated true average change in dielectric constant would be at least -0.05. Does the sample data contradict this belief? Carry out a test of appropriate hypotheses using a significance level of 0.01.

Question 11: How does lateral acceleration—side forces experienced in turns that are largely under driver control—affect nausea as perceived by bus passengers? The article "Motion Sickness in Public Road Transport: The Effect of Driver, Route, and Vehicle" (Ergonomics, 1999: 1646–1664) reported data on $x =$ motion sickness dose (calculated in accordance with a British standard for evaluating similar motion at sea) and $y =$ reported nausea (%). Relevant summary quantities are: $n = 17$, $\sum x_i = 222.1$, $\sum y_i = 193$, $\sum x_i^2 = 3056.69$, $\sum x_i y_i = 2759.6$, $\sum y_i^2 = 2975$

Values of dose in the sample ranged from 6.0 to 17.6.

(a) Assuming that the simple linear regression model is valid for relating these two variables (this is supported by the raw data), calculate and interpret an estimate of the slope parameter that conveys information about the precision and reliability of estimation.

(b) Does it appear that there is a useful linear relationship between these two variables? Test appropriate hypotheses using $\alpha = 0.01$.

(c) Would it be sensible to use the simple linear regression model as a basis for predicting % nausea when dose = 5.0? Explain your reasoning.