### 7.4 Problems

1. In Problem 4.4/29, it was shown that important relations exist between population means, variances, and covariance. Specifically, we have the formulas that appear below left.

   ![Formulas](image)

   In this problem, we verify that these properties are also true for sample means, variances, and covariance in two examples. For data values \(\{x_1, x_2, \ldots, x_n\}\) and \(\{y_1, y_2, \ldots, y_n\}\), recall that:

   \[
   \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
   \]

   \[
   \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2
   \]

   Now suppose that each value \(x_i\) from the first sample is paired with exactly one corresponding value \(y_i\) from the second sample. That is, we have the set of \(n\) ordered pairs of data \(\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}\), with sample covariance given by

   \[
   s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
   \]

   Furthermore, we can label the pairwise sum “\(x + y\)” as the dataset \((x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)\), and likewise for the pairwise difference “\(x - y\).” It can be shown (via basic algebra, or Appendix A2), that for any such dataset of ordered pairs, the formulas that appear above right hold. (Note that these formulas generalize the properties found in Problem 2.5/4.)

   For the following ordered data pairs, verify that the formulas in I and II hold. (In \(\mathbb{R}\), use \texttt{mean}, \texttt{var}, and \texttt{cov}.) Also, sketch the scatterplot.

   ![Data Table](image)

   Repeat for the following dataset. Notice that the values of \(x_i\) and \(y_i\) are the same as before, but the correspondence between them is different!

   ![Data Table](image)
2. Expiration dates that establish the shelf lives of pharmaceutical products are determined from stability data in drug formulation studies. In order to measure the rate of decomposition of a particular drug, it is stored under various conditions of temperature, humidity, light intensity, etc., and assayed for intact drug potency at FDA-recommended time intervals of every three months during the first year. In this example, the assay $Y$ (mg) of a certain 500 mg tablet formulation is determined at time $X$ (months) under ambient storage conditions.

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>500</td>
<td>490</td>
<td>470</td>
<td>430</td>
<td>350</td>
</tr>
</tbody>
</table>

(a) Graph these data points $(x_i, y_i)$ in a scatterplot, and calculate the sample correlation coefficient $r = \frac{s_{xy}}{s_x s_y}$. Classify the correlation as positive or negative, and as weak, moderate, or strong.

(b) Determine the equation of the least squares regression line for these data points, and include a 95% confidence interval for the slope $\beta_1$.

(c) Sketch a graph of this line on the same set of axes as part (a); also calculate and plot the fitted response values $\hat{y}_i$ and the residuals $e_i = y_i - \hat{y}_i$ on this graph.

(d) Complete an ANOVA table for this linear regression, including the $F$-ratio and corresponding $p$-value.

(e) Calculate the value of the coefficient of determination $r^2$, using the two following equivalent ways (and showing agreement of your answers), and interpret this quantity as a measure of fit of the regression line to the data, in a brief, clear explanation.

- via squaring the correlation coefficient $r = \frac{s_{xy}}{s_x s_y}$ found in (a),
- via the ratio $r^2 = \frac{SS_{\text{Regression}}}{SS_{\text{Total}}}$ of sums of squares found in (d).

(f) Test the null hypothesis of no linear association between $X$ and $Y$, either by using your answer in (a) on $H_0: \rho = 0$, or equivalently, by using your answers in (b) and/or (d) on $H_0: \beta_1 = 0$.

(g) Calculate a point estimate of the mean potency when $X = 6$ months. Judging from the data, is this realistic? Determine a 95% confidence interval for this value.

(h) The FDA recommends that the expiration date should be defined as that time when a drug contains 90% of the labeled potency. Using this definition, calculate the expiration date for this tablet formulation. Judging from the data, is this realistic?

(i) The residual plot of this model shows evidence of a nonlinear trend. (Check this!) In order to obtain a better regression model, first apply the linear transformations $\tilde{X} = X / 3$ and $\tilde{Y} = 510 - Y$, then try fitting an exponential curve $\tilde{Y} = \alpha e^{\beta \tilde{X}}$. Use this model to determine the expiration date. Judging from the data, is this realistic?
(j) Redo this problem using the following R code:

```r
# See help(lm) or help(lsfit), and help(plot.lm) for details.

# Compute Correlation Coefficient and Scatterplot
X <- c(0, 3, 6, 9, 12)
Y <- c(500, 490, 470, 430, 350)
corr(X, Y)
plot(X, Y, xlab = "X = Months", ylab = "Y = Assay (mg)", pch=19)

# Least Squares Fit, Regression Line Plot, ANOVA F-test
regline <- lm(Y ~ X)
summary(regline)
abline(regline, col = "blue")

# Exercise: Why does the p-value of 0.02049 appear twice?

# Estimate Mean Potency at 6 Months
new <- data.frame(X = 6)
predict(regline, new, interval = "confidence")

# Residual Plot
resids <- round(resid(regline), 2)
plot(regline, which = 1, id.n = 5, labels.id = resids, pch=19)

# Log-Transformed Linear Regression
Xtilde <- X / 3
Ytilde = 510 - Y
V <- log(Ytilde)
plot(Xtilde, V, xlab = "Xtilde", ylab = "ln(Ytilde)", pch=19)
regline.transf <- lm(V ~ Xtilde)
summary(regline.transf)
abline(regline.transf, col = "red")

# Plot Transformed Model
coeffs <- coefficients(regline.transf)
scale <- exp(coeffs[1])
shape <- coeffs[2]
Yhat <- function(X)(510 - scale * exp(shape * X / 3))
plot(X, Y, xlab = "X = Months", ylab = "Y = Assay (mg)", pch=19)
curve(Yhat, col = "red", add = TRUE)
```
3. **A Third Transformation.** Suppose that two continuous variables $X$ and $Y$ are negatively correlated via the nonlinear relation $Y = \frac{1}{\alpha X + \beta}$, for some parameters $\alpha$ and $\beta$. This is algebraically equivalent to the relation $\frac{1}{Y} = \alpha X + \beta$, which can then be solved via simple linear regression. Use this **reciprocal transformation** on the data and corresponding scatterplot below, to sketch a new scatterplot, and solve for sample-based estimates of the parameters $\alpha$ and $\beta$. (Hint: Finding the parameter values in this example should be straightforward, and not require any “least squares” regression formulas.) Express the original response $Y$ in terms of $X$.

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>60</td>
<td>30</td>
<td>20</td>
<td>15</td>
<td>12</td>
<td>10</td>
</tr>
</tbody>
</table>

4. For this problem, recall that in simple linear regression, we have the following definitions:

\[
b_1 = \frac{s_{xy}}{s_x^2}, \quad MS_{Err} = \frac{SS_{Err}}{n-2}, \quad r^2 = \frac{SS_{Reg}}{SS_{Tot}} = 1 - \frac{SS_{Err}}{SS_{Tot}}, \quad SS_{Tot} = (n-1)s_y^2, \quad S_{xx} = (n-1)s_x^2.
\]

(a) Formally prove that the $T$-score $= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ for testing the null hypothesis $H_0: \rho = 0$,

is equal to the $T$-score $= \left( \frac{b_1 - \beta_1}{\sqrt{MS_{Err}}} \right) \sqrt{S_{xx}}$ for testing the null hypothesis $H_0: \beta_1 = 0$.

(b) Formally prove that, in **simple** linear regression (where $df_{Reg} = 1$), the square of the $T$-score $= \left( \frac{b_1 - \beta_1}{\sqrt{MS_{Err}}} \right) S_{xx}$ is equal to the $F$-ratio $= \frac{MS_{Reg}}{MS_{Err}}$ for testing the null hypothesis $H_0: \beta_1 = 0$. 

\[
\frac{1}{Y} = 0 1 2 3 4 5
\]
5. In a study of “binge eating” disorders among dieters, the average weights \( Y \) of a group of overweight women of similar ages and lifestyles are measured at the end of every two months \( X \) over an eight month period. The resulting data values, some accompanying summary statistics, and the corresponding scatterplot, are shown below.

<table>
<thead>
<tr>
<th>( X )</th>
<th>0</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>200</td>
<td>190</td>
<td>210</td>
<td>180</td>
<td>220</td>
</tr>
</tbody>
</table>

\( \bar{X} = 4 \quad s_x^2 = 10 \)
\( \bar{Y} = 200 \quad s_y^2 = 250 \)

(a) Compute the **sample covariance** \( s_{xy} \) between the variables \( X \) and \( Y \).

(b) Compute the **sample correlation coefficient** \( r \) between the variables \( X \) and \( Y \). Use it to classify the linear correlation as positive or negative, and as strong, moderate, or weak.

(c) Determine the equation of the **least squares regression line** for these data. Sketch a graph of this line on the scatterplot provided above. **Please label clearly!**

(d) Also calculate the **fitted response values** \( \hat{y}_i \), and plot the **residuals** \( e_i = y_i - \hat{y}_i \), on this same graph. **Please label clearly!**

(e) Calculate the **coefficient of determination** \( r^2 \), and interpret its value in the context of evaluating the fit of this linear model to the sample data. Be as clear as possible.

(f) **Interpretation:** Evaluate the overall adequacy of the linear model to these data, using as much evidence as possible. In particular, refer to at least **two** formal “linear regression” assumptions which may or may not be satisfied here, **and why.**
6. A pharmaceutical company wishes to evaluate the results $Y$ of a new drug assay procedure, performed on $n = 5$ drug samples of different, but known potency $X$. In a perfect error-free assay, the two sets of values would be identical, thus resulting in the ideal calibration line $Y = X$, i.e., $Y = 0 + 1X$. However, experimental variability generates the results shown below, along with some accompanying summary statistics: the sample means, variances, and covariance, respectively.

<table>
<thead>
<tr>
<th>$X$ (mg)</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$ (mg)</td>
<td>32</td>
<td>39</td>
<td>53</td>
<td>65</td>
<td>71</td>
</tr>
</tbody>
</table>

- $\bar{x} = 50 \quad s_x^2 = 250 \quad s_{xy} = 260$
- $\bar{y} = 52 \quad s_y^2 = 275$

(a) Graph these data points $(x_i, y_i)$ in a scatterplot.

(b) Compute the sample correlation coefficient $r$. Use it to determine whether or not $X$ and $Y$ are linearly correlated; if so, classify as positive or negative, and as weak, moderate, or strong.

(c) Determine the equation of the least squares regression line for these data. Sketch a graph of this line on the same set of axes as part (a). Also calculate and plot the fitted response values $\hat{y}_i$ and the residuals $e_i = y_i - \hat{y}_i$, on this same graph.

(d) Using all of this information, complete the following ANOVA table for this simple linear regression model. (Hints: SSTotal and dfTotal can be obtained from $s_y^2$ given above; SSError = “residual sum of squares,” and dfError = $n - 2$.) Show all work.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F-ratio</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(e) Construct a 95% confidence interval for the slope $\beta_1$.

(f) Use the $p$-value in (d) and the 95% confidence interval in (e) to test whether the null hypothesis $H_0: \beta_1 = 0$ can be rejected in favor of the alternative $H_A: \beta_1 \neq 0$, at the $\alpha = .05$ significance level. Interpret your answer: What exactly has been demonstrated about any association that might exist between $X$ and $Y$? Be precise.

(g) Use the 95% confidence interval in (e) to test whether the null hypothesis $H_0: \beta_1 = 1$ can be rejected in favor of the alternative $H_A: \beta_1 \neq 1$, at the $\alpha = .05$ significance level. Interpret your answer in context: What exactly has been demonstrated about the new drug assay procedure? Be precise.

7. Refer to the posted Rcode folder for this problem. Please answer all questions.