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.

Mortality rate (M)	102.5	104.5	100.4	95.9	87.0	95.0	88.6	89.2
Temperature (T)	51.3	49.9	50.0	49.2	48.5	47.8	47.3	45.1
Mortality rate (M)	78.9	84.6	81.7	72.2	65.1	68.1	67.3	52.5
Temperature (T)	46.3	42.1	44.2	43.5	42.3	40.2	31.8	34.0

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

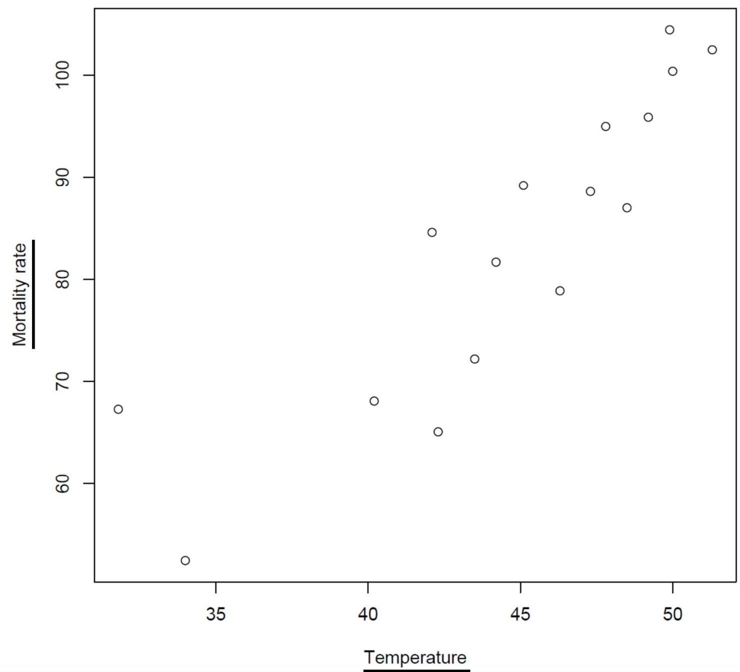


Figure: Scatter Plot of Temperature versus Mortality Rate, Breast Cancer Data.

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

$$\sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2.$$

- **Least Squares Estimates**

  Estimated Coefficients :

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

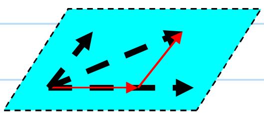
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2)} \right),$$

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

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

ANOVA Table for Simple Linear Regression

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	1	$\hat{\beta}_1^2 \sum(x_i - \bar{x})^2$	$\hat{\beta}_1^2 \sum(x_i - \bar{x})^2$
residual	$N-2$	$\sum_{i=1}^N (y_i - \hat{y}_i)^2$	$\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{(N-2)}$
total (corrected)	$N-1$	$\sum_{i=1}^N (y_i - \bar{y})^2$	

ANOVA Table for Breast Cancer Example

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	1	2599.53	2599.53
residual	14	796.91	56.92
total (corrected)	15	3396.44	

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

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 $=$ 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 α.

Confidence Interval

$100(1-\alpha)\%$ confidence interval for β_j is given by

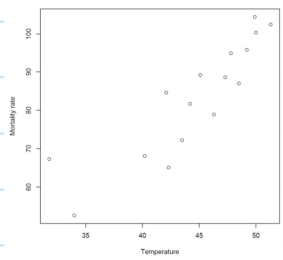
$$\hat{\beta}_j \pm t_{N-2,\frac{\alpha}{2}} \times s.d.(\hat{\beta}_j),$$

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 β_j does not contain 0, then H_0 is rejected.

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

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

Predictor        Coef    SE Coef       T       P
Constant       -21.79      15.67   -1.39   0.186
T              2.3577     0.3489    6.76   0.000

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

Analysis of Variance

Source          DF       SS       MS       F       P
Regression       1   2599.5   2599.5   45.67   0.000
Residual Error  14    796.9     56.9
Total           15   3396.4

Unusual Observations
Obs       T       M     Fit   SE Fit   Residual   St Resid
 15    31.8   67.30   53.18     4.85      14.12     2.44RX

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

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

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

Predictor         Coef    SE Coef        T        P
Constant        -52.62      15.82    -3.33    0.005
T               3.0152     0.3466     8.70    0.000

S = 5.93258     R-Sq = 85.3%      R-Sq(adj) = 84.2%

Analysis of Variance

Source           DF         SS         MS        F        P
Regression        1     2664.3     2664.3    75.70    0.000
Residual Error   13      457.5       35.2
Total            14     3121.9

Unusual Observations
Obs        T        M       Fit      SE Fit     Residual     St Resid
 15     34.0    52.50     49.90        4.25         2.60       0.63 X

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

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

Residual Plots After Outlier Removal

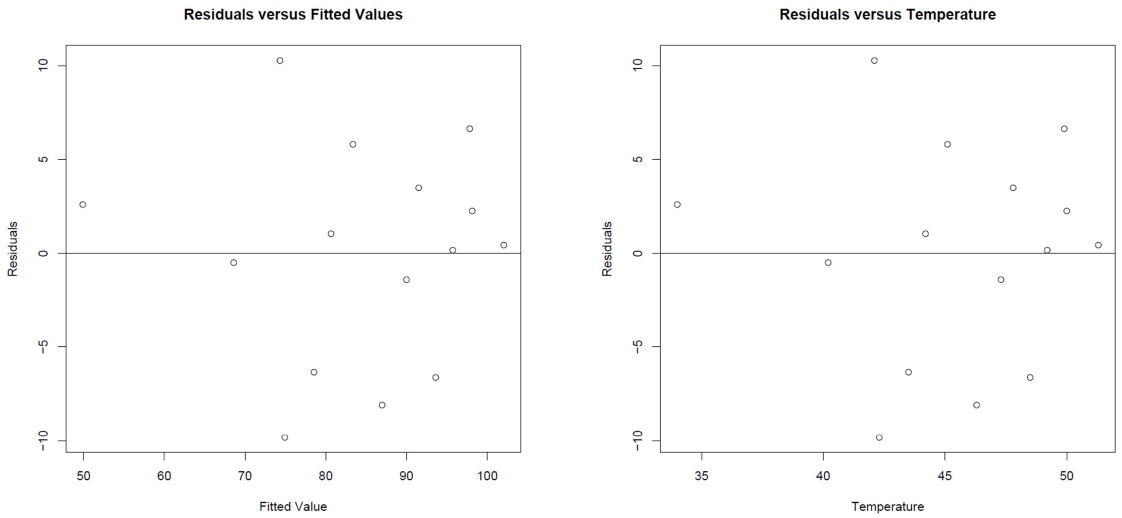


Figure 3: Residual Plots

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 σ^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

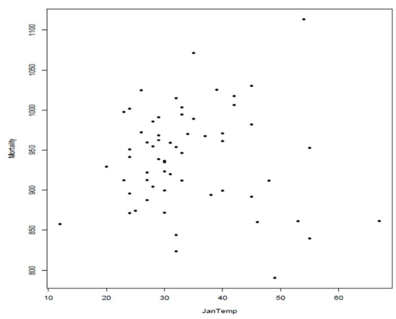
- There are 60 data points.
- Pollution variables are highly skewed, log transformation makes them nearly symmetric. The variables HCPot, NOxPot and SO2Pot are replaced by log(HCPot), log(NOxPot) and log(SO2Pot).
- Observation 21 (Fort Worth, TX) has two missing values, so this data point will be discarded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

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

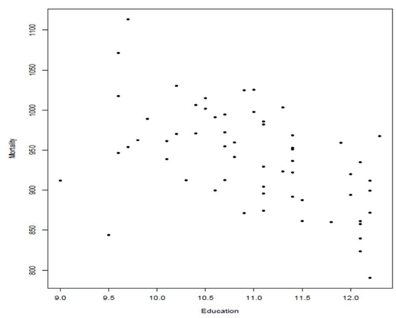
Scatter Plots

Figure 4: Scatter Plots of mortality against selected predictors

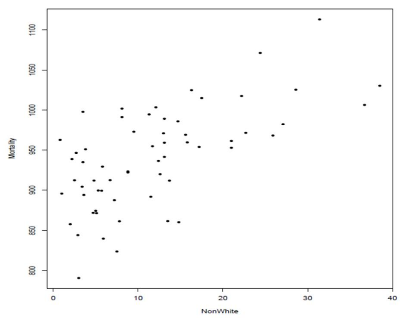
(a) JanTemp



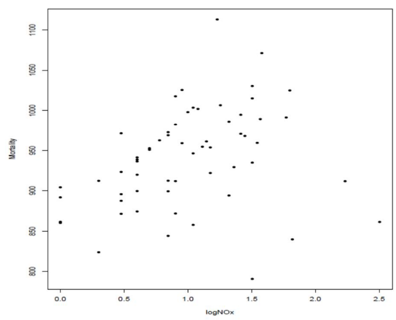
(b) Education



(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

$$\sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \right)^2 = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta).$$

- **Least Squares estimates :**

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

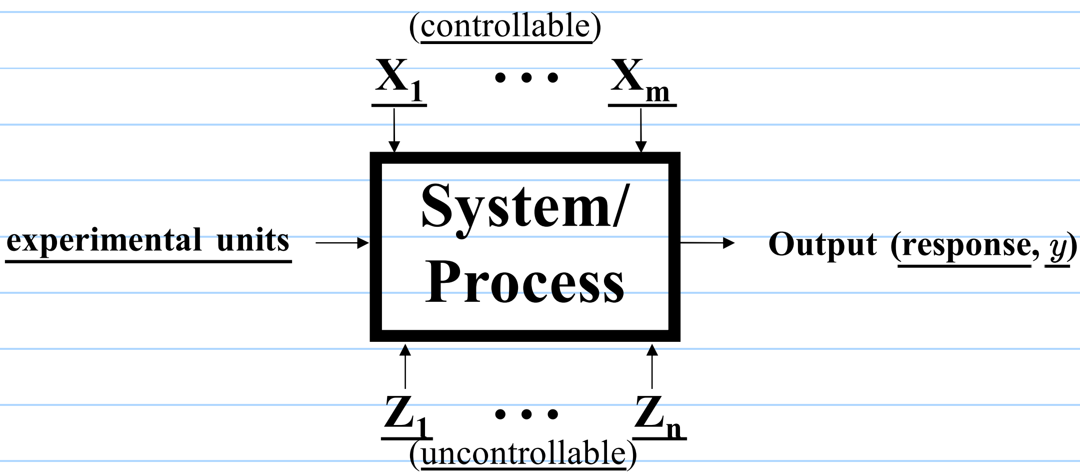
- Variance-Covariance matrix of $\hat{\beta}$: $\Sigma_{\hat{\beta}} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.

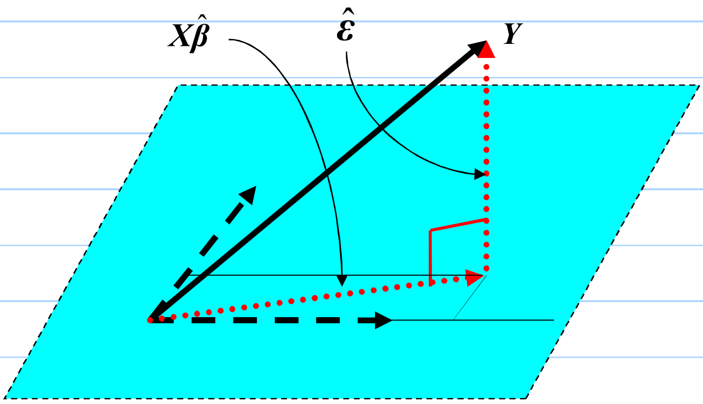
Y	1	g_1	g_2	...	g_p
y_1	1	g_{11}	g_{12}	...	g_{1p}
y_2	1	g_{21}	g_{22}	...	g_{2p}
...			...		
y_n	1	g_{n1}	g_{n2}	...	g_{np}

a row: one group of observations

a column: one effect

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)





Analysis of Variance

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

$$CTSS = RegrSS + RSS,$$

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

ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	k	$\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$	$(\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2)/k$
residual	$N-(k+1)$	$(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})$	$(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})/(N-k-1)$
total	$N-1$	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	

Explanatory Power of the Model

- $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures of 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 β_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 β_j does not contain 0, then H_0 is rejected.

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

Analysis of Air Pollution Data

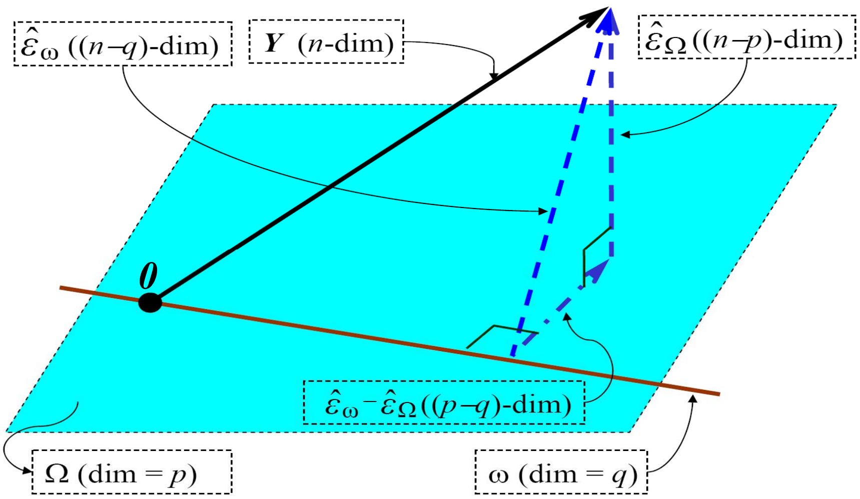
Predictor	Coef	SE Coef	T	P
Constant	1332.7	291.7	4.57	0.000
JanTemp	-2.3052	0.8795	-2.62	0.012
JulyTemp	-1.657	2.051	-0.81	0.424
RelHum	0.407	1.070	0.38	0.706
Rain	1.4436	0.5847	2.47	0.018
Educatio	-9.458	9.080	-1.04	0.303
PopDensi	0.004509	0.004311	1.05	0.301
%NonWhit	5.194	1.005	5.17	0.000
%WC	-1.852	1.210	-1.53	0.133
pop	0.00000109	0.00000401	0.27	0.788
pop/hous	-45.95	39.78	-1.16	0.254
income	-0.000549	0.001309	-0.42	0.677
logHC	-53.47	35.39	-1.51	0.138
logNOx	80.22	32.66	2.46	0.018
logSO2	-6.91	16.72	-0.41	0.681

S = 34.58   R-Sq = 76.7%   R-Sq(adj) = 69.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	14	173383	12384	10.36	0.000
Residual Error	44	52610	1196		
Total	58	225993			

- formulation of hypothesis testing from the view of **comparing models**
  - a model space ≡ the space spanned by columns of some $\boldsymbol{X}$
  - consider a large model space, Ω, and a smaller model space, ω, where $\omega \subset \Omega$, i.e., ω represents a subset/a subspace of Ω. Suppose dimension (# of parameters) of Ω is p and dim(ω)=q, where $p>q$.
  - to answer "which of the model spaces is more adequate" in statistical language
    $\Rightarrow$ perform the test H_0: ω v.s. H_A: Ω\ω



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

- (sequential) ANOVA
  - anova(y~1+A+B+A:B), A: 3 levels, B: 4 levels

    1) test ω:model **1** (y~1) against Ω:model **2** (y~1+A) [$df_\omega - df_\Omega$=2]
    2) test ω:model **2** (y~1+A) against Ω:model **4** (y~1+A+B) [$df_\omega - df_\Omega$=3]
    3) test ω:model **4** (y~1+A+B) against Ω:model **5** (y~1+A+B+A:B) [$df_\omega - df_\Omega$=6]

    - $$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_{\text{model 5}}/df_{\text{model 5}}} \sim F_{df_\omega - df_\Omega, df_{\text{model 5}}}$$
    - invariant to the choice of dummy variables since they generate same ω and Ω
  - ANOVA could have different results when the order of effect sequence is changed, e.g., anova(y~1+B+A+A:B):

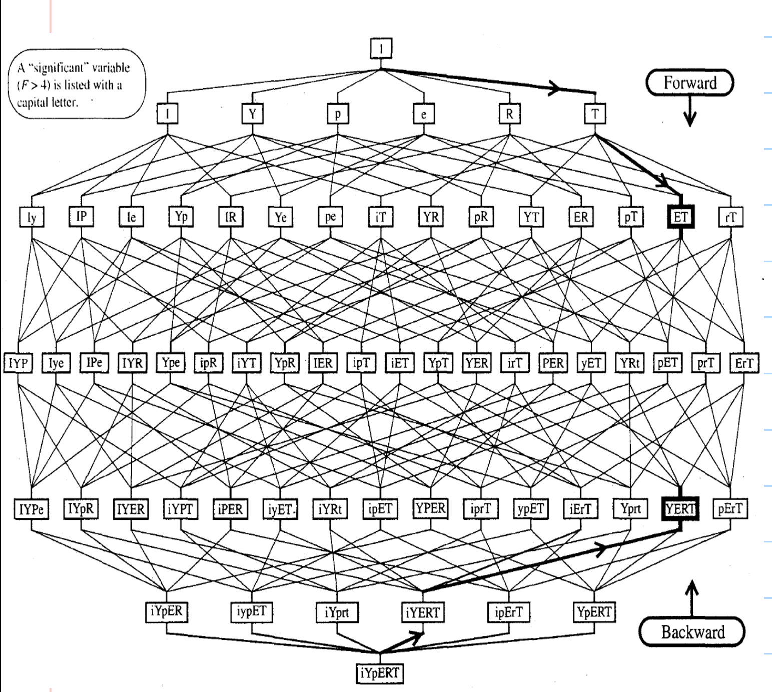
    α) test ω:model **1** (y~1) against Ω:model **3** (y~1+B) [$df_\omega - df_\Omega$=3]
    β) test ω:model **3** (y~1+B) against Ω:model **4** (y~1+B+A) [$df_\omega - df_\Omega$=2]
    χ) test ω:model **4** (y~1+B+A) against Ω:model **5** (y~1+B+A+A:B) [$df_\omega - df_\Omega$=6]
  - anova(y~1+A+B+A:B) and anova(y~1+B+A+A:B) will have **identical** results when orthogonality exists between the three groups of effects: span$\{d_i^A\}$, span$\{d_j^B\}$, span$\{d_{ij}^{A:B}\}$, because in the case, $RSS_\omega - RSS_\Omega$ would equal for 1) and β), 2) and α), 3) and χ)

➢ consider the *full model*:

$$y = \beta_0 + \underline{\beta_1 g_1}(x_1, \ldots, x_m) + \underline{\beta_2 g_2}(x_1, \ldots, x_m) + \cdots + \underline{\beta_k g_k}(x_1, \ldots, x_m) + \epsilon$$

For $1 \le i \le k$, should the term $\beta_i g_i$ be included in the *final fitted model*?

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



➢ (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

Orthogonality

- **Q**: consider the two models:

  model **1**: $y = \beta_0 + \beta_1 x_1 + \varepsilon$, model **2**: $y = \beta_0 + \beta_1 x_1 + \underline{\beta_2 x_2} + \varepsilon$

  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

  fitted model = model **1**,
  true model = model **2**
  $E(\hat{\beta}_1) = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$

- $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$, where $\beta = [\beta_1 \;\; \beta_2]^T$ and $X = [X_1 \;\; X_2]$ with the property $X_1^T X_2 = 0$ ⇒ X_1 and X_2 are orthogonal

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

- Estimation: $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$, $\hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T Y$, and $\hat{\beta}_1, \hat{\beta}_2$ independent

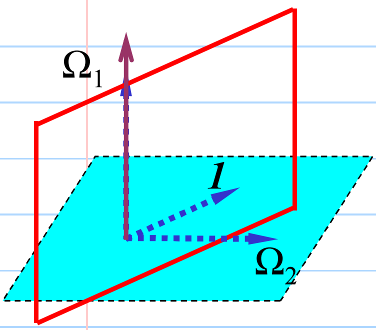
  ⇒ note that $\hat{\beta}_1$ will be the same regardless of whether X_2 is in the model or not (and vise versa).

  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 $Y = X\beta + Z\gamma + \varepsilon$, but Z cannot be measured or may not even be suspected ⇒ $E(\hat{\beta}) = \beta + (X^T X)^{-1} X^T Z \gamma$ ⇒

  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

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

Some Properties of (Multivariate) Normal Distribution

(N1) Linear transformation of normal is still normal

$$\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \;\Rightarrow\; \boldsymbol{AZ} + \boldsymbol{c} \sim N(\boldsymbol{A\mu} + \boldsymbol{c}, \boldsymbol{A\Sigma A}^T).$$

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

(N3) $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{bmatrix} \sim$ normal, and $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ uncorrelated (i.e., $cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2) = \boldsymbol{0}$)
$\Rightarrow \boldsymbol{Z}_1, \boldsymbol{Z}_2$ independent

(N4) $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{W}_1 = \boldsymbol{A}_1\boldsymbol{Z}$, $\boldsymbol{W}_2 = \boldsymbol{A}_2\boldsymbol{Z}$
$\Rightarrow \boldsymbol{W}_1, \boldsymbol{W}_2$ are independent iff $\boldsymbol{A}_1\boldsymbol{\Sigma}\boldsymbol{A}_2^T = \boldsymbol{0}$.

(N5) $\boldsymbol{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{W}_1 = \boldsymbol{A}_1\boldsymbol{Z}$, $\boldsymbol{W}_2 = \boldsymbol{A}_2\boldsymbol{Z}$, $\ldots$, $\boldsymbol{W}_k = \boldsymbol{A}_k\boldsymbol{Z}$, and $cov(\boldsymbol{W}_i, \boldsymbol{W}_j) = \boldsymbol{0}$
for $1 \le i < j \le k$, $\Rightarrow \boldsymbol{W}_1^T\boldsymbol{W}_1, \boldsymbol{W}_2^T\boldsymbol{W}_2, \ldots, \boldsymbol{W}_k^T\boldsymbol{W}_k$ are mutually independent.

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

- if $\boldsymbol{\Sigma}$ is non-singular, $(\boldsymbol{Z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}) \sim \chi_n^2$
- if $\boldsymbol{\Sigma}$ is singular and has rank r $(< n)$,
  let $\boldsymbol{\Sigma}^-$ be a generalized inverse of $\boldsymbol{\Sigma}$ (i.e., $\boldsymbol{\Sigma\Sigma^-\Sigma} = \boldsymbol{\Sigma}$), then

$$(\boldsymbol{Z} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^-(\boldsymbol{Z} - \boldsymbol{\mu}) \sim \chi_r^2$$

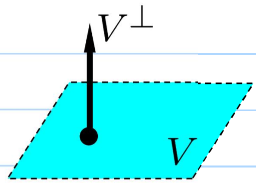
Normal Distribution and Sequential ANOVA

- Consider a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where
  - $\boldsymbol{Y} \in \mathbb{R}^n$: an $n \times 1$ random vector and $\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$
  - $\boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_k \end{bmatrix}$
  - $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_0^T & \boldsymbol{\beta}_1^T & \cdots & \boldsymbol{\beta}_k^T \end{bmatrix}^T$

$$\Rightarrow \boldsymbol{X}\boldsymbol{\beta} = \overbrace{\underbrace{\mathbf{1}}_{\boldsymbol{X}_0}\ \boldsymbol{\beta}_0}^{\boldsymbol{\mu}_0} + \overbrace{\boldsymbol{X}_1\boldsymbol{\beta}_1}^{\boldsymbol{\mu}_1} + \cdots + \overbrace{\boldsymbol{X}_k\boldsymbol{\beta}_k}^{\boldsymbol{\mu}_k}$$

  - $\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$

- Define
  - V_i = the vector space generated by the column vectors of



$$\boldsymbol{A}_i \equiv \begin{bmatrix} \boldsymbol{X}_0 & \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_i \end{bmatrix}, \quad i = 0, 1, \ldots, k.$$

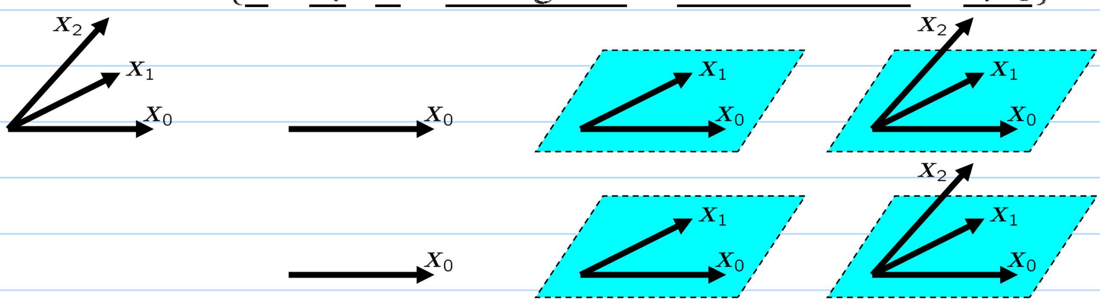
    * V_i is called the *model space* of $\boldsymbol{A}_i$, and denoted by $\text{span}\{\boldsymbol{A}_i\}$
    * $V_0 \subset V_1 \subset \cdots \subset V_k = \text{span}\{\boldsymbol{X}\} \subset \mathbb{R}^n$
  - $V^\perp$ = orthogonal complement of the vector space V, i.e.,

$$V^\perp = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{v} \text{ is orthogonal to all the vectors in } V\}$$

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

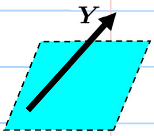
  - $W_0 = \text{span}\{\boldsymbol{X}_0\} = V_0$, and for $i = 1, \ldots, k$,
    W_i = orthogonal complement of V_{i-1} relative to V_i (**note**: $V_{i-1} \subset V_i$), i.e.,

$$\begin{aligned} W_i &= V_i \ominus V_{i-1} = V_i \cap V_{i-1}^\perp \\ &= \{\boldsymbol{v} \in V_i : \boldsymbol{v} \text{ is orthogonal to all the vectors in } V_{i-1}\} \end{aligned}$$

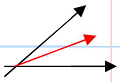


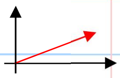
    * $V_i = W_0 \oplus W_1 \oplus \cdots \oplus W_i$
    * $\mathbb{R}^n = W_0 \oplus W_1 \oplus \cdots \oplus W_k \oplus V_k^\perp$ (**note**: $V_k^\perp$ is the residual space)
    * $W_0 \perp W_1 \perp \cdots \perp W_k \perp V_k^\perp$
    * **Note**. In general, $W_i \neq \text{span}\{\boldsymbol{X}_i\}$. However, if $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are mutually orthogonal, then $W_i = \text{span}\{\boldsymbol{X}_i\}$.
  - $r_i \equiv \dim(W_i) = \dim(V_i) - \dim(V_{i-1})$.
    * Let $r = \sum_{i=0}^k r_i = \sum_{i=0}^k \dim(W_i) = \dim(V_k)$. Then, $\dim(V_k^\perp) = n - r$.

- Orthogonal projection of $\boldsymbol{Y}$ onto V_i's and W_i's
  - For a vector space $V \subset \mathbb{R}^n$, denote the *orthogonal projection matrix* of $\boldsymbol{Y}$ onto V by $\boldsymbol{P}_V$. Then, the orthogonal projection of $\boldsymbol{Y}$ onto V is $\boldsymbol{P}_V\boldsymbol{Y}$.



    - ∗ if $V = \text{span}\{\boldsymbol{A}\}$, then $\boldsymbol{P}_V = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$
    - ∗ the orthogonal projection matrix onto $V^\perp$, denoted by $\boldsymbol{P}_{V^\perp}$, is $\boldsymbol{P}_{V^\perp} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T = \boldsymbol{I} - \boldsymbol{P}_V$





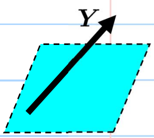
  - Some properties of orthogonal projection matrix
    - ∗ A square matrix $\boldsymbol{P}$ is a *projection matrix* iff $\boldsymbol{P}^2 = \boldsymbol{P}$ (idempotent)
      - · Idempotence implies $\boldsymbol{P}$ is a generalized inverse of $\boldsymbol{P}$ since

$$\boldsymbol{P}\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}^3 = \boldsymbol{P}.$$

    - ∗ A projection matrix $\boldsymbol{P}$ is *orthogonal* iff $\boldsymbol{P}^T = \boldsymbol{P}$ (symmetric)
    - ∗ If $\boldsymbol{P}$ is an orthogonal projection matrix onto V, then
      - · $\boldsymbol{P}$ has $\dim(V)$ eigenvalues equal to 1 and the rest 0
      - · $\boldsymbol{P}$ is diagonalizable, and there exists an orthogonal matrix $\boldsymbol{U}$ ($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$) such that $\boldsymbol{U}^T\boldsymbol{P}\boldsymbol{U} = \boldsymbol{\Lambda}$ is a diagonal matrix. (**Note**. Thus, $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$) Actually,
        - ◇ the columns of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{P}$, and
        - ◇ the diagonal entries of $\boldsymbol{\Lambda}$ are the eigenvalues of $\boldsymbol{P}$

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

- Since $V_i = \text{span}\{\boldsymbol{A}_i\}$, $\boldsymbol{P}_{V_i} = \boldsymbol{A}_i(\boldsymbol{A}_i^T\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i^T$ and $\boldsymbol{P}_{V_i}\boldsymbol{Y} = \boldsymbol{A}_i(\boldsymbol{A}_i^T\boldsymbol{A}_i)^{-1}\boldsymbol{A}_i^T\boldsymbol{Y}$



  - ∗ For the model space V_i,

$$\begin{aligned} RSS_{V_i} &= \|\boldsymbol{Y}\|^2 - \|\boldsymbol{P}_{V_i}\boldsymbol{Y}\|^2 = \boldsymbol{Y}^T\boldsymbol{Y} - (\boldsymbol{P}_{V_i}\boldsymbol{Y})^T\boldsymbol{P}_{V_i}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{P}_{V_i}^T\boldsymbol{P}_{V_i}\boldsymbol{Y} \\ &= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{P}_{V_i}\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{V_i})\boldsymbol{Y} = \|(\boldsymbol{I} - \boldsymbol{P}_{V_i})\boldsymbol{Y}\|^2 \end{aligned}$$

- Since $W_i = V_i \cap V_{i-1}^\perp$ and $V_{i-1} \subset V_i$,

$$\boldsymbol{P}_{W_i} = (\boldsymbol{I} - \boldsymbol{P}_{V_{i-1}})\boldsymbol{P}_{V_i} = \boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}}\boldsymbol{P}_{V_i} = \boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}},$$

and $\boldsymbol{P}_{W_i}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{P}_{V_{i-1}})\boldsymbol{P}_{V_i}\boldsymbol{Y} = (\boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}})\boldsymbol{Y} = \boldsymbol{P}_{V_i}\boldsymbol{Y} - \boldsymbol{P}_{V_{i-1}}\boldsymbol{Y}$.

  - ∗ Since $\mathbb{R}^n = W_0 \oplus W_1 \oplus \cdots \oplus W_k \oplus V_k^\perp$ and $W_0 \perp W_1 \perp \cdots \perp W_k \perp V_k^\perp$,
    - · $\boldsymbol{Y} = \boldsymbol{P}_{W_0}\boldsymbol{Y} + \boldsymbol{P}_{W_1}\boldsymbol{Y} + \cdots + \boldsymbol{P}_{W_k}\boldsymbol{Y} + \boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}$
    - · $\boldsymbol{P}_{W_0}\boldsymbol{Y} \perp \boldsymbol{P}_{W_1}\boldsymbol{Y} \perp \cdots \perp \boldsymbol{P}_{W_k}\boldsymbol{Y} \perp \boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}$
    - · $\|\boldsymbol{Y}\|^2 = \|\boldsymbol{P}_{W_0}\boldsymbol{Y}\|^2 + \|\boldsymbol{P}_{W_1}\boldsymbol{Y}\|^2 + \cdots + \|\boldsymbol{P}_{W_k}\boldsymbol{Y}\|^2 + \|\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}\|^2$
  - ∗ When $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are mutually orthogonal,

$$\boldsymbol{P}_{W_i} = \boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}} = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T.$$

- Consider the sequential ANOVA: for $i = 1, \ldots, k$,

$$H_0^{(i)} : \boldsymbol{\beta}_i = \mathbf{0} \ \ (\omega_i = V_{i-1}) \quad \text{vs.} \quad H_A^{(i)} : \boldsymbol{\beta}_i \neq \mathbf{0} \ \ (\Omega_i = V_i)$$

(P1) $RSS_{\omega_i} - RSS_{\Omega_i} = RSS_{V_{i-1}} - RSS_{V_i} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{V_{i-1}})\boldsymbol{Y} - \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{V_i})\boldsymbol{Y}$

$$= \boldsymbol{Y}^T(\boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}})\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{P}_{W_i}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{P}_{W_i}^T\boldsymbol{P}_{W_i}\boldsymbol{Y} = \|\boldsymbol{P}_{W_i}\boldsymbol{Y}\|^2$$

and $df_{\omega_i} - df_{\Omega_i} = \dim(W_i) = r_i$.

(P2) $\boldsymbol{P}_{W_i}\boldsymbol{Y} = \boldsymbol{P}_{W_i}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_{W_i}\boldsymbol{\varepsilon}$, where

- $\boldsymbol{P}_{W_i}\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}})(\boldsymbol{X}_0\beta_0 + \cdots + \boldsymbol{X}_{i-1}\boldsymbol{\beta}_{i-1} + \boldsymbol{X}_i\boldsymbol{\beta}_i + \cdots + \boldsymbol{X}_k\boldsymbol{\beta}_k)$

$$= (\boldsymbol{P}_{V_i} - \boldsymbol{P}_{V_{i-1}})(\boldsymbol{X}_i\boldsymbol{\beta}_i + \cdots + \boldsymbol{X}_k\boldsymbol{\beta}_k) = \boldsymbol{P}_{W_i}(\boldsymbol{u}_i + \cdots + \boldsymbol{u}_k)$$

  - $*$ If $\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are mutually orthogonal,

$$\boldsymbol{P}_{W_i}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{P}_{W_i}\boldsymbol{u}_i = (\boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T)\boldsymbol{X}_i\boldsymbol{\beta}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i = \boldsymbol{u}_i.$$

- $\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma^2\boldsymbol{P}_{W_i}\boldsymbol{I}\boldsymbol{P}_{W_i}^T)$, where $\boldsymbol{P}_{W_i}\boldsymbol{I}\boldsymbol{P}_{W_i}^T = \boldsymbol{P}_{W_i}$ and $\boldsymbol{P}_{W_i} = \boldsymbol{P}_{W_i}^-$.
- Thus, $\boldsymbol{P}_{W_i}\boldsymbol{Y} \sim \boldsymbol{N}(\boldsymbol{P}_{W_i}(\boldsymbol{u}_i + \cdots + \boldsymbol{u}_k), \sigma^2\boldsymbol{P}_{W_i})$

(P3) $\|\boldsymbol{P}_{W_i}\boldsymbol{Y}\|^2 = \boldsymbol{Y}^T\boldsymbol{P}_{W_i}\boldsymbol{Y} = (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^T\boldsymbol{P}_{W_i}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$

$$= (\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{P}_{W_i}(\boldsymbol{X}\boldsymbol{\beta}) + 2(\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon}$$

$$= \|\boldsymbol{P}_{W_i}\boldsymbol{X}\boldsymbol{\beta}\|^2 + 2(\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}^T\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^T\boldsymbol{P}_{W_i}^T\boldsymbol{P}_{W_i}\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon} = \sigma^2(\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon})^T(\boldsymbol{P}_{W_i}/\sigma^2)(\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon}) \sim \sigma^2\chi^2_{r_i}$

NTHU STAT 5510, 2024, Lecture Notes
jointly made by Jeff Wu (GT, USA) and S.-W. Cheng (NTHU, Taiwan)

(P4) $E(RSS_{V_{i-1}} - RSS_{V_i}) = E(\boldsymbol{Y}^T\boldsymbol{P}_{W_i}\boldsymbol{Y}) = E(\|\boldsymbol{P}_{W_i}\boldsymbol{Y}\|^2)$

$$= \|\boldsymbol{P}_{W_i}\boldsymbol{X}\boldsymbol{\beta}\|^2 + 2(\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{P}_{W_i}E(\boldsymbol{\varepsilon}) + E(\boldsymbol{\varepsilon}^T\boldsymbol{P}_{W_i}\boldsymbol{\varepsilon})$$

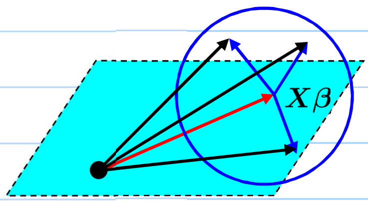
$$= \|\boldsymbol{P}_{W_i}(\boldsymbol{u}_i + \cdots + \boldsymbol{u}_k)\|^2 + r_i\sigma^2$$

- **Note**. If $\boldsymbol{M}$: a symmetric matrix, and $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$E(\boldsymbol{Z}^T\boldsymbol{M}\boldsymbol{Z}) = \boldsymbol{\mu}^T\boldsymbol{M}\boldsymbol{\mu} + \text{trace}(\boldsymbol{M}\boldsymbol{\Sigma}).$$

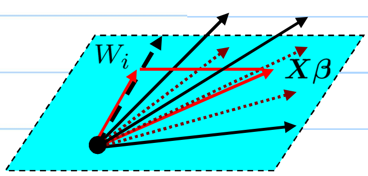
(P5) For the residual space $V_k^\perp$,

- $RSS_{V_k} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{V_k})\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y} = \|\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}\|^2$
- $\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y} = \boldsymbol{P}_{V_k^\perp}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{P}_{V_k^\perp}\boldsymbol{\varepsilon} = \boldsymbol{P}_{V_k^\perp}\boldsymbol{\varepsilon} \ \sim \ \boldsymbol{N}(\mathbf{0}, \sigma^2\boldsymbol{P}_{V_k^\perp})$
- $\|\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}\|^2 = \boldsymbol{\varepsilon}^T\boldsymbol{P}_{V_k^\perp}\boldsymbol{\varepsilon} \ \sim \ \sigma^2\chi^2_{n-r}$
- $E(RSS_{V_k}) = E(\|\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}\|^2) = (n-r)\sigma^2$



(P6) Since $W_0 \perp W_1 \perp \cdots \perp W_k \perp V_k^\perp$,

$\Rightarrow$ $\boldsymbol{P}_{W_0}\boldsymbol{Y}, \ldots, \boldsymbol{P}_{W_k}\boldsymbol{Y}, \boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}$ are *independent* random vectors



$\Rightarrow$ $\|\boldsymbol{P}_{W_0}\boldsymbol{Y}\|^2, \ldots, \|\boldsymbol{P}_{W_k}\boldsymbol{Y}\|^2, \|\boldsymbol{P}_{V_k^\perp}\boldsymbol{Y}\|^2$ are *independent* random variables

❖ **Further reading**: Seber and Lee (2003), *Linear Regression Analysis*, 2nd edition, Chapter 2.