STOR 455
STATISTICAL METHODS I

Jan Hannig

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: $e_i = Y_i - \hat{Y}_i$
  - The book recommends to standardize the residuals.

$$h_{i,i} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX} \qquad r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$$

Residuals

- The e_i should be similar to the ξ_i
  - The model assumes ξ_i iid $N(0, \sigma)$
- Similarly the r_i should be similar to the ξ_i/σ
  - The model assumes ξ_i/σ iid $N(0, 1)$

What do we learn?

- Is the relationship linear?
- Is the variance a constant?
- Are there outliers?
- Are the errors normal?
- Are the errors dependent?

Is the Relationship Linear?

- Scatter plot of Y vs X
- Scatter plot of r vs X
- Scatter plot of r vs $\hat{Y}$
  - Plot of r vs … emphasize deviations from linear pattern
  - Never plot r vs Y

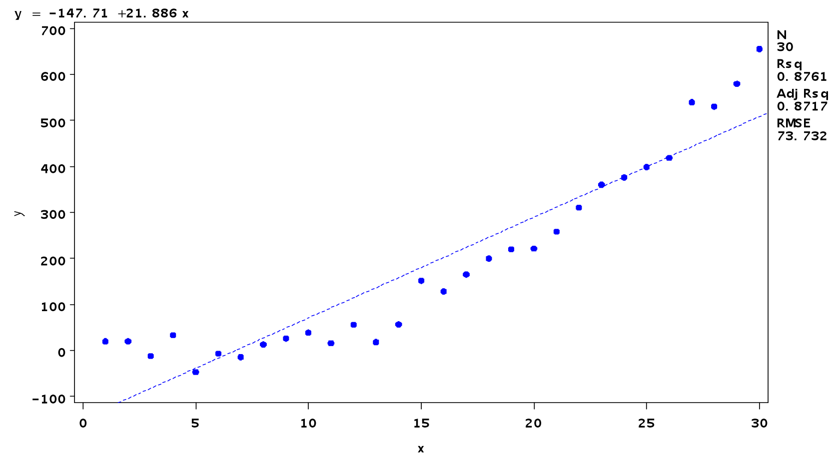
Do it in SAS

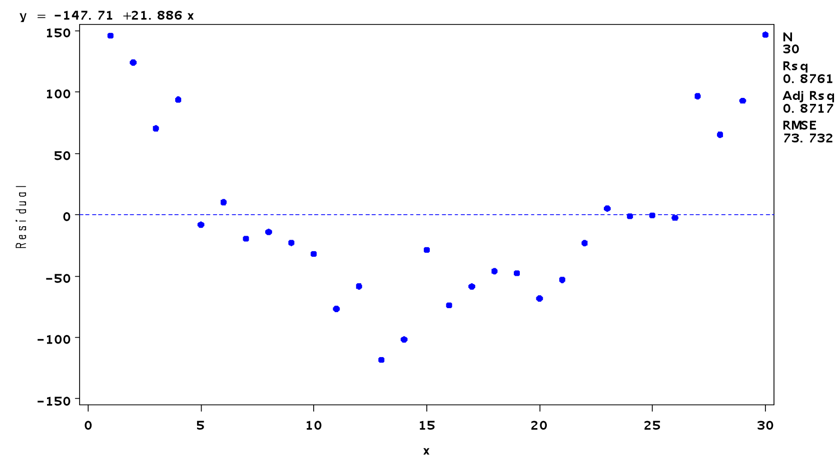
```
/* Residual plots */
symbol1 v=dot h=.8 c=blue;

/* simulated data, is it linear?
   */
Data resid;
  do x=1 to 30;
   y=x*x-10*x+30+25*normal
    (0);
   output;
  end;
proc print data=resid;
run;
```

Obs	x	y
1	1	20.268
2	2	20.292
3	3	-11.644
4	4	33.722
5	5	-46.357
6	6	-6.092
7	7	-13.902
8	8	13.392
9	9	26.473
10	10	39.391
11	11	16.491
12	12	56.712
13	13	18.588
14	14	57.087
…		

Do it in SAS

```
proc reg data=resid
   noprint;
model y=x;
plot y*x student.*x
   student.*p.;
run;
```



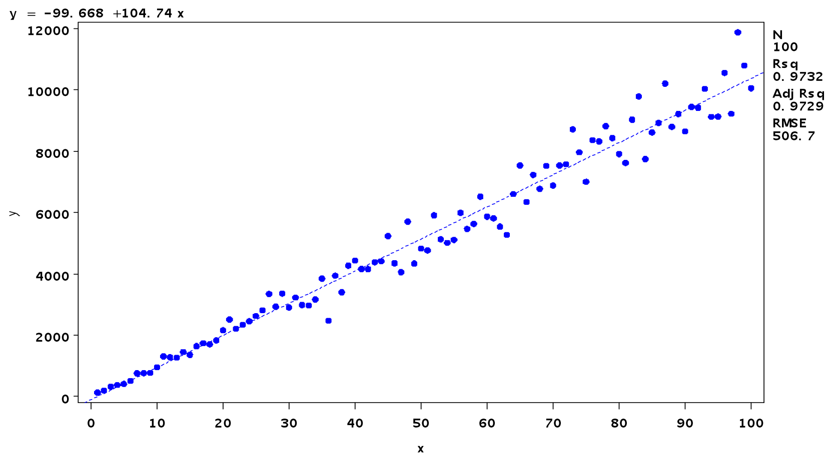


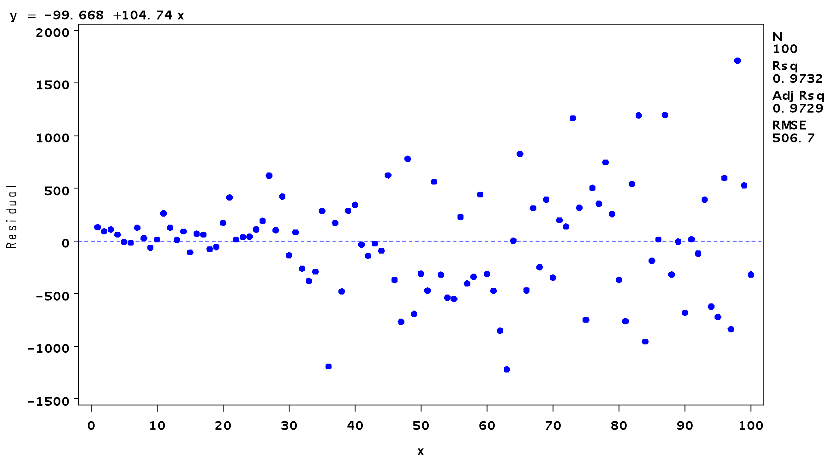
Does the variance depend on X?

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

Do it in SAS

```
/* simulated data, is
  variance constant? */
Data resid2;
   do x=1 to 100;
      y=100*x
  +30+10*x*normal(0);
      output;
   end;
run;

proc reg data=resid2
  noprint;
model y=x;
plot y*x student.*x;
run;
```



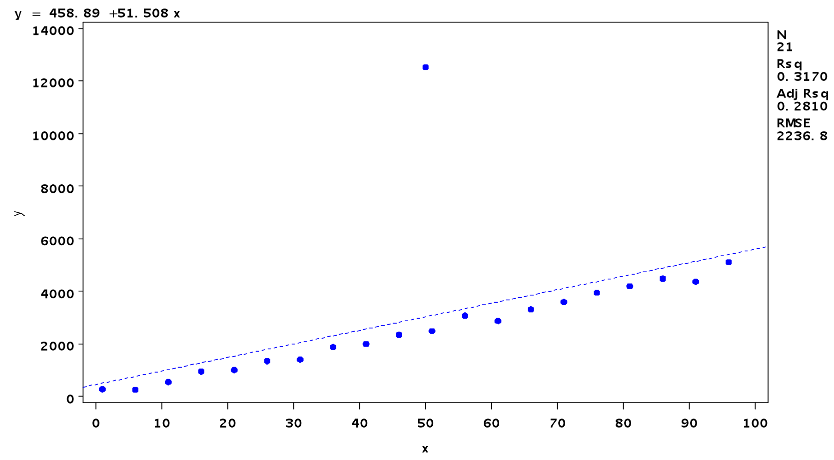


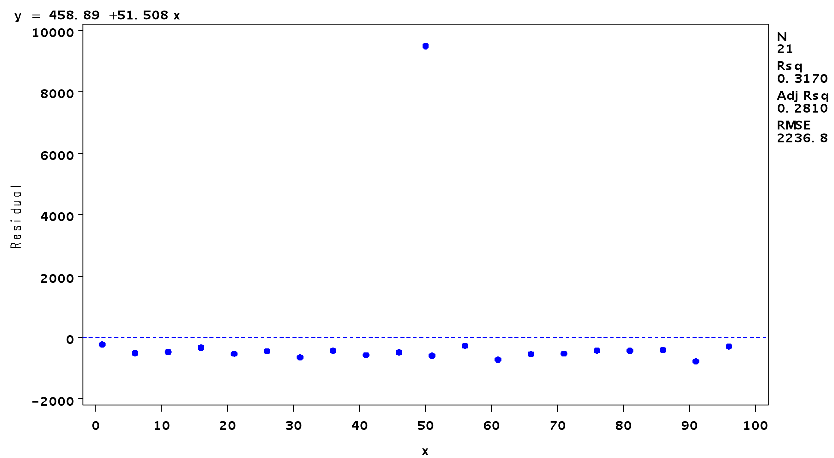
Outliers and Influential Cases

- Outliers inflate the variance and decrease our chances of finding statistically significant results
- Outliers may or may not be influential
- An outlier can be *influential* for some model parameters, and not influential for others
- Influential cases are usually outliers

Do it in SAS

```
/* simulated data, outlier */
Data outlier;
  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 print data=outlier;
run;

proc reg data=outlier;
  model y=x;
  where d ne 'out';
run;
proc reg data=outlier;
  model y=x;
  plot y*x student.*x;
run;
```



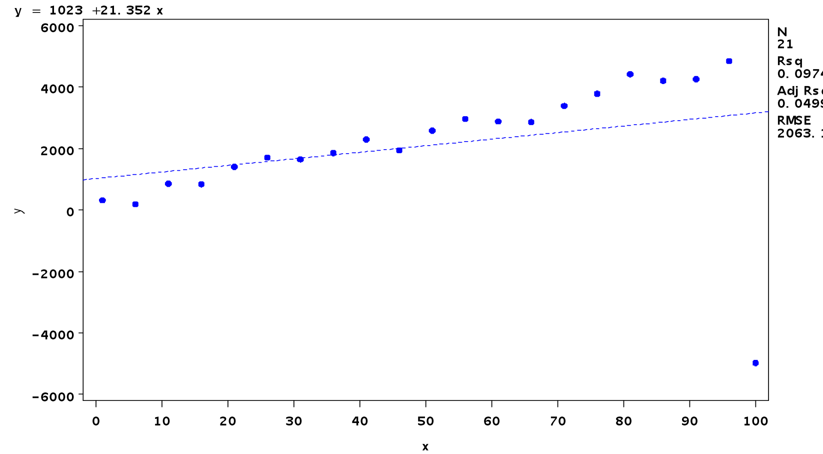


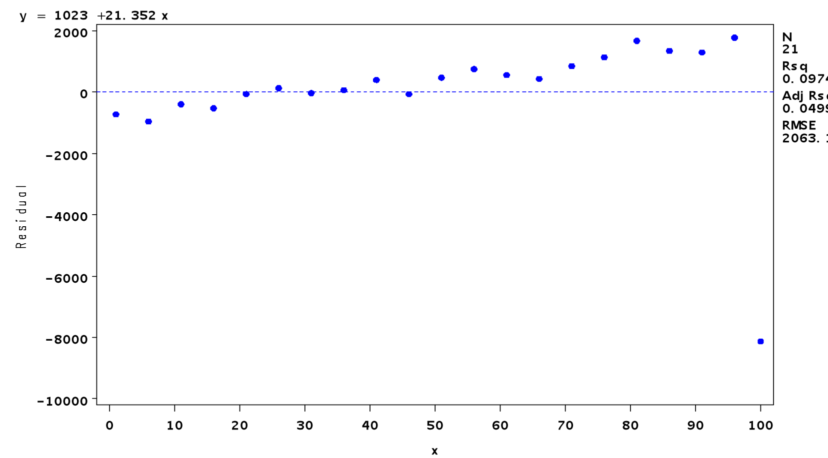
Are there outliers?

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

Do it in SAS

```
/* Simulated data, inf. case */
Data outlier2;
  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=outlier2;
  model y=x;
  plot y*x student.*x;
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

Normal Quantile plots (Rankit plot)

- Consider n=5 observations iid N(0,1)
- From normal table, we find
  - $P(z \leq -.84) = .20$
  - $P(-.84 < z \leq -.25) = .20$
  - $P(-.25 < z \leq .25) = .20$
  - $P(.25 < z \leq .84) = .20$
  - $P(.84 < z) = .20$

Normal Quantile plots (2)

- So we *expect*
  - One observation $\leq$84
  - One observation in (-.84, -.25]
  - One observation in (-.25, .25]
  - One observation in (25, .84]
  - One observation > .84

Normal Quantile plots (3)

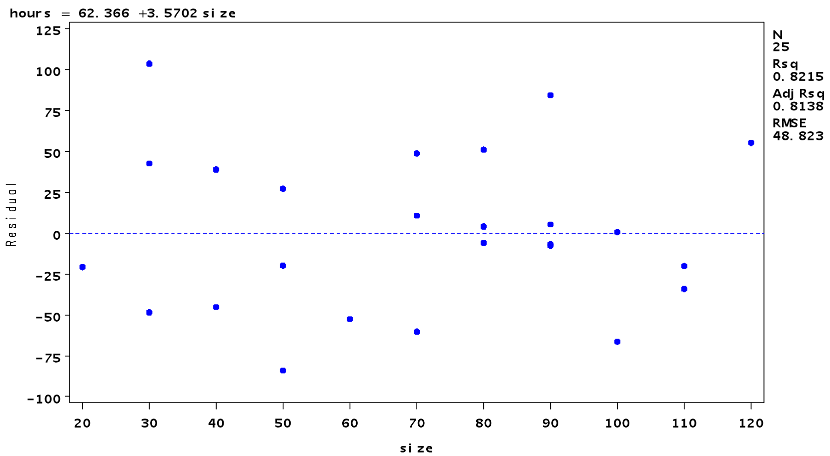
- We use some theory to pick an appropriate value in the interval
  (The book has a nice idea).
- $Znorm_i = \Phi^{-1}((i-.375)/(n+.25))$, i=1 to n
- Plot the order statistics $X_{(i)}$ versus $Znorm_i$

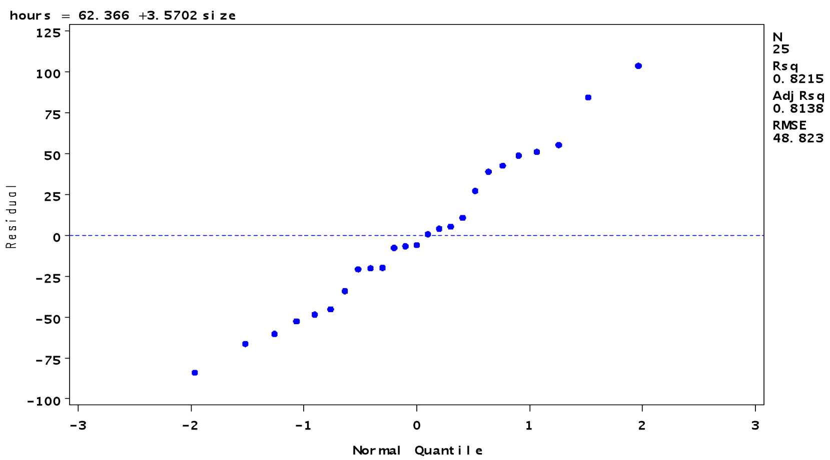
Normal Quantile plots (4)

- The standardized X variable is
  $z = (X - \overline{X})/s$
- So, $X = \overline{X} + s*z$
- If the data are approximately normal, the relationship will be approximately linear with slope s and intercept $\overline{X}$.

Do it in SAS

```
/*lot size example revisited */
data lot2;
  set lot;
  id = _n_;
  group=1;
run;

/* scatter plot and QQ plot of
    residual */
symbol1 v=dot h=.8 c=blue i=none;
proc reg data = lot2 noprint;
  model hours = size;
  output out=temp student=r;
  plot hours*size student.*size
    student.*nqq.;
run;
```





Significance tests for normality

- Many choices for a significance testing procedure
- Proc univariate with the normal option (`proc univariate normal;`) provides four
- Shapiro-Wilk is a good choice

Do it in SAS

```
/* test for normality */
proc univariate normal data=temp;
var r;
run;
```

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.978904	Pr < W	0.8626
Kolmogorov-Smirnov	D	0.09572	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.033263	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.207142	Pr > A-Sq	>0.2500