Statistics & Regression Easier than SAS®

Vincent Maffei, Anthem BC&BS
Michael Davis, Bassett Consulting Svcs. Inc.
NESUG, September 9, 2003

Normal Distributions of:

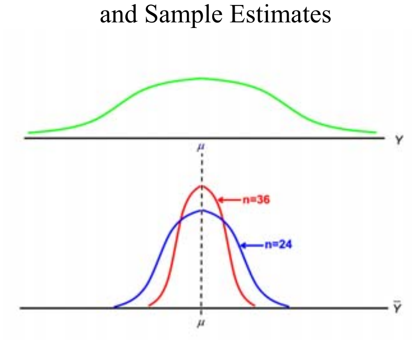
<u>Continuous Variables</u>
Hourly Wage SAS® Consultants

$Y \sim N(\mu, \sigma^2)$
$Y \sim N(150, 324)$

$\bar{Y} \sim N(\mu, \sigma^2/n)$

$\bar{Y} \sim N(150, {}^{324}/_{36})$

Normal Distributions of Populations and Sample Estimates



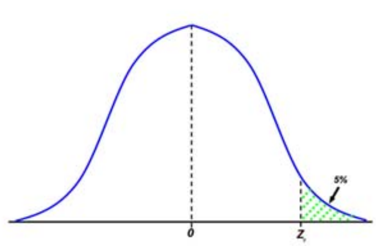
Hypothesis Testing

- Avg. hourly wage SAS® Consultants > $150.00

  Null Hypothesis: H_0: μ = $150
  Alt. Hypothesis: H_A: μ > $150
- Test: Take a sample. Calculate sample mean.
- Check to see if difference between sample mean and hypothesized value too large to be explained by random variation.
- Calculate 'Z' value, compare to critical Z.

  (From 'Z' table)

Unit Normal Distribution



Hypothesis Testing

H_0: μ = $150
H_A: μ > $150

$Z = {}^{\text{observation} - H_0 \text{ value}}/_{\text{std deviation}}$

$z = \dfrac{(x - \mu)}{\sigma/(n)^{1/2}}$

$\alpha = 5\%$ and $n = 36$ ➠ $Z_c = 1.645$

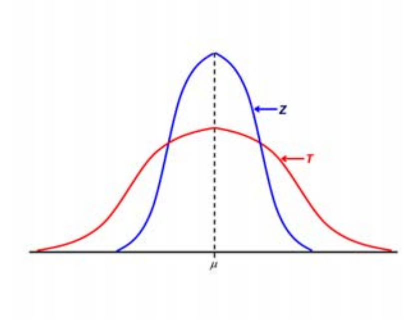
If $Z \leq 1.645$ then accept H_0

$1.33 = \dfrac{(154 - 150)}{{}^{18}/(36)^{1/2}}$

Unknown Population Variance

- σ^2 unknown? Use s^2 as an estimate.
- Introduces error since $s^2 \neq \sigma^2$
- T distribution accommodates error with more area in tails.
- T = $^{\text{observation – mean}}/_{\textbf{estimated}\text{ std deviation}}$

‘*z*’ vs. ‘*t*’ Distributions



Multivariate Statistics (Regression)

$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

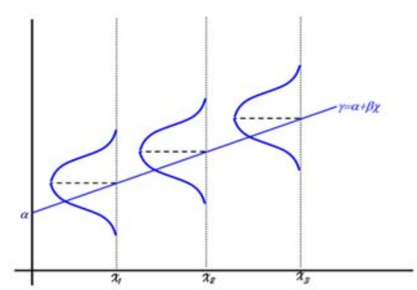
where:

α = Vertical Intercept

β = Change in Y for a unit change in X

ε = Random disturbance term

Classic Regression Description



Regression

- PROC REG;
  MODEL dependant var = explanatory vars ;
- Uses:
  - Prediction
  - Hypothesis Testing
  - Estimation

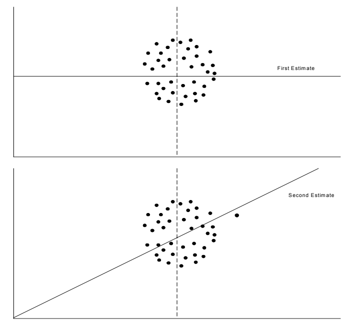
Estimation Error

- True Line: $Y = \alpha + \beta x + \varepsilon$
- Estimated: $y^e = a + bx$
- Error = $|y^e - Y|$
- Sources of error:
  a is an estimate of α
  b is an estimate of β
  $|\varepsilon|$ will be positive

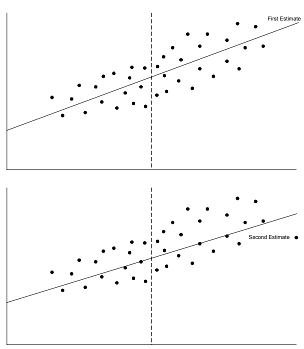
Reliable Estimates & Accurate Predictions

- Large Sample Size
- Large **range** of data in explanatory variables
- Predict for values close to data range

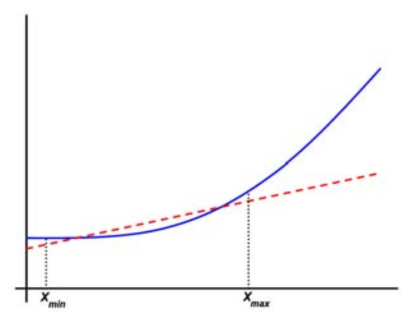
Small Data Ranges



Large Data Ranges



Predict Close to Range



Hypothesis Testing

H_o: $\beta = 0$

H_a: $\beta \neq 0$

$T = {}^{\text{estimate} - \text{mean}}/_{\text{standard error}}$

$T = {}^{(b - \beta)}/\sigma_\beta$

Same rules as before

- Use DF instead of n-1
- DF = n - # parameters estimated

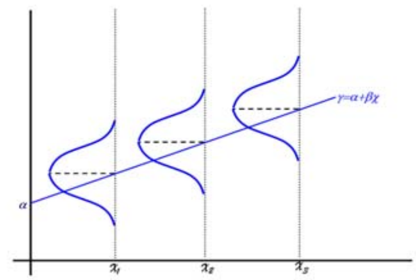
Hypothesis Testing

Variable	Parameter Estimate	Standard Error	T for H_0 Parameter = 0	Prob>|T|
Intercep	23.075672	4.41783713	5.223	0.0001
Income	8.349662	4.5329326	1.842	0.0658
F13-19	1.569866	6.22444137	0.252	0.8009
F20-34	34.131637	4.88649271	6.985	0.0001
F35-49	27.124441	4.90168423	5.534	0.0001
M13-17	-12.316362	5.67993991	-2.168	0.0301
M35-49	9.390942	5.25130049	1.788	0.0737

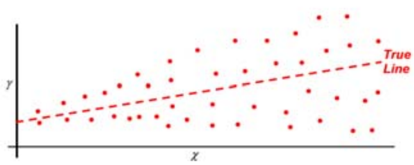
Regression

- Assumptions:
  - Linear Relationship
  - Constant disturbance term variance
  - Independent disturbance terms
  - Independent Explanatory Variables

Description of Classical Assumptions



Heteroscedasticity



Heteroscedasticity

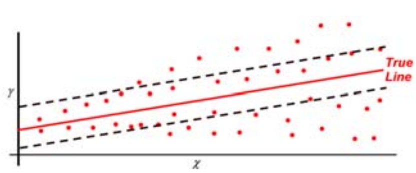
$\sigma^2_\varepsilon = Y \times \sigma^2_u$ or $\sigma^2_\varepsilon = X^2 \times \sigma^2_u$

PROC REG ;
MODEL /SPEC; ➠ χ^2

PROC REG ; WEIGHT w ;
MODEL ;

where: $w = {}^1/Y,\ {}^1/Y^2$, or ${}^1/X^2$

Heteroscedasticity



Serial Correlation

$\varepsilon_t = \lambda\varepsilon_{t-1} + u_t$

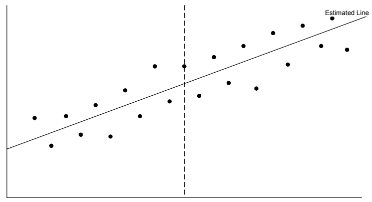
where $0 \le \lambda \le 1$

MODEL {option} /DW ;

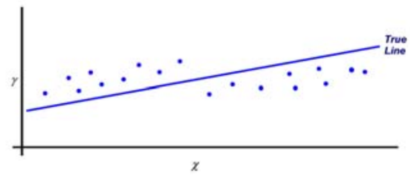
- ✓ $1.8 \le DW \le 2.2$ is good
- ✓ $1.6 \le DW \le 2.4$ is fair
- ✓ $DW > 2.4$ or < 1.6 not good

PROC AUTOREG ;
MODEL {option} /NLAG=n ;

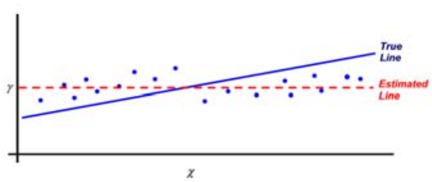
Negative Serial Correlation



Positive Serial Correlation



Positive Serial Correlation



Multicollinearity

- X_1 is a linear combination of the other X's
- Perfect Collinearity ➠ 'Model is not full rank' msg.
- Near perfect ➠ Explanatory variables come up insignificant, unstable parameter estimates.

Model {options}

COLLIN or COLLINOINT;

```
                    Estimate of Pre-existing Parameters:
Model:  MODEL1
Dependent Variable: COSTS

                          Analysis of Variance:

                                Sum of         Mean
         Source          DF    Squares       Square      F Value      Prob>F

         Model           86  2656507331.1 30889620.129    426.670      0.0001
         Error       100722  7291968328.4 72396.977109
         C Total     100808  9948475659.4

           Root MSE     269.06686     R-square        0.2670
           Dep Mean      73.51766     Adj R-sq        0.2664

Note: Model is not full rank. Least-squares solution for the parameters are not
      unique. Some Statistics will be misleading. A reported DF of 0 or B means that the
      estimate is biased. The following parameters have been set to 0, since the
      variables are a linear combination of other variable as shown.

        D07XD20 = 0
        D09XD21 = 0
        D09XD27 = 0
        D09XD29 = 0
```

Dummy (Binary) Variables

$CC = \alpha + \beta GDP + \varepsilon$ Simple

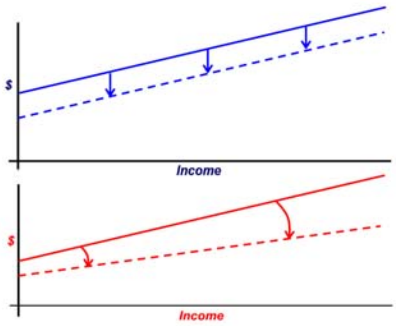
$CC = \alpha + \beta GDP - \delta D + \varepsilon$ Better

where: D = 0 pre Sept 11
D = 1 post Sept. 11

$CC = \alpha + \beta GDP + \varepsilon$ pre Sept 11

$CC = (\alpha - \delta) + \beta GDP + \varepsilon$ post Sept 11

Binary Variables



Dummy (Binary) Variables with Interaction Terms

$CC = \alpha + \beta GDP + \delta D + \lambda D \times GDP + \varepsilon$ Best

$CC = \alpha + \beta GDP + \varepsilon$ pre Sept 11

$CC = (\alpha - \delta) + (\beta - \lambda)GDP + \varepsilon$ post Sept 11

Dummy (Binary) Variables

$CC = \alpha + \beta GDP + \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \varepsilon$

where:

$D_1 = 1$ 1st qtr, 0 otherwise
$D_2 = 1$ 2nd qtr, 0 otherwise
$D_3 = 1$ 3rd qtr, 0 otherwise

Do Not Do:

$CC = \alpha + \beta GDP + \delta D + \varepsilon$

where: $D = 1, 2, \text{ or } 3$

Contact Information

Michael Davis
Bassett Consulting Services, Inc.
North Haven, Connecticut
Tel: 203-562-0640
E-mail: michael@bassettconsulting.com
Web: http://www.bassettconsulting.com

Vincent Maffei
Anthem Blue Cross & Blue Shield
North Haven, Connecticut
Tel: 203-985-7188
E-mail: vincent.maffei@anthem.com