Chapter 4  Regression and Correlation

In this chapter we will explore the relationship between two quantitative variables, X and Y. We will consider n ordered pairs of observations (x,y). Both x and y can be observed (observational study) or y can be observed for specific values of x that are selected by the researcher (experiment). Data must display a linear trend on the scatterplot for us to consider fitting a linear equation that will express y in terms of x. We will use the following vocabulary with respect to x and y:

- y – response variable or dependent variable
- x – predictor variable/explanatory variable or independent variable

### Bivariate Random Sampling Model

assumes each pair (x,y) is sampled randomly from the population of all such pairs.

If our scatterplot indicates existence of a linear trend (visually we can see our data points clustering closely to some line), we would be interested in expressing quantitatively the strength of that linear relationship. The numerical measure we will use is called Linear Correlation Coefficient and we will denote it's value by r.

### Linear Correlation Coefficient of X and Y,  \( r \):

\[
 r = \frac{1}{n-1} \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{1}{n-1} \sum z_x z_y,
\]

where \( s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \), \( s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} \) (standard deviations of x-s and y-s respectively)
Linear Correlation coefficient gives information about strength and direction of a linear trend in our data. Values close to 1 or -1 indicate extremely strong linear association between X and Y.

Following diagrams show various degrees of linear correlation.

Inference concerning correlation.

$r =$ sample linear correlation coefficient can be regarded as an estimate of a population correlation $\rho$ (Greek letter rho). In order to regard $r$ as an estimate of $\rho$ it must be reasonable to assume that both $x$ and $y$ were selected at random from the population. When $x$-s are specified by the researcher, $r$ is not an appropriate estimate of $\rho$. 
Example #1. Consider people on the diet program. Let $X=$ # of times per week person eats out, $Y=$ change in weight (in pounds), (before-after) after three months period of dieting. Compute and interpret linear correlation $r$. (Make sure to plot the data first, scatterplot shows reasonably strong negative linear trend)

First we compute means and standard deviations of our $x$ and $y$ values:

$$
\overline{x} = \frac{25}{5} = 5 \quad s_x = \sqrt{\frac{22}{4}} = 2.3452 \quad \overline{y} = \frac{40}{5} = 8 \quad s_y = \sqrt{\frac{106}{4}} = 5.1478
$$

Now we are ready to compute our z-scores and finally $r$.

$$
r = \frac{-3.6446}{4} = -0.91115 \quad \text{This value indicates very strong negative linear relationship between } x \text{ and } y. \quad \text{Negative trend means that } y \text{ tends to decrease with increasing } x.
$$

Properties of $r$:

- $r$ is unit-free
- $-1 \leq r \leq 1$, the closer is $r$ to -1 or 1 the stronger is linear relationship between $x$ and $y$, $r=1$ or -1 indicates a perfect linear relationship, $r$ close to 0 implies weak relationship
- $r$ is symmetric, if $X$ and $Y$ are reversed, value of $r$ remains the same.
- sign of $r$ matches the sign of the slope of the regression line
- $r$ measures strength of a linear relationship (not for curves)
- $r$ is not resistant and will be affected strongly by outliers
- $r$ is only appropriate to measure relationship for numerical variables, not categorical ones
Correlation is not a complete description of two-variable data, should also use means and standard deviations.

Note: $r=0$ does not necessarily mean that there is no relationship between $x$ and $y$, just that there is no linear relationship.

Warning about correlation: strong (positive or negative) correlation does not have to imply causal relationship between $x$ and $y$.

Example: Researchers found that there is a very strong positive correlation between number of storks present an number of babies born. Does it imply that storks cause babies to be born? Of course not, both can be influenced by some other factors.

Observational studies can't prove causation.

Experiments can prove causation.

Fitting the Regression Line to bivariate data.

We have plotted our data to determine if there appear to be a linear relationship. Now we will discuss how to find and interpret the equation of the line that best summarizes the linear relationship between two quantitative variables.

Our line is called Least Squares Regression Line and will have a form: $\hat{y}=b_0+b_1 x$, where $\hat{y}$ is a predicted value of $y$ for given $x$. $b_0=y$-intercept, $b_1$=slope, a rate of change of $y$ with respect to $x$.

The name of the line reflects the fact that our line has smallest possible sum of squared errors (of all lines that can be possibly fitted to the given data).

Errors or residuals are the vertical deviations of observations (data points) from the line:

$$e = y - \hat{y} \quad SSE=\sum e^2 = \sum (y_i-\hat{y}_i)^2$$

This is known as LEAST SQUARES CRITERION.

Formulas for the slope and $y$-intercept of Least-Squares Regression equation:

$$\hat{y}=b_0+b_1 x \quad b_1 = \frac{S_{xy}}{S_{xx}} \quad b_0=\bar{y}-b_1 \bar{x}$$

Slope can also be computed using: $b_1=r \frac{s_y}{s_x}$
Special summations used in Regression and Correlation:

**Definition:**

\[ S_{xx} = \sum (x_i - \bar{x})^2 \]

\[ S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) \]

\[ S_{yy} = \sum (y_i - \bar{y})^2 \]

**Computational**

\[ S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \]

\[ S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \]

\[ S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \]

**In our Example#1:**

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i^2 )</th>
<th>( x_i y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>9</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>49</td>
<td>14</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>36</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>49</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>25</td>
<td>40</td>
<td>147</td>
<td>156</td>
</tr>
</tbody>
</table>

Then special summations are:

\[ S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 156 - \frac{(25 \times 40)}{5} = -44 \text{ and} \]

\[ S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 147 - \frac{(25)^2}{5} = 22 \]

So, \( b_1 = \frac{-44}{22} = -2 \) \( b_0 = 8 - (-2)(5) = 18 \)

Our equation is: \( \hat{y} = 18 - 2x \)

**Note:** We can also use the second formula for slope:

\( x = \frac{25}{5} = 5 \) \( y = \frac{40}{5} = 8 \)

\( b_1 = \frac{-0.91115 \times 5.1478}{2.3452} = -2 \) \( b_0 = 8 - (-2)(5) = 18 \)

Equation tells us that for every 1 time increase in \( x \) (times to eat out in a week), \( y \) (weight change) decreases by 2 pounds. The more often you go out to eat the less pounds are lost.

Units of the slope are: \( \text{(units of } y \text{/ (units of } x) \)
Notes about the regression line:

- \((\bar{x}, \bar{y})\) lies on the fitted line
- we can use the regression line to predict values of y for given values of x.
- Interpreting the slope: If x increases by 1 unit, y increases (or decreases) by \(b_1\) units (direction of change depends on the sign of \(b_1\).)
- Regression line is based on the assumption that data points are scattered about the line. If data are scattered about the curve, or do not show any trend at all, avoid the linear regression.
- Linear regression can be strongly affected by outliers. Always plot data first.

Residual plot – a scatterplot of the regression residuals against the explanatory variable.

- It helps us to assess the fit of the regression line.
- If plot unstructured and centered about 0, no major problem
- If plot has a curve then a straight line is not the best fit of the data
- If the residuals get bigger as you go from left to right, predictions are more precise on the left than on the right.

Making predictions using our equation:

In Example #1 Predict number of pounds lost if dieting person eats out 5 times per week: \(y=18-10=8\) lb (this prediction is for x within data range)

Warning about making predictions:
It is safe to predict y for x within data range.

Extrapolation - making predictions for values of the predictor variable outside the range of the observed values of the predictor variable (x)
\(\Rightarrow\) Grossly incorrect predictions can result from extrapolation.

In Example #1. Prediction for \(x=15\) is \(y=-12\) (weight gain of 12 lb). One may think that value is reasonable, weight gain is to be expected if you eat out so often, but because we have no data in that range, we may not be sure that our particular trend will hold as far out of the range of x values.

Problems we may encounter with our data:

Outlier – a data point that lies far from the regression line, relative to other data points

Influential observation – a data point whose removal causes the regression to change considerably. It is usually separated in the x-direction from the other data points. It pulls the regression line towards itself.
Important Sums of Squares:

There are three summations that are important in assessment of the fit of our regression line.

Total Sum of Squares: \( SST = \sum (y_i - \bar{y})^2 \) gives total variability in y values

Residual (Error) Sum of Squares: \( SSE = \sum (y_i - \hat{y}_i)^2 \) gives total variability in y values unexplained by the regression equation

Regression (Model) Sum of Squares: \( SSR = \sum (\hat{y}_i - \bar{y})^2 \) gives total variability in y values explained by the regression equation

Regression Identity: \( SST = SSR + SSE \)

In our example \( SST \) was already computed.

We can now obtain residual sum of squares, and regression (model) sum of squares by obtaining predicted values of y for each x, by using our equation.

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( y_i )</th>
<th>((y_i - \bar{y})^2)</th>
<th>( \hat{y}_i )</th>
<th>( y_i - \hat{y}_i )</th>
<th>((y_i - \hat{y}_i)^2)</th>
<th>((\hat{y}_i - \bar{y})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>4</td>
<td>12</td>
<td>-2</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>36</td>
<td>4</td>
<td>-2</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>16</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>49</td>
<td>14</td>
<td>1</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>25</td>
<td>40</td>
<td>106</td>
<td>18</td>
<td></td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>\text{SST}</td>
<td>\text{SSE}</td>
<td>\text{SSR}</td>
<td>\text{SSE}</td>
<td>\text{SSR}</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We can verify that in our example: \( 106 = 88 + 18 \)

We will use the three sums of squares in assessment of the fit of our regression line.

Assessing the fit of our regression line:

Coefficient of Determination: \( r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \) (square of linear corr. coef. r)

This measure gives us proportion of total variability in y values that is explained by our regression line. If this is close to 1 (100%), then that indicates a good fit. \( 0 \leq r^2 \leq 1 \)

In our Example#1 \( r^2 = 1 - \frac{18}{106} = \frac{88}{106} = 0.83 \), it indicates a good fit, 83% of total variability in y-values is explained by the regression line.
Example#2.

<table>
<thead>
<tr>
<th>Age (yr): $x$</th>
<th>5</th>
<th>4</th>
<th>6</th>
<th>5</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>2</th>
<th>7</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price ($100): $y$</td>
<td>85</td>
<td>103</td>
<td>70</td>
<td>82</td>
<td>89</td>
<td>98</td>
<td>66</td>
<td>95</td>
<td>169</td>
<td>70</td>
</tr>
</tbody>
</table>

Plot $x$ against $x$, if the points seem to follow a straight line, then a straight line can be used to approximate the relationship between $x$ and $y$.

**Regression equation** – The equation of the regression line: $\hat{y} = 195.47 - 20.26x$

Following picture illustrates how we compute coefficient of determination $r^2$ and how we interpret it's value.
\[ r^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{8285.0}{9708.5} = 0.853 \ (85.3\%) \]

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>( \hat{y} )</th>
<th>( y - \bar{y} )</th>
<th>( \hat{y} - \bar{y} )</th>
<th>( y - \hat{y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>85</td>
<td>94.16</td>
<td>-3.64</td>
<td>5.53</td>
<td>-9.16</td>
</tr>
<tr>
<td>4</td>
<td>103</td>
<td>114.42</td>
<td>14.36</td>
<td>25.79</td>
<td>-11.42</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>73.90</td>
<td>-18.64</td>
<td>-14.74</td>
<td>-3.90</td>
</tr>
<tr>
<td>5</td>
<td>82</td>
<td>94.16</td>
<td>-6.64</td>
<td>5.53</td>
<td>-12.16</td>
</tr>
<tr>
<td>5</td>
<td>89</td>
<td>94.16</td>
<td>0.36</td>
<td>5.53</td>
<td>-5.16</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>94.16</td>
<td>9.36</td>
<td>5.53</td>
<td>3.84</td>
</tr>
<tr>
<td>6</td>
<td>66</td>
<td>73.90</td>
<td>-22.64</td>
<td>-14.74</td>
<td>-7.90</td>
</tr>
<tr>
<td>6</td>
<td>95</td>
<td>73.90</td>
<td>6.36</td>
<td>-14.74</td>
<td>21.10</td>
</tr>
<tr>
<td>2</td>
<td>169</td>
<td>154.95</td>
<td>80.36</td>
<td>66.31</td>
<td>14.05</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>53.64</td>
<td>-18.64</td>
<td>-35.00</td>
<td>16.36</td>
</tr>
<tr>
<td>7</td>
<td>48</td>
<td>53.64</td>
<td>-40.64</td>
<td>-35.00</td>
<td>-5.64</td>
</tr>
</tbody>
</table>

\[ \sum (\hat{y} - \bar{y})^2 = 8285.0 \land \sum (y - \bar{y})^2 = 9708.5 \]

**CALCULATOR**

1) To use above computational formulas and/or find basic statistics for x and y
   use STAT EDIT place x-s and y-s on L1 and L2 respectively, then
   STAT CALC , select 2-Var Stats <enter>
   2-Var Stats L1, L2 <enter>

2) To find the regression line: STAT EDIT place x-s and y-s on L1 and L2 respectively
   STAT CALC , select LinReg(ax+b) <enter>
   LinReg(ax+b) L1, L2 <enter>
Output should display slope, y intercept, r and r$^2$. If you do not see r and r$^2$, then go to the CATALOG, select **DIAGNOSTICS ON** <enter><enter>, next time you run the regression it will show both values.