Lecture 1

- 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

```
piza<-read.table('piza.txt', header=TRUE)
piza <- data.frame (piza)
reg1<-lm(LEAN~YEAR, piza )
c1<-predict.lm(reg1)
Plot(LEAN~YEAR, piza)
lines(c1~YEAR, piza)
summary.lm(reg1)
new <- data.frame(YEAR = c(100))
u<-predict(reg1, new)
```

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

- $\hat{Y}_i = b_0 + b_1 X_i$
- $e_i = Y_i - \hat{Y}_i$, residual
- $e_i = Y_i - (b_0 + b_1 X_i)$

Least Squares

- minimize $\Sigma(Y_i - (b_0 + b_1 X_i))^2 = \sum e_i^2$
- 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

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

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

- These are also maximum likelihood estimators

Maximum Likelihood

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i$$
$$Y_i \sim N\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$$
$$f_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2}$$
$$L = f_1 \cdot f_2 \cdot \ldots \cdot f_n - \text{likelihood function}$$

Estimation of σ^2

$$s^2 = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$
$$= \frac{\text{SSE}}{\text{dfE}} = \text{MSE}$$
$$\text{s} = \sqrt{\text{s}^2} = \text{Root MSE}$$

Theory for β_1 Inference

- **$b_1 \sim$ Normal(β_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(β_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 $\sigma^2(b_1)=\sigma^2/\Sigma(X_i - \bar{X})^2$, we can make this quantity small by making $\Sigma(X_i - \bar{X})^2$ large.

```
reg1<-lm(LEAN~YEAR, piza )
summary(reg1)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-61.1209	25.1298	-2.432	0.0333	*
YEAR	9.3187	0.3099	30.069	6.5e-12	*

Residual standard error: 4.181 on 11 degrees of freedom confint(reg1)

	2.5 %	97.5 %
(Intercept)	-116.431237	-5.810522
YEAR	8.636565	10.000798

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 β_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

```
n<-25;
sig2<-2500;
ssx<-19800;
alpha<-.05;
sig2b1<-sig2/ssx;
df=n-2;
tc<-qt(1-alpha/2,df);
beta1<-seq(from=-2.0, to= 2.0, by= .05);
delta<-beta1/sqrt(sig2b1);
prob1<-function(delta){pt(tc,df,delta)}
prob2<-function(delta){pt(-tc,df,delta)}
power<-1-prob1(delta)+prob2(delta);
plot(beta1,power,type='l')
```

Power for the slope in simple linear regression