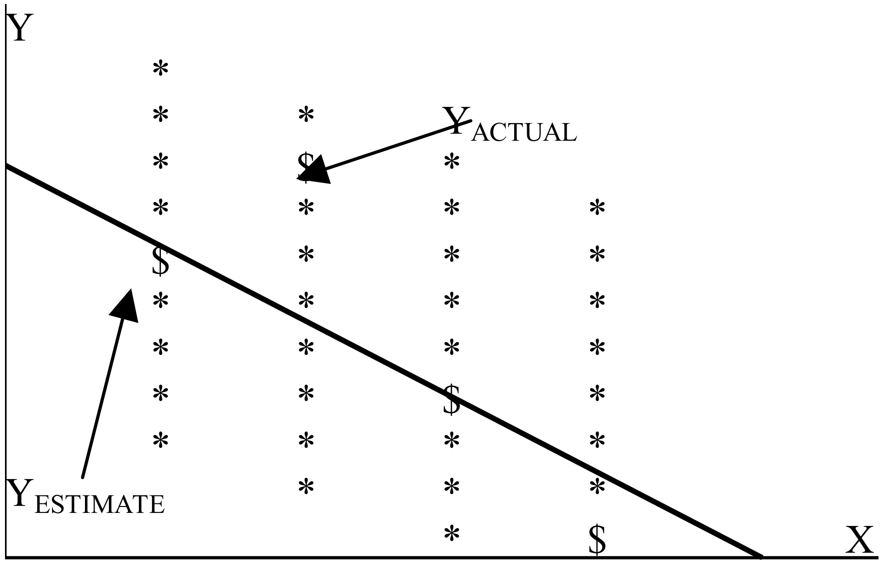
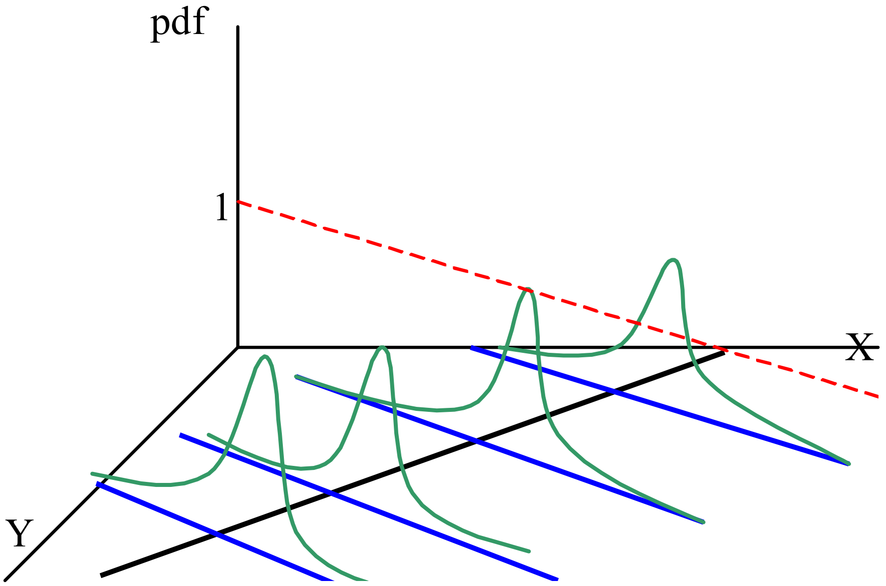
OLS Theory

1. Regression: Curve -Fitting

a. Homoskedastic Data Distribution



b. Probability Distribution of Homoskedstic Sample Distribution



c. Two Properties that Characterize Sample Distribution

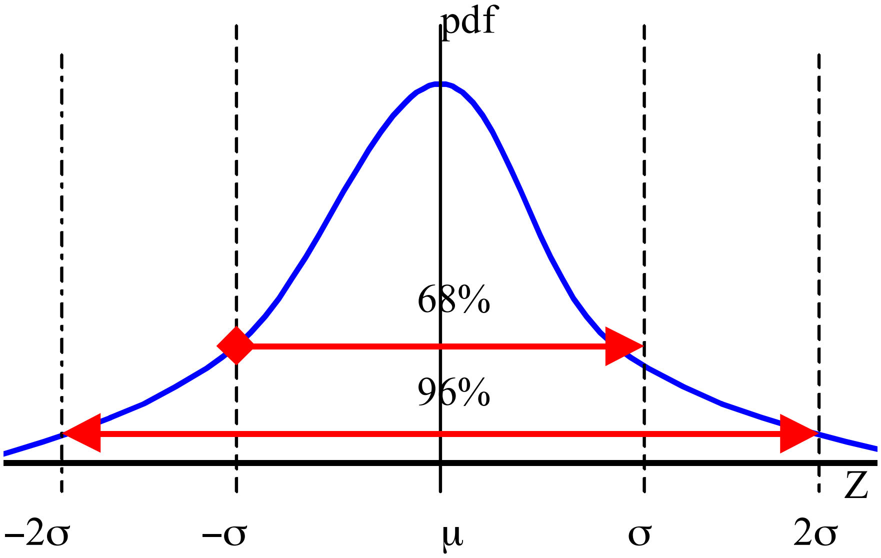
i) First Moment: $E[X] = \sum_{1}^{n} f(X_i)X_i = \dfrac{\sum_{1}^{n} X_i}{n} = \overline{X}$

ii) Second Moment: $VAR[X] = \dfrac{\sum_{1}^{n}(X_i - \overline{X})^2}{n-1} \rightarrow Stdev[X] = \sqrt{\dfrac{\sum_{1}^{n}(X_i - \overline{X})^2}{n-1}}$

iii) Sample & Population

First Moment = E[X]	$\overline{X}$	μ_X
Second Moment = VAR[X]	S_X^2	σ_X^2

d. Characteristics Normal Distribution (n≥30)



e. Sampling Distribution via Law of Large Numbers or Central Limit Theorem

Population (True) Mean μ_X	Mean of Means $E[\overline{X}]$	Sample Means ($\overline{X}_i$)	Samples (i = A...Z))
	Sampling Distribution	$\overline{X}_A = 25$	A = {n = 200}
		$\overline{X}_B = 50$	B = {n = 200}
μ_X =	$E[\overline{X}] = \dfrac{\sum_{A}^{Z} \overline{X}_i}{26}$	$\vdots$	$\vdots$
		$\overline{X}_Z = 65$	Z = {n = 200}

2. Ordinary Least Squares Method

- Assumptions

i) $\varepsilon_i \sim ND(0, \sigma^2)$

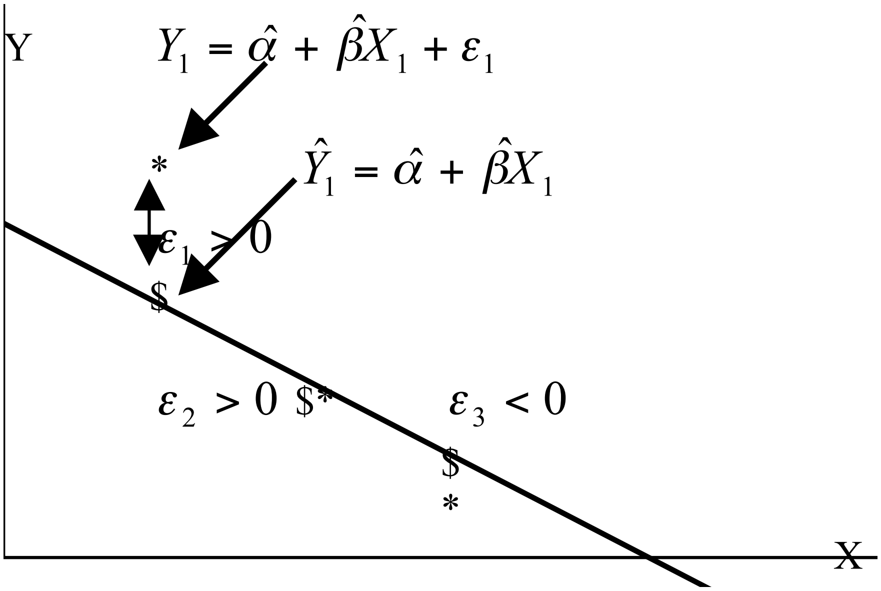
ii) $X_i \neq X_j \quad \forall \quad i \neq j$

iii) $E[X, \varepsilon] = \sigma_{X,\varepsilon} = 0$

iv) $E[\varepsilon_i, \varepsilon_j] = \sigma_{ij} = 0$

v) σ_ε^2 is homoskedastic.

- Estimates & Actual Data Points



1. 1. $\min\limits_{\beta} \sum_{t=1}^{n} \varepsilon_t^2 \rightarrow$

   $i)\quad FOC: \quad \frac{\partial \sum \varepsilon_t^2}{\partial \beta} = \frac{\partial \sum (Y_t - \hat{Y}_t)^2}{\partial \beta} = 0$

   $ii)\quad SOC: \quad \frac{\partial^2 \sum \varepsilon_t^2}{\partial \beta^2} > 0$

2. Solution by Matrix: $X'\varepsilon = 0,$ where $\varepsilon = y - X\hat{\beta}$

$$\therefore X'(y - X\hat{\beta}) = 0$$

$$X'y - X'X\hat{\beta} = 0$$

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

3. Solution by Normal Equation:

$$\sum \varepsilon_t = 0 = \sum (Y_t - \hat{\alpha} - \hat{\beta}X_t) = \sum Y_t - n\hat{\alpha} - \hat{\beta}\sum X_t$$

By definition $\frac{\sum \varepsilon_t}{n} = 0 = \frac{1}{n}\sum Y_t - \hat{\alpha} - \frac{\hat{\beta}}{n}\sum X_t$

$$\frac{\sum Y_t}{n} = \hat{\alpha} + \frac{\hat{\beta}}{n}\sum X_t \rightarrow \overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X}$$

$$\sum X_t\varepsilon_t = \sum X_t(Y_t - \hat{\alpha} - \hat{\beta}X_t) = 0$$
$$= \sum X_tY_t - \hat{\alpha}\sum X_t - \hat{\beta}\sum X_t^2 = 0$$
$$\sum X_tY_t = \hat{\alpha}\sum X_t + \hat{\beta}\sum X_t^2$$

, where $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = \frac{1}{n}\sum Y_t - \frac{\hat{\beta}}{n}\sum X_t$

$$= \left(\frac{\sum Y_t}{n} - \frac{\hat{\beta}}{n}\sum X_t\right)\sum X_t + \hat{\beta}\sum X_t^2$$
$$= \frac{\sum X_t \sum Y_t}{n} + \hat{\beta}\sum\left(X_t^2 - \frac{X_t^2}{n}\right)$$

$$\hat{\beta}\sum\left(X_t^2 - \frac{X_t^2}{n}\right) = \sum X_tY_t - \frac{\sum X_t\sum Y_t}{n} \rightarrow \hat{\beta} = \frac{\sum X_tY_t - \frac{\sum X_t\sum Y_t}{n}}{\sum X_t^2 - \frac{\sum X_t^2}{n}} = \frac{\sigma_{XY}}{\sigma_X^2}$$ since

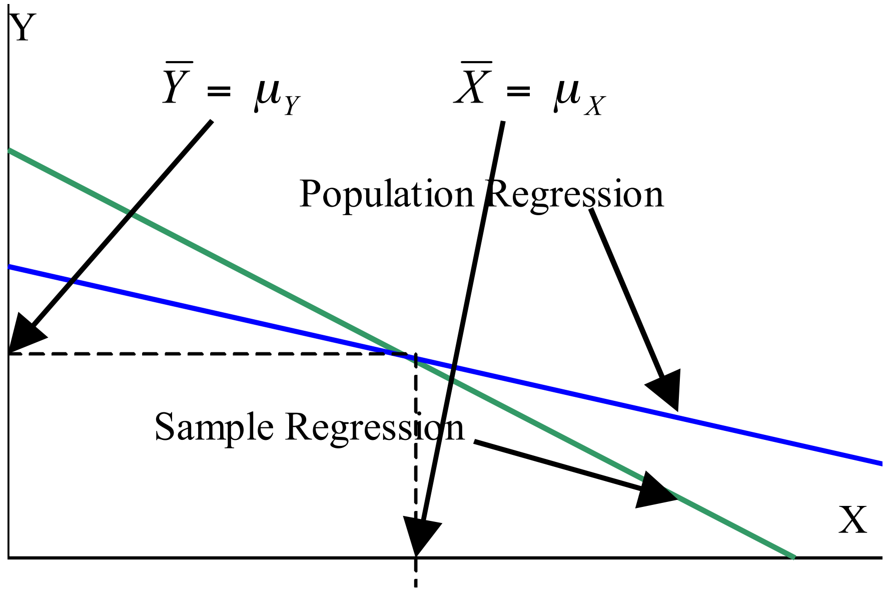
$$\sigma_{XY} = \sum(X_t - \overline{X})(Y_t - \overline{Y})$$
$$= \sum X_tY_t - \sum \overline{X}Y_t - \sum X_t\overline{Y} + \sum \overline{X}\overline{Y}$$
$$= \sum X_tY_t - \overline{X}\sum Y_t - \overline{Y}\sum X_t + n\overline{X}\overline{Y}$$
$$= \sum X_tY_t - \overline{X}n\overline{Y} - \overline{Y}n\overline{X} + n\overline{X}\overline{Y}$$
$$= \sum X_tY_t - 2n\overline{X}\overline{Y} + n\overline{X}\overline{Y}$$
$$= \sum X_tY_t - n\overline{X}\overline{Y} = \sum X_tY_t - \frac{\sum X_tY_t}{n}$$

$$\sigma_X^2 = \sum(X_t - \overline{X})^2$$
$$= \sum X_t^2 - 2\overline{X}\sum X_t - \sum \overline{X}^2$$
$$= \sum X_t^2 - 2\overline{X}n\overline{X} + n\overline{X}^2 = \sum X_t^2 - n\overline{X}^2$$
$$\because \overline{X} = \frac{\sum X_t}{n} \quad or \quad n\overline{X} = \sum X_t$$

$$\therefore \hat{\beta}_{Xy} = \frac{\sigma_{Xy}}{\sigma_X^2}$$

$cf.)\ \hat{\rho}_{Xy} = \frac{\sigma_{Xy}}{\sigma_X\sigma_y}$

- LLN & CLT guarantee $E[Y] = \overline{Y} = \mu_Y$



OLS Code

a. SAS

```
filename chev 'a:\companies\chevron(chv)\chev.prn';
data chev;
infile chev;
input  quarter rit    delta1  delta   chgdivt          rmt     invcons;
proc print data=chev;
proc reg;
model rit = delta1      delta      chgdivt      rmt      invcons;
run;
```

b. Gauss

```
new;
xx={r,c};
xx;
n=rows(xx);
" no. of observations=n= " n;
x=ones(119,1)~xx[.,2:6];
" matrix of regressors ";
x;
bigx=x;
k=cols(x);
"k=cols(x)= no of regressors including one for intercept" k;
y=xx[.,1];
" vector of dep variable data ";
y;
xtx=x'x;
"xtx";
xtx;
invxtx=invpd(xtx);
"invxtx";
invxtx;
xty=x'y;
"xty";
xty;
b=invxtx* xty;
"b";
b;
resid=y-x*b;
"resid";
resid;
"residual sum of squares =rss=resid'resid";
rss=resid'resid;
rss;
df=n-k;
"df=degrees of freedom= n-k" df;
resms=rss/df;
"est of sigma-sq= residual mean square = rss/df " resms;
" standard error of regression " sqrt(resms);
ybar=meanc(y);
"ybar=meanc(y)" ybar;
```

```
TotSS= (y-ybar*ones(n,1))'(y-ybar*ones(n,1));
" total sum of sq= (y-ybar*ones(n,1))'(y-ybar*ones(n,1))" totss;
" R-sq  by first method = 1-(rss/totss)";
rsq=1-(rss/totss);rsq;
"predicted values";
pred=x*b;pred;
ypbar=meanc(pred);
"ypbar=meanc(pred)" ypbar;
RegSS= (pred-ypbar*ones(n,1))'(pred-ypbar*ones(n,1));
"Regr sum of sq" regss;
" R-sq=  RegrSS/totalSS by 2nd method " regss/totss;
AdjRsq= 1-(1-rsq)*(n-1)/(n-k);
" Adjusted R-sq= 1-(1-rsq)*(n-1)/(n-k)" Adjrsq;
" F value = [(regss/(k-1)]/[rss/(n-k)]";
Fval = (regss/(k-1))/(rss/(n-k));fval;
"  ";
"Akaike Information criterion AIC=ln(residual-mean-sq) + 2*k/n";
aic=ln(resms) + (2*k)/n;aic;
"Amemiya prediction criterion AmPC=(residual-mean-sq)(1 + k/n)";
ampc=(resms)*(1 + (k/n));ampc;
"  ";
   { vnam,mean,var,std,min,max,valid,mis } = DSTAT(0,xx);
" ";
__con=1;
_olsres=1;
 {vnam,mom,bols,stb,vc,stderrb,sigmares,corrxy,rsqols,residols,dwols} =
                                        OLS(0,y,bigx);
" ";
varcovb=resms*invxtx;
"varcovb=resms*invxtx";
varcovb;
" ";
"Now standard errors of regr coeff's got directly from cov mtx of b";
se1=sqrt(varcovb[1,1]);
"se1=sqrt(varcovb[1,1])" se1;
se2=sqrt(varcovb[2,2]);
"se2=sqrt(varcovb[2,2])" se2;
se3=sqrt(varcovb[3,3]);
"se3=sqrt(varcovb[3,3])" se3;

"Now t-values as ratios of regr coeff to se's ";
t1=b[1]/se1;t1;
t2=b[2]/se2;t2;
t3=b[3]/se3;t3;
end;
```

Dummy Variable

$y = \alpha + \beta_i X_i + \varepsilon_i$, where $\begin{array}{ll} y = \alpha + \beta_i + \varepsilon_i & if\ X_i = 1 \\ y = \alpha + \varepsilon_i & if\ X_i = 0 \end{array}$

$\therefore H_0 : \beta_i = 0, \quad i.e. \quad \left| t_\beta \right| < 2\ldots$

Interactive Effect

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_i X_i + \gamma Z + \gamma_1 X_1 Z + \gamma_2 X_2 Z + \ldots + \gamma_i X_i Z + \varepsilon$,

where X= degree or qualification
Z= public/private institution
XZ= degree & public/private institution (interactive term)

Dummy Variable Trap

If $X_3 = 1 - (X_1 + X_2)$, specifying X_1, X_2, X_3 is wrong because of singularity.

Seasonality Adjustment (or Deseasonlization)

$y_t = \alpha + \beta X_t + \gamma_2 Q_{t2} + \gamma_3 Q_{t3} + \gamma_4 Q_{t4} + \varepsilon_t$, where Q_{t2} = 1 if t = 2 or 0 otherwise
Q_{t3} = 1 if t = 3 or 0 otherwise
Q_{t4} = 1 if t = 4 or 0 otherwise.

$$\begin{array}{lcll} & & \alpha = \alpha_1 & Y_t = \alpha_1 + \beta X_t + \varepsilon_t \\ H_0 : \gamma_2 = \gamma_3 = \gamma_4 = 0 & \text{If} & \alpha + \gamma_2 & Y_t = \alpha_2 + \beta X_t + \varepsilon_t \\ H_a : H_{0,}^{c} \quad or \quad H_0^{'} & & \alpha + \gamma_3 & Y_t = \alpha_3 + \beta X_t + \varepsilon_t \\ & & \alpha + \gamma_4 & Y_t = \alpha_4 + \beta X_t + \varepsilon_t \end{array}$$

SAS Code

```
data dummy.prn dummy;
infile “c:\***\dummy;
input x1, x2, x3 … Xi;
if x>$1,000, then x=1;
else x=0;
/ * x1=1 if X1=PhD, else Xi=0
x2=1 if Xi=MA, else Xi=0
x3=1 if xi=BA, else x3=0*/
model y=x1, x2,x3, … xi;
run;
```