# STOR 455
# STATISTICAL METHODS I

Jan Hannig

# Multivariate Regression

- $Y = \mathbf{X}\beta + \varepsilon$
  - $\mathbf{X}$ is a regression matrix, $\beta$ is a vector of parameters and $\varepsilon$ are independent $N(0,\sigma)$
- Estimated parameters $b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$
- Predicted responses $\hat{Y} = \mathbf{H}Y$, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- Residuals $e = (\mathbf{I} - \mathbf{H})Y$
- Estimated regression variance $s^2 = e'e/(n-p)$

# Board Example

- Simple linear regression

# Residual Analysis (Section 4.5)

- Recall $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ - matrix $\mathbf{H}=(h_{ij})$
- Standardized residuals

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

- Similarly as with SLR we should look at
  - Plot of r vs predictors $X_i$ (p –plots)
  - Plot of r vs predicted values $\hat{Y}$
  - Gaussian QQ- Plot of r

# Do It in SAS

```
*Data shown on page 237 of the OPTIONAL
  textbook - file CH06FI05.txt;
data studios;
  input x1 x2 y;
  x1x2=x1*x2;
  label x1='targtpop'
        x2='dispoinc';
cards;
  68.5  16.7  174.4
  45.2  16.8  164.4
  91.3  18.2  244.2
  ...

  52.3  16.0  166.5
  ;

run;
```

# Do it in SAS

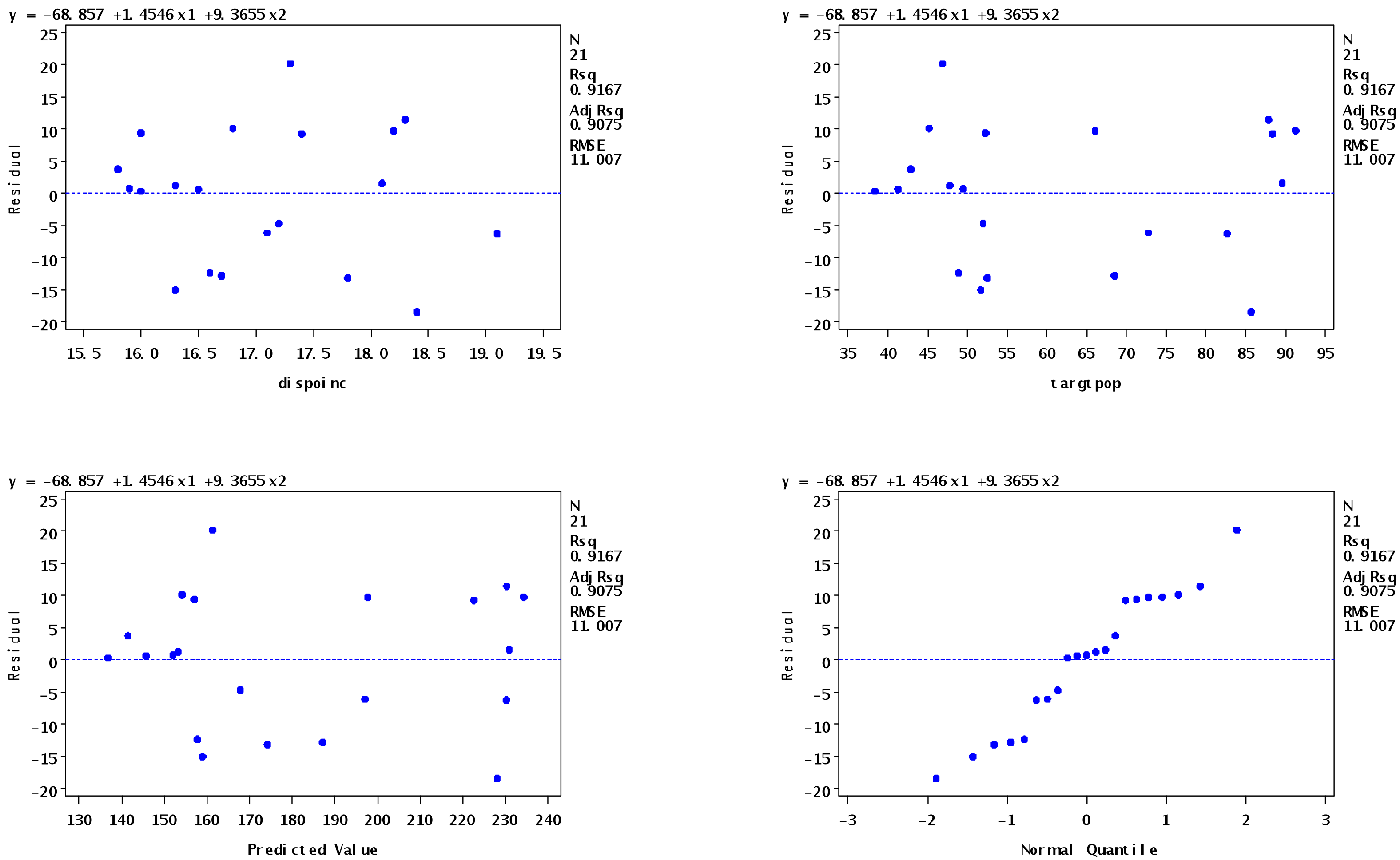
```
* plot residuals, QQ plot;
proc reg data = studios
   noprint;
  model y = x1 x2;
  plot student. * (x1 x2
   p.) ;
  plot student. * nqq.;
run;


*Alternative way of
   plotting;
proc reg data = studios;
  model y = x1 x2;
  output out=output p =
   fitted student =
   residual;
run;
```

```
proc gplot data = output;
  plot residual*fitted;
  plot residual*x1;
  plot residual*x2;
  plot residual*x1x2;
run;

proc univariate data =
   output noprint ;
  qqplot residual / normal;
run;
*End of an alternative way
   of plotting;
```

# Do it in SAS




10/14/10

# Inference for b (Section 4.6+4.7)

- b ~ N($\beta$, $\sigma^2(X'X)^{-1}$)
- Estimate Var($b_i$) by $s^2(b_i)$ =MSE* $(X'X)^{-1}_{i,i}$
- CI: $b_i \pm t^* s(b_i)$
- Significance test for $H_{0i}$: $\beta_i$, = 0 uses the test statistic $t = b_i/s(b_i)$, df=dfE=n-p, and the P-value computed from the t(n-p) distribution
- Studios example (proc reg, option /clb)

# Do it in SAS

```
proc reg data=studios;
  model y=x1 x2/clb clm cli  alpha=0.01;
/*clb CI for b;
clm CI for mean;
cli CI for prediction; */
    model y=x1 x2/xpx i covb corrb;
/* xpx gives the X'X matrix
  i gives the inverse of the X'X matrix
  covb gives the variance covariance matrix
  of b
  corrb gives the correlation matrix of b*/
run;
```

# Do it in SAS

```
                     The REG Procedure
                          Model: MODEL1
                     Dependent Variable: y

                       Output Statistics

     Dependent Predicted    Std Error
Obs   Variable     Value Mean Predict     99% CL Mean          99% CL Predict     Residual

  1   174.4000  187.1841       3.8409  176.1283  198.2400  153.6265  220.7418  -12.7841
  2   164.4000  154.2294       3.5558  143.9944  164.4645  120.9332  187.5257   10.1706
  3   244.2000  234.3963       4.5882  221.1895  247.6032  200.0699  268.7228    9.8037
  4   154.6000  153.3285       3.2331  144.0223  162.6347  120.3060  186.3511    1.2715
  5   181.6000  161.3849       4.4300  148.6334  174.1365  127.2311  195.5388   20.2151
  ...
 20   224.1000  230.3161       5.8120  213.5865  247.0457  194.4864  266.1457   -6.2161
 21   166.5000  157.0644       4.0792  145.3228  168.8060  123.2746  190.8542    9.4356
```

Sum of Residuals 0

Sum of Squared Residuals 2180.92741

Predicted Residual SS (PRESS) 3002.92331

# Do it in SAS

Model Crossproducts X'X X'Y Y'Y

| Variable | Label | Intercept | x1 | x2 | y |
|---|---|---|---|---|---|
| Intercept | Intercept | 21 | 1302.4 | 360 | 3820 |
| x1 | targtpop | 1302.4 | 87707.94 | 22609.19 | 249643.35 |
| x2 | dispoinc | 360 | 22609.19 | 6190.26 | 66072.75 |
| y | | 3820 | 249643.35 | 66072.75 | 721072.4 |

X'X Inverse, Parameter Estimates, and SSE

| Variable | Label | Intercept | x1 | x2 | y |
|---|---|---|---|---|---|
| Intercept | Intercept | 29.728923483 | 0.0721834719 | -1.992553186 | -68.85707315 |
| x1 | targtpop | 0.0721834719 | 0.0003701761 | -0.005549917 | 1.4545595828 |
| x2 | dispoinc | -1.992553186 | -0.005549917 | 0.1363106368 | 9.3655003765 |
| y | | -68.85707315 | 1.4545595828 | 9.3655003765 | 2180.9274114 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 24015 | 12008 | 99.10 | <.0001 |
| Error | 18 | 2180.92741 | 121.16263 | | |
| Corrected Total | 20 | 26196 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 11.00739 | R-Square | 0.9167 |
| Dependent Mean | 181.90476 | Adj R-Sq | 0.9075 |
| Coeff Var | 6.05118 | | |

# Do it in SAS

Parameter Estimates

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|
| Intercept | Intercept | 1 | -68.85707 | 60.01695 | -1.15 | 0.2663 |
| x1 | targtpop | 1 | 1.45456 | 0.21178 | 6.87 | <.0001 |
| x2 | dispoinc | 1 | 9.36550 | 4.06396 | 2.30 | 0.0333 |

Covariance of Estimates

| Variable | Label | Intercept | x1 | x2 |
|---|---|---|---|---|
| Intercept | Intercept | 3602.0346743 | 8.7459395806 | -241.4229923 |
| x1 | targtpop | 8.7459395806 | 0.0448515096 | -0.672442604 |
| x2 | dispoinc | -241.4229923 | -0.672442604 | 16.515755794 |

The SAS System 22:26 Monday, October 24, 2005 7

The REG Procedure
Model: MODEL2
Dependent Variable: y

Correlation of Estimates

| Variable | Label | Intercept | x1 | x2 |
|---|---|---|---|---|
| Intercept | Intercept | 1.0000 | 0.6881 | -0.9898 |
| x1 | targtpop | 0.6881 | 1.0000 | -0.7813 |
| x2 | dispoinc | -0.9898 | -0.7813 | 1.0000 |

# ANOVA (Section 4.8)

- Sources of variation are
  - Model or Regression (SSM or SSR)
  - Error or Residual (SSE)
  - Total (SSTO)
- SS and df add
  - SSM + SSE =SSTO
  - dfM + dfE = dfT

# Sum of Squares

$$\mathrm{SSM} = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

$$\mathrm{SSE} = \Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\mathrm{SST} = \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2$$

# Degree of Freedom and Mean Squares

$$df_M = p-1$$

$$df_E = n-p$$

$$df_T = n-1$$

$$MSM = SSM/df_M$$

$$MSE = SSE/df_E$$

$$MST = SST/df_T$$

# ANOVA Table

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Model | SSM | dfM | MSM | MSM/MSE |
| Error | SSE | dfE | MSE | |
| Total | SST | dfT | (MST) | |

# Do it in SAS

The REG Procedure
Model: MODEL1
Dependent Variable: y

| | |
|---|---|
| Number of Observations Read | 21 |
| Number of Observations Used | 21 |

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 24015 | 12008 | 99.10 | <.0001 |
| Error | 18 | 2180.92741 | 121.16263 | | |
| Corrected Total | 20 | 26196 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 11.00739 | R-Square | 0.9167 |
| Dependent Mean | 181.90476 | Adj R-Sq | 0.9075 |
| Coeff Var | 6.05118 | | |

# ANOVA F test

- $H_0$: $\beta_1 = \beta_2 = \ldots \beta_{p-1} = 0$
- $H_a$: $\beta_k$ neq 0, for at least one *k=1, … , p-1*
- Under $H_0$, $F \sim F(p-1,n-p)$
- Reject $H_0$ if F is large, use P value
- Studios example: see SAS output

# Interpret F-test

- The p-value for the F significance test tells us one of the following:
  - p-value large: there is no evidence to conclude that *any* of our explanatory variables can help us to model the response variable using this kind of model
  - P-value small: one or more of the explanatory variables in our model *is* potentially useful for predicting the response variable in a linear model