Outlier Detection

Recall LM
LNp.7-1~7-7

influential obs

- 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).

dangerous to exclude in an automatic manner

- 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.

report

- 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?)

obs. with large leverage exist in most data

- 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.

overall pattern $\xleftrightarrow{\text{cf.}}$ unusual observations

Regression Results after Removing the Outlier

cf. results in LNp.2-9

```
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
```

$\hat{\beta}_1$ → 3.0152 > 2.3577 (LNp.2-9)

8.70 > 6.76 (LNp.2-9), d.f. −1

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

5.93258 < 7.54466 (LNp.2-9)

85.3% > 76.5% (LNp.2-9)

```
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
```

2664.3 > 2599.5 (LNp.2-9)

3121.9 $\xrightarrow{\text{cf.}}$ < 3396.4 (LNp.2-9)

```
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
```

original 16th obs

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

Residual Plots After Outlier Removal

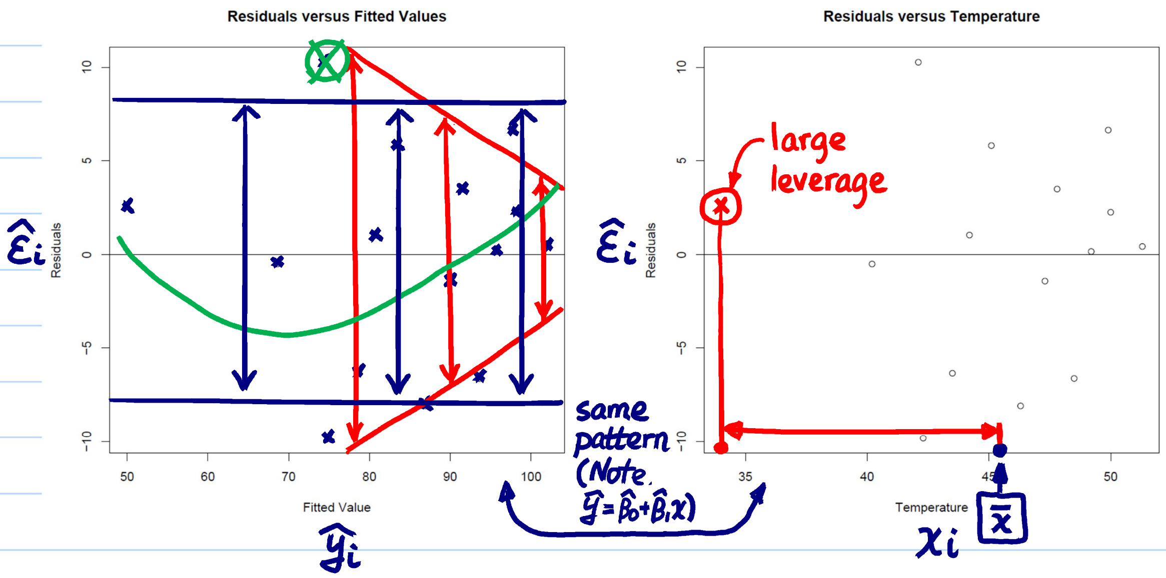


Figure 3: Residual Plots

Comments : No systematic pattern is discerned.

1. non-constant variance
2. curvature in the mean of residuals

---

objective of many regression analysis

① for mean response $E(y_x)=\mu_x$ ② for future obs. $y_x = \mu_x + \varepsilon$

Prediction from the Breast Cancer Data

① interpolation ② extrapolation

- The fitted regression model is $Y = -21.79 + 2.36X$, where Y denotes the mortality rate and X denotes the temperature. ($\hat{\beta}_0$, $\hat{\beta}_1$)
- The predicted mean (future obs.) 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 $y_{x_0} = 93.85$.
  - $\hat{y}_{x_0} \to y_{x_0} = \mu_{x_0} + \varepsilon$
  - $\hat{\mu}_{x_0} \to \mu_{x_0} = \beta_0 + \beta_1 x_0$
- The standard error of $\hat{\mu}_{x_0}$ is given by

$Var(\hat{\mu}_{x_0}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x_0)$ (exercise, note. $cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$)

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

Note1. It's a function of x_i's only

Note2. What happens if x_0 is away from $\bar{x}$?

Note3. It converges to zero when $N \to \infty$

- 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, $s.e.(\hat{\mu}_{x_0}) = 2.432$.

estimate ± (critical value) × s.e.(estimate)

Confidence interval for mean and prediction interval for future observation

μ_{x_0}

$y_{x_0} = \mu_{x_0} + \varepsilon$

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

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

same

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

$E[\hat{\beta}_0 + \hat{\beta}_1 x_0 - (\beta_0 + \beta_1 x_0 + \varepsilon)]^2$

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

Note. It converges to σ^2 when $N \to \infty$

$E(\hat{y}_{x_0} - y_{x_0})^2 = Var(\hat{\mu}_{x_0}) + Var(\varepsilon)$, $Var(\varepsilon) = \sigma^2$

$$\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_0}_{\text{estimate } \hat{y}_{x_0}} \pm \underbrace{t_{N-2,0.025}}_{\text{critical value}} \times \underbrace{\hat{\sigma}\sqrt{1 + \frac{1}{N} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}}_{\text{s.e.(estimate)}},$$

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].

becomes wider

many concept/idea/formula in MLR are similar to those in SLR

Multiple Linear Regression : Air Pollution Data

more than one predictor/effect

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

- Data collected by General Motors.
- Response is age-adjusted mortality.

Why?

- Predictors :
  - Variables measuring demographic characteristics.
  - Variables measuring climatic characteristics.
  - Variables recording pollution potential of 3 air pollutants.

their roles are similar to block factors

* true model

$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

* fitted model

$Y = X_1\beta_1 + \varepsilon'$

- Objective : To determine whether air pollution is significantly related to mortality.

relationship

Q: Why should we put demographic & climatic variables in the model when we are only interested in

Predictors

climate
1. **JanTemp :** Mean January temperature (degrees Farenheit)
2. **JulyTemp :** Mean July temperature (degrees Farenheit)
3. **RelHum :** Relative Humidity
4. **Rain :** Annual rainfall (inches)

demographic
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

pollutant
12. **HCPot :** HC pollution potential
13. **NOxPot :** Nitrous Oxide pollution potential
14. **SO2Pot :** Sulphur Dioxide pollution potential

Initial Data Analysis ← Getting Started →

- data cleaning
- need transformation ?
- scatter plots
- histogram
- descriptive statistics
- ...

- 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. → 59 data points.

Why? Note. response is not skewed in the case.

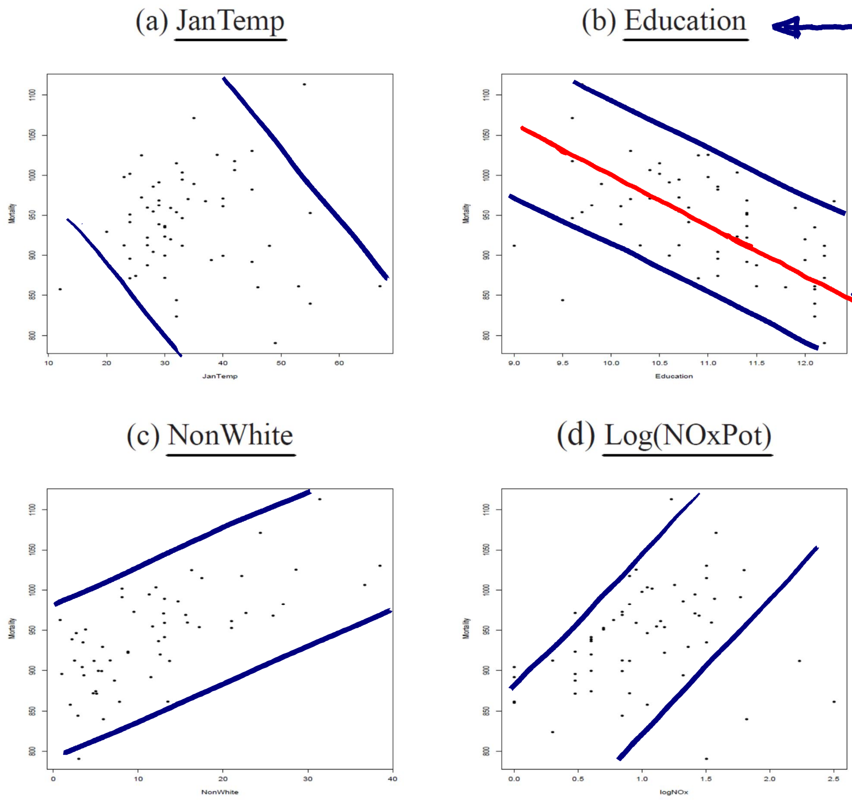
Scatter Plots

"rain" 4 significant (LNp.2-24)

cf., why? check ✩ in LNp.2-15

Figure 4: Scatter Plots of mortality against selected predictors

not significant (LNp.2-24)



1 5 7 13

simple regression: $y=\beta_0+\beta_1 x_5+\varepsilon$

Fitting the Multiple Regression Equation

estimate β

unknown parameter — known function ($g_0=1$, intercept)

$$y_{\underset{\sim}{x}} = \underbrace{\sum_{j=0}^{k} \beta_j \times g_j(x_1,\cdots,x_m)}_{E(y_{\underset{\sim}{x}})} + \varepsilon$$

functional form

cf. matrix form: $\underline{Y}=\underline{X}\underline{\beta}+\varepsilon$, $\underline{\beta}=\begin{bmatrix}\beta_0\\ \vdots \\ \beta_k\end{bmatrix}$

- Underlying Model :

$$y = \underbrace{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k}_{E(y_{\underset{\sim}{x}})}+\varepsilon, \quad \varepsilon \sim N(0,\sigma^2).$$

- Coefficients are estimated by minimizing

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

$\sum_i \beta_i g_i = X\beta$

$\partial S/\partial\beta = 0$

- **Least Squares estimates :**

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

random

$\Rightarrow E(\hat{\beta})=\beta$ if model correct

Data matrix

y	x_1	$\cdots$	x_k
y_1	x_{11}	$\cdots$	x_{1k}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_n	x_{n1}	$\cdots$	x_{nk}

$g_{ij} = g_j(x_{i1},\cdots,x_{ik})$

model matrix X

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

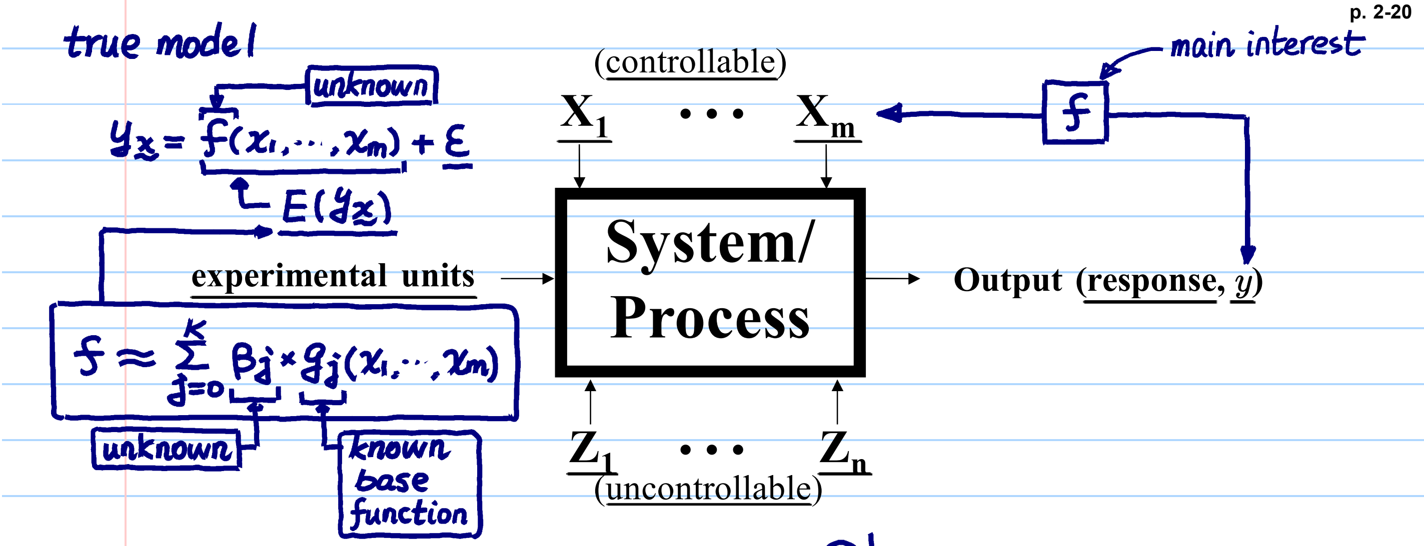
depend only on x_i's

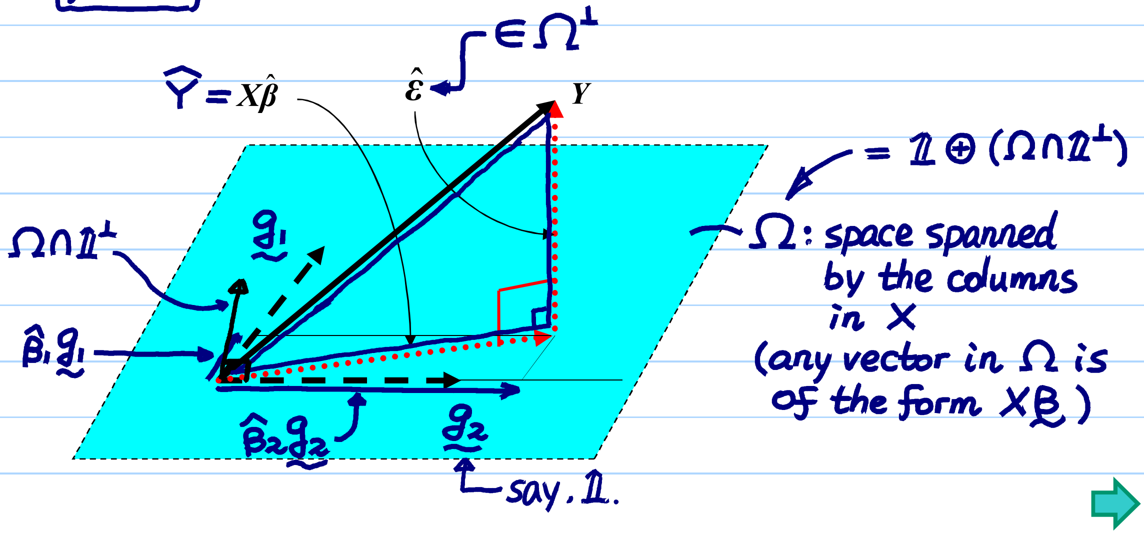
$(X^TX)_{jl}$: inner product of the jth & lth columns in X.

a row: one group of observations

a column: one effect

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





Analysis of Variance

overall F-test
(ω) $H_0: y = \beta_0 + \varepsilon$
(Ω) $H_A: y = \beta_0 + \sum_{j=1}^{k} \beta_j g_j + \varepsilon$

- The total variation in y, i.e., corrected total sum of squares, $\|y - \bar{y}\mathbb{1}\|^2$

$CTSS = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \mathbf{y}^T\mathbf{y} - N\bar{y}^2$, can be decomposed into two parts (Analysis of Variance (ANOVA)), $-\hat{y}_i + \hat{y}_i$

規律 隨機 RSS under Ω

$$CTSS = RegrSS + RSS,$$

RSS under ω

check the graph in LNp.2-4

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 $= \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$.

mean of $\hat{y}_i$'s $= \bar{y}$

y_i's $\to Y$, $\hat{y}_i$'s $\to X\hat{\beta}$, $\bar{y} \to \bar{y}\mathbb{1}$

ANOVA Table

source of variation

$y - \bar{y}\mathbb{1}$, $\hat{\varepsilon}$, $\hat{y} - \bar{y}\mathbb{1}$

Source	Degrees of Freedom	Sum of Squares	Mean Squares
regression	k ($\beta_1, \cdots, \beta_k$)	$\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2$	$(\hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} - N\bar{y}^2)/k$
residual	$N - (k+1)$ (β)	$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$	$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})/(N - k - 1)$
total	$N - 1$ (β_0)	$\mathbf{y}^T\mathbf{y} - N\bar{y}^2$	

F-statistic $\div$

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.

規律 隨機

cor(y, $\hat{y}$) ← check the graph in LNp.2-20

- **Adjusted R^2 :**

$$R_a^2 = 1 - \frac{\frac{RSS}{N-(k+1)}}{\frac{CTSS}{N-1}} = 1 - \left(\frac{N-1}{N-k-1}\right)\frac{RSS}{CTSS} \le R^2$$

df df ≥ 1

∵ RSS always descreases

- 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*.

Q: When can R^2 reach 1?
Ans. $\hat{y}_i = y_i$, only when data have no replicates & # of parameters β = # of obs

total df available N (sample size) → df available for studying β; df available for estimating σ^2

Want to have a final fitted model such that
1. # of parameters : small
2. RSS : small

Testing significance of coefficients : t-Statistic

Examine whether $\beta_j = 0$ when other effects β_i's (x_i's), $i \ne j$, are still in the model

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

Note. collinearity

$$t_j = \frac{\hat{\beta}_j - 0}{s.d.(\hat{\beta}_j)}.$$

$t_j^2 = F$

$\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$

null model
(ω) H_0 : $y = \beta_0 + \cdots + \beta_{j-1}x_{j-1} + \beta_{j+1}x_{j+1} + \cdots + \varepsilon$
(Ω) H_A : $y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon$
alternative model

- 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.

null dist., under H_0, $t_j \sim t_{N-(k+1)}$

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

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

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.

df_{RSS}

If the confidence interval for β_j does not contain 0, then H_0 is rejected.

$H_0 : \beta_j = \beta_j^*$

acceptance region $\left|\frac{\hat{\beta}_j - \beta_j^*}{s.d.(\hat{\beta}_j)}\right| \le t_{N-(k+1),\frac{\alpha}{2}}$

check graphs in LNp.2-18

Analysis of Air Pollution Data

	Predictor	Coef ($\hat{\beta}_j$'s)	SE Coef ($s.d(\hat{\beta}_j)$'s)	T (t-test)	P
β_0	Constant	1332.7	291.7	4.57	0.000 ✓
1	JanTemp	-2.3052	0.8795	-2.62	0.012 ✓
2	JulyTemp	-1.657	2.051	-0.81	0.424
3	RelHum	0.407	1.070	0.38	0.706
4	Rain	1.4436	0.5847	2.47	0.018 ✓
5	Educatio	-9.458	9.080	-1.04	0.303
6	PopDensi	0.004509	0.004311	1.05	0.301
7	%NonWhit	5.194	1.005	5.17	0.000 ✓
8	%WC	-1.852	1.210	-1.53	0.133
9	pop	0.00000109	0.00000401	0.27	0.788
10	pop/hous	-45.95	39.78	-1.16	0.254
11	income	-0.000549	0.001309	-0.42	0.677
12	logHC	-53.47	35.39	-1.51	0.138
13	logNOx	80.22	32.66	2.46	0.018 ✓
14	logSO2	-6.91	16.72	-0.41	0.681

Q: Can we remove all insignificant effects simultaneously to simplify the model? Ans. No, in general, but OK if orthogonality exists.

how to interpret it? It becomes insignificant when other effects are in the model (cf graph in LNp 2-18). Possible reason: collinearity

S = 34.58 ($\hat{\sigma}$) R-Sq = 76.7% R-Sq(adj) = 69.3% (smaller)

$$R^2 = 1 - \frac{1}{1 + \frac{k}{N-(k+1)}F}$$

Analysis of Variance

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

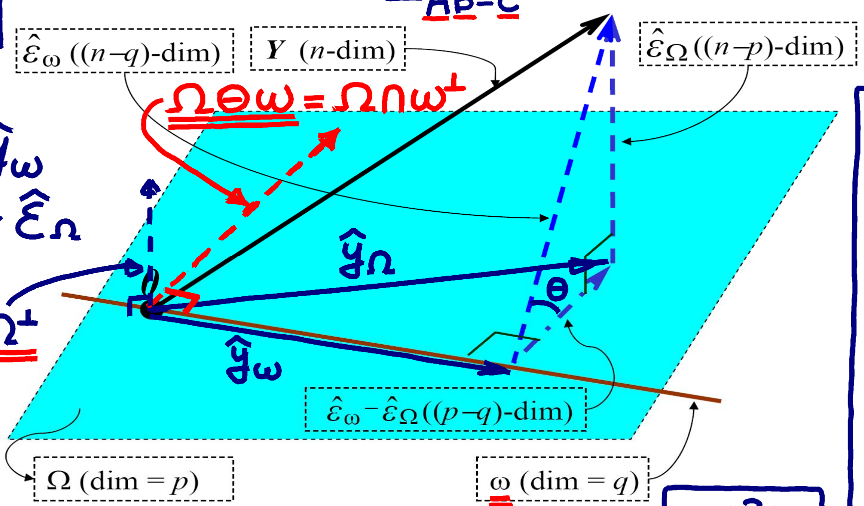
59 obs. meaning

- formulation of hypothesis testing from the view of **comparing models** ← L.M. LNp.4-6~16
  - a model space ≡ the space spanned by columns of some $\boldsymbol{X}$ (effects)
  - consider a large model space, Ω, and a smaller model space, ω, where $\omega \subset \Omega$, (nested) i.e., ω represents a subset/a subspace of Ω. Suppose dimension (# of parameters) of Ω is p and dim(ω)=q, where $p>q$. → $df_\Omega = n-p$, $df_\omega = n-q$
  - to answer "which of the model spaces is more adequate" in statistical language

$\Rightarrow$ perform the test H_0: ω v.s. H_A: $\Omega \backslash \omega$ ← check examples in LNp.2-21&23

H_0: $\beta_j = \beta_j^* = 0$

AB=C



$\hat{y}_\Omega - \hat{y}_\omega = \hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega$

- θ large ⇒ $\hat{\varepsilon}_\omega \approx \hat{\varepsilon}_\Omega$ ⇒ $\hat{y}_\Omega \approx \hat{y}_\omega$ ⇒ prefer ω
- θ small ⇒ $\hat{\varepsilon}_\omega$ quite different from $\hat{\varepsilon}_\Omega$ ⇒ $\hat{y}_\Omega$ quite different from $\hat{y}_\omega$ ⇒ prefer Ω

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

$(RSS_\omega - RSS_\Omega) = \|\hat{\varepsilon}_\omega - \hat{\varepsilon}_\Omega\|^2$, $RSS_\Omega = \|\hat{\varepsilon}_\Omega\|^2$, ratio $\cot^2(\theta)$; $p-q = df_\omega - df_\Omega$, $n-p = df_\Omega$

a special case: t-test