Topic 3: Simple Linear Regression

Outline

- Simple linear regression model
  - Model parameters
  - Distribution of error terms
- Estimation of regression parameters
  - Method of least squares
  - Maximum likelihood

Data for Simple Linear Regression

- Observe i=1,2,...,n pairs of variables
- Each pair often called a case
- Y_i = i^th response variable
- X_i = i^th explanatory variable

Simple Linear Regression Model

- $Y_i = b_0 + b_1 X_i + e_i$
- b_0 is the intercept
- b_1 is the slope
- e_i is a random error term
  - $E(e_i)=0$ and $s^2(e_i)=s^2$
  - e_i and e_j are uncorrelated

Simple Linear Normal Error Regression Model

- $Y_i = b_0 + b_1 X_i + e_i$
- b_0 is the intercept
- b_1 is the slope
- e_i is a Normally distributed random error with mean 0 and variance σ^2
- e_i and e_j are uncorrelated → indep

Model Parameters

- β_0 : the intercept
- β_1 : the slope
- σ^2 : the variance of the error term

Features of Both Regression Models

- $Y_i = \beta_0 + \beta_1 X_i + e_i$

- $E(Y_i) = \beta_0 + \beta_1 X_i + E(e_i) = \beta_0 + \beta_1 X_i$
- $Var(Y_i) = 0 + var(e_i) = \sigma^2$
  - Mean of Y_i determined by value of X_i
  - All possible means fall on a line
  - The Y_i vary about this line

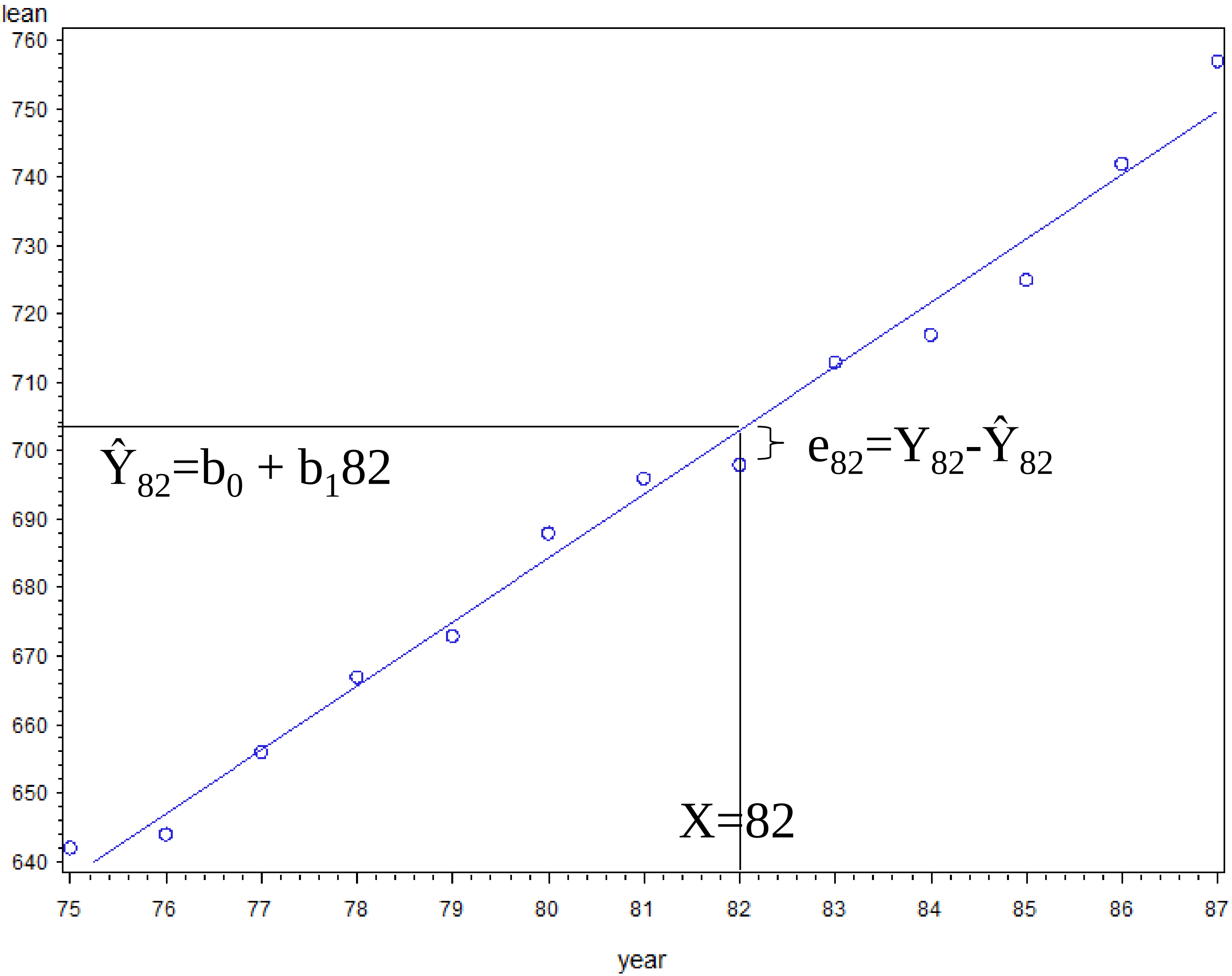
Features of Normal Error Regression Model

- $Y_i = \beta_0 + \beta_1 X_i + e_i$

- If e_i is Normally distributed then Y_i is $N(\beta_0 + \beta_1 X_i , \sigma^2)$ (A.36)
- Does <u>not</u> imply the collection of Y_i are Normally distributed

Fitted Regression Equation and Residuals

- $\hat{Y}_i = b_0 + b_1X_i$
  - b_0 is the estimated intercept
  - b_1 is the estimated slope
- e_i : residual for i^th case
- $e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1X_i)$


lean
760
750
740
730
720
710
700
690
680
670
660
650
640
75
76
77
78
79
80
81
82
83
84
85
86
87
year
$\hat{Y}_{82}=b_0 + b_1 82$
$e_{82}=Y_{82}-\hat{Y}_{82}$
X=82

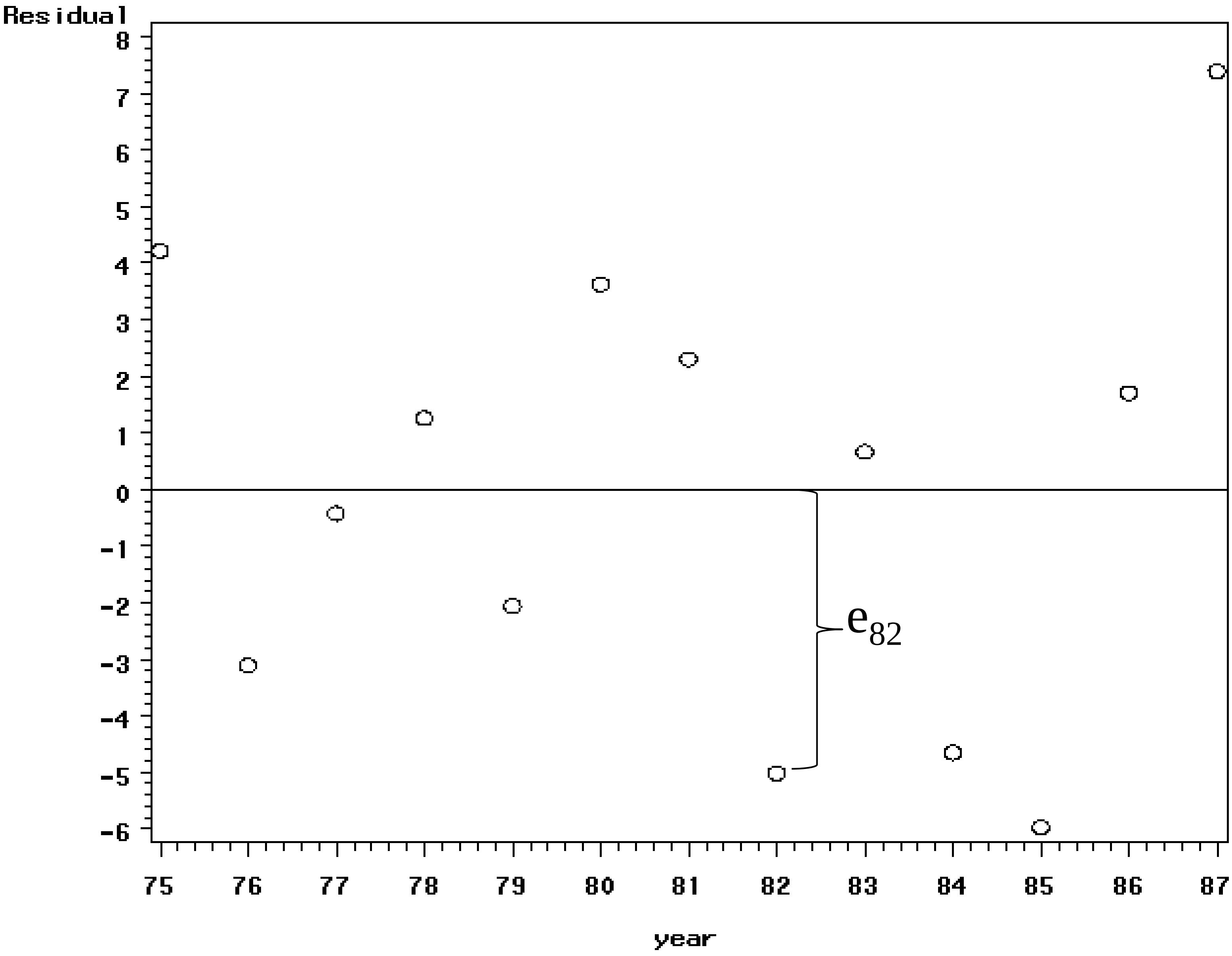
Plot the residuals

Continuation of pisa.sas

Using data set from output statement

```
proc gplot data=a2;
    plot resid*year vref=0;
    where lean ne .;
    run;
```

vref=0 adds horizontal line to plot at zero



Residual
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
75
76
77
78
79
80
81
82
83
84
85
86
87
year
e_{82}

Least Squares

- Want to find "best" b_0 and b_1
- Will minimize $\Sigma(Y_i - (b_0 + b_1X_i))^2$
- Use calculus: take derivative with respect to b_0 and with respect to b_1 and set the two resulting equations equal to zero and solve for b_0 and b_1
- See KNNL pgs 16-17

Least Squares Solution

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

- **These are also maximum likelihood estimators for Normal error model, see KNNL pp 30-32**

Maximum Likelihood

$$\mathrm{Y_i} \sim N\left(\beta_0 + \beta_1 \mathrm{X_i}, \sigma^2\right)$$

$$f_\mathrm{i} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mathrm{Y_i} - \beta_0 - \beta_1 \mathrm{X_i}}{\sigma}\right)^2}$$

$L = f_1 \cdot f_2 \cdot \ldots \cdot f_\mathrm{n}$ (likelihood function)

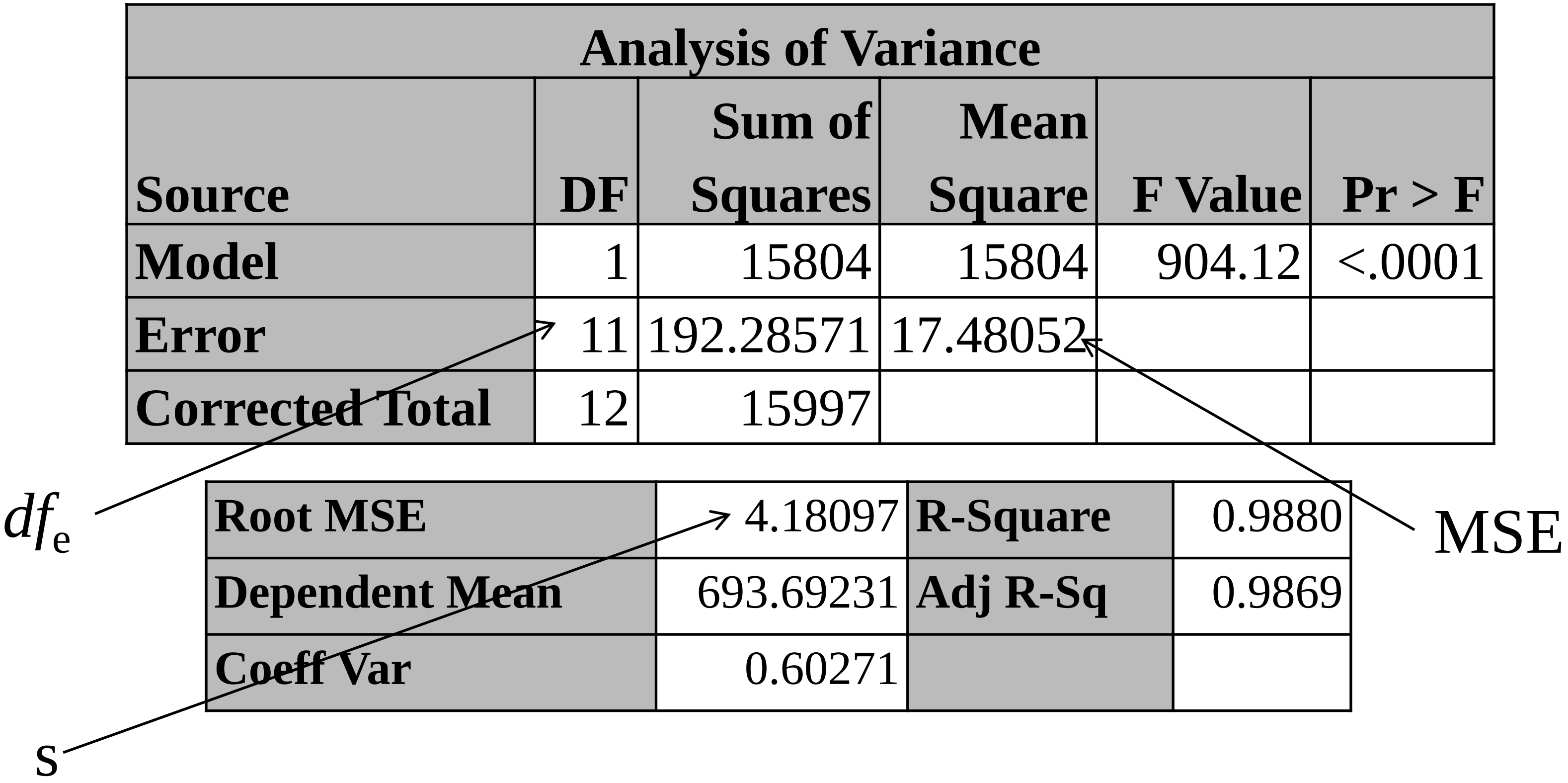
Find β_0 and β_1 which maximizes L

Estimation of σ²

$$s^2 = \frac{\sum\left(\mathrm{Y_i} - \mathrm{\hat{Y}_i}\right)}{\mathrm{n}-2} = \frac{\sum \mathrm{e_i}^2}{\mathrm{n}-2}$$

$$= \frac{SSE}{df_E} = MSE$$

$$s = \sqrt{s^2} = \mathrm{Root}\,MSE$$



Analysis of Variance					
		Sum of	Mean		
Source	DF	Squares	Square	F Value	Pr > F
Model	1	15804	15804	904.12	<.0001
Error	11	192.28571	17.48052		
Corrected Total	12	15997			

Root MSE	4.18097	R-Square	0.9880
Dependent Mean	693.69231	Adj R-Sq	0.9869
Coeff Var	0.60271		

Standard output from Proc REG

Properties of Least Squares Line

- The line always goes through $(\overline{\mathrm{X}}, \overline{\mathrm{Y}})$
- $$\begin{aligned} \sum e_i &= \sum (Y_i - (b_0 + b_1 X_i)) \\ &= \sum Y_i - \sum b_0 - \sum b_1 X_i \\ &= n\overline{Y} - nb_0 - nb_1\overline{X} = n((\overline{Y} - b_1\overline{X}) - b_0) \\ &= 0 \end{aligned}$$
- Other properties on pgs 23-24

Background Reading

- Chapter 1
  - 1.6 : Estimation of regression function
  - 1.7 : Estimation of error variance
  - 1.8 : Normal regression model
- Chapter 2
  - 2.1 and 2.2 : inference concerning β ’s
- Appendix A
  - A.4, A.5, A.6, and A.7