1) Introduction to Conditional Distributions

- Correlation measures the degree of association between two variables. Suppose however, that we want to predict one variable based on known values of some other variable. Simple linear regression is one technique to do so. In regression, let $Y$ denote the response (also called dependent) variable and $X$ denote the explanatory (or independent) variable. Our objective is to predict $Y$ based on $X$.

- For example, Figures 18.1 and 18.2 show that head circumference of children appears to increase linearly with the age of children from 2 to 18 yrs.

- As a specific example, the book considers the head circumference, say $Y$, at birth of low birth weight infants born in Boston. Suppose $Y$ is approximately normal, with $\mu_y = 27$ and $\sigma_y = 2.5$, i.e.

$$y \sim N(27, 2.5)$$

It follows that approximately 95% of such newborn head circumference would be

$$\mu \pm 1.96\sigma$$

$$= 27 \pm 1.96(2.5)$$

$$= (22.1, 31.9)$$

Is this a confidence interval?

- Suppose it is known that the head size tends to increase with gestational age. Consider notation $\mu_{y|x}$ and $\sigma_{y|x}$ for the mean and standard deviation of $y$ for given $x$. For example, suppose that for $x = 26$, one has

$$\mu_{y|26} = 24$$

and that

$$\sigma_{y|26} = 1.6$$

This would imply that for the population with $x = 26$, the mean and standard deviation would be lower than for the whole population. Is this reasonable? For $x = 26$, approximately 95% of head sizes are in the interval $(20.9, 27.1)$.

- We could assume similar values for other $x$. For example, for $x = 29$, one might have

$$y \sim N(26.5, 1.6)$$

and for $x = 32$,

$$y \sim N(29, 1.6)$$

Note, we are assuming that $\sigma_{y|x}$ is the same for all $x$. Is this reasonable? One could show in general that
\[ \sigma_{y|x}^2 = (1 - \rho^2)\sigma_y^2 \]

where \( \rho \) is the correlation coefficient.

2) Simple linear regression model

2.1 Regression assumptions

• Recall in algebra, there were simple math models. One such is the straight line

\[ y = a + bx. \]  

(1)

This model has two coefficients, namely \( a \), the intercept, and \( b \), the slope. In this simple (deterministic) model, \( y \) is totally determined from \( x \).

• Real life is usually more complicated, with uncertainties. This leads to the probability model

\[ y = \alpha + \beta x + \varepsilon, \]  

(2)

where \( \alpha \) and \( \beta \), the intercept and slope, are unknown parameters. The term \( \varepsilon \) is called the random error term. It represents all the uncertainties in predicting \( y \) from \( x \). The usual assumption is that

\[ \varepsilon \sim IN(0, \sigma). \]  

(3)

This means that the error terms have a normal distribution. Their mean and standard deviation, at any \( x \), are 0 and \( \sigma \), respectively. Based on (2), \( y \) has two parts, one part predicted from \( x \) and the other a random error, i.e.

\[ y = \text{fit (based on } x) + \text{error} \]

• One way to quantify the first part is to assume that the means of \( y \) for given \( x \)'s lie on a straight line, i.e.

\[ \mu_{y|x} = \alpha + \beta x. \]  

(4)

Such a relationship is called the population regression line. One such relationship where \( \mu_{y|x} \) is the conditional mean head circumference and \( x \) is gestational age is illustrated in Figure 18.3.

• The complete set of assumptions for our regression analysis are:

1) At any specified \( x \), the values of \( y \) have a normal distribution with mean \( \mu_{y|x} \) and standard deviation \( \sigma_{y|x} \).

2) The means \( \mu_{y|x} \) lie on the straight line

\[ \mu_{y|x} = \alpha + \beta x \]

3) The standard deviation \( \sigma_{y|x} \) is the same for all \( x \).

4) The \( y \) observations are independent.

This model is illustrated in Figure 18.4.
2.2 Least squares regression analysis

- As an illustration, consider the following data, where the wing length, $y$, (in cm) were measured for $n = 13$ sparrows of different age, $x$, in days. The data are:

<table>
<thead>
<tr>
<th>Obs</th>
<th>$x_i$</th>
<th>$y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2.2</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>2.4</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>3.1</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>3.2</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>3.2</td>
</tr>
<tr>
<td>8</td>
<td>11</td>
<td>3.9</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>4.1</td>
</tr>
<tr>
<td>10</td>
<td>14</td>
<td>4.7</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>4.5</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>5.2</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Is it reasonable to assume that the observations are independent? The data are plotted, and a simple linear regression model seems plausible. The objective is to predict $y$ based on $x$.

- Let $\hat{y}$ denote a predicted value, and let the best fitting line be denoted

$\hat{y} = \hat{\alpha} + \hat{\beta}x$ \hspace{1cm} (5)

Let $e_i$ denote the $i$th residual, i.e.

$e_i = y_i - \hat{y}_i$. \hspace{1cm} (6)

- The error sum of squares, also called residual sum of squares, is

$\sum e_i^2$. \hspace{1cm} (7)

The least squares estimates, $\hat{\alpha}$ and $\hat{\beta}$, in (5) are the values which make the error sum of squares as small as possible. Is this a reasonable thing to do? Why or why not?

- If we substitute (5) and (6) into (7), one has

$\sum(y_i - \hat{\alpha} - \hat{\beta}x_i)^2$

Using calculus, one can show that (7) is minimized for

$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
For our example, note

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
x_i & y_i & x_i - \bar{x} & y_i - \bar{y} & (x_i - \bar{x})^2 & (x_i - \bar{x})(y_i - \bar{y}) & (y_i - \bar{y})^2 \\
\hline
3 & 1.4 & -7 & -2.02 & 49 & 14.14 & 4.0804 \\
4 & 1.5 & -6 & -1.92 & 36 & 11.52 & 3.6864 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
17 & 5.0 & 7 & 1.58 & 49 & 11.06 & 2.4964 \\
\hline
\text{Total} & 130 & 44.4 & 0 & 0 & 262 & 70.80 & 19.6569 \\
\text{Mean} & 10 & 3.42 & & & & & \\
\hline
\end{array}
\]

Therefore, with \( \bar{x} = 10 \) and \( \bar{y} = 3.42 \)

\[
\hat{\beta} = \frac{70.80}{262} = 0.270 \text{ cm/day}
\]

\[
\hat{\alpha} = 3.42 - 0.270(10) = 0.72 \text{ cm.}
\]

The estimated linear regression equation in (5) is

\[
\hat{y} = 0.72 + 0.27 \hat{x}
\]

How would you interpret this?

2.3 Inferences for regression coefficients

- Our assumed regression model has three parameters, \( \alpha, \beta \) and \( \sigma_{y|x} \). The best estimates of the first two are \( \hat{\alpha} \) and \( \hat{\beta} \).

- To estimate \( \sigma_{y|x} \), first find mean squared error, \( MSE \), as

\[
MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2}
\]

This is difficult computationally, hence an equivalent formula is

\[
MSE = \frac{\sum(y_i - \bar{y})^2 - \hat{\beta} \sum(x_i - \bar{x})(y_i - \bar{y})}{n - 2}
\]

Then the standard deviation from regression, \( s_{y|x} \) which estimates \( \sigma_{y|x} \) is

\[
s_{y|x} = \sqrt{MSE}
\]

- For example, for our data

\[
MSE = \frac{19.6569 - 0.27(70.80)}{11} = 0.0477 \text{ (approx.)}
\]

\[
s_{y|x} = \sqrt{0.0477} = 0.22 \text{ cm}
\]

This estimates the variability of the \( y \) values at a given \( x \).

- To estimate the variability of the \( \hat{\alpha} \) and \( \hat{\beta} \) estimates, one has
\[
\hat{s}_\beta = \frac{s_{y|x}}{\sqrt{\sum(x_i - \bar{x})^2}} \\
\hat{s}_{\alpha} = s_{y|x} \sqrt{\frac{1}{n} \frac{x^2}{\sum(x_i - \bar{x})^2}}
\]

For example, for our data
\[
\hat{s}_\beta = 0.22 / \sqrt{262} = 0.0135 \text{ cm/day} \\
\hat{s}_{\alpha} = 0.22 \sqrt{\frac{1}{13} \frac{10^2}{262}} = 0.1490 \text{ cm}
\]

- 100(1 - \alpha)% confidence intervals for \(\beta\) and \(\alpha\) are
\[
\hat{\beta} \pm t_{\alpha/2} \hat{s}_\beta \\
\hat{\alpha} \pm t_{\alpha/2} \hat{s}_{\alpha}
\]

where the \(t\) has the \(n - 2\) df.

For example, 95% confidence intervals, with \(t_{11,025} = 2.201\) are
\[
0.270 \pm 2.201 (0.0135) = (0.24, 0.30) \\
0.72 \pm 2.201 (0.1490) = (.39, 1.05)
\]

How would you interpret these?

- In many problems, it is important to test

\[H_0 : \beta = 0,\]

which would imply that there is no linear relationship between \(x\) and \(y\). This has test statistic, with \(n - 2\) df, of
\[
t = \frac{\hat{\beta} - \beta_0}{\hat{s}_\beta}
\]

where \(\beta_0\) is the hypothesized value. For example, to test
\[H_0 : \beta = 0 \quad vs \quad H_A : \beta > 0\]

the \(RR\) is \(t > 1.796\).

Our test statistic is
\[
t = \frac{0.270}{0.0135} = 20.
\]

Hence we have strong statistical evidence that \(\beta > 0\), as one would expect. This is also clear from the confidence interval.
2.4 Inferences for conditional mean, $\mu_{y|x}$

- By assumption, our conditional means lie on the straight line
  
  $$\mu_{y|x} = \alpha + \beta x.$$  

  The best estimate of this mean for given $x$ is
  
  $$\hat{y} = \hat{\alpha} + \hat{\beta} x.$$  

  (8)

  The standard error of $\hat{y}$ is
  
  $$s_{\hat{y}} = s_{\hat{y|x}} \sqrt{\frac{1}{n} + \frac{(x-x)^2}{\Sigma (x-x)^2}}.$$  

  (9)

- Note that $s_{\hat{y}}$ varies with $x$. It is obviously smallest at $x = \bar{x}$, and increases the further away the given $x$ is from $\bar{x}$.

- Confidence intervals under our assumptions are
  
  $$\hat{y} \pm t_{n-2}s_{\hat{y}}$$

  For example, for $x = 7.0$, using (8),
  
  $$\hat{y} = 0.72 + 0.270(7) = 2.16 \text{ cm}$$

  and from (9),
  
  $$s_{\hat{y}} = 0.22 \sqrt{\frac{1}{13} + \frac{(7-10)^2}{262}} = 0.073 \text{ cm}$$

  Hence a 95% confidence interval for $\mu_{y|x=7}$, using $t$ with 11 df, i.e. $t_{0.025} = 2.201$, is
  
  $$2.16 \pm 2.201(0.073)$$

  $$= 2.16 \pm 0.16$$

  $$= (2.00, 2.32)$$

- We could construct confidence bands by calculating confidence intervals at various $x$'s and then graphing the results. For example, note that a 95% ci for $x = 10$ is
  
  $$[0.72 + 0.27(10)] \pm 2.201(0.22)\sqrt{\frac{1}{13}}$$

  $$= 3.42 \pm 0.13$$

  $$= (3.29, 3.55)$$

  Confidence bands are illustrated in the attached printout.

- Note we could also test hypotheses, say that $\mu_{y|x} = \mu_0$ using statistic
  
  $$t_{n-2} = \frac{\hat{y} - \mu_0}{s_{\hat{y}}}.$$
2.5 Inferences for future $y$ value

- Suppose we want to predict the value of a new observation at some given $x$. The prediction is

$$\tilde{y} = \hat{\alpha} + \hat{\beta}x,$$

i.e. the same as $\hat{y}$ which predicted the mean. However, the standard error changes to

$$s_\tilde{y} = s_{\hat{y}|x} \sqrt{1 + \frac{1}{n} + \frac{\overline{x}^2}{\sum(x-x)\overline{y}}}$$

This is larger than $s_\hat{y}$, reflecting the fact that we have one additional source of variability, namely that of an observation about its mean.

- Prediction intervals are

$$\tilde{y} \pm t_{n-2}s_\tilde{y}$$

For example, the predicted $y$ for a new sparrow with $x = 7$ is

$$\tilde{y} = 2.16 \text{ cm}$$
as before. The standard error is

$$s_\tilde{y} = 0.22 \sqrt{1 + \frac{1}{13} + \frac{(7-10)^2}{262}} = 0.232$$

which is substantially larger than $s_\hat{y}$.

A 95% prediction interval is

$$2.16 \pm 2.20(0.232) = 2.16 \pm 0.51 = (1.65, 2.67)$$

- We would also obtain prediction bands using multiple $x$ values, but not tests of hypothesis. Why not?

3. Model evaluation

3.1 Coefficient of determination

- The coefficient of determination, $R^2$, is a useful measure on how well the model fits the data. It can be calculated as

$$R^2 = r^2$$

where $r$ is the correlation coefficient. Clearly

$$0 \leq R^2 \leq 1.$$ Why?

- One can show that $R^2$ can be defined as

$$1 - \frac{\text{variability remaining after regression}}{\text{total variability in } y}$$

Symbolically, this implies
\[ R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}, \]

which is also easy to display visually. This definition is often stated as \( R^2 \) is “the proportion of total variability among the \( y \) values that is explained by the linear regression of \( y \) on \( x \).”

• One can also show that
  \[ s^2_{y|x} = (1 - R^2)s^2_y, \]
  which is analogous to the relationship
  \[ \sigma^2_{y|x} = (1 - \rho^2)\sigma^2_y. \]

Using (11), one way to calculate \( R^2 \) is
  \[ R^2 = 1 - \frac{s^2_{y|x}}{s^2_y}. \]

• For example, with our sparrow data,
  \[ s^2_{y|x} = 0.0477. \]

Also, the sample variance is
  \[ s^2_y = \frac{\sum(y - \bar{y})^2}{n-1} = \frac{19.6569}{12} = 1.638 \]

Therefore
  \[ R^2 = 1 - \frac{0.0477}{1.638} = 0.971, \]

i.e. the regression equation explains 97% of the variability in using lengths, \( y \). Less than 3% remains unexplained.

3.2 Residual plots

• Residual plots give insight not only into the data set at hand, but also serve as a diagnostic test to see whether our regression assumptions hold. The residual
  \[ e_i = y_i - \hat{y}_i \]

would be calculated for each observation, and then plotted vs. \( \hat{y}_i \).

• For example, for the sparrow data one has

<table>
<thead>
<tr>
<th>( i )</th>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( \hat{y}_i = 0.72 + 0.27x )</th>
<th>( e_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>1.4</td>
<td>1.53</td>
<td>-0.13</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1.5</td>
<td>1.80</td>
<td>-0.30</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>5.0</td>
<td>5.31</td>
<td>-0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>
The residuals will sum to 0, and would be plotted vs. \( \hat{y} \).

- To gain some perspective, let’s reconsider the prediction of head circumference based on gestational age. The book claims that the least squares prediction equation is
  \[
  \hat{y} = 3.9143 + 0.7801x
  \]
  with (p. 426)
  \[
  s_{y|x} = 1.5904.
  \]
  This is illustrated in Figure 18.7.

- Also,
  \[
  s_\beta = 0.0631 \text{ and } s_\alpha = 1.8291,
  \]
  from where a 95% ci for \( \beta \) is
  \[
  0.7801 \pm 1.98(0.0631) = (0.656, 0.904).
  \]
  How would you interpret this? Would a confidence interval for \( \alpha \) be meaningful?

- Confidence bands for the conditional mean \( \mu_{y|x} \) and prediction bands for future \( y \) are displayed in Figures 18.8 and 18.9. For these data \( R^2 = 0.61 \).

- The residual plot for the 100 observations is given in Figure 18.10. Are there any peculiarities?

- Residual plots can be used for three purposes.
  1) They can help detect outliers. For example, the child with \( x = 31 \) and \( y = 35 \) has a large residual. The residuals should appear normally distributed under our assumptions.
  2) They indicate whether the assumption of constant conditional variance, \( s_{y|x} \), is tenable. Figure 18.11 shows a case where this assumption is violated.
  3) The residuals should not have any trend vs. \( x \). Any such trend would violate the simple linear regression line assumption. What might a trend look like?

3.3 Transformations

- Obviously, there are many possible mathematical relationships between two variables besides the simple linear one. For example, a common one in physiological and pharmacokinetic modeling is the exponential decay model
  \[
  y = \alpha e^{-\beta x}.
  \]
  Note that this may be “linearized” by taking the natural log of \( y \) i.e.
  \[
  y' = \ln(y) = \ln \alpha - \beta x, \text{ or } y' = \alpha' - \beta x
  \]
One could use simple linear regression on $\ln(y)$ vs. $x$ to estimate $\alpha'$ and $\beta$.

- In general, if a scatter-plot does not indicate a straight line relationship, one may be able to transform $y$ and/or $x$ to yield the desired straight line. One could then use the simple linear regression analysis with the transformed data. This is illustrated in a homework problem. Though less commonly used, one could also transform to satisfy the assumption of normality and/or constant variances.

3.4 Further topics

- Regression is a widely used tool. Those who have projects dealing with multiple regression should investigate Chapter 19. A frequent problem is the prediction of $y$ on the basis of many independent variables, say $x_1, x_2, \ldots, x_\rho$.

- Also, there are other powerful regression diagnostic tools besides the residual plots. These too are left for another day.