Class 6: Diagnostics

- Acknowledgement:
- Lectures on Regression and ANOVA are based on the notes of Professor George McCabe for the course on ``Applied Regression Analysis'' given at Purdue University
- Diagnostics: look at the data to diagnose situations where the assumptions of our model are violated

Look at the data

- Before trying to describe the relationship between a response variable (Y) and an explanatory variable (X), we should look at the distributions of these variables
- We should always look at X
- If Y depends on X, looking at Y alone may not be very informative

Diagnostics for X

- If X has many values, use Proc Univariate
- If X has only a few values, use Proc Freq or the Freq option in Proc Univariate

Diagnostics for X (2)

- Examine the distribution of X
  - Is it skewed?
  - Are there outliers?
- Do the values of X depend on time (order in which the data were collected)?

```
data a1;
infile 'h:/STAT512/ch01ta01.txt';
   input lotsize workhrs;
   seq=_n_;
proc print data=a1;
   run;
```

Obs	lotsize	workhrs	seq
1	80	399	1
2	30	121	2
3	50	221	3
4	90	376	4
5	70	361	5

```
proc univariate data=a1 plot;
   var lotsize workhours;
   run;
```

Variable: lotsize

N	25
Mean	70
Std Deviation	28.7228132
Skewness	-0.1032081
Uncorrected SS	142300
Coeff Variation	41.0325903

Variable: lotsize

Sum Weights	25
Sum Observations	1750
Variance	825
Kurtosis	-1.0794107
Corrected SS	19800
Std Error Mean	5.74456265

Basic Statistical Measures

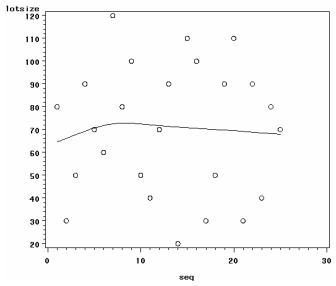
Location	
Mean	70.00000
Median	70.00000
Mode	90.00000

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
20	14	100	9
30	21	100	16
30	17	110	15
30	2	110	20
40	23	120	7

```
Stem Leaf                     #  Boxplot
   12 0                       1     |
   11 00                      2     |
   10 00                      2     |
    9 0000                    4  +-----+
    8 000                     3  |     |
    7 000                     3  *--+--*
    6 0                       1  |     |
    5 000                     3  +-----+
    4 00                      2     |
    3 000                     3     |
    2 0                       1     |
      ----+----+----+----+
  Multiply Stem.Leaf by 10**+1
```

```
symbol1 v=circle i=sm70;
proc gplot data=a1;
plot lotsize*seq;
run;
```



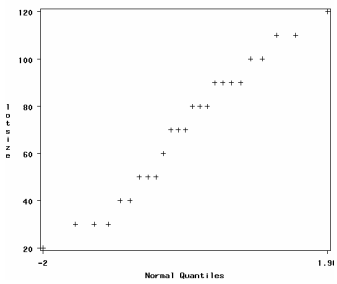
Normal distributions

- Our model does *not* state that X comes from a single normal population
- Same comment applies to Y
- In some cases, X and/or Y may be normal and it can be useful to know this

Normal quantile plots

- $Znorm_i = \Phi^{-1}((i-.375)/(n+.25))$, i=1 to n
- Plot the order statistics $X_{(i)}$ versus $Znorm_i$

```
proc univariate data=a1 plot;
var lotsize workhrs;
qqplot lotsize workhrs ;
run;
```



Diagnostics for residuals

- Model: $Y_i = \beta_0 + \beta_1 X_i + \xi_i$
- Predicted values: $\hat{Y}_i = b_0 + b_1 X_i$
- Residuals: $e_i = Y_i - \hat{Y}_i$
- So, $Y_i = \hat{Y}_i + e_i$
- e_i should be similar to the ξ_i
- The model assumes ξ_i iid $N(0, \sigma^2)$

Plot **Plot**

PLOT
PLOT
PLOT

Plot **Plot**

Questions addressed by diagnostics for residuals

- Is the relationship linear?
- Does the variance depend on X?
- Are there outliers?
- Do the errors depend on order (_n_)
- Are the errors normal?
- Are the errors dependent?

Is the Relationship Linear?

- Plot Y vs X
- Plot e vs X
- Plot of e vs X will emphasize deviations from linear pattern

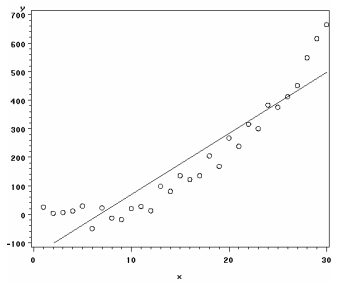
```
libname xxx 'c:/ECMI/data';

Data xxx.a100;
   do x=1 to 30;
    y=x*x-10*x+30+25*normal(0);
    output;
   end;
run;
```

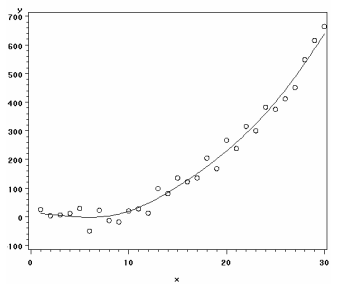
```
proc reg data=xxx.a100;
   model y=x;
   output out=a2 r=resid;
run;
```

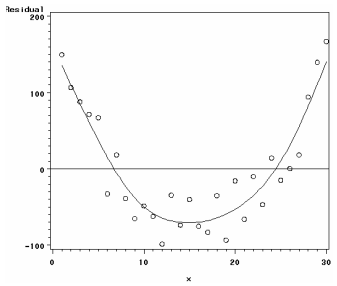
Source	DF	Sum of Squares	Mean Square
Model	1	1032098	1032098
Error	28	169048	6037
C Tot	29	1201145	

F Value	Pr > F
170.95	<.0001

```
symbol1 v=circle i=rl;
proc gplot data=a2;
   plot y*x;
   run;

symbol1 v=circle i=sm60;
proc gplot data=a2;
   plot y*x;
proc gplot data=a2;
   plot resid*x/vref=0;
run;
```







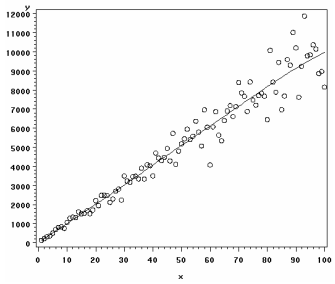
Does the variance depend on X?

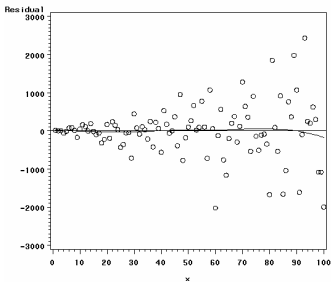
- Plot Y vs X
- Plot e vs X
- Plot of e vs X will emphasize problems with the variance assumption

```
libname xxx 'h:/STAT512/data';

Data xxx.a100a;
   do x=1 to 100;
      y=100*x+30+10*x*normal(0);
      output;
   end;
run;
```

```
proc reg data=xxx.a100a;
   model y=x;
   output out=a2 r=resid;
run;
```

```
symbol1 v=circle i=sm60;
proc gplot data=a2;
   plot y*x;

proc gplot data=a2;
   plot resid*x vref=0;
run;
```

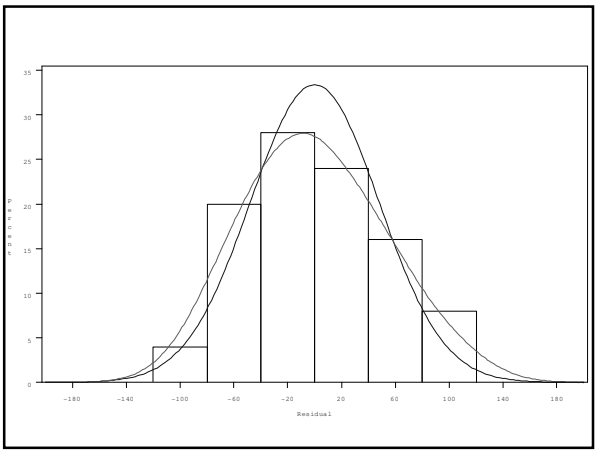


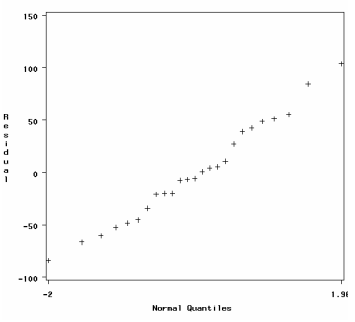


Are the errors normal?

- The *real* question is whether the distribution of the errors is far enough away from normal to invalidate our confidence intervals and significance tests
- Look at the distribution of the residuals
- Use a normal quantile plot

```
data a1;
  infile `../data/ch01ta01.dat';
  input lotsize workhrs;
proc reg data=a1;
  model workhrs=lotsize;
  output out=a2 r=resid;
proc univariate data=a2 plot
      normal;
  var resid;
  histogram resid/ normal kernel
 (color=green);
  qqplot resid; run;
```





Dependent Errors

- Usually we see this in a plot of residuals vs time order or seq (our SAS variable)
- We can have trends and/or cyclical effects

Are there outliers?

- Plot Y vs X
- Plot e vs X
- Plot of e vs X should emphasize an outlier

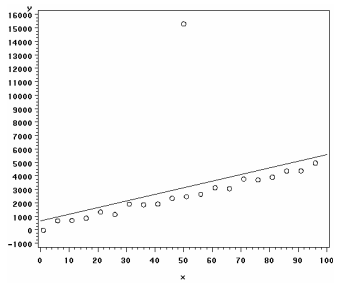
```
Data xxx.a100b1;
   do x=1 to 100 by 5;
      y=30+50*x+200*normal(0);
      output;
   end;
   x=50; y=30+50*50 +10000;
   d='out'; output;
run;
```

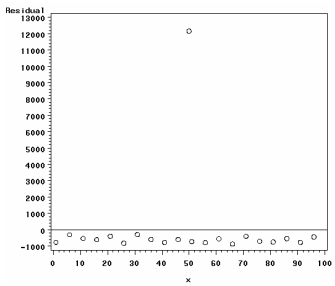
```
proc reg data=xxx.a100b1;
   model y=x;
   where d ne 'out';
run;
proc reg data=xxx.a100b1;
   model y=x;
   output out=a2 r=resid;
run;
```

Var	Par Est	St Error	t	Pr>|t|
Int	-2.54	95.29		
x	50.51	1.68	29.91	<.0001
Int	432.20	979.57		
x	51.37	17.45	2.94	.0083

Root MSE	217.77514
Root MSE	2250.22683

```
symbol1 v=circle i=rl;
proc gplot data=a2;
   plot y*x;
proc gplot data=a2;
   plot resid*x/ vref=0;
run;
```





Different kinds of outliers

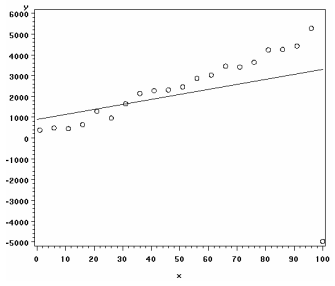
- The outlier in the last example *influenced* the intercept
- but not the slope
- It inflated all of our standard errors
- Here is an example of an outlier that *influences* the slope

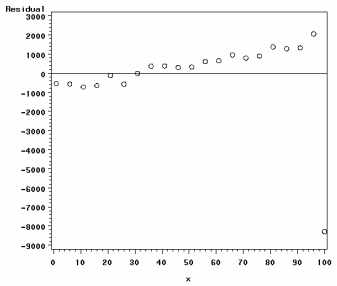
```
Data xxx.a100c1;
   do x=1 to 100 by 5;
      y=30+50*x+200*normal(0);
      output;
   end;
   x=100; y=30+50*100 -10000;
   d='out'; output;
run;
```

```
proc reg data=xxx.a100c1;
   model y=x;
   where d ne 'out';
run;
proc reg data=xxx.a100c1;
   model y=x;
   output out=a2 r=resid;
run;
```

```
      Par   St
Var   Est   Err      t   Pr>|t|
Int   73.2 93.6
x     49.8  1.6   30.02   <.0001

Int 903.9 899.3
x     24.1  15.1   1.59   0.1285
```

```
symbol1 v=circle i=rl;
proc gplot data=a2;
   plot y*x;
proc gplot data=a2;
   plot resid*x/ vref=0;
run;
```




Residual
3000
2000
1000
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
-8000
-9000
0
10
20
30
40
50
60
70
80
90
100
x