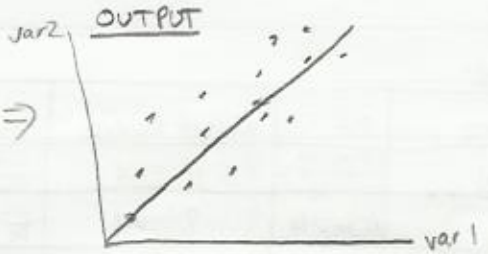
CH 1

- To see the statistical relationship of 2 variables we can look at a scatterplot

CODE

```
PROC SGSCATTER;
   Plot var1*var2;
run;
```

⇒ OUTPUT



- Basic Regression Equation: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

NOTE: $E(Y_i) = \beta_0 + \beta_1 X_i$

$\text{var}(Y_i) = \sigma^2$ as $\epsilon_i \sim N(0, \sigma^2)$

CODE

```
PROC GLM;
   Model Y = X1;
run;
```

OUTPUT

Tests that all model coefficients = 0

Source	Df	Sum of Squares	Mean Square	F Value	Pr > F
Model	# of estimated coefficients = 0 initialized, $2_0 - 1_n$	$\sum(\hat{y}_i - \bar{y})^2 = SSR$	$\frac{\text{Sum of Squares - Model}}{Dm} = MSR$	$\frac{MSR}{MSE}$	p-value
Error	$n - Dm - 1 = DE$	$\sum(y_i - \hat{y})^2 = SSE$	$\frac{\text{Sum of Squares - Error}}{DE} = MSE$		
Corrected total	$n - 1$	$\sum(y_i - \bar{y})^2 = SSTO$			

R-Square	Coeff Var	Root MSE	Y mean
$1 - \frac{SSE}{SSTO}$ = variation explained by model	$\frac{\text{root MSE}}{\text{Y mean}} \times 100$	$\sqrt{MSE}$ = standard deviation of ϵ_i	$\frac{1}{n}\sum Y_i$

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X_1					

Source	DF	Type II SS	Mean Square	F Value	Pr > F
X_1					

Parameter	Estimate	Standard error	t Value	Pr > |t|
intercept	$\hat{\beta}_0$	$se(\hat{\beta}_0)$	$\frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$	
X_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$	

Tests that each coefficient = 0

CODE

```
Proc reg;
   model y = x1;
run;
```

Output

Tests whether all estimated coefficients are 0

Source	Df	sum of squares	mean square	F value	Pr > F
Model	# of coefficients estimated (which = # of variables including the intercept)	$\sum_i (\hat{y}_i - \bar{y})^2 = SSR$	$\frac{SSR}{Dm} = MSR$	$\frac{MSR}{MSE}$	
Error	$n - Dm - 1 = De$	$\sum_i (y_i - \hat{y})^2 = SSE$	$\frac{SSE}{De} = MSE$		
Corrected total	$n-1$	$\sum_i (y_i - \bar{y})^2 = SSTO$			

Root MSE	$\sqrt{MSE}$ ≈ standard deviation of ε	R-Square	$1 - \frac{SSE}{SSTO}$
Y Mean	$\frac{1}{n}\sum_i Y_i$	Adj-R-Sq	$1 - \frac{(n-1)(1-R^2)}{n-Dm-1}$
Coeff Var	$\frac{\sqrt{MSE}}{Ymean} \times 100$		

Tests whether the coefficients are 0

Variable	Df	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1 – constant c-1 – for class	$\hat{\beta}_0$	$se(\hat{\beta}_0)$	$\frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$	
X_1	"	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$	

Method of Least Squares

- Sum of Squares : $Q = \sum_i^n (Y_i - \hat{Y}_i)^2 = \sum_i^n (Y_i - \beta_0 - \beta_1 X_i)^2$
- Our least squares estimates minimize Q
- Normal equations

$$\sum_i^n Y_i = nb_0 + b_1 \Sigma X_i$$

$$\Sigma X_i Y_i = b_0 \sum_i^n X_i + b_1 \sum_i^n X_i^2$$

* see page 17 for derivation

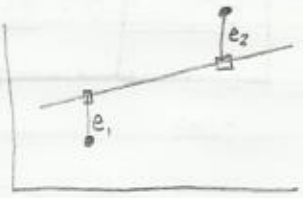
- solving simultaneously:

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

$$b_0 = \frac{1}{n}\left(\sum_i^n Y_i - b_1 \sum_i^n X_i\right) = \bar{Y} - b_1 \bar{X}$$

- Gauss-Markov Theorem : Least squares estimators are unbiased and have minimum variance among unbiased estimators

- Residuals $e_i = Y_i - \hat{Y}_i$



① $\sum_i^n e_i = 0$

② $\sum_i^n e_i^2$ is a min w/ least squares

③ $\sum_i^n Y_i = \sum_i^n \hat{Y}_i$

④ $\sum_i^n X_i e_i = 0$

⑤ $\sum_i^n \hat{Y}_i e_i = 0$

⑥ Regression line goes through $(\bar{X}, \bar{Y})$

- S^2 estimates σ^2 : $S^2 = \frac{1}{n-1} \sum_i^n (Y_i - \bar{Y})^2$
- MSE estimates σ^2 : $MSE = \frac{SSE}{n-2}$ (for a model w/ 1 ind. var)

Method of maximum Likelihood

Example: • Given: $\sigma = 10$ $\mu = ?$

• A random sample of $n=3$: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$

• Assuming $Y \sim N(\mu, 100)$ we wish to maximize

$$L(\mu) = \prod \frac{1}{10\sqrt{2\pi}} e^{\frac{-1}{200}(Y_i - \mu)^2}$$

Answer $\mu = 258$

For Regression

$$L(\beta_0, \beta_1, \sigma^2) = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$$

β_0	$\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$
β_1	$\frac{1}{n}(\sum Y_i - b\Sigma X_i)$
σ^2	$\hat{\sigma}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n}$ ← biased, so we use MSE