STAT 5302  Applied Regression Analysis

Sample solution to Homework 2

Do-by-hand portion:

4.1 a. True. The bivariate normal implies constant conditional variance

b. True. If the regression function is a constant and the variance is constant (as implied by the information that the data are bivariate normal) then \( x \) and \( y \) are independent.

c. False. We predict \( x = y/2 \), but we predict \( y = x/2 \). This is the ‘two regression’ paradox that the regression of \( y \) on \( x \) uses a different line that the regression of \( x \) on \( y \).

d. True. \( \text{Var}(x+y) = \text{Var}(x) + \text{Var}(y) + 2 \text{Covar}(x,y) \) and Covar has the same sign as \( \rho \).

e. False. See answer to part d.

f. True. See answer to part d.

Lab portion

3.1.1 Here is the scatter plot, with \textit{lowess} smooth and 45° reference line. If I compare this plot with Figure 3.10 of the book (which deals with Observer 2), I see that this observer pretty consistently underestimates the size of the flock. For Observer 2, the \textit{lowess} smooth and the 45° line interlaced more.

3.1.2 If we compare this figure with 3.11 of the book, it looks as though Observer 2 scores on variance as well as mean, with a generally tighter pair of SD bands.

3.1.3 The plot of \textit{obs1 vs h_obs1} captures the curvature in the relationship between \textit{photo} and \textit{Obs1}. That is rises on the right is an indication that the relationship is not linear; it is more or less linear in the early part of the range, but the steep rise on the right corresponds to the visual flattening in the lowess plot. The graph of \textit{photo vs h_obs1} should center on a straight line of slope 1. It does reasonably seem to be doing that. We notice that the spread of the points widens a lot to the right.
3.2 We start with a lowess using a small value of the smoothing constant (0.2). Visually, this is rather rough, and we go to a plot with the constant set at 0.6, which seems to do the trick quite well. Values not quite as big—for example 0.4 or 0.5 though are also highly defensible.

The relationship looks quite nasty. At low temperatures, the ozone level decreases quite slowly with temperature decrease, but around 60 degrees the relationship seems to get a whole lot steeper.

The variance function is not entirely obvious. It looks to be very narrow over at low temperatures, but this can be expected on the basis that if a necessarily non-negative number is centered around zero, it does not have any room to vary a lot. Once we get to some 50 degrees, the variance function looks to be more stable.

One of the tricks that often works on variables like these ozone measurements is a transformation. To illustrate this, try a log transformation. Following standard practice in the book, we will make the logs to base 2, though this does not affect anything.
fundamental. The corresponding plot and lowess smooth seem to have traded one set of unusual features for another; the lowess smooth is quite wavy, and the variance function looks to be narrow at the top where it used to narrow at the bottom. Maybe this means it is time to quit on this relationship.

4.4 We are told that \((C, \text{Over})\) is bivariate normal. So to estimate these two mean functions, we need to estimate the means, variances and covariances of the two variables. Arc gives us the summary numbers

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Average</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Median</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>120</td>
<td>69.09</td>
<td>5.1274</td>
<td>58.</td>
<td>69.</td>
<td>81.</td>
</tr>
</tbody>
</table>

Data set = Haystacks, Sample Correlations

\[
\begin{array}{ccc}
C & 1.0000 & 0.6386 \\
Over & 0.6386 & 1.0000 \\
C & Over \\
\end{array}
\]

4.4.1 So we estimate that

\[
\begin{align*}
E(C|\text{Over}) &= 69.09 + 0.6386 \times 5.1271 \times (\text{Over} - 36.12) / 3.8412 = 38.3 + 0.85 \times \text{Over} \\
E(\text{Over}|C) &= 36.12 + 0.6383 \times 3.8412 \times (C - 69.06) / 5.1271 = 3.07 + 0.48 \times C
\end{align*}
\]

4.4.2 Doesn’t look like it to me; that coefficient of 0.85 looks quite a distances from 2, and the intercept of 38.3 looks quite a distance from 0. To make this more formal though, we’d need to do some formal testing.

Curiously (in the face of this), it does look as though the Over measurement is close to one-half the circumference. That this does not translate into ‘circumference=2*Over’ is a manifestation of the two-regression situation that the regressions of \(x\) on \(y\) and of \(y\) on \(x\) are totally different lines.

4.4.3 The correlation is 0.6386, so \(E(\text{Over}|C)\) is 0.6385*1.5 = 0.96 standard deviations above \(E(\text{Over})\).