# MTH499/599 Lecture Notes 06

Donghui Yan

Department of Math, Umass Dartmouth

# Outline

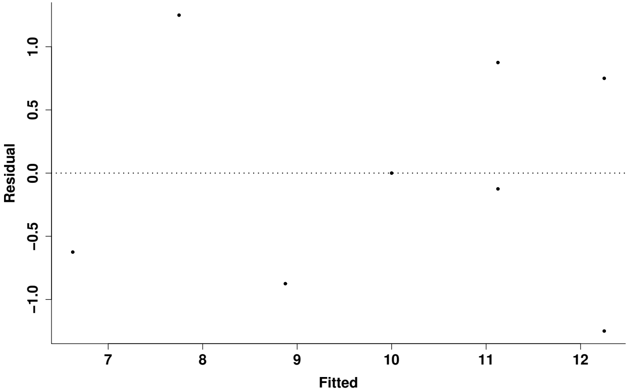
- Constant variance assumption

# Regression diagnosis

- Is the linear model appropriate (goodness-of-fit)
- Normality assumption
- Homoscedasticity (constant variance)
- Leverages of individual data points.

# Testing of constant variance

- Visual inspection by residual plot



# Testing of constant variance

- The Cook-Weisberg's score test
- Function *ncvTest()* in R package *library{car}*
  - install.packages("car")

*R. Dennis Cook and Sanford Weisberg. Diagnostics for heteroscedasticity in regression. Biometrika, Vol 70(1): 1-10, 1983*

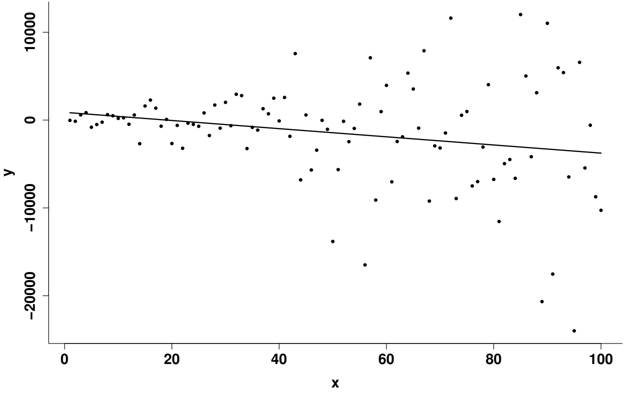
- The ncvTest() on the car data

```
>ncvTest(mylm)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chi-square = 4.115226e-5    Df = 1    p = 0.9948816
```

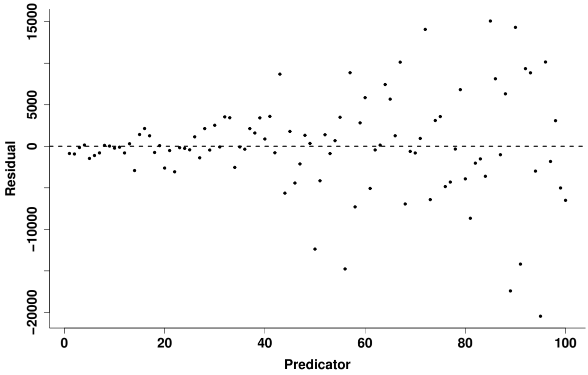
# Example of non-constant variance

- Let the data be generated by

$$Y = 30 + 100X + \mathcal{N}(0, X^2)$$



# Example of non-constant variance (continued)

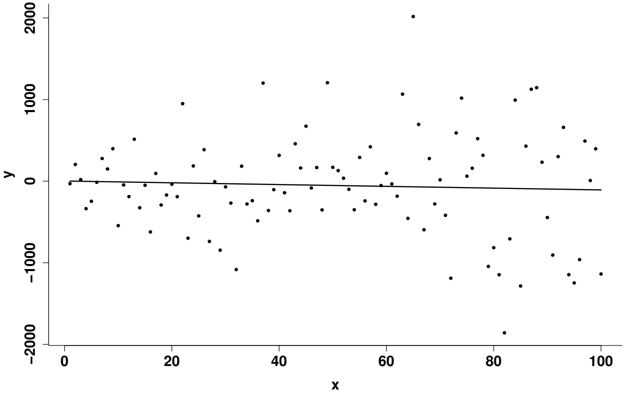



```
>ncvTest(mylm2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chi-square = 41.39011    Df = 1     p = 1.246864e-10
```

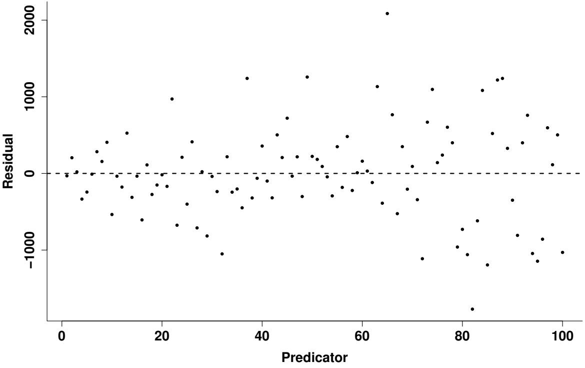
# A weaker example

- Let the data be generated by

$$Y = 30 + 100X + \mathcal{N}(0, \mid X \mid)$$



# A weaker example (continued)




```
>ncvTest(mylm3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chi-square = 16.78339    Df = 1     p = 4.189835e-05
```

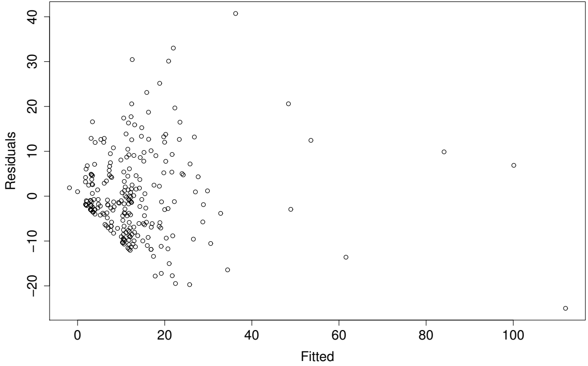
# A real example of non-constant variance

- The Ornstein data in the "car" package
  - J. Fox (2008) *Applied Regression Analysis and Generalized Linear Models*
  - Interlocking directorates among major Canadian firms in mid-70s
  - 248 observations and 4 columns
    - Assets – assets in millions of dollars
    - Sector – industrial sector
    - Nation – nation of control
    - Interlocks – number interlocking director and executive positions.

# lm(formula = interlocks assets + sector + nation)

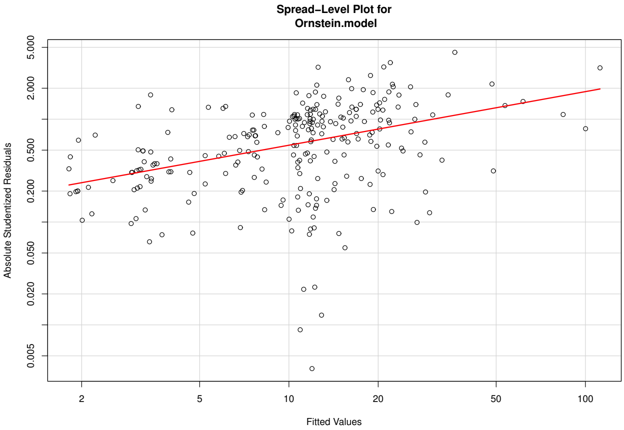
```
Coefficients  Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.027e+01  1.561e+00   6.575 3.14e-10 ***
assets       8.096e-04  6.119e-05  13.231  < 2e-16 ***
sectorBNK   -1.781e+01  5.906e+00  -3.016  0.00284 **
sectorCON   -4.709e+00  4.728e+00  -0.996  0.32034
sectorFIN    5.153e+00  2.646e+00   1.948  0.05266 .
sectorHLD    8.777e-01  4.004e+00   0.219  0.82669
sectorMAN    1.149e+00  2.065e+00   0.556  0.57849
sectorMER    1.491e+00  2.636e+00   0.566  0.57206
sectorMIN    4.880e+00  2.067e+00   2.361  0.01905 *
sectorTRN    6.171e+00  2.760e+00   2.236  0.02629 *
sectorWOD    8.228e+00  2.679e+00   3.072  0.00238 **
nationOTH   -1.241e+00  2.695e+00  -0.461  0.64555
nationUK    -5.775e+00  2.674e+00  -2.159  0.03184 *
nationUS    -8.618e+00  1.496e+00  -5.760 2.64e-08 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

# The Ornstein data (continued)




```
>ncvTest(Ornstein.model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 46.98537    Df = 1     p = 7.151848e-12
```

# The spread-level plot ($\log |E_i|$ vs $\log \hat{Y}_i$)




```
>myspread<-spread.level.plot(Ornstein.model);
Suggested power transformation:  0.4788627
```

# Power-transforming the Ornstein data

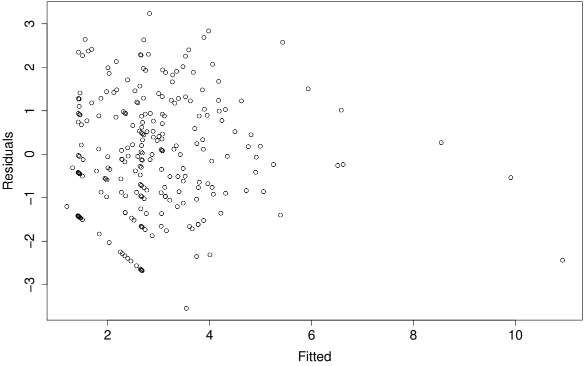
- Power-transforming $Y$ by

$$Y \to Y^{1-b}$$

  where $b$ is the slope of the spread-level plot
  - Variance-stabilizing transform
  - Why does this work? Requires delta-method.

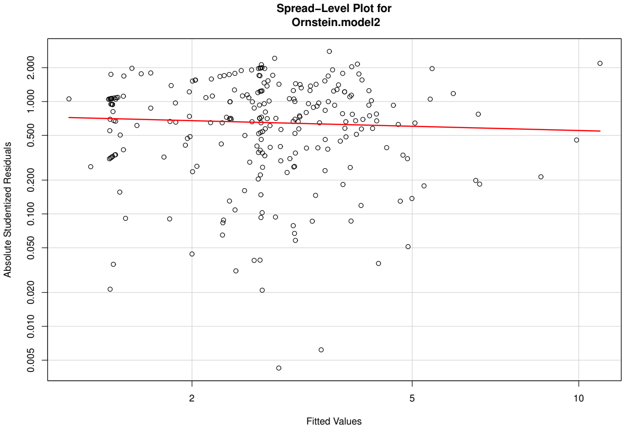
```
> z<-interlocks^(myspread$PowerTransformation);
> Ornstein.model2<-lm(z ~ assets + sector+nation);
```

# The power-transformed Ornstein data




```
>ncvTest(Ornstein.model2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.00353606    Df = 1     p = 0.9525819
```

# The spread-level plot




```
>spread.level.plot(Ornstein.model2);
Suggested power transformation:  1.125167
```