Handout and Project on Correlation and Regression (Unit 11)

This material is self contained. It is an introduction to regression that will help you in MSC 317 where you will study the subject in more detail. You may find portions of chapter 10, particularly figures 10.1, 10.7, 10.16 helpful, but it is not necessary for you to read chapter 10.

I. Population correlation $\rho$ and sample correlation $r$.

**Correlation** attempts to measure (by means of a single number) the "straight line dependence" between scores $x, y$. Correlation is always a number in the interval range $[-1, +1]$. For a generally upward sloping plot of $(x, y)$ values correlation is positive. For a downward sloping plot correlation is negative. Correlation $= +1$ if and only if the plot is a perfect upward sloping line. Correlation $= -1$ if and only if the scatterplot is a perfect downward sloping line. The population correlation $\rho$ (pronounced "rho") and sample correlation $r$ are defined as follows:

- **population correlation**
  \[
  \rho = \frac{(E XY - EX EY)}{(\sigma_x \sigma_y)}
  \]

- **sample correlation (calculating formula)**
  \[
  r = \frac{(\sum xy - \sum x \sum y / n)}{(n-1) s_x s_y}
  \]

*Correlation is not changed by a location or (positive) scale change in either of $x, y$ or both.* For example, correlation between
- $x =$ Monday noon temperature
- $y =$ Tuesday noon temperature
(over several weeks say) is not changed by converting either Monday (or Tuesday or both) from Fahrenheit to Celsius.

Here are example plots and their correlations (downward sloping counterparts of these plots have the same correlations, only negative):
Here is a correlation calculation for a sample of \( n = 3 \) pairs \((x, y)\):

\[
x \quad y \quad x^2 \quad y^2 \quad x \ y \\
-2 \quad 3 \quad 4 \quad 9 \quad -6 \\
0 \quad 5 \quad 0 \quad 25 \quad 0 \\
4 \quad 0 \quad 16 \quad 0 \quad 0 \\
\text{total} \quad 2 \quad 8 \quad 20 \quad 34 \quad -6
\]

\[
r = \frac{(\Sigma xy - \Sigma x \Sigma y / n)}{(n-1) s_x s_y} \\
= \frac{(-6 - (2)(8) / 3)}{[(3 - 1) \sqrt{(20 - 2^2 / 3) / 2}] \sqrt{(34 - 8^2 / 3) / 2}] \\
= -0.737043
\]
II. Example Population. Class of 600 students of a large lecture course.
\[ x = \text{score on pre-midterm} \]
\[ y = \text{score on midterm}. \]

Example Population statistics.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_x$</td>
<td>67</td>
</tr>
<tr>
<td>$\mu_y$</td>
<td>73</td>
</tr>
<tr>
<td>$\sigma_x$</td>
<td>10.4</td>
</tr>
<tr>
<td>$\sigma_y$</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Population correlation $\rho = 0.79$

III. Population scatterplot and regression line of $y$ on $x$. Here is a scatterplot of the population $(x, y)$ scores with the population regression line drawn in. The population regression line of $y$ on $x$ is the line passing through the point $x = 67, y = 73$ (the point of means) and having the slope $\rho \frac{\sigma_y}{\sigma_x} = 0.66$, where $\rho$ (pronounced "rho") is the population correlation.

Another name for the regression line is the least squares line since it is the unique line that minimizes the sum of the squares of the vertical discrepancies, called residuals, between the plot and line.
As is often the case with other population statistics, we typically do not get to see the population scatterplot or population regression line unless we perform a census of the population. Therefore we turn to a sample.

**IV. Sample statistics.** For illustration, we have selected an equal-probability and with-replacement sample of only \( n = 60 \) from the example population above. For the sample we find

\[
\begin{align*}
\bar{x} &= 68.88 & \bar{y} &= 73.19 \\
\sigma_x &= 9.92 & \sigma_y &= 6.96 \\
\text{sample correlation} r &= 0.78
\end{align*}
\]

**V. Sample scatterplot and sample regression line of \( y \) on \( x \).** Here is a scatterplot of the sample \((x, y)\) scores with the sample regression line drawn in. The sample regression line of \( y \) on \( x \) is the line passing through the point \( \bar{x} = 68.88, \bar{y} = 73.19 \) (the point of sample means) and having the slope \( r \frac{\sigma_{\hat{y}}}{\sigma_{\hat{x}}} = 0.78 \frac{(6.96)}{9.92} = 0.547 \) where \( r \) is the sample correlation.
Here are the population and sample regression lines drawn together. The sample line happens to be the more horizontal of the two since the slope of the sample regression line $r_{sY}/s_x = 0.547$ is closer to zero than the slope of the population regression line $\rho \sigma_y / \sigma_x = 0.66$.

![Graph showing population and sample regression lines](image)

VI. Galton’s observations on scatterplots, "naive" line, regression line.

To continue with the example, say you score one s.d. above the average on the pre-midterm. Then it is only natural that you expect to score around one s.d. above average on the midterm. Galton observed that many scatterplots do not support this reasoning. He found, for the elliptically contoured (i.e. football shaped) scatterplots often seen, that the vertical strip averages tend to fall on the regression line, which has a more horizontal slope than the naive line.

Galton termed this the regression effect (regression toward mediocrity). In practical terms, as a group, those who score one $\sigma_x$ above the mean $\mu_x$ on the pre-midterm tend on average to score only around $\rho \sigma_y$ above the mean $\mu_y$ on the midterm. Likewise, those who score one $\sigma_x$ below the mean $\mu_x$ on the pre-midterm tend on average to score only around
\( \rho \sigma_y \) below the mean \( \mu_y \) on the midterm. There is a tendency to fall toward the mean \( \mu_y \). This applies equally to those falling any amount \( \# \sigma_x \) above \( \mu_x \) who tend to fall around \( \rho \# \sigma_y \) above \( \mu_y \). Here are some illustrations from the book *Statistics* by Freedman, Pisani and Purves, Norton Publ.

*Figure 7. The regression effect. The scatter diagram shows the heights of 1,078 fathers and sons. The sons averaged 1 inch taller than the fathers. If the son was 1 inch taller than his father, this family is plotted along the dashed line. The points in the strip over 72 inches correspond to the families where the father was 72 inches tall, to the nearest inch. Most of these points are below the dashed line, and the average height of these sons is only 71 inches. The points in the strip over 64 inches correspond to families where the father was 64 inches tall, to the nearest inch. Most of these points are above the dashed line, and the average height of these sons is 67 inches. The solid regression line picks off the centers of all the vertical strips, and is much flatter than the dashed line.*
Figure 8. The regression effect. The SD line is dashed, the regression line is solid. The dots show the average height of the sons, for each value of father's height. They rise much less steeply than the SD line. This is the regression effect. The regression line follows the dots very well indeed.

Figure 2. Regression estimate. When the independent variable goes up by an SD, the average value of the dependent variable only goes up by $r$ SDs.
VII. Utilizing the sample regression line for the purpose of improving upon yBAR as an estimator of \( \mu_y \).

In MSC 317 you will see many different uses of regression. We now show you one of these. The sample regression line can be regarded as our guess at the population regression line. Since the point \( x = \mu_x, y = \mu_y \) lies on the population regression line, it might make sense to estimate \( \mu_y \) by inserting the value of \( \mu_x \) (if it is known!) into the sample regression line. This will work well if the lines are close. That is, estimate unknown \( \mu_y \) by:

\[
\hat{\mu}_{y, \text{regr}} = \bar{y} + r \left( \frac{s_y}{s_x} \right) (\mu_x - \bar{x})
\]

provided the population xmean \( \mu_x \) is known! Notice that the estimator above is the same as:

\[
\hat{\mu}_{y, \text{regr}} = \bar{y} + r \left( \frac{\sigma_{\text{HAT}, y}}{\sigma_{\text{HAT}, x}} \right) (\mu_x - \bar{x})
\]

since whether we use n-1 divisor or n divisor affects both s.d. equally.

Calculating the regression estimator for our example. For our Example population, we have the average pre-midterm score of \( \mu_x = 67 \) for the entire class of 600. Perhaps we know this number because it is in the computer since the pre-midterm has already been entirely graded. But we only have y scores (midterm scores) from the 60 sample people (of course we also have their x scores too). Then instead of estimating \( \mu_y = 73.19 \) (the yBAR of the sample of 60) we could use the sample regression based estimator:

\[
\hat{\mu}_{y, \text{regr}} = \bar{y} + r \left( \frac{s_y}{s_x} \right) (\mu_x - \bar{x})
\]

\[
= 73.19 + 0.78 \left( \frac{6.96}{9.92} \right) (67 - 68.8) = 72.20
\]

How well did the regression estimator do compared with just using yBAR? The true value of the population \( \mu_y \) is actually 73. So in this case (for this sample of 60) we would have been better off using \( \bar{y} = 73.19 \) than the fancy regression estimator which turned out in the end to be 72.20, even further from the true value of 73. Of course from merely looking at the sample we would not have known this had happened to us since the true value of 73 would not be known until we grade all of the midterms.

How well does the regression estimator do in general? In general, the
regression based estimator $\mu_{HAT, y, regr}$ will improve on $yBAR$ in a statistical sense. $yBAR$ is unbiased and $\mu_{HAT, y, regr}$ nearly unbiased

$$E \ yBAR = \mu_y \quad E \ \mu_{HAT, y, regr} \approx \mu_y.$$ 

But $\mu_{HAT, y, regr}$ has a smaller s.d. by the approximate factor

$$\sqrt{(1 - \rho^2)} = \sqrt{(1 - 0.79^2)} = \sqrt{0.3759} = 0.613.$$

In practice the 95% C.I. for $\mu_y$ based on $\mu_{HAT, y, regr}$ cannot use $\rho$, since it is not known, but uses $r$ instead:

$$\mu_{hat, y, regr} \pm (\sqrt{(1 - r^2)} \ 1.96 \ \frac{s_y}{\sqrt{n}}$$

which is around 0.63 as wide as for $yBAR$ since $\sqrt{(1 - r^2)} = \sqrt{(1 - 0.78^2)} = 0.63$. To gain equivalent precision using C. I. $yBAR \pm 1.96 \ \frac{s_y}{\sqrt{n}}$ would require a larger sample of $n = 60 / (0.63)^2 = 151$. Gaining this much additional statistical precision using regression can result in big cost savings!

Our regression estimator of $\mu_y$ just lost out when we compared it with $yBAR$ above. Let’s try it all again to see what happens with a different sample of 60.

$$xBAR = 69.17 \quad s_x = 9.17$$
$$yBAR = 74.85 \quad s_y = 7.71$$
$$r = 0.72$$
\[ \mu_{\text{hat, regr}} = y\text{BAR} + r \left( \frac{s_y}{s_x} \right) (\mu_x - x\text{BAR}) \]
\[ = 74.85 + 0.72 \left( \frac{7.71}{9.17} \right) (67 - 69.17) = 73.54 \]

For this (second) sample of n = 60 the regression estimator 73.54 has indeed come closer to the correct population value \( \mu_y = 73 \) than has yBAR = 74.85. Sometimes you win, sometimes you lose, but mostly you will improve matters by using the regression estimator provided the true correlation is near \( \pm 1 \) (so that \( \sqrt{1 - r^2} \) tends to be near zero).
STT 315 Project 7.

Hand in your written solution to this project at the end of recitation Thursday, April 23, 1998.

Be sure to also complete and submit at that time the bubble sheet whose answers will be directly drawn from this assignment.

The bubble will be circulated in recitation.

0. Cartoon an elliptical scatterplot and the least squares line for that plot. Illustrate the "regression effect."
An equal probability with-replacement sample of 100 properties is selected from the tax rolls of a city. For each of these 100 sample properties the score $x = \text{1997 tax paid}$ can be gotten directly from the tax rolls. However the 1998 taxes have not yet been filed. The city pays $200 each to audit the score $y = \text{1998 tax}$ that will be due for each of these 100 sample properties. From the 100 data pairs $(x, y) = (\text{1997 tax, 1998 tax})$ the city finds (in thousands of dollars)

$$
\begin{align*}
\text{xBAR} &= 1.57 & \hat{\sigma}_x &= 2.44 \\
\text{yBAR} &= 1.88 & \hat{\sigma}_y &= 2.61 \\
\end{align*}
$$

$$
r = 0.86
$$

1. What would you estimate to be the population mean $\mu_x$ of all 1997 property taxes in the city?

2. What would you estimate to be the population mean $\mu_y$ of all 1998 property taxes in the city?

3. Ignoring the $x = 1997$ data altogether, give a 0.95 confidence interval for the unknown population mean $\mu_y$ of all 1998 property taxes in the city.

4. 
   a. How much did the sample leading to (3) cost the city?

   b. How much would it cost the city in additional sampling charges to cut the width of the confidence interval approximately in half by taking a larger sample?
c. Had the sample been without replacement we could have used the FPC and enjoyed a narrower confidence interval. Around what fraction of the population must a sample represent in order for the FPC to cut the width of the CI in half?

5. Since the city knows the total property tax it collected in 1997, and also knows the number of properties, it will also know the mean 1997 property tax $\mu_x$. Suppose it knows that $\mu_x$ is really equal to 1.63 (not the 1.57 shown by the sample).

    a. Plot the sample regression line, illustrating how the estimator $\mu_{\text{HAT}_{\text{regr}}}$ can be read off as the y-score which results from inserting $x = \mu_x$ (known!) into the sample regression line.

    b. Calculate the regression estimator $\mu_{\text{HAT}_{\text{regr}}}$, of the 1998 population average tax $\mu_y$, according to the formula

    $$\mu_{\text{HAT}_{\text{regr}}} = \bar{y} + r \left( \frac{\sigma_{\text{HAT}_y}}{\sigma_{\text{HAT}_x}} \right) (\mu_x - \bar{x}).$$
6. Give the 0.95 confidence interval for $\mu_y$ based upon the estimator $\mu_{HAT_{regr}}$ instead of $y_{BAR}$.

7. Refer to questions (3) (6). In view of the relative widths of the two confidence intervals, how much cost savings does the CI (6) represent over the CI (3)?

8. For the data below, calculate

\[
\begin{align*}
    x_{BAR} &= \\
    \sigma_{HAT_x} &= \\
    y_{BAR} &= \\
    \sigma_{HAT_y} &= \\
    r &= \\
\end{align*}
\]

\[
\begin{array}{cc}
    x & y \\
    2 & 7 \\
    1 & 8 \\
    4 & 7 \\
    9 & 14 \\
\end{array}
\]

9. For the data of exercise (8) plot the sample regression line. Clearly indicate the point ($x_{BAR}$, $y_{BAR}$) on this line and draw the little triangle which gives the slope of the line. Also include the 4 data points in your plot.