Lecture 4

- SIMPLE LINEAR REGRESSION

Leaning tower of Pisa

Example (1)

- response variable *the lean (Y)*
- explanatory variable *time (X)*
- plot
- fit a line
- predict the future

Example (2)

- Basic SAS, see program p1.sas on class website
- Simple linear regression model
- Estimation of regression parameters

SAS Data Step

```
data a1;
input year lean @@;
cards;
75 642 76 644 77 656 78 667 79 673 80 688
81 696 82 698 83 713 84 717 85 725 86 742
87 757 100 .
;
data a1p; set a1; if lean ne .;
```

SAS Proc Print

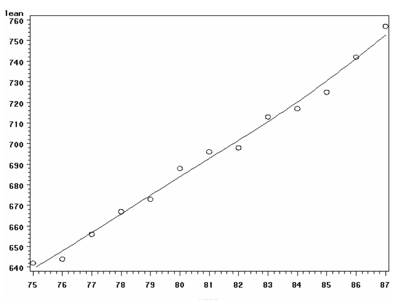
```
proc print data=a1; run;
```

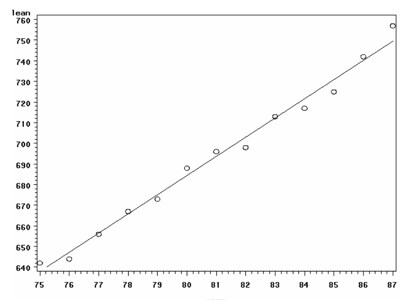
OBS	YEAR	LEAN
1	75	642
2	76	644
3	77	656
4	78	667
5	79	673
6	80	688
7	81	696
8	82	698
9	83	713
10	84	717
11	85	725
12	86	742
13	87	757
14	100	.

SAS Proc Gplot

```
symbol1 v=circle i=sm70s;
proc gplot data=a1p; plot lean*year;
run;

symbol1 v=circle i=rl;
proc gplot data=a1p; plot lean*year;
run;
```





SAS Proc Reg

```
proc reg data=a1;
   model lean=year/p r;
   output out=a2  p=pred r=resid;
   id year;
```

Variable	DF	Parameter Estimate	Standard Error
INTERCEP	1	-61.120879	25.12981850
YEAR	1	9.318681	0.30991420

T for H0: Parameter=0	Prob > |T|
-2.432	0.0333
30.069	0.0001

Obs	YEAR	Dep Var LEAN	Predict Value	Residual
1	75	642.0	637.8	4.2198
2	76	644.0	647.1	-3.0989
3	77	656.0	656.4	-0.4176
4	78	667.0	665.7	1.2637
5	79	673.0	675.1	-2.0549
6	80	688.0	684.4	3.6264
7	81	696.0	693.7	2.3077
8	82	698.0	703.0	-5.0110
9	83	713.0	712.3	0.6703
10	84	717.0	721.6	-4.6484
11	85	725.0	731.0	-5.9670
12	86	742.0	740.3	1.7143
13	87	757.0	749.6	7.3956
14	100	.	870.7	

Data for Simple Linear Regression

- Y_i the response variable
- X_i the explanatory variable
- for cases *i = 1 to n*

Simple Linear Regression Model

- $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
- Y_i is the value of the response variable for the i^{th} case
- β_0 is the intercept
- β_1 is the slope

Simple Linear Regression Model (2)

- X_i is the value of the explanatory variable for the i^{th} case
- ξ_i is a normally distributed random error with mean 0 and variance σ^2

Simple Linear Regression Model (3) Parameters

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

Features of Simple Linear Regression Model

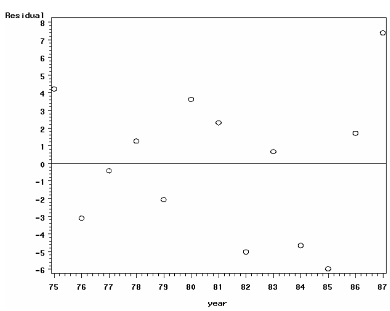
- $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
- $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$
- $Var(Y_i|X_i) = var(\xi_i) = \sigma^2$

Fitted Regression Equation and Residuals

- $\hat{Y}_i = b_0 + b_1X_i$
- $e_i = Y_i - \hat{Y}_i$, residual
- $e_i = Y_i - (b_0 + b_1X_i)$

Plot the residuals

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



Least Squares

- minimize $\Sigma(Y_i - (b_0 + b_1X_i))^2 = \sum e_i^2$
- use calculus
- take derivative with respect to b_0 and with respect to b_1
- set the two resulting equations equal to zero and solve for b_0 and b_1

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

Maximum Likelihood

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

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

$$L = f_1 \cdot f_2 \cdot \ldots \cdot f_n - \text{likelihood function}$$

Estimation of σ^2

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

$$= \frac{\text{SSE}}{\text{dfE}} = \text{MSE}$$

$$\text{s} = \sqrt{\text{s}^2} = \text{Root MSE}$$

```
                    Parameter      Standard
Variable  DF         Estimate         Error
INTERCEP   1       -61.120879   25.12981850
YEAR       1         9.318681    0.30991420

                          Sum of          Mean
Source          DF       Squares        Square
Model            1   15804.48352   15804.48352
Error           11     192.28571      17.48052
C Total         12   15996.76923

    Root MSE        4.18097
    Dep Mean      693.69231
    C.V.            0.60271
```

Theory for β_1 Inference

- $b_1 \sim$ Normal$(\beta_1, \sigma^2(b_1))$
- where $\sigma^2(b_1) = \sigma^2 / \Sigma(X_i - \overline{X})^2$
- $t = (b_1 - \beta_1)/s(b_1)$
- where $s^2(b_1) = s^2 / \Sigma(X_i - \overline{X})^2$
- $t \sim t(n-2)$

Confidence Interval for β_1

- $b_1 \pm t_c s(b_1)$
- where $t_c = t(1-\alpha/2, n-2)$, the upper
- $(1-\alpha/2)100$ percentile of the t distribution with n-2 degrees of freedom
- $1-\alpha$ is the confidence level

Significance tests for β_1

- H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$
- $t = (b_1 - 0)/s(b_1)$
- reject H_0 if $|t| \geq t_c$, where
- $t_c =$ $t(1-\alpha/2, n-2)$
- $P = \text{Prob}(|z| \geq |t|)$, where $z \sim t(n-2)$

Theory for β_0 Inference

- $b_0 \sim$ Normal$(\beta_0, \sigma^2(b_0))$
- where $\sigma^2(b_0)=$

$$\sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X_i - \overline{X}\right)^2}\right]$$

- $t = (b_0 - \beta_0)/s(b_0)$
- for $s(b_0)$, replace σ^2 by s^2
- $t \sim t(n-2)$

Confidence Interval for β_0

- $b_0 \pm t_c s(b_0)$
- where $t_c = t(1-\alpha/2, n-2)$, the upper
- $(1-\alpha/2)100$ percentile of the t distribution with n-2 degrees of freedom
- $1-\alpha$ is the confidence level

Significance tests for β_0

- H_0: $\beta_0 = \beta_{00}$, H_a: $\beta_0 \neq \beta_{00}$
- $t = (b_0 - \beta_{00})/s(b_0)$
- reject H_0 if $|t| \geq t_c$, where
- $t_c = t(1-\alpha/2, n-2)$
- $P = \text{Prob}(|z| \geq |t|)$, where $z \sim t(n-2)$

Notes (1)

- The normality of b_0 and b_1 follows from the fact that each of these is a linear combination of the Y_i, which are independent normal variables

Notes (2)

- Usually the CI and significance test for β_0 are not of interest
- If the ξ_i are not normal but are approximately normal, then the CIs and significance tests are generally reasonable approximations

Notes (3)

- These procedures can easily be modified to produce one-sided significance tests
- Because $\sigma^2(b_1) = \sigma^2 / \Sigma(X_i - \overline{X})^2$, we can make this quantity small by making $\Sigma(X_i - \overline{X})^2$ large.

SAS Proc Reg

```
proc reg data=a1;
   model lean=year/clb;
```

Variable	DF	Parameter Estimate	Standard Error
Intercept	1	-61.12088	25.12982
year	1	9.31868	0.30991

t Value	Pr > |t|	95% Confidence Limits	
-2.43	0.0333	-116.43124	-5.81052
30.07	<.0001	8.63656	10.00080

Power

- The *power* of a significance test is the probability that the null hypothesis is to be rejected when, in fact, it is false.
- This probability depends on the particular value of the parameter in the alternative space.

Power for β_1 (1)

- H_0: $\beta_1 = 0$, H_a: $\beta_1 \neq 0$
- $t = b_1/s(b_1)$
- $t_c = t(1-\alpha/2, n-2)$
- for α=.05 , we reject H_0 when $|t| \geq t_c$
- so we need to find $P(|t| \geq t_c)$ for arbitrary values of $\beta_1 \neq 0$
- when $\beta_1 = 0$, the calculation gives ?

Power for β_1 (2)

- $t \sim t(n-2,\delta)$ – noncentral *t* distribution
- $\delta = \beta_1/\sigma(b_1)$ – noncentrality parameter
- We need to assume values for
- $\sigma^2(b_1) = \sigma^2/\Sigma(X_i - \bar{X})^2$ and n

Example of Power Calculations for β_1

- we assume σ^2=2500 , n=25
- and $\Sigma(X_i - \bar{X})^2$ =19800
- so we have $\sigma^2(b_1) = \sigma^2/\Sigma(X_i - \bar{X})^2$= 0.1263

Example of Power (2)

- consider $\beta_1 = 1.5$
- we now can calculate $\delta = \beta_1/\sigma(b_1)$
- $t \sim t(n-2,\delta)$, we want to find $P(|t| \geq t_c)$
- we use a function that calculates the cumulative distribution function for the noncentral t distribution

```
data a1;
   n=25; sig2=2500; ssx=19800;
   alpha=.05; sig2b1=sig2/ssx; df=n-2;
   beta1=1.5;
   delta=beta1/sqrt(sig2b1);
   tc=tinv(1-alpha/2,df);
   power=1-probt(tc,df,delta)
          +probt(-tc,df,delta);
   output;
proc print data=a1;
run;
```

```
Obs     n    sig2     ssx     alpha
 1     25    2500    19800     0.05

sig2b1    df    beta1     delta
0.12626   23     1.5     4.22137

tc         power
2.06866    0.98121
```

```
data a2;
   n=25; sig2=2500; ssx=19800;
   alpha=.05; sig2b1=sig2/ssx; df=n-2;
   tc=tinv(1-alpha/2,df);
   do beta1=-2.0 to 2.0 by .05;
      delta=beta1/sqrt(sig2b1);
      power=1-probt(tc,df,delta)
             +probt(-tc,df,delta);
      output;
   end;
```

```
title1 'Power for the slope in
   simple linear regression';
symbol1 v=none i=join;
proc gplot data=a2;
   plot power*beta1;
proc print data=a2;
run;
```