Simple Linear Regression: Mortality Data

The data, taken from certain regions of Great Britain, Norway, and Sweden contains the mean annual temperature (in degrees F) and mortality index for neoplasms of the female breast.

<table>
<thead>
<tr>
<th>Mortality rate (M)</th>
<th>102.5</th>
<th>104.5</th>
<th>100.4</th>
<th>95.9</th>
<th>87.0</th>
<th>95.0</th>
<th>88.6</th>
<th>89.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (T)</td>
<td>51.3</td>
<td>49.9</td>
<td>50.0</td>
<td>49.2</td>
<td>48.5</td>
<td>47.8</td>
<td>47.3</td>
<td>45.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mortality rate (M)</th>
<th>78.9</th>
<th>84.6</th>
<th>81.7</th>
<th>72.2</th>
<th>65.1</th>
<th>68.1</th>
<th>67.3</th>
<th>52.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (T)</td>
<td>46.3</td>
<td>42.1</td>
<td>44.2</td>
<td>43.5</td>
<td>42.3</td>
<td>40.2</td>
<td>31.8</td>
<td>34.0</td>
</tr>
</tbody>
</table>

Objective: Obtaining the relationship between mean annual temperature and the mortality rate for a type of breast cancer in women.

Website of my LM course
http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat5410/

Getting Started

Figure: Scatter Plot of Temperature versus Mortality Rate, Breast Cancer Data.
Fitting the Regression Line

- Underlying Model:

\[ y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \]

- Coefficients are estimated by minimizing

\[ \sum_{i=1}^{N} \left( y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2. \]

- **Least Squares Estimates**

  Estimated Coefficients:

  \[ \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}, \]

  \[ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad var(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{N} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right), \]

  \[ \bar{x} = \frac{1}{N} \sum x_i, \quad \bar{y} = \frac{1}{N} \sum y_i. \]

---

Explanatory Power of the Model

- The total variation in \( y \) can be measured by corrected total sum of squares

\[ CTSS = \sum_{i=1}^{N} (y_i - \bar{y})^2. \]

- This can be decomposed into two parts (Analysis of Variance (ANOVA)):

\[ CTSS = RegrSS + RSS, \]

where

\[ RegrSS = \text{Regression sum of squares} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2, \]

\[ RSS = \text{Residual sum of squares} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2. \]

\( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \) is called the predicted value of \( y_i \) at \( x_i \).

- \( R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS} \) measures the proportion of variation in \( y \) explained by the fitted model.
ANOVA Table for Simple Linear Regression

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression</td>
<td>1</td>
<td>$\hat{\beta}_1^2 \sum (x_i - \bar{x})^2$</td>
<td>$\hat{\beta}_1^2 \sum (x_i - \bar{x})^2$</td>
</tr>
<tr>
<td>residual</td>
<td>$N - 2$</td>
<td>$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$</td>
<td>$\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$</td>
</tr>
<tr>
<td>total (corrected)</td>
<td>$N - 1$</td>
<td>$\sum_{i=1}^{N} (y_i - \bar{y})^2$</td>
<td></td>
</tr>
</tbody>
</table>

ANOVA Table for Breast Cancer Example

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression</td>
<td>1</td>
<td>2599.53</td>
<td>2599.53</td>
</tr>
<tr>
<td>residual</td>
<td>14</td>
<td>796.91</td>
<td>56.92</td>
</tr>
<tr>
<td>total (corrected)</td>
<td>15</td>
<td>3396.44</td>
<td></td>
</tr>
</tbody>
</table>

$t$-Statistic

- To test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_A : \beta_j \neq 0$ under the full model, use the test statistic

$$t_j = \frac{\hat{\beta}_j}{s.d.(\hat{\beta}_j)}.$$  

- The higher the value of $|t_j|$, the more significant is the coefficient.

- For 2-sided alternatives, $p$-value $= \text{Prob} \left( |t_{df}| > |t_{obs}| \right)$, $df =$ degrees of freedom for the $t$-statistic, $t_{obs} =$ observed value of the $t$-statistic. If $p$-value is very small, then either we have observed something which rarely happens, or $H_0$ is not true. In practice, if $p$-value is less then $\alpha = 0.05$ or 0.01, $H_0$ is rejected at level $\alpha$. 
Confidence Interval

100(1 − α)% confidence interval for $\beta_j$ is given by

$$\hat{\beta}_j \pm t_{N-2, \frac{\alpha}{2}} \times s.d. (\hat{\beta}_j),$$

where $t_{N-2, \frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the $t$ distribution with $N-2$ degrees of freedom.

If the confidence interval for $\beta_j$ does not contain 0, then $H_0$ is rejected.

Predicted Values and Residuals

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of $y_i$ at $x_i$.
- $r_i = y_i - \hat{y}_i$ is the corresponding residual.
- Standardized residuals are defined as $\frac{r_i}{s.d.(r_i)}$.
- Plots of residuals are extremely useful to judge the “goodness” of fitted model.
  - Normal probability plot (will be explained in Unit 3).
  - Residuals versus predicted values.
  - Residuals versus covariate $x$. 
Analysis of Breast Cancer Data

The regression equation is
M = -21.79 + 2.36 T

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-21.79</td>
<td>15.67</td>
<td>-1.39</td>
<td>0.186</td>
</tr>
<tr>
<td>T</td>
<td>2.3577</td>
<td>0.3489</td>
<td>6.76</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 7.54466  R-Sq = 76.5%  R-Sq(adj) = 74.9%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2599.5</td>
<td>2599.5</td>
<td>45.67</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>14</td>
<td>796.9</td>
<td>56.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>3396.4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>T</th>
<th>M</th>
<th>Fit</th>
<th>SE Fit</th>
<th>Residual</th>
<th>St Resid</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>31.8</td>
<td>67.3</td>
<td>53.18</td>
<td>4.85</td>
<td>14.12</td>
<td>2.44RX</td>
</tr>
</tbody>
</table>

R denotes an observation with a large standardized residual. X denotes an observation whose X value gives it large leverage.

Outlier Detection

- Minitab identifies two types of outliers denoted by R and X:
  - R: its standardized residual \((y_i - \hat{y}_i)/se(\hat{y}_i)\) is large.
  - X: its X value gives large leverage (i.e., far away from majority of the X values).

- For the mortality data, the observation with \(T = 31.8, M = 67.3\) (i.e., left most point in plot on LNP.2-2) is identified as both R and X.

- After removing this outlier and refitting the remaining data, the output is given on LNP.2-11. There is still an outlier identified as X but not R. This one (second left most point on LNP.2-2) should not be removed (why?)

- Residual plots on LNP.2-12 show no systematic pattern.

Notes: Outliers are not discussed in the book, see standard regression texts. Residual plots will be discussed in unit 3.
Regression Results after Removing the Outlier

The regression equation is
\[ M = -52.62 + 3.02 \, T \]

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-52.62</td>
<td>15.82</td>
<td>-3.33</td>
<td>0.005</td>
</tr>
<tr>
<td>T</td>
<td>3.0152</td>
<td>0.3466</td>
<td>8.70</td>
<td>0.000</td>
</tr>
</tbody>
</table>

\[ s = 5.93258 \quad R^2 = 85.3\% \quad R^2(\text{adj}) = 84.2\% \]

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2664.3</td>
<td>2664.3</td>
<td>75.70</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>13</td>
<td>457.5</td>
<td>35.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>3121.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unusual Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>T</th>
<th>M</th>
<th>Fit</th>
<th>SE Fit</th>
<th>Residual</th>
<th>Std Resid</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>34.0</td>
<td>52.50</td>
<td>49.90</td>
<td>4.25</td>
<td>2.60</td>
<td>0.63 X</td>
<td></td>
</tr>
</tbody>
</table>

X denotes an observation whose X value gives it large leverage.

Residual Plots After Outlier Removal

![Residual Plots](image)

Figure 3: Residual Plots

Comments: No systematic pattern is discerned.
Prediction from the Breast Cancer Data

- The fitted regression model is $Y = -21.79 + 2.36X$, where $Y$ denotes the mortality rate and $X$ denotes the temperature.

- The predicted mean of $Y$ at $X = x_0$ can be obtained from the above model. For example, prediction for the temperature of 49 is obtained by substituting $x_0 = 49$, which gives $y_{x_0} = 93.85$.

- The standard error of $\hat{\mu}_{x_0}$ is given by

$$s.e.(\hat{\mu}_{x_0}) = \hat{\sigma} \sqrt{\frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}}.$$ 

- Here $x_0 = 49$, $1/N + (\bar{x} - x_0)^2/\sum_{i=1}^{N} (x_i - \bar{x})^2 = 0.1041$, and $\hat{\sigma} = \sqrt{MSE} = 7.54$. Consequently, $s.e.(\hat{\mu}_{x_0}) = 2.432$.

Confidence interval for mean and prediction interval for future observation

- A 95% confidence interval for the mean response $\mu_{x_0} = \beta_0 + \beta_1 x_0$ at $x = x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{N-2,0.025} \times s.e.(\hat{\mu}_{x_0}).$$

- Here the 95% confidence interval for the mean mortality corresponding to a temperature of 49 is [88.63, 99.07].

- A 95% prediction interval for an individual observation $y_{x_0}$ corresponding to $x = x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{N-2,0.025} \times \hat{\sigma} \sqrt{1 + \frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}},$$

where 1 under the square root represents $\sigma^2$, variance of the new observation $y_{x_0}$.

- The 95% prediction interval for the predicted mortality of an individual corresponding to the temperature of 49 is [76.85, 110.85].
Multiple Linear Regression: Air Pollution Data

http://lib.stat.cmu.edu/DASL/Stories/AirPollutionandMortality.html

- Data collected by General Motors.
- Response is age-adjusted mortality.
- Predictors:
  - Variables measuring demographic characteristics.
  - Variables measuring climatic characteristics.
  - Variables recording pollution potential of 3 air pollutants.
- Objective: To determine whether air pollution is significantly related to mortality.

Predictors

1. JanTemp: Mean January temperature (degrees Farenheit)
2. JulyTemp: Mean July temperature (degrees Farenheit)
3. RelHum: Relative Humidity
4. Rain: Annual rainfall (inches)
5. Education: Median education
6. PopDensity: Population density
7. %NonWhite: Percentage of non whites
8. %WC: Percentage of white collar workers
9. pop: Population
10. pop/house: Population per household
11. income: Median income
12. HCPot: HC pollution potential
13. NOxPot: Nitrous Oxide pollution potential
14. SO2Pot: Sulphur Dioxide pollution potential
Getting Started

- There are 60 data points.
- Pollution variables are highly skewed, log transformation makes them nearly symmetric. The variables HCPot, NOxPot and SO2Pot are replaced by log(HCPot), log(NOxPot) and log(SO2Pot).
- Observation 21 (Fort Worth, TX) has two missing values, so this data point will be discarded from the analysis.

Scatter Plots

Figure 4: Scatter Plots of mortality against selected predictors

(a) JanTemp

(b) Education

(c) NonWhite

(d) Log(NOxPot)
Fitting the Multiple Regression Equation

- Underlying Model:
  \[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \]

- Coefficients are estimated by minimizing
  \[ \sum_{i=1}^{N} \left( y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \right)^2 = (y - X\hat{\beta})' (y - X\hat{\beta}). \]

- Least Squares estimates:
  \[ \hat{\beta} = (X'X)^{-1}X'y. \]

- Variance-Covariance matrix of \( \hat{\beta} \):
  \[ \Sigma_{\hat{\beta}} = \sigma^2(X'X)^{-1}. \]

- A row: one group of observations
- A column: one effect

\[ \begin{array}{ccc}
Y & g_1 & g_2 & \ldots & g_p \\
y_1 & g_{11} & g_{12} & \ldots & g_{1p} \\
y_2 & g_{21} & g_{22} & \ldots & g_{2p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y_n & g_{n1} & g_{n2} & \ldots & g_{np}
\end{array} \]

Experimental units \( \rightarrow \) System/Process \( \rightarrow \) Output (response, \( y \))

 uncontrollable

controllable

\( X_1 \rightarrow \ldots \rightarrow X_m \)

\( Z_1 \rightarrow \ldots \rightarrow Z_n \)

\( (\hat{X}\hat{\beta}, \hat{\varepsilon}, \hat{Y}) \)
Analysis of Variance

- The total variation in $y$, i.e., corrected total sum of squares,
  \[ CTSS = \sum_{i=1}^{N} (y_i - \bar{y})^2 = y^T y - N\bar{y}^2, \]
  can be decomposed into two parts
  (Analysis of Variance (ANOVA)):

  \[ CTSS = \text{RegrSS} + \text{RSS}, \]

  where $\text{RSS} = \text{Residual sum of squares} = \sum (y_i - \hat{y}_i)^2 = (y - X\hat{\beta})^T (y - X\hat{\beta}),$
  $\text{RegrSS} = \text{Regression sum of squares} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 = \hat{\beta}^T X^T X\hat{\beta} - N\bar{y}^2.$

  **ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression</td>
<td>$k$</td>
<td>$\hat{\beta}^T X^T X\hat{\beta} - N\bar{y}^2$</td>
<td>$(\hat{\beta}^T X^T X\hat{\beta} - N\bar{y}^2)/k$</td>
</tr>
<tr>
<td>residual</td>
<td>$N - (k + 1)$</td>
<td>$(y - X\hat{\beta})^T (y - X\hat{\beta})$</td>
<td>$(y - X\hat{\beta})^T (y - X\hat{\beta})/(N - k - 1)$</td>
</tr>
<tr>
<td>total</td>
<td>$N - 1$</td>
<td>$y^T y - N\bar{y}^2$</td>
<td></td>
</tr>
</tbody>
</table>

Explanatory Power of the Model

- $R^2 = \frac{\text{RegrSS}}{CTSS} = 1 - \frac{\text{RSS}}{CTSS}$ measures the proportion of variation in $y$
  explained by the fitted model. $R$ is called the multiple correlation coefficient.

- Adjusted $R^2$:

  \[ R_a^2 = 1 - \frac{\frac{RSS}{N-(k+1)}}{\frac{CTSS}{N-1}} = 1 - \left( \frac{N - 1}{N - k - 1} \right) \frac{RSS}{CTSS}. \]

- When an additional predictor is included in the regression model, $R^2$ always
  increases. This is not a desirable property for model selection. However, $R_a^2$
  may decrease if the included variable is not an informative predictor.
  Usually $R_a^2$ is a better measure for comparing different model fits.
Testing significance of coefficients: $t$-Statistic

- To test the null hypothesis $H_0: \beta_j = 0$ against the alternative hypothesis $H_A: \beta_j \neq 0$ under the full model, use the test statistic

$$t_j = \frac{\hat{\beta}_j}{s.d.(\hat{\beta}_j)}.$$

- The higher the value of $|t_j|$, the more significant is the coefficient.
- In practice, if $p$-value is less then $\alpha = 0.05$ or 0.01, $H_0$ is rejected.
- **Confidence Interval**: $100(1 - \alpha)\%$ confidence interval for $\beta_j$ is given by

$$\hat{\beta}_j \pm t_{N-(k+1), \frac{\alpha}{2}} \times s.d.(\hat{\beta}_j),$$

where $t_{N-k-1, \frac{\alpha}{2}}$ is the upper $\alpha/2$ point of the $t$ distribution with $N - k - 1$ degrees of freedom.

If the confidence interval for $\beta_j$ does not contain 0, then $H_0$ is rejected.

---

Analysis of Air Pollution Data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1332.7</td>
<td>291.7</td>
<td>4.57</td>
<td>0.000</td>
</tr>
<tr>
<td>JanTemp</td>
<td>-2.3052</td>
<td>0.8795</td>
<td>-2.62</td>
<td>0.012</td>
</tr>
<tr>
<td>JulyTemp</td>
<td>-1.657</td>
<td>2.051</td>
<td>-0.81</td>
<td>0.424</td>
</tr>
<tr>
<td>RelHum</td>
<td>0.407</td>
<td>1.070</td>
<td>0.38</td>
<td>0.706</td>
</tr>
<tr>
<td>Rain</td>
<td>1.4436</td>
<td>0.5847</td>
<td>2.47</td>
<td>0.018</td>
</tr>
<tr>
<td>Educatio</td>
<td>-9.458</td>
<td>9.080</td>
<td>-1.04</td>
<td>0.303</td>
</tr>
<tr>
<td>PopDenai</td>
<td>0.004509</td>
<td>0.004311</td>
<td>1.05</td>
<td>0.301</td>
</tr>
<tr>
<td>%NonWhit</td>
<td>5.194</td>
<td>1.005</td>
<td>5.17</td>
<td>0.000</td>
</tr>
<tr>
<td>%WC</td>
<td>-1.852</td>
<td>1.210</td>
<td>-1.53</td>
<td>0.133</td>
</tr>
<tr>
<td>pop</td>
<td>0.00000109</td>
<td>0.000000401</td>
<td>0.27</td>
<td>0.788</td>
</tr>
<tr>
<td>pop/hous</td>
<td>-45.95</td>
<td>39.78</td>
<td>-1.16</td>
<td>0.254</td>
</tr>
<tr>
<td>income</td>
<td>-0.000549</td>
<td>0.001309</td>
<td>-0.42</td>
<td>0.677</td>
</tr>
<tr>
<td>logHC</td>
<td>-53.47</td>
<td>35.39</td>
<td>-1.51</td>
<td>0.138</td>
</tr>
<tr>
<td>logNOx</td>
<td>80.22</td>
<td>32.66</td>
<td>2.46</td>
<td>0.018</td>
</tr>
<tr>
<td>logSO2</td>
<td>-6.91</td>
<td>16.72</td>
<td>-0.41</td>
<td>0.681</td>
</tr>
</tbody>
</table>

$S = 34.58$, $R$-Squ = 76.7%, $R$-Squ(adj) = 69.3%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>14</td>
<td>173383</td>
<td>12834</td>
<td>10.36</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>44</td>
<td>52610</td>
<td>1196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>225993</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• formulation of hypothesis testing from the view of \textbf{comparing models}

\begin{itemize}
  \item a model space \( \equiv \) the space spanned by columns of some \( X \)
  \item consider a large model space, \( \Omega \), and a smaller model space, \( \omega \), where \( \omega \subset \Omega \), i.e., \( \omega \) represents a subset/a subspace of \( \Omega \). Suppose dimension (# of parameters) of \( \Omega \) is \( p \) and \( \text{dim}(\omega)=q \), where \( p>q \).
  \item to answer “which of the model spaces is more adequate” in statistical language
  \( \Rightarrow \) perform the test \( H_0: \omega \) v.s. \( H_A: \Omega \backslash \omega \)
\end{itemize}

\[ F = \frac{(RSS_{\omega} - RSS_{\Omega})}{RSS_{\Omega}} / (p-q)
\sim F_{p-q,n-p} \text{ (under } \omega) \]

\[ \Omega \text{ (dim } p) \]
\[ \omega \text{ (dim } q) \]

\[ \hat{Y} \text{ (n-dim)} \]

\[ \hat{\epsilon}_{\omega} \text{ (n-q-dim)} \]

\[ \hat{\epsilon}_{\Omega} \text{ (n-p-dim)} \]

\[ \hat{\epsilon}_{\Omega} - \hat{\epsilon}_{\omega} \text{ (p-q-dim)} \]

(sequential) \textbf{ANOVA}

- \text{anova}(y\sim I+A+B+A:B), A: 3 levels, B: 4 levels

1) test \( \omega \text{-model 1 (y} \sim I) \) against \( \Omega \text{-model 2 (y} \sim I+A) \) \( [df_{\omega} - df_{\Omega} =2] \)
2) test \( \omega \text{-model 2 (y} \sim I+A) \) against \( \Omega \text{-model 4 (y} \sim I+A+B) \) \( [df_{\omega} - df_{\Omega} =3] \)
3) test \( \omega \text{-model 4 (y} \sim I+A+B) \) against \( \Omega \text{-model 5 (y} \sim I+A+B+A:B) \) \( [df_{\omega} - df_{\Omega} =6] \)

\[ F = \frac{(RSS_{\omega} - RSS_{\Omega})}{RSS_{\text{model 5}} / df_{\text{model 5}}} \sim F_{df_{\omega} - df_{\Omega}, df_{\text{model 5}}} \]

- invariant to the choice of dummy variables since they generate same \( \omega \) and \( \Omega \)

- \text{ANOVA could have different results when the order of effect sequence is changed, e.g., anova}(y\sim I+ B+A+A:B)\):

\[ \alpha) \text{ test } \omega \text{-model 1 (y} \sim I) \text{ against } \Omega \text{-model 3 (y} \sim I+B) \text{ [df}_{\omega} - df_{\Omega} =3] \]
\[ \beta) \text{ test } \omega \text{-model 3 (y} \sim I+B) \text{ against } \Omega \text{-model 4 (y} \sim I+B+A) \text{ [df}_{\omega} - df_{\Omega} =2] \]
\[ \chi) \text{ test } \omega \text{-model 4 (y} \sim I+B+A) \text{ against } \Omega \text{-model 5 (y} \sim I+B+A+A:B) \text{ [df}_{\omega} - df_{\Omega} =6] \]

\text{anova}(y\sim I+A+B+A:B) \text{ and anova}(y\sim I+B+A+A:B) \text{ will have \textit{identical} results when orthogonality exists between the three groups of effects: span}\{d_{i}^{A}\}, \text{ span}\{d_{j}^{B}\}, \text{ span}\{d_{ij}^{A:B}\}, \text{ because in the case, } RSS_{\omega} - RSS_{\Omega} \text{ would equal for 1) and } \beta), 2) \text{ and } \alpha), 3) \text{ and } \chi)
Consider the full model:

\[ y = \beta_0 + \beta_1 g_1(x_1, \ldots, x_m) + \beta_2 g_2(x_1, \ldots, x_m) + \cdots + \beta_k g_k(x_1, \ldots, x_m) + \epsilon \]

For \(1 \leq i \leq k\), should the term \(\beta_i g_i\) be included in the final fitted model?

- (sub-)model: a model with a subset of all \(k\) terms, e.g., \(\{1, g_1, g_2\}\), \(\{1, g_2, g_4, g_5, g_k\}\), ...

- hierarchical structure of all sub-models (see graph)
  - \(p = \# \) of terms in a sub-model
  - \# of different sub-models = \(2^k\)
  - connecting line: model nesting

Example: 6 effects, I, Y, p, e, R, T

Orthogonality

- **Q**: consider the two models:
  - model 1: \(y = \beta_0 + \beta_1 x_1 + \epsilon\)
  - model 2: \(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon\)

  In general, \(\hat{\beta}_1\), in the 2 models are not identical (of course, test \(H_0: \beta_1 = 0\) not identical neither)

  an exception: when \(x_1\) and \(x_2\) are orthogonal

  \[ Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon, \text{ where } \beta = [\beta_1, \beta_2]^T \text{ and } X = [X_1, X_2] \] with the property \(X_1^T X_2 = 0 \Rightarrow X_1\) and \(X_2\) are orthogonal

  \[ X^T X = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{bmatrix} = \begin{bmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{bmatrix} \Rightarrow \left( X^T X \right)^{-1} = \begin{bmatrix} (X_1^T X_1)^{-1} & 0 \\ 0 & (X_2^T X_2)^{-1} \end{bmatrix} \]

  \[ \hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y, \quad \hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T Y, \text{ and } \hat{\beta}_1, \hat{\beta}_2 \text{ independent} \]

  
  \(\Rightarrow\) note that \(\hat{\beta}_1\) will be the same regardless of whether \(X_2\) is in the model or not (and vice versa).

  **Q**: what if only two predictors, say some \(x_i\) in \(X_1\) and some \(x_j\) in \(X_2\), are orthogonal?

  - Randomization: In an exp’t, suppose that true model is \(Y = X\beta + Z\gamma + \epsilon\), but \(Z\) cannot be measured or may not even be suspected \(\Rightarrow E(\hat{\beta}) = \beta^\top (X^T X)^{-1} X^T Z \gamma \Rightarrow\)

  **Q**: what’s the best way of controlling \(X\) to make \(X\) and \(Z\) as orthogonal as possible?
• Generalization.

**Reading:** Textbook, 1.4–1.6, 1.8

Some Properties of (Multivariate) Normal Distribution

(N1) Linear transformation of normal is still normal

\[ Z \sim N(\mu, \Sigma) \Rightarrow AZ + c \sim N(A\mu + c, A\Sigma A^T). \]

(N2) When 1st and 2nd moments are given, the normal distribution is specified.

(N3) \( Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \) \( \sim \) normal, and \( Z_1, Z_2 \) uncorrelated (i.e., \( \text{cov}(Z_1, Z_2) = 0 \))

\( \Rightarrow Z_1, Z_2 \) independent

(N4) \( Z \sim N(\mu, \Sigma), W_1 = A_1 Z, W_2 = A_2 Z \)

\( \Rightarrow W_1, W_2 \) are independent iff \( A_1 \Sigma A_2^T = 0. \)

(N5) \( Z \sim N(\mu, \Sigma), W_1 = A_1 Z, W_2 = A_2 Z, \ldots, W_k = A_k Z, \) and \( \text{cov}(W_i, W_j) = 0 \)

for \( 1 \leq i < j \leq k \), \( \Rightarrow W_1^T W_1, W_2^T W_2, \ldots, W_k^T W_k \) are mutually independent.

(N6) \( Z \): an \( n \times 1 \) random vector and \( Z \sim N(\mu, \Sigma) \), then

- if \( \Sigma \) is non-singular, \( (Z - \mu)^T \Sigma^{-1} (Z - \mu) \sim \chi_n^2 \)

- if \( \Sigma \) is singular and has rank \( r < n \),

let \( \Sigma^- \) be a generalized inverse of \( \Sigma \) (i.e., \( \Sigma \Sigma^- \Sigma = \Sigma \)), then

\( (Z - \mu)^T \Sigma^- (Z - \mu) \sim \chi_r^2 \)
Normal Distribution and Sequential ANOVA

- Consider a linear model $Y = X\beta + \varepsilon$, where
  - $Y \in \mathbb{R}^n$: an $n \times 1$ random vector and $Y \sim N(X\beta, \sigma^2 I)$
  - $X = \begin{bmatrix} 1 & X_1 & \cdots & X_k \end{bmatrix}$
  - $\beta = \begin{bmatrix} \beta_0^T & \beta_1^T & \cdots & \beta_k^T \end{bmatrix}^T$
  - $\varepsilon \sim N(0, \sigma^2 I)$

- Define
  - $V_i$ = the vector space generated by the column vectors of $A_i = \begin{bmatrix} X_0 & X_1 & \cdots & X_i \end{bmatrix}$, $i = 0, 1, \ldots, k$.
  - $V_i$ is called the model space of $A_i$, and denoted by $\text{span}\{A_i\}$
  - $V_0 \subset V_1 \subset \cdots \subset V_k = \text{span}\{X\} \subset \mathbb{R}^n$
  - $V_i^\perp$ = orthogonal complement of the vector space $V_i$, i.e.,
    $$V_i^\perp = \{v \in \mathbb{R}^n : v \text{ is orthogonal to all the vectors in } V_i\}$$

- $W_0 = \text{span}\{X_0\} = V_0$, and for $i = 1, \ldots, k$,
  - $W_i$ = orthogonal complement of $V_{i-1}$ relative to $V_i$ (note: $V_{i-1} \subset V_i$), i.e.,
    $$W_i = V_i \ominus V_{i-1} = V_i \cap V_{i-1}^\perp$$
    $$= \{v \in V_i : v \text{ is orthogonal to all the vectors in } V_{i-1}\}$$
  - $V_i = W_0 \oplus W_1 \oplus \cdots \oplus W_i$
  - $\mathbb{R}^n = W_0 \oplus W_1 \oplus \cdots \oplus W_k \oplus V_k^\perp$ (note: $V_k^\perp$ is the residual space)
  - $W_0 \perp W_1 \perp \cdots \perp W_k \perp V_k^\perp$
  - Note. In general, $W_i \neq \text{span}\{X_i\}$. However, if $X_0, X_1, \ldots, X_k$ are mutually orthogonal, then $W_i = \text{span}\{X_i\}$.
  - $r_i = \dim(W_i) = \dim(V_i) - \dim(V_{i-1})$.
    - Let $r = \sum_{i=0}^k r_i = \sum_{i=0}^k \dim(W_i) = \dim(V_k)$. Then, $\dim(V_k^\perp) = n - r$. 

NTHU STAT 5510, 2023. Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)
• Orthogonal projection of $Y$ onto $V_i$’s and $W_i$’s
  - For a vector space $V \subset \mathbb{R}^n$, denote the orthogonal projection matrix of $Y$ onto $V$ by $P_V$. Then, the orthogonal projection of $Y$ onto $V$ is $P_V^*Y$.
  * if $V = \text{span}\{A\}$, then $P_V = A(A^T A)^{-1} A^T$
  * the orthogonal projection matrix onto $V^\perp$, denoted by $P_{V^\perp}$, is $P_{V^\perp} = I - A(A^T A)^{-1} A^T = I - P_V$
  - Some properties of orthogonal projection matrix
    * A square matrix $P$ is a projection matrix iff $P^2 = P$ (idempotent)
      · Idempotence implies $P$ is a generalized inverse of $P$ since $PPP = P^3 = P$.
    * A projection matrix $P$ is orthogonal iff $P^T = P$ (symmetric)
      * If $P$ is an orthogonal projection matrix onto $V$, then
        · $P$ has $\dim(V)$ eigenvalues equal to 1 and the rest 0
        · $P$ is diagonalizable, and there exists an orthogonal matrix $U$ $(U^T U = I)$ such that $U^T P U = \Lambda$ is a diagonal matrix.
        (Note. Thus, $P = U \Lambda U^T$) Actually,
          ◦ the columns of $U$ are orthonormal eigenvectors of $P$, and
          ◦ the diagonal entries of $\Lambda$ are the eigenvalues of $P$.

NTHU STAT 5510, 2023, Lecture Notes

jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

- Since $V_i = \text{span}\{A_i\}$, $P_{V_i} = A_i (A_i^T A_i)^{-1} A_i^T$ and $P_{V_i} Y = A_i (A_i^T A_i)^{-1} A_i^T Y$
  - For the model space $V_i$,
    \[
    \text{RSS}_{V_i} = \|Y\|^2 - \|P_{V_i} Y\|^2 = Y^T Y - (P_{V_i} Y)^T P_{V_i} Y = Y^T Y - Y^T P_{V_i}^T P_{V_i} Y
    \]
    \[= Y^T Y - Y^T P_{V_i} Y - Y^T (I - P_{V_i}) Y - \| (I - P_{V_i}) Y \|^2 \]

- Since $W_i = V_i \cap V_{i-1}^\perp$ and $V_{i-1} \subset V_i$,
  \[ P_{W_i} = (I - P_{V_{i-1}}) P_{V_i} = P_{V_i} - P_{V_{i-1}} P_{V_i} = P_{V_i} - P_{V_{i-1}}, \]
  and $P_{W_i} Y = (I - P_{V_{i-1}}) P_{V_i} Y = (P_{V_i} - P_{V_{i-1}}) Y = P_{V_i} Y - P_{V_{i-1}} Y$.
  * Since $\mathbb{R}^n = W_0 \oplus W_1 \oplus \cdots \oplus W_k \oplus V_k^\perp$ and $W_0 \perp W_1 \perp \cdots \perp W_k \perp V_k^\perp$,
    
    · $Y = P_{W_0} Y + P_{W_1} Y + \cdots + P_{W_k} Y + P_{V_k} Y$
    · $P_{W_0} Y \perp P_{W_1} Y \perp \cdots \perp P_{W_k} Y \perp P_{V_k} Y$
    · $\|Y\|^2 = \|P_{W_0} Y\|^2 + \|P_{W_1} Y\|^2 + \cdots + \|P_{W_k} Y\|^2 + \|P_{V_k} Y\|^2$

* When $X_0, X_1, \ldots, X_k$ are mutually orthogonal,
  \[ P_{W_i} = P_{V_i} - P_{V_{i-1}} = X_i (X_i^T X_i)^{-1} X_i^T. \]