Lecture 28  Linear Regression 1

Lecture 6 concerned the plotting of X-Y variables to show how the Y-variable changes as X changes.
The shape of the plot (i.e. trend) provided understanding of the association between X and Y.
This lecture introduces the topic of regression which is the calculation of the lines of best fit
that summarize the response of a Y variable to an X variable.
The Y variable is termed the dependent variable and the X is the independent or predictor variable.
The term regression was coined by Galton in 1874 when he observed the relationship between
heights of fathers (X) and their sons (Y).

\[ Y_{i,j} | X_i = \beta_0 + \beta_1 X_i + \epsilon_{i,j} \sim N(0, \sigma^2) \] (28.1)

where the left side of the equation reads as \( Y \) given \( X \).
The estimates of \( \beta_0 \) (the intercept) and \( \beta_1 \) (the slope) which minimise the Error SS (\( \sum_{i,j} \epsilon_{i,j}^2 \))
in Example ?? are

\[ \hat{\beta}_0 = -2.04 \]
\[ \hat{\beta}_1 = 0.025 \]

and the equation of best fit is \( \hat{y}|x = \hat{\beta}_0 + \hat{\beta}_1 \times x \).
The model is fitted in Rcmdr using the menus

- **Data** → **Import data from text file**. Name the data set `CO2uptake` and import `CO2uptake.txt`.

- **Statistics** → **Fit models** → **Linear regression**
  You must be clear about what variable is the Response variable (i.e. \( Y \)) and which is the Explanatory variable (\( X \)). In this case the response is `Uptake` and the explanatory variables is `Conc`.

- The output window contains the *coefficients*,

- You can produce a plot of the “fitted” line with
  **Models** → **Graphs** → **Component + residual**
  and save this graph.

<table>
<thead>
<tr>
<th>Response: Uptake</th>
</tr>
</thead>
<tbody>
<tr>
<td>Df</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Figure 28.1: CO₂ (Y) uptake as a function of Concentration (X) - data and fitted regression

\[
\hat{y} = -2.04 + 0.025 \times x
\]

We can get fitted values at intermediate values of \(x\) by plugging those values into the equation that has been derived,

\[
\hat{y}|x = -2.04 + 0.025 \times x
\]

and this is done in R using the function `predict()`. As of 2007, this function was not available in Rcmdr so we need to provide the script ourselves.

Suppose that we wanted to get the estimate of Uptake when Conc is 150, 180 and 220 (say). Other values work the same.

The script and output are:-

```R
> newConc <- data.frame(Conc=c(150,180,220))
> preds <- predict(CO2uptake.lm,newdata=newConc,se=T)
> preds
$fit
   1    2    3
1.699064 2.447411 3.445206

$se.fit
   1    2    3
0.06068255 0.06423381 0.09288595

> cbind(newConc,preds$fit,preds$se.fit)
   Conc    preds$fit     preds$se.fit
1  150 1.699064 0.06068255
2  180 2.447411 0.06423381
3  220 3.445206 0.09288595
```

Confidence intervals are also produced by the `predict()` function,
Observe that for the `predict()` function, we supply just the `x` because after fitting the model, we know \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) and the determination of \( \hat{y}|x \) follows. The standard errors of the fitted values are also calculated but the mechanics of this operation is not covered in stat100.

### 28.1 Interpreting the AOV and coefficients

The AOV is calculated as before, Models → Hypothesis tests → ANOVA.

<table>
<thead>
<tr>
<th>Response: Uptake</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conc</td>
<td>1</td>
<td>26.90</td>
<td>26.90</td>
<td>446.0</td>
<td>1.4e-12***</td>
</tr>
<tr>
<td>Residuals</td>
<td>15</td>
<td>0.91</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| (Intercept) | -2.043 | 0.1979 | -10.203 | 3.3e-08 |
| Conc       | 0.025     | 0.0012  | 21.04   | 1.4e-12 |

- The AOV indicates that the response of CO\(_2\) uptake due to increasing Concentration is unlikely to be due to chance because the F test indicates that the amount of information which can be attributed to the systematic component of Concentration, outweighs that for the random component which is the sum of squares of residuals or \( \sum(y_i - \hat{y}_i)^2 \).

- There is 1 degree of freedom in the AOV for Conc. This represents the linear effect of Conc on Uptake.

- The item denoted Intercept is the estimate of \( Y \) when \( X = 0 \) and has the same role as the base level for a linear model with factors.

- This accounts for the 1 degree of freedom subtracted from the sample size, i.e. \( (n - 1) \) and gives a reference value for the relationship.
### Example - Hanford

<table>
<thead>
<tr>
<th>Sample</th>
<th>County</th>
<th>Exposure</th>
<th>Mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Umatilla</td>
<td>2.5</td>
<td>147.1</td>
<td></td>
</tr>
<tr>
<td>2 Morrow</td>
<td>2.6</td>
<td>130.1</td>
<td></td>
</tr>
<tr>
<td>3 Gilliam</td>
<td>3.4</td>
<td>129.9</td>
<td></td>
</tr>
<tr>
<td>4 Sherman</td>
<td>1.2</td>
<td>113.5</td>
<td></td>
</tr>
<tr>
<td>5 Wasco</td>
<td>1.6</td>
<td>137.5</td>
<td></td>
</tr>
<tr>
<td>6 HoodRiver</td>
<td>3.8</td>
<td>162.3</td>
<td></td>
</tr>
<tr>
<td>7 Portland</td>
<td>11.6</td>
<td>207.5</td>
<td></td>
</tr>
<tr>
<td>8 Columbia</td>
<td>6.4</td>
<td>177.9</td>
<td></td>
</tr>
<tr>
<td>9 Clatsop</td>
<td>8.3</td>
<td>210.3</td>
<td></td>
</tr>
</tbody>
</table>

### Analysis of Variance Table

**Response: Mortality**

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposure</td>
<td>1</td>
<td>8310</td>
<td>8310</td>
<td>42.3</td>
<td>0.00033 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>7</td>
<td>1374</td>
<td>196</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Estimate Std. Error t value Pr(>|t|)**

- (Intercept) 114.7 8.0 14.3 2.0e-06
- Exposure 9.2 1.4 6.5 3.3e-04

- 2.5 % 97.5 %
- (Intercept) 95.7 134
- Exposure 5.9 13
Lecture 29 Covariance and Correlation

Covariance

If random variables $Y_1$ and $Y_2$ are independent,
\[
\text{var}(Y_1 + Y_2) = \text{var}(Y_1) + \text{var}(Y_2)
\]
\[
\text{var}(Y_1 - Y_2) = \text{var}(Y_1) + \text{var}(Y_2)
\]

If the random variables are NOT independent or that there is an association between $Y_1$ and $Y_2$, the equality of does not hold and we would need to take account of a measure of that association.

One measure of linear association between 2 random variables is covariance.

- $\text{var}(Y) = E( (Y - \mu)^2 )$
- the covariance between 2 random variables $Y_1$ and $Y_2$ is
  \[
  \text{cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \tag{29.1}
  \]
- If $Y_1$ and $Y_2$ are independent, there is no association between them and $\text{cov}(Y_1, Y_2) = 0$.
- In R, covariance is calculated by $\text{var}(Y1,Y2)$.

Example

These data are yields of white clover obtained by cuts from 5 plots in November and April.

<table>
<thead>
<tr>
<th>Nov</th>
<th>20</th>
<th>83</th>
<th>136</th>
<th>46</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr</td>
<td>650</td>
<td>1587</td>
<td>866</td>
<td>1170</td>
<td>576</td>
</tr>
</tbody>
</table>

The total yield (Nov+Apr) is of interest.

- Enter the data, call it clover.
- Calculate a new variable
  
  Data → Manage variables → Compute new variable

\[
\text{Nov} + \text{Apr}
\]
• Make the script `var(clover)` and Submit this is the output,

```
> var(clover)

<table>
<thead>
<tr>
<th></th>
<th>Nov</th>
<th>Apr</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov</td>
<td>2776.50</td>
<td>9488.75</td>
<td>12265.25</td>
</tr>
<tr>
<td>Apr</td>
<td>9488.75</td>
<td>172285.20</td>
<td>181773.95</td>
</tr>
<tr>
<td>Total</td>
<td>12265.25</td>
<td>181773.95</td>
<td>194039.20</td>
</tr>
</tbody>
</table>
```

and it means

\[
\begin{align*}
\text{var}(\text{Nov}) &= 2776.5 \\
\text{var}(\text{Apr}) &= 172285.2 \\
\text{cov}(\text{Apr}, \text{Nov}) &= 9488.8
\end{align*}
\]

The difference \( \text{var}(\text{Apr}+\text{Nov}) - (\text{var}(\text{Nov}) + \text{var}(\text{Apr})) = 194039 - 175062 = 18977 \) and it appears that Nov and Apr are (as expected) not independent random variables. The difference between the variance of the total and the sum of the variances is due to the association between the variables or the covariance.

Were we interested in the Apr - Nov difference, make a new variable `Difference` and repeat the command `var(clover)`,

```
> var(clover)

<table>
<thead>
<tr>
<th></th>
<th>Nov</th>
<th>Apr</th>
<th>Total</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nov</td>
<td>2776.50</td>
<td>9488.75</td>
<td>12265.25</td>
<td>6712.25</td>
</tr>
<tr>
<td>Apr</td>
<td>9488.75</td>
<td>172285.20</td>
<td>181773.95</td>
<td>162796.45</td>
</tr>
<tr>
<td>Total</td>
<td>12265.25</td>
<td>181773.95</td>
<td>194039.20</td>
<td>169508.70</td>
</tr>
<tr>
<td>Difference</td>
<td>6712.25</td>
<td>162796.45</td>
<td>169508.70</td>
<td>156084.20</td>
</tr>
</tbody>
</table>
```

In this matrix, the variances are on the diagonal and the covariance is off-diagonal. We ignore irrelevant numbers such as the Nov, Difference element.

We can now write the general rule,

\[
\begin{align*}
\text{var}(Y_1 + Y_2) &= \text{var}(Y_1) + \text{var}(Y_2) + 2 \cdot \text{cov}(Y_1, Y_2) \\
\text{var}(Y_1 - Y_2) &= \text{var}(Y_1) + \text{var}(Y_2) - 2 \cdot \text{cov}(Y_1, Y_2)
\end{align*}
\]

Observe that

\[
\begin{align*}
194039.2 &= 2776.5 + 172285.2 + 2 \times 9488.75 \\
156084.2 &= 2776.5 + 172285.2 - 2 \times 9488.75
\end{align*}
\]

**Correlation**

The correlation coefficient is a measure of the strength of the linear association between two variables.

The population and sample correlations for 2 random variables \( Y_1 \) and \( Y_2 \) are defined in Table 29.1.

<table>
<thead>
<tr>
<th>( \rho_{Y_1,Y_2} )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_{Y_1,Y_2} = \frac{\text{cov}(Y_1,Y_2)}{\sigma_{Y_1} \sigma_{Y_2}} ) (29.2)</td>
<td>( r = \frac{\sum_{i=1}^{n}(Y_{1,i} - \bar{Y}<em>1)(Y</em>{2,i} - \bar{Y}<em>2)}{\sqrt{S^2</em>{Y_1} \cdot S^2_{Y_2}}} ) (29.3)</td>
</tr>
</tbody>
</table>

Properties of \( r \):-
- $-1 \leq r \leq +1$
- The magnitude of $r$ indicates the strength of the linear relationship.
- The sign of $r$ indicates whether the relationship is direct or inverse. If $r$ is positive, $y_2$ tends to increase linearly as $y_1$ increases and if $r$ is negative, $y_2$ decreases as $y_1$ increases.
- A values of $r$ close to +1 or −1 indicates a strong relationship.
- A value of $r$ close to 0 indicates a weak linear relationship.
The \texttt{cor()} function in R is similar to \texttt{cov()}. For instance,

\begin{verbatim}
print(cor(cbind(Nov,Apr)))
   Nov Apr
Nov 1.00 0.43
Apr 0.43 1.00
\end{verbatim}

In Rcmdr, the menu is \textit{Statistics} \textasciitilde \textit{Summaries} \textasciitilde \textit{Correlation matrix}. You must have the data as an active data set.

\begin{verbatim}
To select \texttt{Apr} and \texttt{Nov}, click on one of them then hold \texttt{Ctrl} whilst you click on the other.

\begin{verbatim}
> cor(clover[,c("Apr","Nov")], use="complete.obs")
   Apr Nov
Apr 1.0000000 0.4338466
Nov 0.4338466 1.0000000
\end{verbatim}
\end{verbatim}

\textbf{Example 29.1}

Government statisticians in England conducted a study of the relationship between smoking and lung cancer. The data concern 25 occupational groups and are condensed from data on thousands of individual men. The explanatory variable is the number of cigarettes smoked per day by men in each occupation relative to the number smoked by all men of the same age. This smoking ratio is 100 if men in an occupation are exactly average in their smoking, it is below 100 if they smoke less than average, and above 100 if they smoke more than average. The response variable is the standardized mortality ratio for deaths from lung cancer. It is also measured relative to the entire population of men of the same ages as those studied, and is greater or less than 100 when there are more or fewer deaths from lung cancer than would be expected based on the experience of all English men.

\begin{verbatim}
#_________ smoking.R _____________
options(digits=2)
smoking <- c(77,137,117,94,116,102,111,93,88,102,91,104,
         107,112,113,110,125,133,115,105,87,91,100,76,66)
mortality <- c(84,116,123,128,155,101,118,113,104,88,104,
        129,86,96,144,139,113,146,128,115,79,85,120,60,51)
SM <- data.frame(Smoking=smoking,Mortality=mortality)
plot(Mortality ~ Smoking,data=SM)
print(cor(SM))
\end{verbatim}

\begin{verbatim}
   Smoking Mortality
Smoking   1.00   0.72
Mortality 0.72   1.00
\end{verbatim}
Example 29.2

These data were collected by Professor D Cottle at UNE. Fleece measurements were made on 100 sheep (hoggets100.txt on the web page). There are many missing variables and the NA's are overcome by using the argument
use="pairwise.complete.obs".

```r
#_____ hoggets.R _______
options(digits=2)
hoggets <- read.table("hoggets100.txt",header=T)
Rmat <- cor(hoggets, ,use="pairwise.complete.obs")
print(Rmat)

MicAve MicDev GFW SFMic Slmm
MicAve 1.00 1.00 0.50 0.98 0.24
MicDev 1.00 1.00 0.50 0.98 0.24
GFW 0.50 0.50 1.00 0.49 0.44
SFMic 0.98 0.98 0.49 1.00 0.23
Slmm 0.24 0.24 0.44 0.23 1.00
```

Example 29.3

Example ?? concerned the difference in wear of boys’ shoes. The paired t-test used the variance of the difference $A - B$ as the measure of randomness.

The data are used here to demonstrate covariance.

We name the data shoes and enter the data in columns like:-

```
<table>
<thead>
<tr>
<th>Copy</th>
<th>Paste</th>
<th>var3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.2</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>8.2</td>
<td>6.8</td>
</tr>
<tr>
<td>3</td>
<td>10.9</td>
<td>11.2</td>
</tr>
<tr>
<td>4</td>
<td>14.3</td>
<td>14.2</td>
</tr>
<tr>
<td>5</td>
<td>10.7</td>
<td>11.5</td>
</tr>
<tr>
<td>6</td>
<td>6.6</td>
<td>6.4</td>
</tr>
<tr>
<td>7</td>
<td>9.6</td>
<td>5.8</td>
</tr>
<tr>
<td>8</td>
<td>10.8</td>
<td>11.3</td>
</tr>
<tr>
<td>9</td>
<td>8.8</td>
<td>5.3</td>
</tr>
<tr>
<td>10</td>
<td>13.3</td>
<td>13.3</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

The variance of A, B and the covariance of A & B are calculated as below

```
> var(shoes)
      A     B
A 6.009000 6.100889
B 6.100889 6.342667
```
That is

\[ \text{var}(A) = 6 \]
\[ \text{var}(B) = 6.3 \]
\[ \text{cov}(A, B) = 6.1 \]