Chapter 6
Assessing the Assumptions of the Regression Model (Part 2)

Terry Dielman

Applied Regression Analysis

for Business and Economics

Asset Risk
= Systematic Risk + Nonsystematic Risk

Systematic Risk : 증권시장의 불확실성에 따른 위험. 분산투자로 제거되지 않는 위험이므로 Undiversifiable Risk 라고 함.

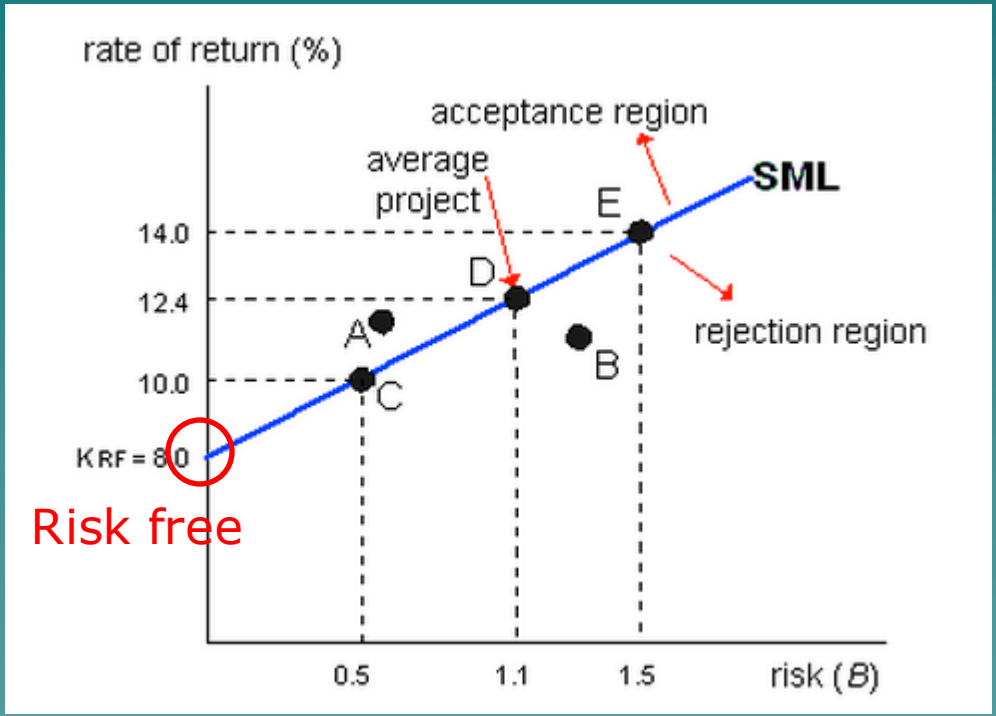
Nonsystematic Risk : 증권시장의 전반적인 움직임과 관계없는 기업특정 요인 (노사문제, 매출, 기업이미지) 등에 의한 위험. 분산투자로 제거가능하므로 Diversifiable Risk 라고 함.

베타 & 시그마

- 베타 (Beta) : 체계적 위험에 대한 측정변수의 하나로서, 증권시장의 수익률 변동에 대한 특정기업의 수익률 변동 비율

- 시그마 (Sigma) : 주식의 총 변동으로 전체 위험도를 나타내는 측정변수

SML(Security Market Line)



The higher risk is, the higher the return rate is required

Example 6.7 S&L Rate of Return

Data set SL6

n =35 Saving and Loans stocks

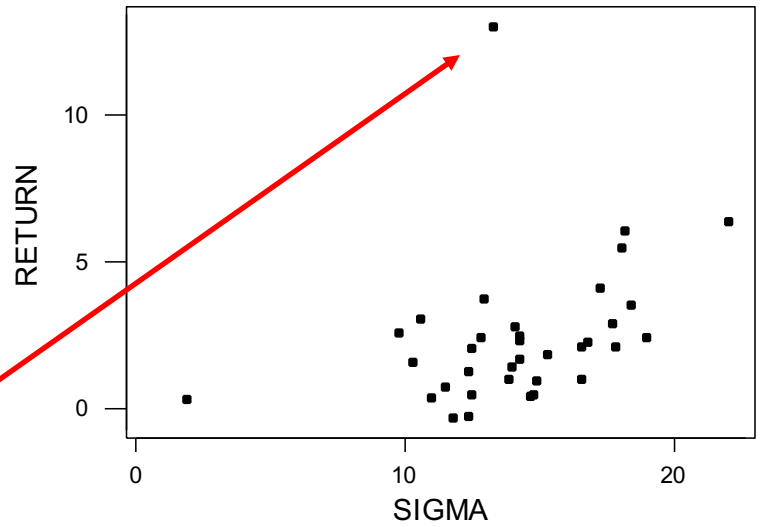
Y = rate of return for 5 years ending 1982

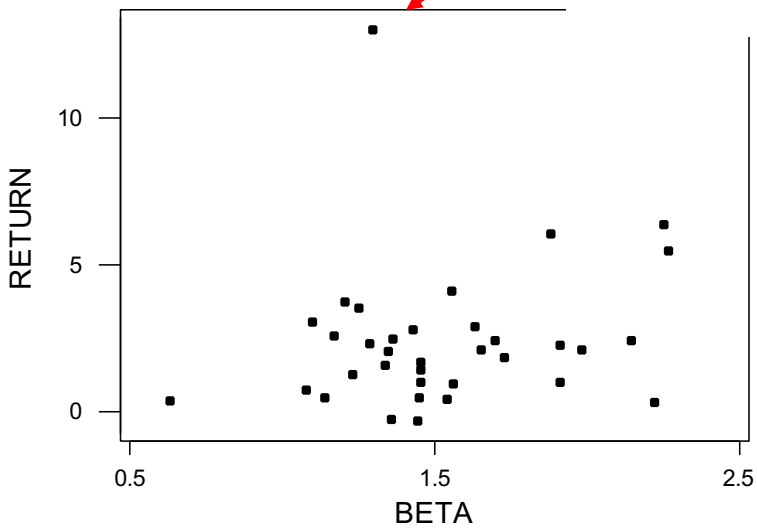
X_1 = the "Beta" of the stock

X_2 = the "Sigma" of the stock

Beta is a measure of nondiversifiable risk and Sigma a measure of total risk

Basic exploration





Correlations: RETURN, BETA, SIGMA

	RETURN	BETA
BETA	0.180	
SIGMA	0.351	0.406

Not much explanatory power

The regression equation is
RETURN = - 1.33 + 0.30 BETA + 0.231 SIGMA

Predictor	Coef	SE Coef	T	P
Constant	-1.330	2.012	-0.66	0.513
BETA	0.300	1.198	0.25	0.804
SIGMA	0.2307	0.1255	1.84	0.075

S = 2.377 R-Sq = 12.5% R-Sq(adj) = 7.0%

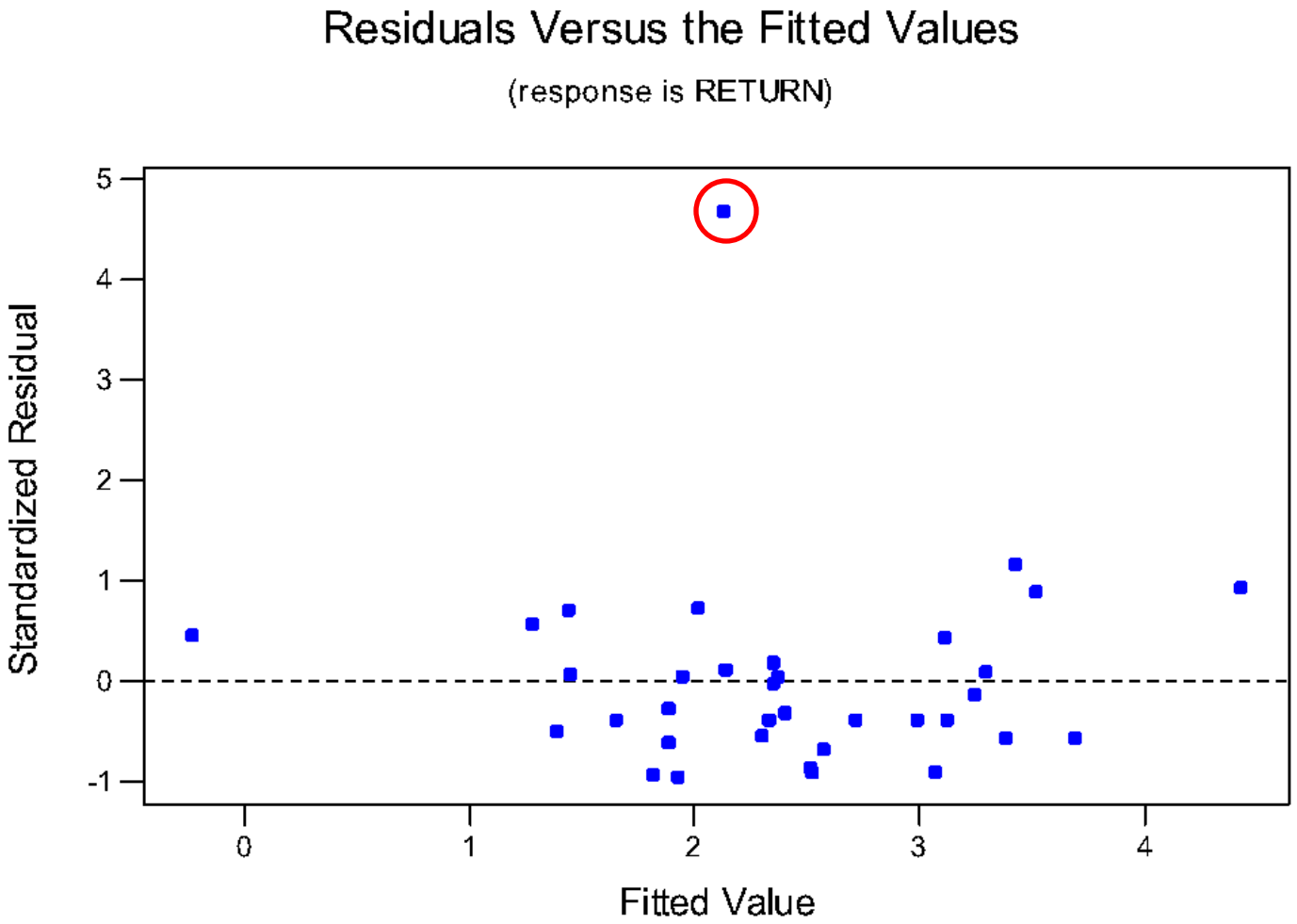
Analysis of Variance
(deleted)

Unusual Observations

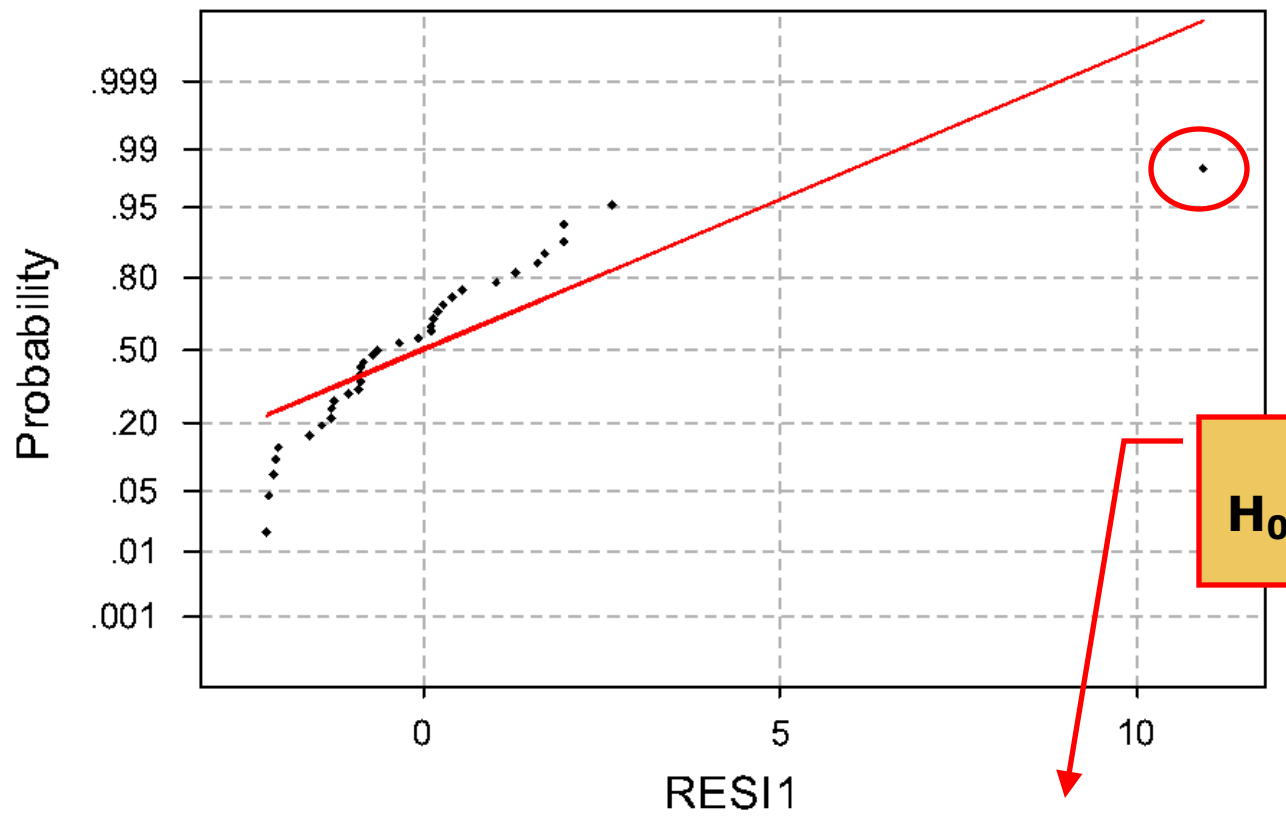
Obs	BETA	RETURN	Fit	SE Fit	Residual	St Resid
19	2.22	0.300	-0.231	2.078	0.531	0.46 X
29	1.30	13.050	2.130	0.474	10.920	4.69R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

One in every crowd? (outlier / normality)



Normality Test



Reject H_0: Normality

6.6.3 Corrections for Nonnormality

- Normality is not necessary for making inference with large samples.
- It is required for inference with small samples.
- The remedies are similar to those used to correct for nonconstant variance.

Influential Point (영향점)

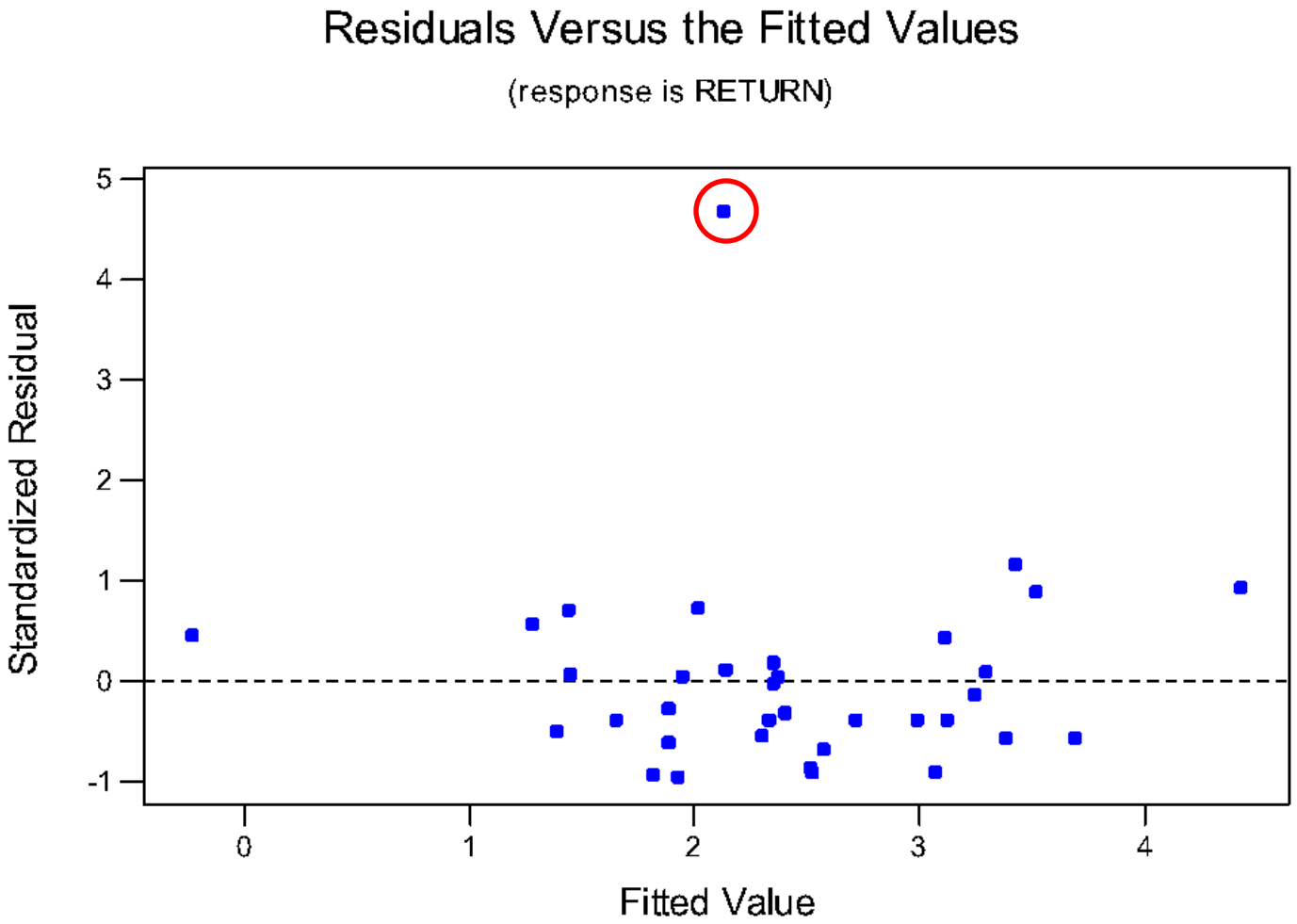
6.7 Influential Observations

- In minimizing SSE, the least squares procedure tries to avoid large residuals.
- It thus "pays a lot of attention" to y values that don't fit the usual pattern in the data. Refer to the example in Figures 6.42(a) and 6.42(b).
- That probably also happened in the S&L data where the one very high return masked the relationship between rate of return, beta and sigma for the other 34 stocks.

6.7.2 Identifying outliers

- Minitab flags any residual bigger than 2 in absolute value as a potential outlier.
- A boxplot of the residuals uses a slightly different rule, but should give similar results.
- There is also a third type of residual that is often used for this purpose.

One in every crowd? (outlier / normality)



Deleted residuals (skip !)

- If you (temporarily) eliminate the i^{th} observation from the data set, it cannot influence the estimation process.
- You can then compute a "deleted" residual to see if this point fits the pattern in the other observations.

Deleted Residual Illustration (skip !!)

```
The regression equation is
ReturnWO29 = - 2.51 + 0.846 BETA + 0.232 SIGMA

34 cases used 1 cases contain missing values

Predictor        Coef     SE Coef           T         P
Constant       -2.510       1.153       -2.18     0.037
BETA           0.8463      0.6843        1.24     0.225
SIGMA         0.23220     0.07135        3.25     0.003

S = 1.352        R-Sq = 37.2%     R-Sq(adj) = 33.1%
```

Without observation 29, we get a much better fit.

Predicted Y_{29} = -2.51 + .846(1.2973) + .232(13.3110) = 1.678

Prediction SE is 1.379

Deleted residual$_{29}$ = (13.05 – 1.678)/1.379 = 8.24

The influence of observation 29 (skip !!)

- When it was temporarily removed, the R^2 went from 12.5% to 37.2% and we got a very different equation

- The deleted residual for this observation was a whopping 8.24, which shows it had a lot of weight in determining the original equation.

Influential Point (영향점)

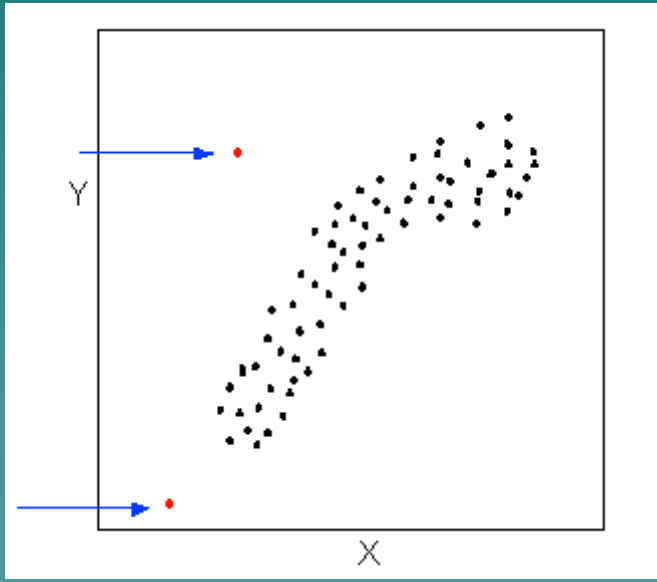
Outlier (이상점) : Y-direction

Leverage Point (지렛대점): X-direction

Influential Point (영향점)

Outlier

Leverage point



6.7.3 Identifying Leverage Points

- Outliers have unusual y values; data points with unusual X values are said to have *leverage*. Minitab flags these with an X.
- These points can have a lot of influence in determining the Yhat equation, particularly if they don't fit well. Minitab would flag these with both an R and an X.

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X\left(X^\top X\right)^{-1} X^\top \mathbf{y}.$$

Therefore the hat matrix is given by

$$H = X\left(X^\top X\right)^{-1} X^\top.$$

$$H_{p\times p}, \quad \text{where } p = k+1$$

$$h_i = (H)_{(i,i)}$$

$$V(\hat{e}_i) = (1-h_i)\sigma^2$$

$$\hat{V}(\hat{e}_i) = (1-h_i)MSE$$

$$\hat{e}_{SD,i} = \frac{\hat{e}_i}{\sqrt{(1-h_i)MSE}}$$

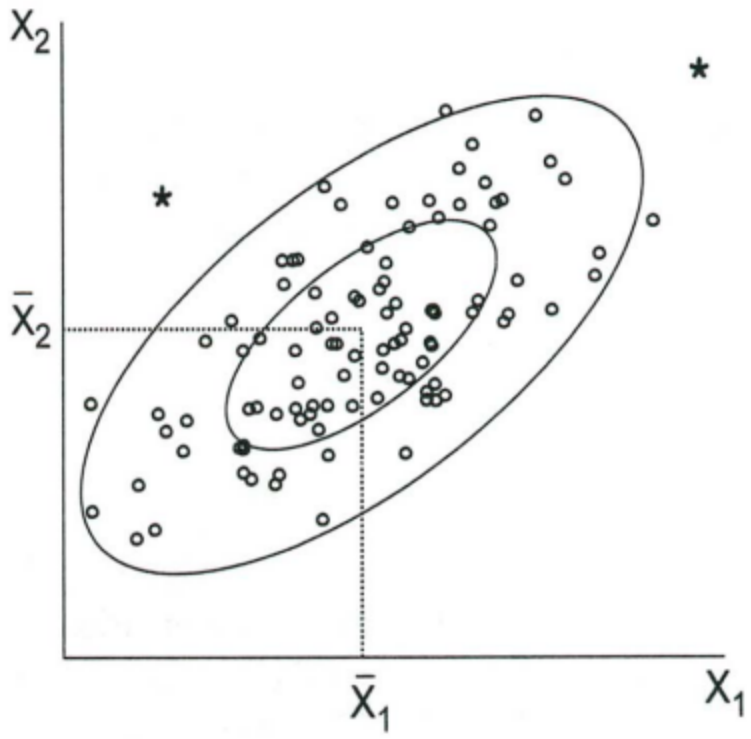


Figure 11.3 from Fox (1997)

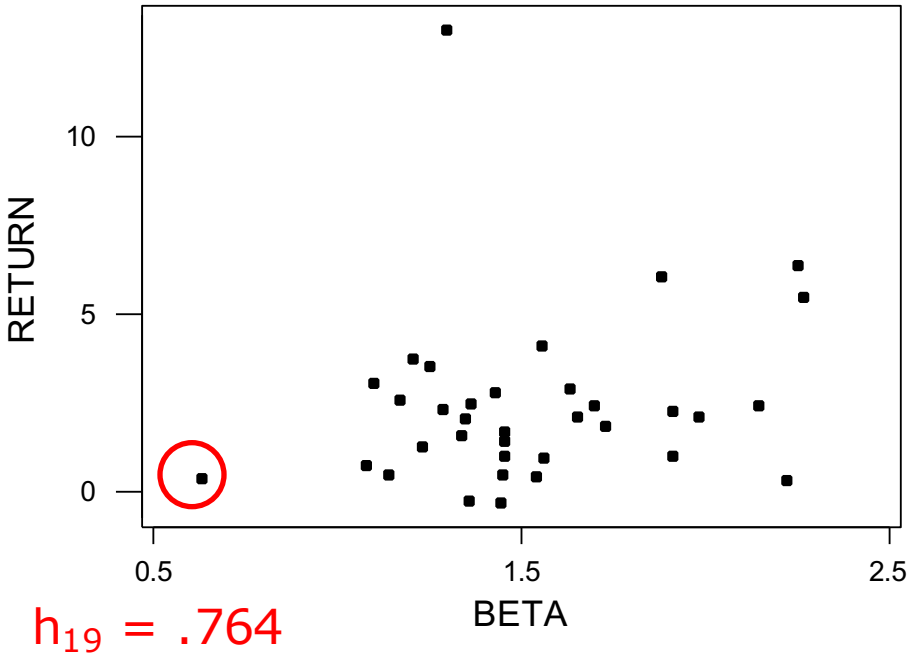
Contour Plot for the h_i

Leverage

- The leverage of the i^{th} observation is h_i (it is hard to show where this comes from without matrix algebra).
- If $h > 2(K+1)/n$ it has high leverage.
- For S&P returns, $k = 2$ and $n = 35$ so the benchmark is $2(3)/35 = .171$
- Observation 19 has a very small value for Sigma, this is the reason why it has $h_{19} = .764$

Leverage Point

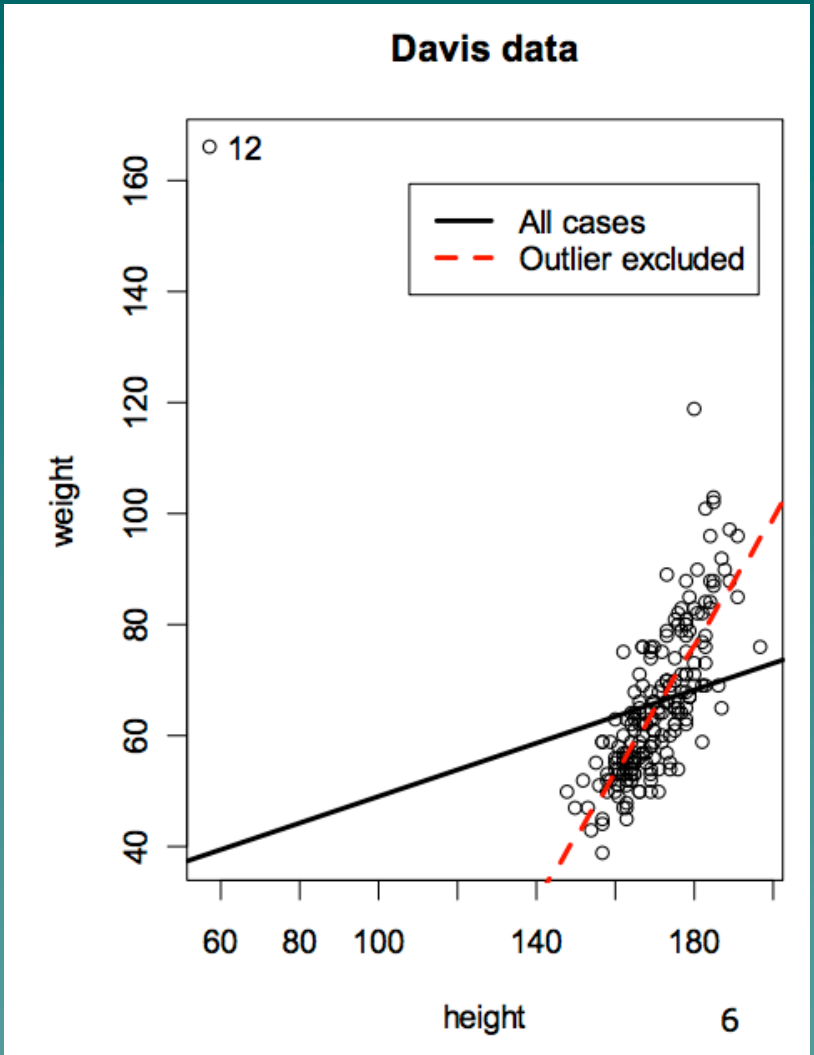
$h > 2(K+1)/n = 0.171$



$h_{19} = .764$

6.7.4 Combined Measures

- The effect of an observation on the regression line is a function of both the y and X values.
- Several statistics have been developed that attempt to measure combined influence.
- The DFIT statistic and Cook's D are two more-popular measures.

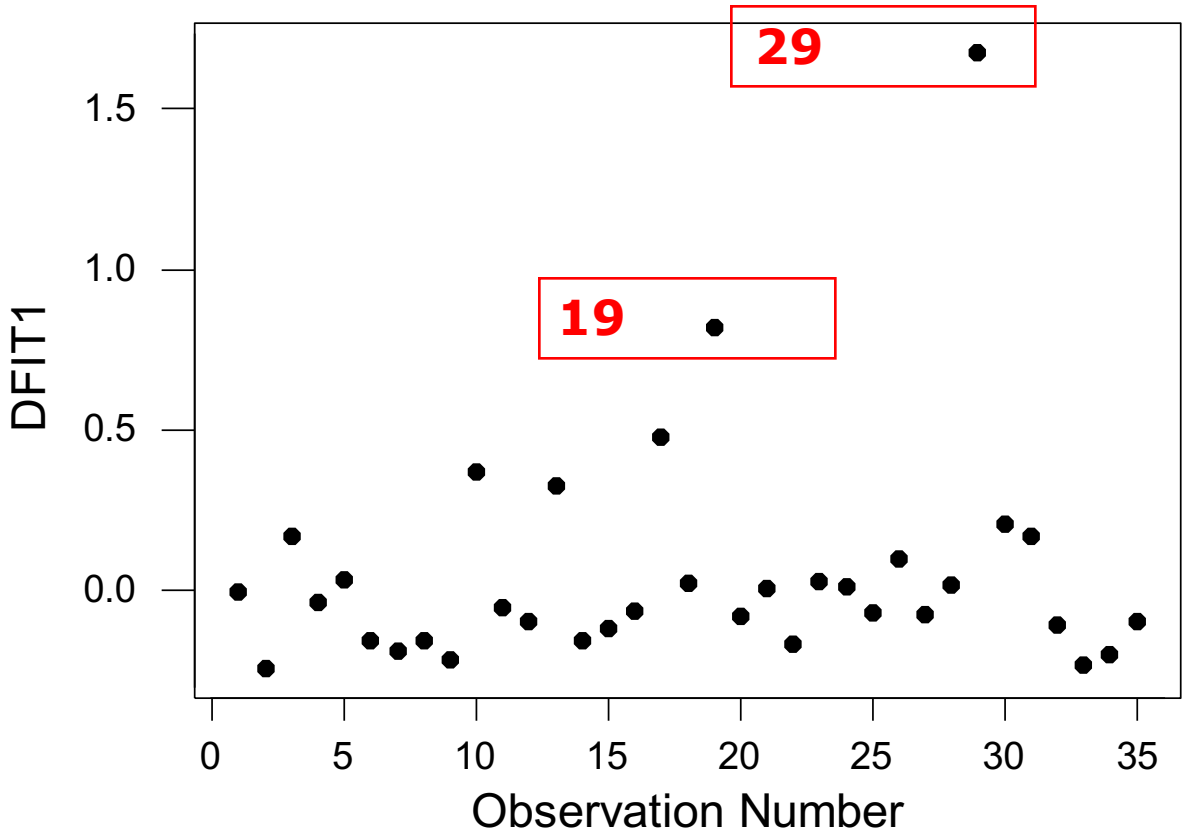


- Leverage & Outlier
- Is there any Combined Measure?

The DFIT statistic

- The DFIT statistic is a function of both the residual and the leverage.
- Minitab can compute and save these under "Storage".
- Sometimes a cutoff is used, but it is perhaps best just to look for values that are high.

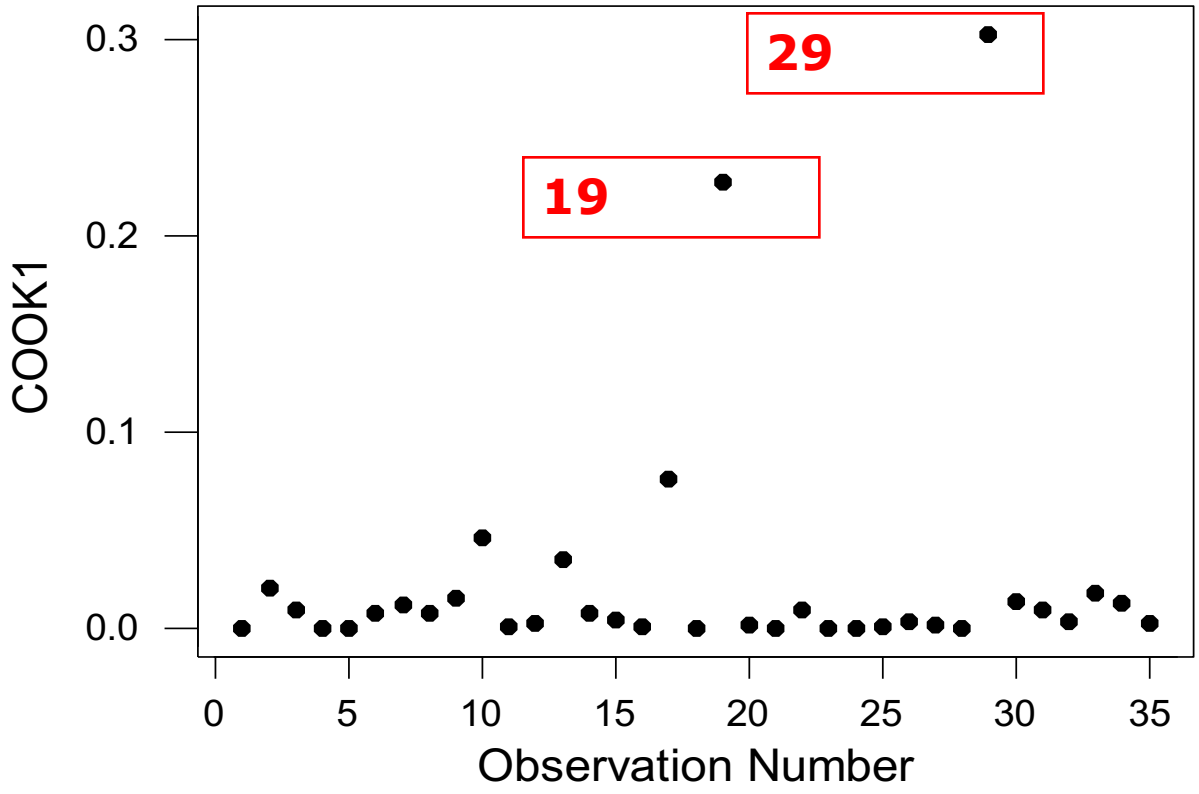
DFIT Graphed



Cook's D

- Often called Cook's *Distance*
- Minitab also will compute these and store them.
- Again, it might be best just to look for high values rather than use a cutoff.

Cook's D Graphed



6.7.5 What to do with Unusual Observations

- Observation 19 (First Lincoln Financial Bank) has high influence because of its very low Sigma.
- Observation 29 (Mercury Saving) had a very high return of 13.05 but its Beta and Sigma were not unusual.
- Since both values are out of line with the other S&L banks, they may represent data recording errors.

Eliminate? Adjust?

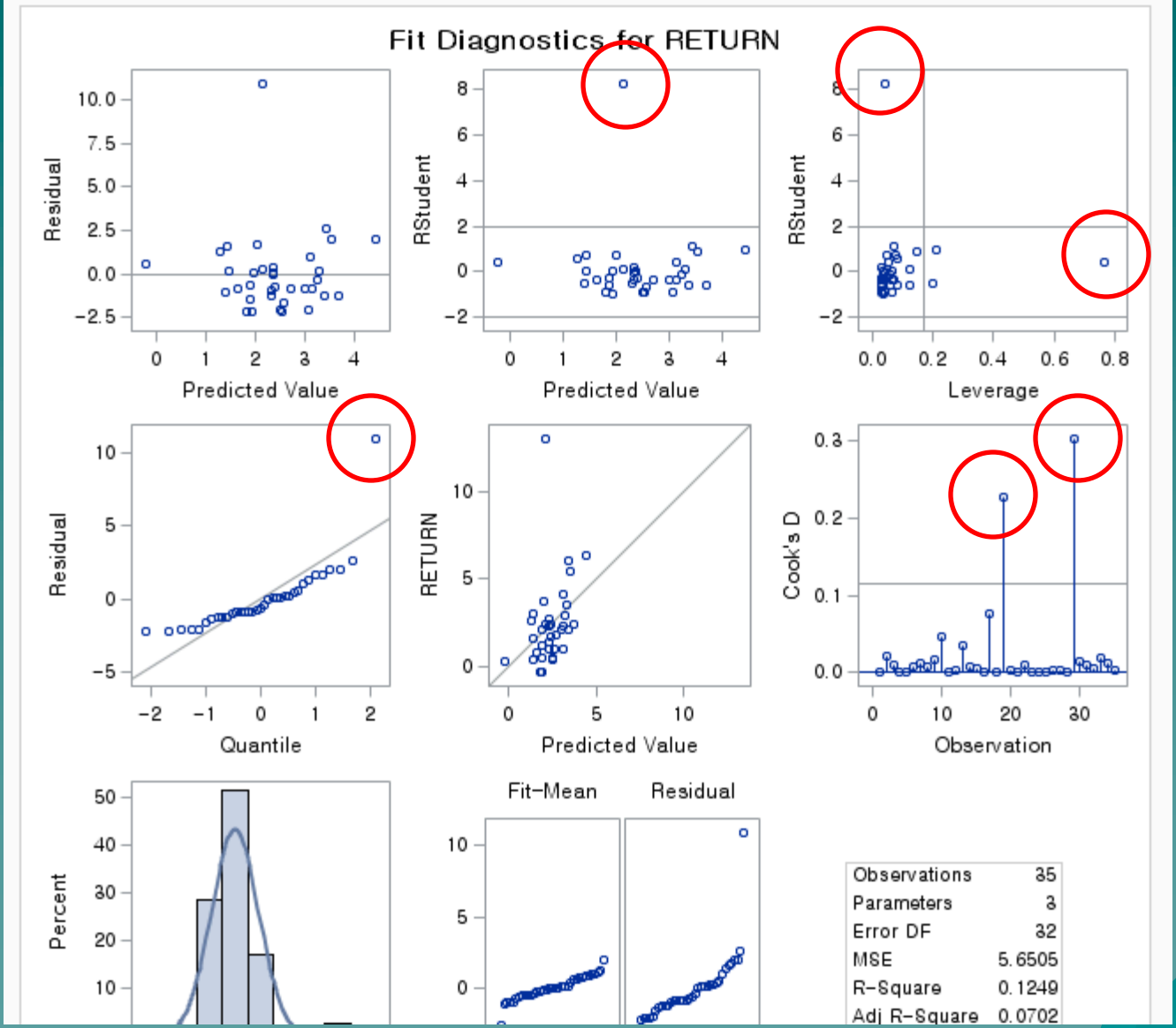
- If you can do further research you might find out the true story.
- You should eliminate an outlier data point only when you are convinced it does not belong with the others (for example, if Mercury was speculating wildly).
- An alternative is to keep the data point but add an indicator variable to the model that signals there is something unusual about this observation.

```
proc reg ;
  model return = beta sigma / influence ;
run;
```

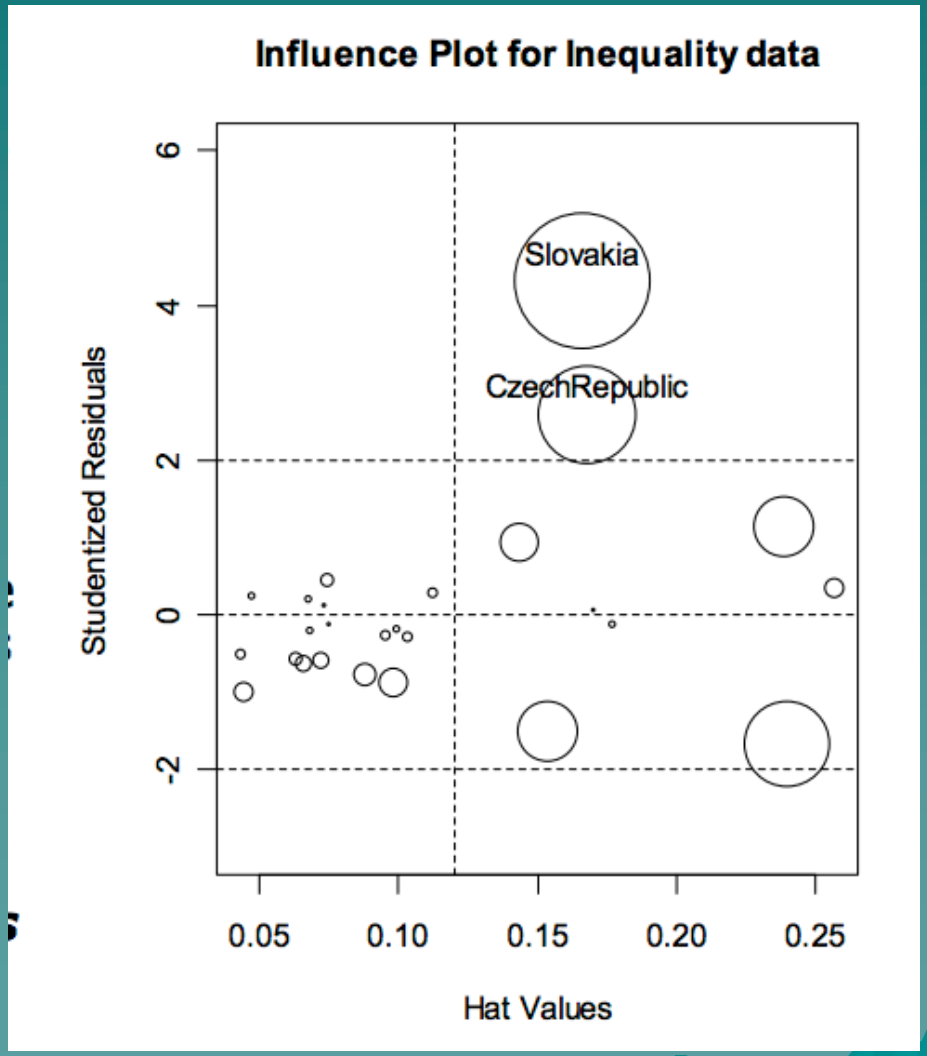
The REG Procedure
Model: MODEL1
Dependent Variable: RETURN

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS		
						Intercept	BETA	SIGMA
1	-0.0623	-0.0264	0.0434	1.1498	-0.0056	-0.0027	0.0033	-0.0014
2	-1.0504	-0.4884	0.2009	1.3451	-0.2449	-0.2034	0.2098	-0.0064
3	1.2900	0.5602	0.0818	1.1622	0.1672	0.1486	-0.0490	-0.0949
4	-0.3379	-0.1442	0.0571	1.1642	-0.0355	0.0135	0.0031	-0.0240
5	0.2002	0.0885	0.1223	1.2523	0.0330	-0.0005	-0.0217	0.0263
18	0.2493	0.1058	0.0471	1.1531	0.0235	0.0026	0.0128	-0.0119
19	0.5309	0.4538	0.7638	4.5659	0.8162	0.1689	0.5750	-0.7427
20	-0.8993	-0.3812	0.0414	1.1314	-0.0792	0.0254	-0.0056	-0.0377
28	0.1416	0.0606	0.0648	1.1758	0.0168	0.0121	-0.0006	-0.0106
29	10.9202	8.2445	0.0397	0.0352	1.6763	1.0320	-0.8025	-0.0212
30	1.6012	0.6968	0.0792	1.1402	0.2042	0.1841	-0.1009	-0.0764



Bubble Plot for Cook's D



6.8 Assessing the Assumption That the Disturbances are Independent

- If the disturbances are independent, the residuals should not display any patterns.
- One such pattern was the curvature in the residuals from the linear model in the telemarketing example.
- Another pattern occurs frequently in data collected over time.

6.8.1 Autocorrelation

- In time series data we often find that the disturbances tend to stay at the same level over consecutive observations.
- If this feature, called *autocorrelation*, is present, all our model inferences may be misleading.

First-order autocorrelation

If the disturbances have first-order autocorrelation, they behave as:

$$e_i = \rho\, e_{i-1} + \mu_i$$

where μ_i is a disturbance with expected value 0 and independent over time.

The effect of autocorrelation

If you knew that e_{56} was 10 and ρ was .7, you would expect e_{57} to be 7 instead of zero.

This dependence can lead to high standard errors for the b_j coefficients and wider confidence intervals.

6.8.2 A Test for First-Order Autocorrelation

Durbin and Watson developed a test for positive autocorrelation of the form:

$$H_0: \rho = 0$$
$$H_a: \rho > 0$$

Their test statistic d is scaled so that it is 2 if no autocorrelation is present and near 0 if it is very strong.

A Three-Part Decision Rule

The Durbin-Watson test distribution depends on n and K. The tables (Table B.7) list two decision points d_L and d_U.

If $d < d_L$ reject H_0 and conclude there is positive autocorrelation.

If $d > d_U$ accept H_0 and conclude there is no autocorrelation.

If $d_L \leq d \leq d_U$ the test is inconclusive.

Example 6.10 Sales and Advertising

n = 36 years of annual data

Y = Sales (in million $)

X = Advertising expenditures ($1000s)

Data in Table 6.6

The Test

n = 36 and K = 1 X-variable

At a 5% level of significance, Table B.7 gives $d_L = 1.41$ and $d_U = 1.52$

Decision Rule:

Reject H_0 if $d < 1.41$

Accept H_0 if $d > 1.52$

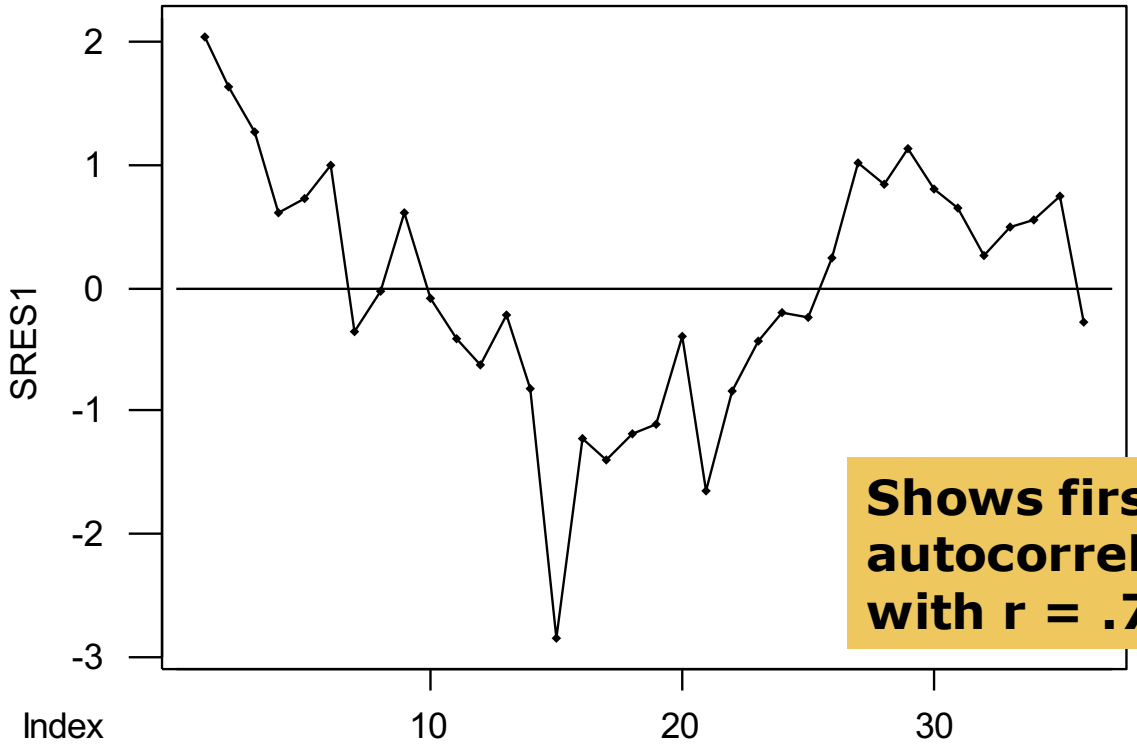
Inconclusive if $1.41 \leq d \leq 1.52$

Regression With DW Statistic

```
The regression equation is
Sales = - 633 + 0.177 Adv

Predictor         Coef      SE Coef           T         P
Constant       -632.69        47.28      -13.38     0.000
Adv           0.177233     0.007045       25.16     0.000

S = 36.49        R-Sq = 94.9%      R-Sq(adj) = 94.8%

Analysis of Variance

Source              DF          SS          MS          F         P
Regression           1      842685      842685     632.81     0.000
Residual Error      34       45277        1332
Total               35      887961

Unusual Observations
Obs         Adv       Sales          Fit       SE Fit     Residual     St Resid
  1        5317      381.00       309.62        11.22        71.38         2.06R
 15        6272      376.10       478.86         6.65      -102.76        -2.86R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 0.47
```

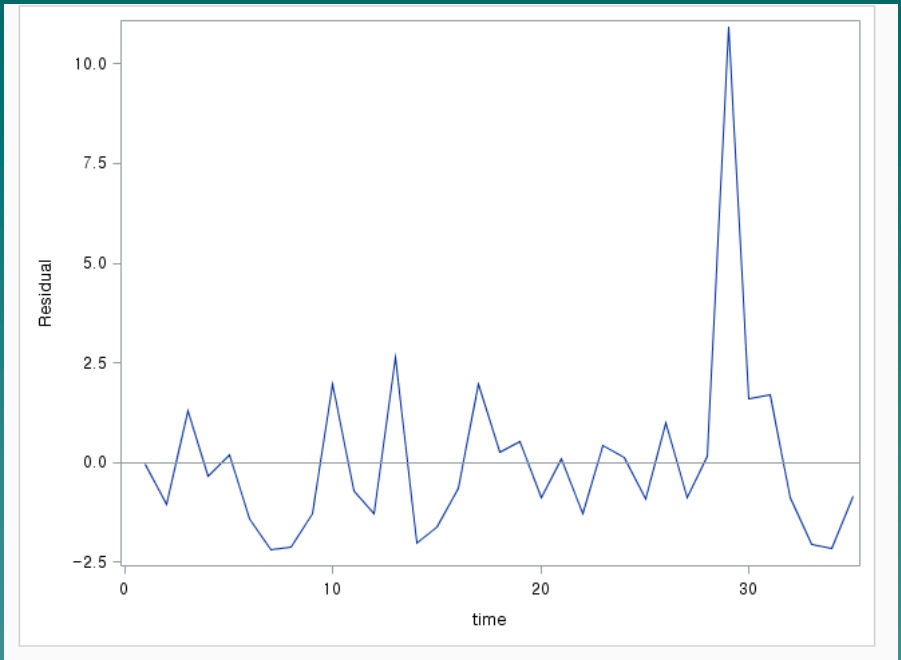
Significant autocorrelation

Plot of Residuals over Time



Shows first-order autocorrelation with r = .71

```
proc  reg  ;
   model  return  =  beta  sigma / dwprob ;
   output  out=res_out  r=res ;
run;

proc sgplot data=res_out ;
   series  x=time  y=res;
   refline  0.0 ;
run;
```



Durbin-Watson D	1.712
Pr < DW	0.1809
Pr > DW	0.8191
Number of Observations	35
1st Order Autocorrelation	0.142

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

6.8.3 Correction for First-Order Autocorrelation

One popular approach creates a new y and x variable.

First, obtain an estimate of ρ. Here we use $r = .71$ from Minitab's Autocorrelation analysis.

Then compute $y_i^* = y_i - r\, y_{i-1}$

and $x_i^* = x_i - r\, x_{i-1}$

First Observation Missing

Because the transformation depends on lagged y and x values, the first observation requires special handling.

The text suggests $y_1^* = \sqrt{1 - r^2}\ y_1$

and a similar computation for x_1^*

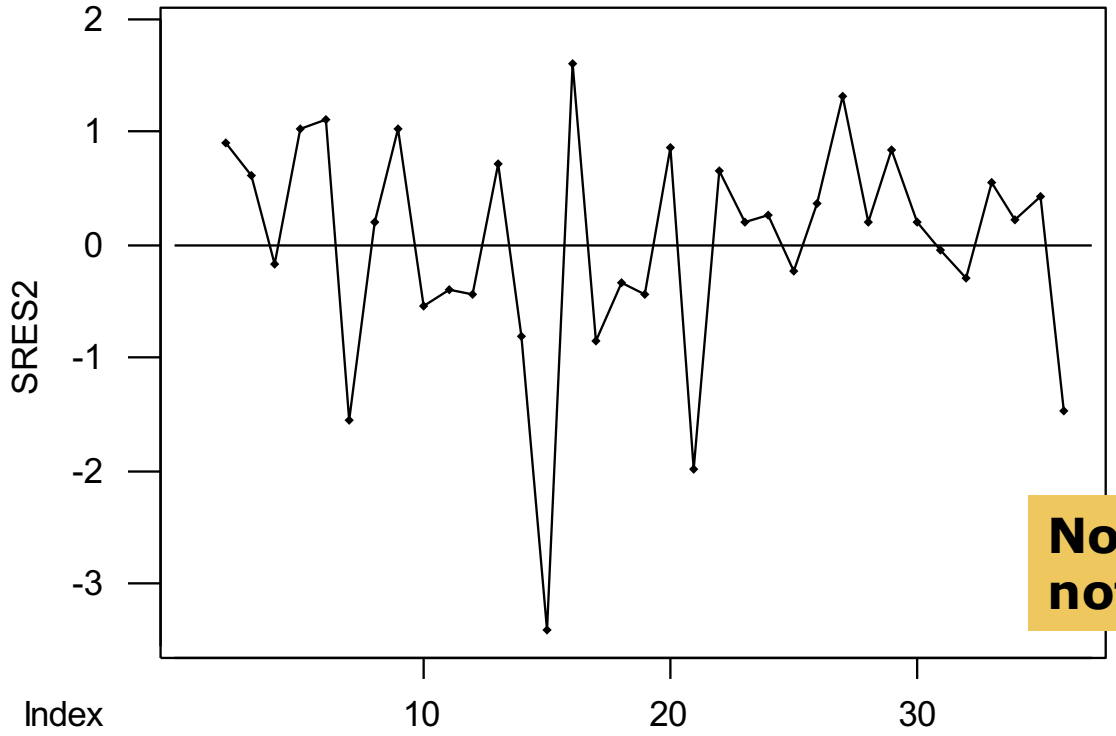
Other Approaches

- An alternative is to use an estimation technique (such as SAS's Autoreg procedure) that automatically adjusts for autocorrelation.
- A third option is to include a lagged value of *y* as an explanatory variable. In this model, the DW test is no longer appropriate.

Regression With Lagged Sales as a Predictor

```
The regression equation is
Sales = - 234 + 0.0631 Adv + 0.675 LagSales

35 cases used 1 cases contain missing values

Predictor         Coef      SE Coef           T         P
Constant       -234.48        78.07       -3.00     0.005
Adv            0.06307      0.02023        3.12     0.004
LagSales        0.6751       0.1123        6.01     0.000

S = 24.12        R-Sq = 97.8%     R-Sq(adj) = 97.7%

Analysis of Variance

(deleted)

Unusual Observations
Obs         Adv       Sales          Fit      SE Fit     Residual     St Resid
 15        6272      376.10       456.24        5.54       -80.14        -3.41R
 16        6383      454.60       422.02       12.95        32.58         1.60 X
 21        6794      512.00       559.41        4.46       -47.41        -2.00R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

Residuals From Model With Lagged Sales



Now r = -.23 is not significant