**What is Statistics?**

<table>
<thead>
<tr>
<th>馨伊波特</th>
<th>Real Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>占卜學</td>
<td>Statistics</td>
</tr>
<tr>
<td>崔老妮</td>
<td>Statisticians</td>
</tr>
<tr>
<td>水晶球</td>
<td>Data</td>
</tr>
<tr>
<td>未來的資訊</td>
<td>Information</td>
</tr>
</tbody>
</table>

**aim of statistics:** provide *insight* by means of *data*

---

**Basic Procedures of Statistics**

- Statistics divides the study of data into *five* steps:
  - Problem Formulation & Modeling (conceptual)
  - Data Collection
  - Statistical Modeling (empirical)
  - Data Analysis
  - Decision Making
When to use regression analysis?

- Regression: a statistical tool for investigating the “linearity relationship” between $x$ and $y$.

Example (cont.)

- data type in regression analysis and some terminologies

A \{response, output, dependent\} variable $Y$ is modeled or explained by $p$ \{predictor, input, independent, regressor\} variables that are functions of $X_1, \ldots, X_m$

\[ Y : \text{“approximately” continuous} \]
\[ X_1, \ldots, X_m : \text{continuous and discrete (quantitative), categorical (qualitative)} \]  
(Q: example?)
\[ p=1, \text{ simple regression}; p>1, \text{ multiple regression} \]
\[ X_1, \ldots, X_m \]
all quantitative $\Rightarrow$ multiple regression
quantitative+qualitative $\Rightarrow$ analysis of covariance
all qualitative $\Rightarrow$ analysis of variance (ANOVA)
\[ \text{more than one } Y, \text{ multivariate regression} \]
linear model: \[ Y = \sum_{i=1}^{p} \beta_i \cdot g(X_1, \ldots, X_m) + \epsilon \]

- Deterministic component
- Mean function
- Error: Random component
- Variance function

\[ Y \] \( X_1, \ldots, X_m \) are regarded as deterministic, i.e., no random phenomenon (when they are random variables, regard the linear model as conditional on \( X_1, \ldots, X_m \))

\[ g_1(X_1, \ldots, X_m), \ldots, g_p(X_1, \ldots, X_m) \] are known transformation of \( X_1, \ldots, X_m \)

Parameters \( \beta_1, \ldots, \beta_p \) enter linearly

Variation due to random error only appears on y-axis

Rationale: a general model for the relationship between \( y \) and \( x_1, \ldots, x_m \), is:

\[ Y = f(X_1, X_2, \ldots, X_m) + \epsilon \] where \( f \) is unknown and arbitrary

Local approximation of \( f \) may be achievable by a linear model

Note: Because the predictors can be transformed and combined in any way, linear models are actually very flexible.

Matrix representation

- Given the data,

\[
\begin{array}{c|cccc}
Y & X_1 & X_2 & \ldots & X_m \\
\hline
y_1 & x_{11} & x_{12} & \ldots & x_{1m} \\
y_2 & x_{21} & x_{22} & \ldots & x_{2m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y_n & x_{n1} & x_{n2} & \ldots & x_{nm} \\
\end{array}
\]

- a row: one group of observations
- a column: one variable

\[ y_i = \beta_0 + \beta_1 g_1(x_{i1}, \ldots, x_{im}) + \beta_2 g_2(x_{i1}, \ldots, x_{im}) + \ldots + \beta_p g_p(x_{i1}, \ldots, x_{im}) + \epsilon_i, \quad i = 1, 2, \ldots, n, \]

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

\( g_{ij} = g_j(x_{ij}, \ldots, x_{im}) \)

- a row: one group of observations
- a column: response or predictor

- we may write the model as:

\( y_i \) the expression is (i) ugly notation (ii) conceptually awkward

\( y \) matrix/vector notation is more elegant
Matrix notation:

\[ Y = X\beta + \varepsilon, \]

where

\[ Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & g_{11} & \cdots & g_{1p} \\ 1 & g_{21} & \cdots & g_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & g_{n1} & \cdots & g_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \]

and \( \text{E}(\varepsilon) = 0 \) and \( \text{var}(\varepsilon) = \sigma^2 I \) (Note: the assumption that errors are normal distributed is not required at the estimation stage)

**Estimating \( \beta \)**

- (ordinary) least square estimator
  - \( \hat{\beta} \) assume \( \varepsilon \) are (i) uncorrelated (ii) equal variance
  - \( \hat{\beta} \) define the best \( \hat{\beta} \) as that minimizes sum of squared error \( \varepsilon^T \varepsilon = \sum_{i=1}^{n} \varepsilon_i^2 \)
    \[ \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta) = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta \]  \( \Rightarrow \) a second-order polynomial of \( \beta \)
  - \( \hat{\beta} \) One method of finding the minimizer is to differentiate \( \varepsilon^T \varepsilon \) w.r.t. \( \beta \) and set the derivatives equal to zero \( \Rightarrow -2X^TY + 2X^TX\beta = 0 \)
  - \( \hat{\beta} \) By calculus, \( \hat{\beta} \) is the solution of
    \[ X^TX\hat{\beta} = X^TY \quad \Leftarrow \text{called normal equation} \]
  - \( \hat{\beta} \) assume \( X^TX \) is non-singular,
    \[ \hat{\beta} = (X^TX)^{-1}X^TY \quad \Rightarrow \quad X\hat{\beta} = X \hat{\beta} = X (X^TX)^{-1}X^TY = HY \]
  - \( \hat{\beta} \) predicted values: \( \hat{Y} = X\hat{\beta} = HY \)
  - \( \hat{\beta} \) residuals: \( \hat{\varepsilon} = Y - X\hat{\beta} = Y - \hat{Y} = (I-H)Y \)
  - \( \hat{\beta} \) residual sum of squares (RSS): \( \hat{\varepsilon}^T \hat{\varepsilon} = [Y^T(I-H)][(I-H)Y] = Y^T(I-H)Y \)
• mean and covariance matrix of OLS estimator $\hat{\beta}$

$\hat{\beta} = (X^TX)^{-1}X^TY$ is a $p \times 1$ vector of random variables, so

$\bar{y}$ mean: $E(\hat{\beta}) = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T\beta = \beta$ (i.e., unbiased)

$\bar{y}$ var($\hat{\beta}$) = $(X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} = (X^TX)^{-1}\sigma^2$ ($\Rightarrow$ irrelevant to $Y$ and $\beta$)

Note: when you can control $X$, you can decide the var-cov matrix before observing $Y$)

Since $\hat{\beta}$ is a vector of random variables, $(X^TX)^{-1}\sigma^2$ is a variance-covariance matrix, i.e., $se(\hat{\beta}_i) = \sqrt{(X^TX)^{-1}i\sigma}$

$\bar{y}$ how to calculate the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$?

Estimating $\sigma^2$

• estimate $\sigma^2$ by $\hat{\epsilon}^T \hat{\epsilon} / (n-p) = RSS / (n-p)$ $\Rightarrow$ an unbiased estimate

• actually, it has the minimum variance among all quadratic unbiased estimators of $\sigma^2$

• $\hat{\sigma} = \sqrt{RSS / (n-p)}$

• the maximum likelihood estimator of $\sigma^2$ is $\hat{\epsilon}^T \hat{\epsilon} / n = RSS / n$

goodness-of-fit: how well does the model fit the data?

• $R^2$, coefficient of determination or percentage of variance explained

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \left( \frac{\sum (y_i - \hat{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{y})^2}} \right)^2$$

$RSS$ is calculated from model with all independent variables,
$TSS$ from model without any independent variables

Interpretation of $R^2$: “proportion of total variation in $y$ that can be explained by the independent variables”

$\bar{y}$ $R$=correlation between $\hat{y}$ and $y$; for simple regression, $R$=correlation between $x$ and $y$ (from the geometry viewpoint, ...)

$\bar{y}$ 0 $\leq R^2 \leq 1$, values closer to 1 indicate better fits. (what if $n$ $\approx$ $p$?)

$\bar{y}$ What is a good value of $R^2$?

• alternative measure for goodness of fit: $\hat{\sigma}$

$\bar{y}$ it’s related to standard error of estimates of $\beta$ and prediction
$\bar{y}$ it’s measured in the unit of the response (cf: $R^2$ is free of unit)
Normality assumption

- **Note**: If we want to perform any hypothesis tests or make any confidence intervals, we will need to assume a distributional form for $\varepsilon$. The usual assumption is:

  $$\varepsilon \sim N(0, \sigma^2 I)$$

  Model: $Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$

  - **Q**: what does the model describe?
    - e.g., $y = \beta_0 + \beta_1 x + \varepsilon$ (see graph)
    - $E(y) = \beta_0 + \beta_1 x$
    - $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ at $x = x_i$

- Some properties of linear models when $\varepsilon \sim N(0, \sigma^2 I)$:

  - Distribution of $Y = X\beta + \varepsilon \sim N(X\beta, \sigma^2 I)$
  - Distribution of $\hat{\beta} = (X^T X)^{-1} X^T Y \sim N(\beta, (X^T X)^{-1}\sigma^2)$
  - Distribution of $\hat{\varepsilon} = (I - H)\varepsilon \sim N(0, (I - H)\sigma^2)$, which has a singular covariance matrix $I - H$ with rank $n - p$ (Note: dim($\hat{\varepsilon}$) = $n - p$)
  - Distribution of $RSS = (n - p)\hat{\sigma}^2 = \hat{\varepsilon}^T \hat{\varepsilon} = \varepsilon^T (I - H)\varepsilon \sim \sigma^2 \chi^2_{n - p}$
  - Distribution of $\hat{Y} = X\hat{\beta} = HY \sim N(X\beta, H\sigma^2)$, which has a singular covariance matrix with rank $p$ (Note: dim($\hat{Y}$) = $p$)
  - $\hat{\beta}$ is independent of $\hat{\sigma}^2$ (Note: cov($(X^T X)^{-1}X^T Y, (I - H)Y$) = 0)
  - $\hat{Y}$ is independent of $\hat{\varepsilon}$ (Note: cov($HY, (I - H)Y$) = 0)
  - Distribution of prediction for a new set of predictors, $x_0 = (g_1(x_{10}, \ldots, x_{m0}), \ldots, g_p(x_{10}, \ldots, x_{m0}))^T$
    - Mean response v.s. future observation (Q: what different?)
    - Example: average yield when $x = x_0^T$ and tomorrow's yield when $x = x_0^T$?
    - Same predicted value $x_0^T \hat{\beta}$, but their distributions are different
    - Distribution of prediction for mean response at $x_0$
      $$x_0^T \hat{\beta} \sim N(x_0^T \beta, (x_0^T (X^T X)^{-1} x_0)\sigma^2)$$
    - Distribution of prediction for future observations at $x_0$
      $$x_0^T \hat{\beta} + \varepsilon \sim N(x_0^T \beta, (x_0^T (X^T X)^{-1} x_0 + 1)\sigma^2)$$
**hypothesis testings** (for $\beta$)

- formulation of hypothesis testing from the view of comparing models

  Ġ y a model space $\equiv$ the space spanned by columns of some $X$

  Ġ y 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 $\dim(\omega)=q$, where $p > q$.

  Ġ y to answer “which of the model spaces is more adequate” in statistical language $\Rightarrow$ perform the test $H_0: \omega$ v.s. $H_1: \Omega \setminus \omega$

  a geometric view:

<table>
<thead>
<tr>
<th>$\hat{y}$</th>
<th>suppose dimension (# of parameters) of $\Omega$ is $p$ and $\dim(\omega)=q$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under the null $H_0: \omega$,</td>
<td></td>
</tr>
<tr>
<td>$\frac{(\text{RSS}<em>\omega - \text{RSS}</em>\Omega)}{(p-q)} \sim \sigma^2 \chi^2_{p-q}$,</td>
<td></td>
</tr>
<tr>
<td>$\frac{\text{RSS}<em>\Omega}{(n-p)} \sim \sigma^2 \chi^2</em>{n-p}$,</td>
<td></td>
</tr>
<tr>
<td>and they are independent.</td>
<td></td>
</tr>
<tr>
<td>So, we have $F = \frac{(\text{RSS}<em>1 - \text{RSS}</em>\Omega) = (p-q)}{\text{RSS}<em>\Omega = (n-p)} \sim F</em>{p-q,n-p}$,</td>
<td></td>
</tr>
<tr>
<td>therefore, reject if $F &gt; F_{p-q,n-p}^{(\alpha)}$ (usually check if $p$-value &lt; $\alpha$)</td>
<td></td>
</tr>
<tr>
<td>Ġ y general form: because the degree of freedom of residuals in a model is the number of observations minus the number of parameter so this test statistics can be written as:</td>
<td></td>
</tr>
<tr>
<td>$F = \frac{(\text{RSS}<em>1 - \text{RSS}</em>\Omega) = (d_f_1 - d_f_\Omega)}{\text{RSS}<em>\Omega = d_f</em>\Omega} \sim F_{d_f_1 - d_f_\Omega, d_f_\Omega}$,</td>
<td></td>
</tr>
<tr>
<td>where $d_f_\omega = \dim(\omega^\prime) = n-q$ and $d_f_\Omega = \dim(\Omega^\prime) = n-p$.</td>
<td></td>
</tr>
<tr>
<td>Ġ y The test is widely used in regression and ANOVA. The beauty of this approach is you only need to know the general form.</td>
<td></td>
</tr>
</tbody>
</table>
• Example 1: test of all predictors

\[ \mathbf{Q}: \text{are any of the predictors useful in predicting the response?} \]

\[ \begin{align*}
\Omega: Y &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \text{dim}(\Omega) = p, \quad df_{\Omega} = n-p \\
\omega: Y &= \beta_0 + \varepsilon, \quad \text{dim}(\omega) = 1, \quad df_{\omega} = n-1 \\
H_0: \beta_1 = \cdots = \beta_p = 0 & \quad H_1: \text{at least one is non-zero} \\
\text{RSS}_\Omega: \frac{1}{2} \sum (\mathbf{y} - \mathbf{\hat{y}}_\Omega)^2 & \quad \text{RSS}_\omega: (\bar{Y} - \bar{\hat{Y}})^2 = \frac{1}{n} \sum (\mathbf{y} - \bar{Y})^2 \\
\text{(the overall F)}: F &= \frac{\text{RSS}_\omega - \text{RSS}_\Omega}{\text{RSS}_\omega} \frac{1}{p-1} \frac{(\mathbf{y} - \mathbf{\hat{y}}_\Omega)^2}{(\mathbf{y} - \bar{Y})^2} \\
\end{align*} \]

• Example 2: testing just one predictor

\[ \mathbf{Q}: \text{Can one particular predictor, say } X_i, \text{ be dropped from the model?} \]

\[ \begin{align*}
\Omega: Y &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \text{dim}(\Omega) = p, \quad df_{\Omega} = n-p \\
\omega: Y &= \beta_0 + \varepsilon, \quad \text{dim}(\omega) = p-1, \quad df_{\omega} = n-p+1 \\
H_0: \beta_i = 0, \quad \beta_j \neq 0 & \quad H_1: \beta_i \neq 0, \quad \beta_j \neq 0 \\
F &= \left( \frac{(\text{RSS}_\omega - \text{RSS}_\Omega)}{(\mathbf{y} - \mathbf{\hat{y}}_\Omega)^2} \right) \left( \frac{1}{p-1} \frac{(\mathbf{y} - \bar{Y})^2}{(\mathbf{y} - \bar{Y})^2} \right) \\
\text{alternative method } t\text{-test: } t_i &= \frac{\hat{\beta}_i / \text{se}(\hat{\beta}_i)}{t_{n-p}} \sim F_{1, n-p} \\
& \quad \text{RSS}_\omega = (n-p) \hat{\sigma}_\omega^2 \\
\end{align*} \]

\[ \mathbf{Y} \text{ ANOVA} \]

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

\[ \begin{align*}
F &= \frac{(\text{RSS}_\omega - \text{RSS}_\Omega)(\mathbf{d}_\omega - \mathbf{d}_\Omega)}{\text{RSS}_{\text{model 5}} \mathbf{d}_\text{model 5}} \sim F_{1, n-p} \mathbf{d}_{\omega} - \mathbf{d}_\Omega \mathbf{d}_{\text{model 5}} \\
\text{invariant to the choice of dummy variables if they generate same } \omega \text{ and } \Omega \\
\end{align*} \]

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

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

- Model: \( Y = X\beta + \varepsilon; \quad \varepsilon \sim N(0, \sigma^2 I) \); \( \hat{\beta} : \) OLS estimator \( \Rightarrow \hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2) \)

\( \bar{y} \) Confidence region for \( A\beta \), where \( A \) is a full rank \( d \times p \) matrix and \( d \leq p \)

\[ A\hat{\beta} \sim N(A\beta, A(X'X)^{-1}A'\sigma^2) \Rightarrow [(A\hat{\beta} - A\beta)'(A(X'X)^{-1}A)\hat{\beta} - A\beta)]/\sigma^2 \sim \chi^2_{d}, \quad (n-p) \hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}, \]

and they are **independent**.

\[ [(A\hat{\beta} - A\beta)'(A(X'X)^{-1}A)\hat{\beta} - A\beta)]/d \hat{\sigma}^2 \sim F_{d,n-p} \]

\( \bar{y} \) 100(1-\( \alpha \))% confidence region: collection of \( \beta \)'s that satisfy

\[ [(A\hat{\beta} - A\beta)'(A(X'X)^{-1}A)\hat{\beta} - A\beta)] \leq d \hat{\sigma}^2 F_{d,n-p}(\alpha) \]

The regions are often ellipsoidally shaped.

- Examples:

\( \bar{y} \) confidence region for \( \beta \), i.e., \( A=I_{p\times p} \)

\[ (\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \leq p \hat{\sigma}^2 F_{p,n-p}(\alpha) \]

\( \bar{y} \) confidence region of \( \beta_i, \beta_p \) i.e., \( A = \begin{pmatrix} 0; \cdots ;0;1;0;\cdots ;0;0;\cdots ;0 \\ 0;\cdots ;0;0;0;\cdots ;0;1;0;\cdots ;0 \end{pmatrix} \)

\[ [(A\hat{\beta} - A\beta)'(A(X'X)^{-1}A)\hat{\beta} - A\beta)] \leq 2 \hat{\sigma}^2 F_{2,n-p}(\alpha) \]

\( \bar{y} \) confidence interval for \( \beta_i \), i.e., \( A=(0,\ldots,0,1,0,\ldots,0) \)

\[ (\hat{\beta}_i - \beta_i)^2/(X'X)^{-1}_{ii} \leq \hat{\sigma}^2 F_{1,n-p}(\alpha) \]

alternative method:

\[ \hat{\beta}_i \sim N(\beta_i, \sigma^2(X'X)^{-1}_{ii}), \quad (n-p) \hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}, \quad \text{and they are independent} \]

\[ \Rightarrow (\hat{\beta}_i - \beta_i)/\hat{\sigma} \sqrt{(X'X)^{-1}_{ii}} \sim t_{n-p} \Rightarrow \text{C.I.}: \quad \hat{\beta}_i \pm t_{n-p}(\alpha/2) \hat{\sigma} \sqrt{(X'X)^{-1}_{ii}} \]

\( \bar{y} \) confidence interval for prediction of mean response at \( x_0 \)

\[ x_0^T\hat{\beta} \sim N(x_0^T\beta, (x_0^T(X'X)^{-1}x_0)\sigma^2) \Rightarrow (x_0^T\hat{\beta} - x_0^T\beta)/\hat{\sigma} \sqrt{x_0^T(X'X)^{-1}x_0} \sim t_{n-p} \]

\[ \Rightarrow \text{C.I.}: \quad x_0^T\hat{\beta} \pm t_{n-p}(\alpha/2) \hat{\sigma} \sqrt{x_0^T(X'X)^{-1}x_0} \]

\( \bar{y} \) C.I. for prediction of future observation at \( x_0 \)

\[ x_0^T\hat{\beta} + \varepsilon \sim N(x_0^T\beta, (x_0^T(X'X)^{-1}x_0 + 1)\sigma^2) \]

\[ \Rightarrow \text{C.I.}: \quad x_0^T\hat{\beta} \pm t_{n-p}(\alpha/2) \hat{\sigma} \sqrt{1 + x_0^T(X'X)^{-1}x_0} \]

\( \bar{y} \) a general form for confidence interval:

\[ \text{estimate} \pm \text{(critical value)} \times \text{(standard error of estimate)} \]
Interpreting parameter estimates

- Q: $Y = X\beta + \epsilon$, what does $\hat{\beta}$ mean?
  
  Some matters needing attention about $\hat{\beta}$:
  
  - $\hat{\beta}$ have units [example, fuel consumption data, fitted model:
    
    $\text{fuel} = 154.19 - 4.23 \text{ Tax} + 0.47 \text{ Dlic} - 6.14 \text{ Income} + 18.54 \log_2(\text{Miles})$
  
  - sign of $\hat{\beta}$: direction of the relationship between the term and the response
  
  - interpretation of estimated value (see next two slides)
  
  - better to also consider values contained in its confidence interval
  
  - causality or association
  
  - some $\beta_i$'s have physical interpretation, especially those from a conceptual model [example: attach weights $x$ to a spring and measure the extension $y$]
    
    $\Rightarrow$ unfortunately, such cases are rare
  
  - usually, $\beta_i$'s do not have such physical interpretation
  
    $\Rightarrow$ in the case, the model $Y = X\beta + \epsilon$ is only an empirical model, i.e., a convenience for representing a complex reality within the range of $X$
    
    $\Rightarrow$ the real meaning of a particular $\beta_i$ is not obvious, interpretation is difficult

- Some interpretations of parameter estimates
  
  - a naive interpretation:
    
    “A unit change in $X_i$ will produce an average change of $\hat{\beta}_i$ in $Y$” $\Rightarrow$ causality assignment
    
    [example, $Y$ is annual income and $X$ is years of education]

  - Q: what if there exist lurking variables?
    
    [example, $X$: shoe size, $Y$: reading abilities, $Z$: age of child]
    
    $\Rightarrow$ causal conclusion is doubtful

  - Q: what if some important predictors are not included in model?
    
    - $X$ fixed. $E(\hat{\beta}_1) = \beta_1 + (X^T_1 X_1)^{-1} X^T_1 X_2 \beta_2$
    
    - $X$ random. true model: $E(Y | X_1, X_2) = X_1 \beta_1 + X_2 \beta_2$
    
      fitted model: $E(Y | X_1) = X_1 \beta_1$
    
    $\text{Var}(Y | X_1) = \sigma^2 + \beta_2^T \text{Var}(X_2 | X_1) \beta_2$

    even though we have all important predictors in the model and no lurking variables, there still are problems. example:
    
    $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon = \beta_0 + (\beta_1 - \beta_2) X_1 + \beta_2 (X_1 + X_2) + \epsilon$

    in a properly designed experiment, the naive interpretation is more reasonable (because of its use of orthogonal designs and randomization); but for observational data, it's often not true.
an alternative interpretation

“\( \hat{\beta}_i \) is the effect of \( X_i \) when all the other (specified) predictors held constant.”

Q: can other predictors be held constant? e.g.

- \( X_1 \) and \( X_2 \) are highly correlated
- consider the model \( \text{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \)

it requires the specification of the other variables.

(Q: what if strong collinearity exists between variables?

\Rightarrow \) estimates and tests of \( \hat{\beta}_i \)'s may significantly change according to what other predictors are included in the model. It can make the interpretation almost impossible (check lab). In some cases, the problem can be removed by redefining the terms into new linear combination that are easier to interpret.)

an interpretation from prediction viewpoint

regarding the parameters and their estimates as fictional quantities, and concentrating on prediction enable a rather cautious interpretation of \( \hat{\beta} \):

given \((x_1^0, \ldots, x_i^0, \ldots, x_{p-1}^0) \rightarrow \hat{y}, \) observe \((x_1^0, \ldots, x_i^0+1, \ldots, x_{p-1}^0) \rightarrow \hat{y} + \hat{\beta}_i \)

prediction is more stable than parameter estimation (check lab)

directly interpretable and success may be measured in future

dangers of extrapolation, be cautious when \( x^0 \) is outside the range of \( X \)

Mean structure

- idea: data are generated from an underlying system, which is assumed to have the form: \( y = f(x_1, \ldots, x_m) + \epsilon \), where \( f \) is unknown.

- regression approximates the mean structure \( f \) by a linear combination of (known) base functions \( g_i(x_1, \ldots, x_m) \)'s, \( i=1, \ldots, p \), i.e.,

\[
f = \sum_{i=1}^{p} \beta_i \cdot g_i(x_1, \ldots, x_m)
\]

\( \hat{y} \) when the structure of \( f \) is simple and almost linear, it can be approximated by a simple structure with fewer terms, e.g.,

\[
E(y) = f \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m
\]

Q: nature is simple?

Q: are there sufficient data to support/fit a complex model?

\( \hat{y} \) when \( f \) is complex and non-linear \( \Rightarrow \) need more terms to get a good approximation

- more parameters, need more degrees of freedom, i.e., more data
- two levels, only linear effects; three levels, linear and quadratic effects

Q: what other complex models?

- base functions for quantitative and qualitative predictors are defined in different ways
Polynomial regression

- one predictor case: \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d + \epsilon \)
- two predictor case: \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \) (2\textsuperscript{nd}-order model)

\( \hat{y} \) the cross product term \( x_1 x_2 \) can be interpreted as an "interaction" effect, e.g.,
\[
E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \text{ where } x_1, x_2 \in \{-1, 1\}
\]
\[ x_1 = -1 \implies y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) x_2 \]
\[ x_1 = 1 \implies y = (\beta_0 - \beta_1) + (\beta_2 - \beta_3) x_2 \]

\( \hat{y} \) models for more predictors can be similarly extended
\[
y = \beta_0 + \sum_{i=1}^{m} \beta_{1,i} x_i + \sum_{i=1}^{m} \beta_{2,i} x_i^2 + \sum_{1 \leq i < j \leq m} \beta_{3,i,j} x_i x_j + \epsilon
\]

- orthogonal polynomials

\( \hat{y} \) polynomial terms can cause numerical instability (especially when \( d \) large) and collinearity

broken stick regression (segmented regression)

- Recall. polynomial regression: suitable for smooth mean structure, but cannot capture local abrupt change
- suppose the break occurs at the known value \( c \), define the base function (where \( c \) is called a knot):
\[
d_c(x) = \begin{cases} 1, & \text{if } x > c, \\ 0, & \text{if } x \leq c. \end{cases}
\]

- model: \( y = \beta_0 + \beta_1 x + \beta_2 (x-c) d_c(x) + \epsilon \)
\[
E(y) = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x \leq c, \\ (\beta_0 - \beta_2 c) + (\beta_1 + \beta_2) x, & \text{if } x > c, \end{cases}
\]

\( \hat{y} \) the two lines meet at \( c \) \( \Rightarrow \) continuous fit
\( \hat{y} \) notice only 3 parameters in the model \( \Rightarrow \) one degree of freedom is saved because of the continuity restriction
**dummy variable (indicator variable, coding)**

- categorical (qualitative) predictor
  - nominal v.s. ordinal
  - examples: male/female, treatment/control, eye colors, blocks, ...
  - qualitative in nature:
    - values are symbols, no quantitative meaning
    - no value exist between categories
  - **Q**: what properties can we explore for qualitative predictor?
    - category $i \rightarrow y_{ij}$, $\mu_i = E(y_{ij})$ ⇒ can only study difference between $\mu_i$’s
      (cf., quantitative predictor)
  - **Q**: how to fit these predictors into the format of linear regression model
    - $Y = X\beta + \epsilon$? ⇒ Ans: dummy variables

- one dichotomous predictor: two categories
  - for a dichotomous predictor $C$ with two categories $c_1$ and $c_2$, define a dummy variable $d$:
    $$d(C) = \begin{cases} 0, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases}$$
  - for a data set with response $y$, one quantitative predictor $x$, and one qualitative predictor $C$ (dummy variable $d$), possible models are:
    - model 1: $y = \beta_0 + \beta_1 d + \epsilon$
    - model 2: $y = \beta_0 + \beta_1 x + \epsilon$
    - model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon$
    - model 4: $y = \beta_0 + \beta_1 x + \beta_2 xd + \epsilon$
    - model 5: $y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 xd + \epsilon$

- **Q**: how to interpret $\beta_i$’s in models 1~5?
  - model 1: $y = \beta_0 + \beta_1 d + \epsilon$
    - $C = c_1$ : $\mu_1 = E(y|d = 0) = \beta_0 \implies \beta_0 = \mu_1$
    - $C = c_2$ : $\mu_2 = E(y|d = 1) = \beta_0 + \beta_1 \implies \beta_1 = \mu_2 - \mu_1$
  - model 2: $y = \beta_0 + \beta_1 x + \epsilon$
  - model 3: $y = \beta_0 + \beta_1 d + \beta_2 x + \epsilon$
    - $C = c_1$ : $\mu_{1,x} = E(y|d = 0, x) = \beta_0 + \beta_2 x$
    - $C = c_2$ : $\mu_{2,x} = E(y|d = 1, x) = (\beta_0 + \beta_1) + \beta_2 x$
    - $\beta_0 = \mu_{1,0}$ (intercept in $c_1$ group)
    - $\beta_1 = \mu_{2,x} - \mu_{1,x}$ (difference of intercepts)
    - $\beta_2$ = slope (same slope in two categories)
  - model 4: $y = \beta_0 + \beta_1 x + \beta_2 (d \cdot x) + \epsilon$
    - $C = c_1$ : $\mu_{1,x} = E(y|d = 0, x) = \beta_0 + \beta_1 x$
    - $C = c_2$ : $\mu_{2,x} = E(y|d = 1, x) = \beta_0 + (\beta_1 + \beta_2)x$
    - $\beta_0 = \mu_{1,0} = \mu_{2,0}$ (same intercept in two categories)
    - $\beta_1$ = slope of category $c_1$
    - $\beta_2$ = difference in slopes
model 5: \( y = \beta_0 + \beta_1 d + \beta_2 x + \beta_3 (d \cdot x) + \epsilon \)

\[ C = c_1 : \quad \mu_{1,x} = E(y|d = 0, x) = \beta_0 + \beta_2 x \]
\[ C = c_2 : \quad \mu_{2,x} = E(y|d = 1, x) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x \]

\( \beta_0 = \mu_{1,0} \) (intercept of category \( c_1 \))
\( \beta_2 = \) slope of category \( c_1 \)
\( \beta_1 = \) difference in intercepts
\( \beta_3 = \) difference in slopes

\( d(C) = \begin{cases} -1, & \text{if } C = c_1, \\ 1, & \text{if } C = c_2. \end{cases} \)

**Q:** how to interpret \( \beta_i \)'s in models 1~5 under this coding?

```
model 1: \( y = \beta_0 + \beta_1 d + \epsilon \)
\[ C = c_1 : \quad \mu_1 = E(y|d = -1) = \beta_0 - \beta_1 \quad \Rightarrow \quad \beta_0 = \frac{\mu_1 + \mu_2}{2} \]
\[ C = c_2 : \quad \mu_2 = E(y|d = 1) = \beta_0 + \beta_1 \quad \Rightarrow \quad \beta_1 = \frac{\mu_2 - \mu_1}{2} \]
```

**analysis of covariance:** testing model 3 (\( \Omega \)) against model 2 (\( \omega \)) (more than 2 categories and more quantitative predictors allowed). The quantitative predictor is called **covariate** and is expected to have the same effect in all categories. The difference between categories is assumed to be an additive effect.

- **one polytomous predictor**: more than two categories
  - for \( k \) categories, \( k-1 \) dummy variables are needed to depict the variation between categories (one parameter is used to represent constant term)

**various coding of dummy variables:** 4 categories \( c_1, c_2, c_3, c_4 \) example

<table>
<thead>
<tr>
<th>treatment coding</th>
<th>helmert coding</th>
<th>sum coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_1 )</td>
<td>( d_1 )</td>
<td>( d_1 )</td>
</tr>
<tr>
<td>( c_2 )</td>
<td>( -1 )</td>
<td>( -1 )</td>
</tr>
<tr>
<td>( c_3 )</td>
<td>( 1 )</td>
<td>( -1 )</td>
</tr>
<tr>
<td>( c_4 )</td>
<td>( 0 )</td>
<td>( 2 )</td>
</tr>
<tr>
<td>( c_5 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
</tbody>
</table>

**consider the model:** \( y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_3 d_3 + \epsilon \)

**properties of treatment coding:**
\[ C = c_1 : \quad \mu_1 = E(y|d_1 = 0, d_2 = 0, d_3 = 0) = \beta_0 \]
\[ C = c_2 : \quad \mu_2 = E(y|d_1 = 1, d_2 = 0, d_3 = 0) = \beta_0 + \beta_1 \quad \Rightarrow \quad \beta_0 = \mu_1 \]
\[ C = c_3 : \quad \mu_3 = E(y|d_1 = 0, d_2 = 1, d_3 = 0) = \beta_0 + \beta_2 \quad \Rightarrow \quad \beta_1 = \mu_2 - \mu_1 \]
\[ C = c_4 : \quad \mu_4 = E(y|d_1 = 0, d_2 = 0, d_3 = 1) = \beta_0 + \beta_3 \quad \Rightarrow \quad \beta_2 = \mu_3 - \mu_1 \]
\[ C = c_5 : \quad \mu_5 = E(y|d_1 = 0, d_2 = 0, d_3 = 0) = \beta_0 + \beta_3 \quad \Rightarrow \quad \beta_3 = \mu_4 - \mu_1 \]
treats \( c_1 \) as a reference

it's convenient if a "standard" categories exists

\( d_1, d_2, \) and \( d_3 \) are mutually orthogonal, but not orthogonal to constant term

properties of helmert coding:

\[
\begin{align*}
C &= c_1 : & \mu_1 &= E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\
C &= c_2 : & \mu_2 &= E(y|d_1 = 1, d_2 = -1, d_3 = -1) = \beta_0 + \beta_1 - \beta_2 - \beta_3 \\
C &= c_3 : & \mu_3 &= E(y|d_1 = 0, d_2 = 2, d_3 = -1) = \beta_0 + 2\beta_2 - \beta_3 \\
C &= c_4 : & \mu_4 &= E(y|d_1 = 0, d_2 = 0, d_3 = 3) = \beta_0 + 3\beta_3 \\
\beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \equiv \bar{\mu} \\
\beta_1 &= \frac{\mu_2 - \mu_1}{2} \\
\beta_2 &= \frac{\mu_3 - ((\mu_1 + \mu_2)/2)}{3} \\
\beta_3 &= \frac{\mu_4 - ((\mu_1 + \mu_2 + \mu_3)/3)}{4}
\end{align*}
\]

constant term, \( d_1, d_2, \) and \( d_3 \) are orthogonal when there are equal # of observations in each categories

hard to interpret parameters

may suitable for ordinal qualitative predictor

properties of sum coding:

\[
\begin{align*}
C &= c_1 : & \mu_1 &= E(y|d_1 = -1, d_2 = -1, d_3 = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3 \\
C &= c_2 : & \mu_2 &= E(y|d_1 = 1, d_2 = 0, d_3 = 0) = \beta_0 + \beta_1 \\
C &= c_3 : & \mu_3 &= E(y|d_1 = 0, d_2 = 1, d_3 = 0) = \beta_0 + \beta_2 \\
C &= c_4 : & \mu_4 &= E(y|d_1 = 0, d_2 = 0, d_3 = 1) = \beta_0 + \beta_3 \\
\beta_0 &= \frac{\mu_1 + \mu_2 + \mu_3 + \mu_4}{4} \equiv \bar{\mu} \\
\beta_1 &= \mu_2 - \bar{\mu} \\
\beta_2 &= \mu_3 - \bar{\mu} \\
\beta_3 &= \mu_4 - \bar{\mu}
\end{align*}
\]

\( \bar{\mu} \) represent overall mean

compare each category with the overall mean

lesser orthogonal

Note: the choice of coding does not affect the \( R^2, \hat{\Sigma} \) and overall \( F \)-test (to test \( H_0: \beta_1 = \beta_2 = \beta_3 = 0 \), the three codings have same \( \omega \) and \( \Omega \))

the overall \( F \)-test is one-way ANOVA (ANalysis Of VAriance)

Q: how to work with quantitative predictors? \( \Rightarrow \) identical methodology as in 2 categories case. Q: how to interpret parameters in the case?
• two qualitative predictors (say, \(A\): 3 categories \(a_1, a_2, a_3\); \(B\): 4 categories, \(b_1, b_2, b_3, b_4\))

\(\hat{y}\) number of different category combinations = \(3 \times 4 = 12\),
denote their means as \(\mu_{ij}\), \(i=1, 2, 3\) and \(j=1, 2, 3, 4\), i.e.,

\[ y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad k = 1, 2, \ldots, n_{ij}, \]

\(n_{ij}\) = number of observations in category \(A=a_i\) and \(B=b_j\)

\(\hat{y}\) \(Q\): how to depict the difference between \(\mu_{ij}\)'s?

consider the following linear models:

\(n\) model 1: \(E(y_{ijk}) = \beta_0\)

\(n\) model 2: \(E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A\)

\(n\) model 3: \(E(y_{ijk}) = \beta_0 + \beta_1 d_1^B + \beta_2 d_2^B + \beta_3 d_3^B\)

\(n\) model 4: \(E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B\)

\(Q\): how to perform interaction coding? what is interaction?

\(\text{interaction plot: replace } \mu_{ij}\text{'s by cell means }\)

\[ \hat{y}_{ij} = \sum_{k=1}^{n_{ij}} y_{ijk} / n_{ij} \]

\(n\) model 5:

\[ E(y_{ijk}) = \beta_0 + \beta_1 d_1^A + \beta_2 d_2^A + \beta_3 d_1^B + \beta_4 d_2^B + \beta_5 d_3^B + \sum_{i=1}^{2} \sum_{j=1}^{3} \beta_{ij} d_{ij} \]

\# of parameters: \(1 + 2 + 3 + 6 = 12\)

• identical methodology applies for more qualitative (3-factor interaction, 4-factor interaction, …) and quantitative predictors

---

**Transformation**

• transformation of response

\(\hat{y}\) Box-Cox transformation family: \(t_\lambda(y) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log(y), & \text{if } \lambda = 0. \end{cases}\)

\(n\) \(t_\lambda(y)\) is continuous in \(\lambda\) for fixed \(y>0\),

\[ \lim_{\lambda \to 0} t_\lambda(y) = \lim_{\lambda \to 0} (y^\lambda - 1)/\lambda = \lim_{\lambda \to 0} (y^\lambda \log(y))/1 = \log(y) \]

\(n\) \(\lambda=1 \Rightarrow\) no transformation, \(\lambda=0 \Rightarrow\) log, \(\lambda \neq 0\) or \(1 \Rightarrow\) power transformation

\(n\) model: \(t_\lambda(y) = X\beta + \varepsilon, \varepsilon \sim \text{N}(0, \sigma^2 I)\)

\(p\) parameters: \(\lambda, \beta, \sigma\)

\(p\) can write down likelihood for estimation and testing of \(\lambda\)

\(p\) choice of transformation becomes a estimation/test problem

\(n\) the log-likelihood is

\[ L(\lambda) = (-n/2) \log(\text{RSS}_\lambda/n) + (\lambda - 1) \sum \log(y_i) \]

goodness of fit adjustment

where RSS\(_\lambda\) is the residual sum of square when using \(t_\lambda(y)\) as response, i.e.,

\[ \text{RSS}_\lambda = [t_\lambda(y)]^T (I - H) t_\lambda(y) \]
estimation of \( \lambda \): choose \( \lambda \) to fit data well using maximum likelihood.

- can compute \( L(\lambda) \), for various values of \( \lambda \) and compute \( \hat{\lambda} \) exactly to maximize \( L(\lambda) \)
- but usually \( \hat{\lambda} \) is not a nice round number, e.g., \( \hat{\lambda} = -0.17 \). It would be hard to explain what this new response means.
- to avoid this, maximize \( L(\lambda) \) over a grid of values, such as \{ -2, -1, -1/2, 0, 1/2, 1, 2 \}. This helps with interpretation.
- for \( \lambda \) outside \([-2, 2]\), pay more attention on whether such transformation is required

test of \( \lambda \): is the transformation really necessary?
- we can answer the question form a C.I. for \( \lambda \)
- likelihood ratio test:
  \[
  -2(L(\lambda) - L(\hat{\lambda})) \sim \chi^2_l
  \]
- a 100(1-\(\alpha\))% C.I for \( \lambda \) can be formed by:
  \[
  \{ \lambda \mid L(\lambda) > L(\hat{\lambda}) - (1/2) \chi^2_l(1-\alpha) \}
  \]
- is \( \lambda = 1 \) in the interval? if so, may as well stay with no transformation.
- if rounding \( \hat{\lambda} \), check that rounded value is in the C.I.

Generalized Least Square (GLS)

- model: \( Y = X\beta + \varepsilon, E(\varepsilon) = 0 \) and \( \text{var}(\varepsilon) = \sigma^2 I \Rightarrow \varepsilon: \) uncorrelated and constant variance
- Consider the case \( \text{var}(\varepsilon) = \sigma^2 \Sigma \), where \( \Sigma (\neq I) \) is known but \( \sigma^2 \) is unknown, i.e., we know the correlation and relative variance between the errors but we don't know the absolute scale
- Because \( \Sigma_{\text{null}} \) is symmetric and positive definite, we can write \( \Sigma = SS^T \), where \( S \) is an \( n \times n \) nonsigular matrix (by Choleski decomposition or spectral decomposition)

\[
Y = X\beta + \varepsilon \Rightarrow S^{-1} Y = S^{-1} X\beta + S^{-1} \varepsilon \Rightarrow Y' = X'\beta + \varepsilon', \text{ where}
\]

\[
Y' = S^{-1} Y, \quad X' = S^{-1} X, \quad \varepsilon' = S^{-1} \varepsilon, \text{ and}\\
E(\varepsilon') = 0 \text{ and } \text{var}(\varepsilon') = S^{-1} \text{var}(\varepsilon) S^{-T} = S^{-1} \sigma^2 SS^T S^{-T} = \sigma^2 I
\]

\( \Rightarrow \) For \( Y' \) and \( X' \), the assumption in ordinary least square is satisfied
- GLS: find \( \hat{\beta} \) that minimize
  \[
  \varepsilon'^T \varepsilon' = (Y' - X'\hat{\beta})^T (Y' - X'\hat{\beta}) = (Y - X\beta)^T S^{-T} S^{-1} (Y - X\beta) = (Y - X\beta)^T \Sigma^{-1} (Y - X\beta)
  \]
  \[
  \Rightarrow \hat{\beta} = (X'^T X')^{-1} X'^T Y' = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y
  \]
  \[
  \Rightarrow \text{var}(\hat{\beta}) = \sigma^2 (X'^T X')^{-1} = \sigma^2 (X^T \Sigma^{-1} X)^{-1}
  \]
- GLS is like regressing \( X' = S^{-1} X \) on \( Y' = S^{-1} Y \)
- The practical problem is that \( \Sigma \) may not be known. It's usually necessary to make some assumptions and examine the residuals to estimate \( \Sigma \)
**Weighted Least Square (WLS)**  
- Sometimes, the errors are uncorrelated, but have unequal variance where the form of the inequality is known ($\Rightarrow \Sigma$ is diagonal, it's a special case of GLS), example:

  - $\varepsilon$: uncorrelated, but not constant variance $\Rightarrow \Sigma$ is diagonal. Write

    $$
    \Sigma = \begin{pmatrix}
    1/w_1 & 0 & \cdots & 0 \\
    0 & 1/w_2 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & 1/w_n
    \end{pmatrix} \Rightarrow \Sigma^{-1} = \begin{pmatrix}
    w_1 & 0 & \cdots & 0 \\
    0 & w_2 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & w_n
    \end{pmatrix}
    $$

  - where $w_i$'s ($=1/\text{var}(\varepsilon_i)$) are called *weights*.

  - low weight $\Leftrightarrow$ high variance; high weight $\Leftrightarrow$ low variance

- $S = \text{diag}(1/\sqrt{w_1}, \ldots, 1/\sqrt{w_n})$, then $\Sigma = SS^T$

  $\Rightarrow$ regress $S^{-1}X$ (i.e., $\sqrt{w_i}x_i$) on $S^{-1}Y$ ($\sqrt{w_i}y_i$) (Note. the column of ones needs to be replaced with $\sqrt{w_i}$)

  $\Rightarrow$ convenient for regression package without a weighted options

---

**iteratively re-weighted least squares (IRWLS):** In all the previous examples, weights are assumed known. **Q:** what if $\text{var}(\varepsilon_i)$ is not completely known, what weights should we use? where can you find the information of weights?

- $\ddot{Y}$ model the mean response for $Y$, $E(Y) = X\beta$

- $\ddot{Y}$ model the variance in $Y$, $\text{var}(Y) = f(X, \rho)$, where $\rho$ are parameters for the variance model

  - residuals $\Rightarrow$ fit the mean model to estimate $\beta$
  - weights $\Rightarrow$ fit the variance model to estimate $\rho$

- $\dddot{Y}$ Example: $\text{var}(\varepsilon_i) = (\rho_0 + \rho_1 x_{ij})\sigma^2$

  1. start with $w_i = 1$
  2. use weighted least square to estimate $\beta$
  3. use the residuals to estimate $\rho_0$ and $\rho_1$, perhaps by regressing residuals on $x_i$
  4. re-compute the weights and go to 2. Continue until convergence

**Problems:** converge? how is the inference about $\beta$ affected? d.f.? ...etc

- $\dddot{Y}$ alternative approach: jointly estimate the regression and weighting parameters using likelihood based method (in R, use `gls()` function in the `nlme` library)
Diagnostics

- **regression diagnostics**: check model assumptions to suggest further improvement after fitting. *The building of an empirical model is an iterative process.* During the process, it is required to check whether the current fitted model is consistent with data.

- **Q**: what assumptions needed to be checked?

  model: \( Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I) \)

  - error structure: errors independent, equal variance, normally distributed
  - mean structure: whether \( E(Y) = X\beta \) is a correct structure
  - unusual observations: whether some observations do not fit the model

- two types of diagnostic techniques: numerical and graphical

Residual

- residuals are useful for detecting lack of fit and checking model assumptions

  (Q: Why residuals can do the works?)

  \[ Y = X\beta + \varepsilon = \hat{Y} + \hat{\varepsilon} \]

  \[ Y = X_1\beta_1 + X_2\beta_2 + \varepsilon = (X_1\beta_1 + H_1X_2\beta_2) + ((I-H_1)X_2\beta_2 + \varepsilon) = \hat{Y}_{x_1} + \hat{\varepsilon}_{x_1} \]

Leverage

- leverage: \( H_{ii} \equiv h_i \quad \text{(Note 1. var(\hat{\varepsilon}_i) = (1-h_i)\sigma^2. Note 2. \ h_i \text{ is known before observing } Y)} \)
  - \( x_i \) whose \( h_i \) is large \( \Rightarrow \) var(\hat{\varepsilon}_i) small \( \Rightarrow \) fitted model has to force to fit close to \( y_i \)
  - \( x_i \) whose \( h_i \) is small \( \Rightarrow \) var(\hat{\varepsilon}_i) large \( \Rightarrow \) in this \( x_i \), model cannot fit so well
  - \( \hat{y}_i \) roughly determines how close \((x_i, y_i)\) to the regression surface (i.e., \((x_i, \hat{y}_i)\))
  - \( \hat{y}_i \) observation with large \( h_i \)'s should be paid more attention.

- (internally) studentized residuals \( r_i \)'s:

  because \( \text{var}(\hat{\varepsilon}_i) = (1-h_i)\sigma^2 \), let \( r_i = \frac{\hat{\varepsilon}_i}{\sigma \sqrt{1-h_i}} \), then \( \text{var}(r_i) \approx 1 \) (if model assumptions are correct)

  - \( \hat{y} \) non-constant variance removed
  - \( \hat{y} \) dependence is very small in practice
  - \( \hat{y} \) sum of \( r_i \)'s is not zero
  - \( \hat{y} \) \( r_i \) is slightly correlated with \( \hat{y}_i \)
  - \( \hat{y} \) studentized residuals are preferred in residual plots
  - \( \hat{y} \) if there is some underlying heteroscedasticity (i.e., violation of \( \text{var}(\varepsilon) = \sigma^2 I \)) in the errors, studentization cannot correct it
Outlier

- an outlier is a point that does not fit the current model
  ⇒ usually, large residual

- \( \textbf{Q:} \) is there a problem if (studentized) residuals are used to detect outliers?
  ⇒ outliers may affect the fit (see plot)

- idea: exclude \( i \)th observation and re-compute the estimates to get \( \hat{\beta}_{(i)} \) and \( \hat{\sigma}^2_{(i)} \), where \( (i) \) denotes that the \( i \)th case has been excluded. Then, consider \( y_i - \hat{y}_{(i)} \), where \( \hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)} \). (\( \textbf{Q:} \) why is it better in detecting outliers?)

\[
\text{var}(y_i - \hat{y}_{(i)}) = \sigma^2(1 + x_i^T (X_{(i)}^T X_{(i)}) x_i)
\]

(Hint: prediction of future observation)

- jacknife (or externally studentized, or crossvalidated) residuals

\[
t_i = (y_i - \hat{y}_{(i)}) / \left( (1 + x_i^T (X_{(i)}^T X_{(i)}) x_i) \right)^{1/2} \hat{\sigma}_{(i)}
\]

which are distributed as \( t_{(n-1)-p} \) under null, if model is correct and \( \varepsilon \sim \mathcal{N}(\theta, \sigma^2 I) \)

- a simpler way to calculate \( t_i \) (avoid doing \( n \) regression)

\[
t_i = \hat{\varepsilon}_i / (1-h_i)^{1/2} \hat{\sigma}_{(i)} = r_i ((n-p-1)/(n-p-r_i^2))^{1/2}
\]

- test for outliers

\( \hat{\varepsilon}_i \) given a specific case \( i \), conclude an outlier if \( |t_i| > t_{n-p-1}(\alpha/2) \)

\( \text{in practice, a few (or all) } t_i \)'s will be tested ⇒ problem of \textit{multiple testing}: need to adjust the level of the test accordingly

\( H_0: \) no outlier in the \( n \) observations against \( H_i: \) at least one outlier

\[
1 - \alpha^* = 1 - \text{Prob(Type I error } | H_0) = \text{Prob(all tests accept } | H_0) = 1 - \text{Prob(at least one rejected } | H_0) \geq 1 - \Sigma_i \text{Prob(test } i \text{ rejects } | H_0) = 1 - n\alpha
\]

⇒ conclude an outlier if \( |t_i| > t_{n-p-1}(\alpha/2n) \)

⇒ it's conservative, tends \textit{not} to label points as outliers (especially when \( n \) large)

Influential observation

- Each observations have different influence/contribution to the fitted model. Our fitted model should not change too much (i.e., robust) just because of adding/dropping a specific observation.

- influential point: one whose removal from data would \textit{cause large change in the fit}.

- an influential point \textit{may or may not} be an outlier and \textit{may or may not} have large leverage but it will tend to have at least one of those two properties.

- measures of influence (\( \textbf{Q:} \) how to numerically characterize “large change in fit”?)

\( \hat{y} \) change in coefficients: \( \hat{\beta} - \hat{\beta}_{(i)} \) (\( \textbf{Q:} \) how large is large?)

\( \hat{y} \) change in fit: \( \hat{y} - \hat{y}_{(i)} \) (\( \textbf{Q:} \) how large is large?)
Q1: what kinds of points are more influential?

Q2: what information should be included in the detection of influential points?

\[ \hat{y} \quad \text{outlier} \quad \hat{y} \quad \text{not outlier} \quad \hat{y} \quad \text{low leverage} \quad \hat{y} \quad \text{high leverage} \Rightarrow \text{high influence} \]

\( \hat{y} \) Cook statistics/distances (scale and unit free): 

\[
D_i = \left( \hat{\beta} - \beta_0 \right)^T (X^T X)(\hat{\beta} - \beta_0) / (p \hat{\sigma}^2) \\
= \left( \hat{y} - \hat{y}_0 \right)^T \left( \hat{y} - \hat{y}_0 \right) / (p \hat{\sigma}^2) \\
= \left( 1/p \right) r_i^2 \left( h_i / (1 - h_i) \right)
\]

⇒ it’s a combination of residual and leverage. (Q: what are the effects of residual and leverage on Cook’s statistic?)

Residual plots

- residual plots: plot residuals (or absolute values of residuals) against (i) \( \hat{y} \), (ii) \( x_k \) (for predictors in model and not in model), (iii) combination of \( x_k \)'s, (iv) time order (if available), (v) any other quantities that we think relevant to residuals

- in residual plots,
  - \( \hat{y} \) find overall patterns from the shape of all points (cf., residuals used in checking outliers or influential points ⇒ identifying individually unusual point)
  - \( \hat{y} \) check assumptions: (i) non-constant variance; (ii) incorrectly specified mean structure (i.e., \( E(Y) = X\beta \))
  - \( \hat{y} \) rather subjective
  - a satisfactory residual plot (null plot)
    - \( \hat{y} \) constant variance
    - \( \hat{y} \) no curvature in the mean of residuals
  - Note: one satisfactory residual plot cannot guarantee the residual plots for other variables will be satisfactory

- some possible patterns in unsatisfactory residual plots:
  - evidence of non-constant variance
  - curvature in the mean of residuals
    ⇒ evidence of incorrectly specified mean structure
  - evidence of non-constant variance and incorrectly specified mean structure
• possible remedies for unsatisfactory residual plots

<table>
<thead>
<tr>
<th>unsatisfactory residual plot</th>
<th>plot residuals against ...</th>
<th>( x_k )</th>
<th>time order</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-constant variance</td>
<td>1. weighted least square</td>
<td>1. weighted least square</td>
<td>weighted least square</td>
</tr>
<tr>
<td>curvature in mean structure</td>
<td>2. transform ( y )</td>
<td>2. transform ( y )</td>
<td>add term of time in model</td>
</tr>
</tbody>
</table>

| Various plots and tests for diagnostics |

- **Q-Q plot**
  - \( \hat{Y} \): we often see the statement “\( z_1, z_2, ..., z_m \) are i.i.d. from a cdf \( F \)”, how to examine if \( F \) is an appropriate distribution assumption for \( z_i \)’s? (Hint: examine the similarity btw cdf and empirical cdf)
  - \( \hat{Y} \): normal (probability) plot: assessing normality assumption of \( \varepsilon \)
    - (Note: tests and C.I. depend on normality assumption)
    - 1. sort the data \( \hat{\varepsilon}_{(1)} \leq \hat{\varepsilon}_{(2)} \leq \ldots \leq \hat{\varepsilon}_{(n)} \)
    - 2. plot \( \hat{\varepsilon}_{(i)} \) against \( F^{-1}(i/(n+1)) \), where \( F \) is the cdf of N(0, 1)
    - If the residuals are normally distributed, an approximately straight-line relationship will be observed (null plot)

- **non-normality**: long-tail, short-tail, asymmetric
  - worst case is long-tail; mild non-normality can safely be ignored; the larger the sample size, the less troublesome the non-normality
  - for long-tail, (i) use test based on other distributions, or bootstrap, or permutation tests (ii) for estimation, use robust methods (e.g., least absolute deviation instead of least square)
  - asymmetric, transform \( Y \) (e.g., Box-Cox method)
  - short-tail can be reasonably ignored
- formal tests exists (such as Kolmogorov-Smirnov test), but not as flexible as the Q-Q plot
  - normal plot can be applied to identify extreme values (e.g., in residuals, leverages, Cook’s statistics, …): in the case, not interested in a straight line relationship, but rather looking for points that depart from the straight line
Analysis strategy and model uncertainty

- you have learned
  - Parameter estimation and testing: LS estimator, generalized LS, $t$-test, $F$-tests, lack-of-fit, C.I., $R^2$, prediction, ...
  - Diagnostics: checking assumptions, such as constant variance, linearity, normality, outliers, influential points, serial correlation, collinearity, ...
  - Transformation: transforming the response and/or the predictors, Box-Cox, polynomial regression, broken line, dummy variables, ...

- Q: what order should these be done? should procedures be repeated at later stage? when should we stop?

- a recommended analysis strategy:

  ![Diagram]

  a reasonable regression model $\rightarrow$ Diagnostics $\rightarrow$ Transformation $\rightarrow$ Variable Selection $\rightarrow$ Diagnostics $\rightarrow$ Stop

Note: there is no hard-and-fast rules about how it should be done. Regression analysis is a search for structure in data. Better to try a variety of orders.

v Reading: Faraway (2006), chapter 1