

REGRESSION

Introduction

- “Data version” of best linear prediction.
- Very widely used.

Available data

- Y_i = value of response variable for i th observation
- X_{i1}, \dots, X_{ip} = values of predictor variables 1 through p for the i th observation

Goals:

- to understand how Y is related to X_1, \dots, X_p .
- to model the conditional expectation of Y given X_1, \dots, X_p
- to predict future Y values from X_1, \dots, X_p

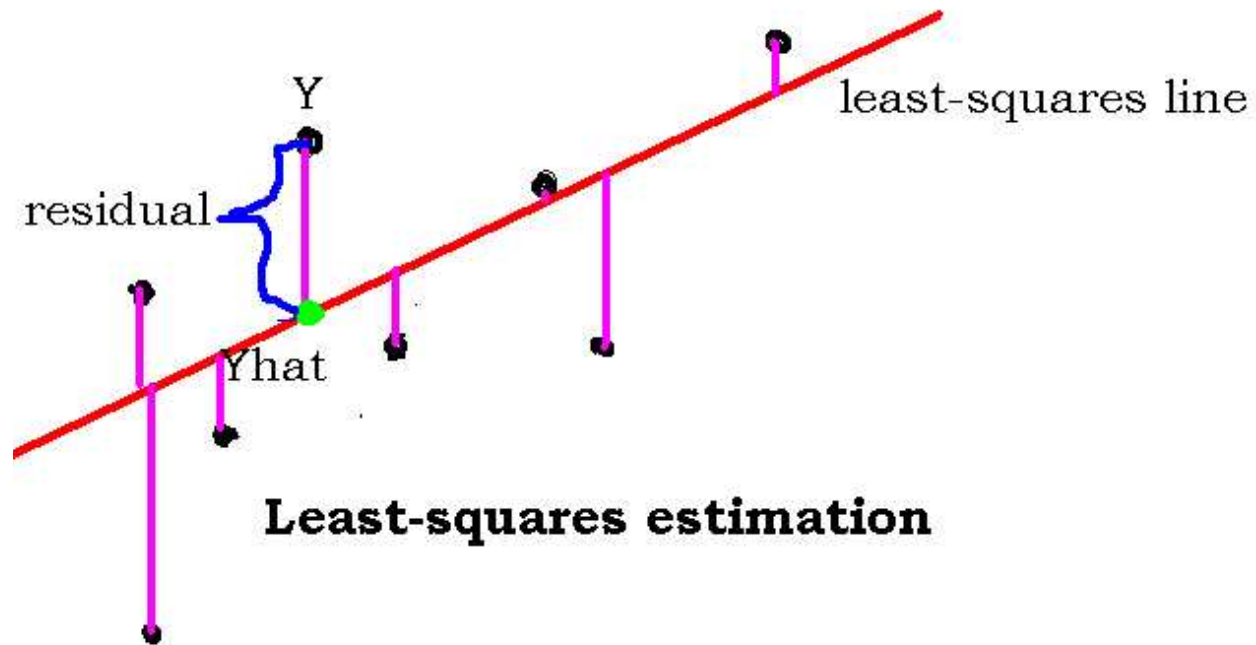
Straight Line Regression

- only one predictor variable
- model is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- β_0 and β_1 are the unknown intercept and slope of the line
- $\epsilon_1, \dots, \epsilon_n$ are iid with mean 0 and constant variance σ^2
- often the ϵ_i 's are assumed to be normally distributed

Least-squares estimation



- least-squares estimate finds b_0 and b_1 to minimize

$$\sum_{i=1}^n \{Y_i - (b_0 + b_1 X_i)\}^2$$

- using calculus, one can show that

$$b_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

and

$$b_0 = \bar{Y} - b_1 \bar{X}$$

- the least-squares line is

$$\begin{aligned}\hat{Y} &= b_0 + b_1 X = \bar{Y} + b_1(X - \bar{X}) \\ &= \bar{Y} + \left\{ \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \right\} (X - \bar{X}) \\ &= \bar{Y} + \frac{s_{xy}}{s_x^2} (X - \bar{X}),\end{aligned}$$

where

$$s_{xy} = (n - 1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})$$

and s_x^2 is the sample variance of the X_i 's, that is,

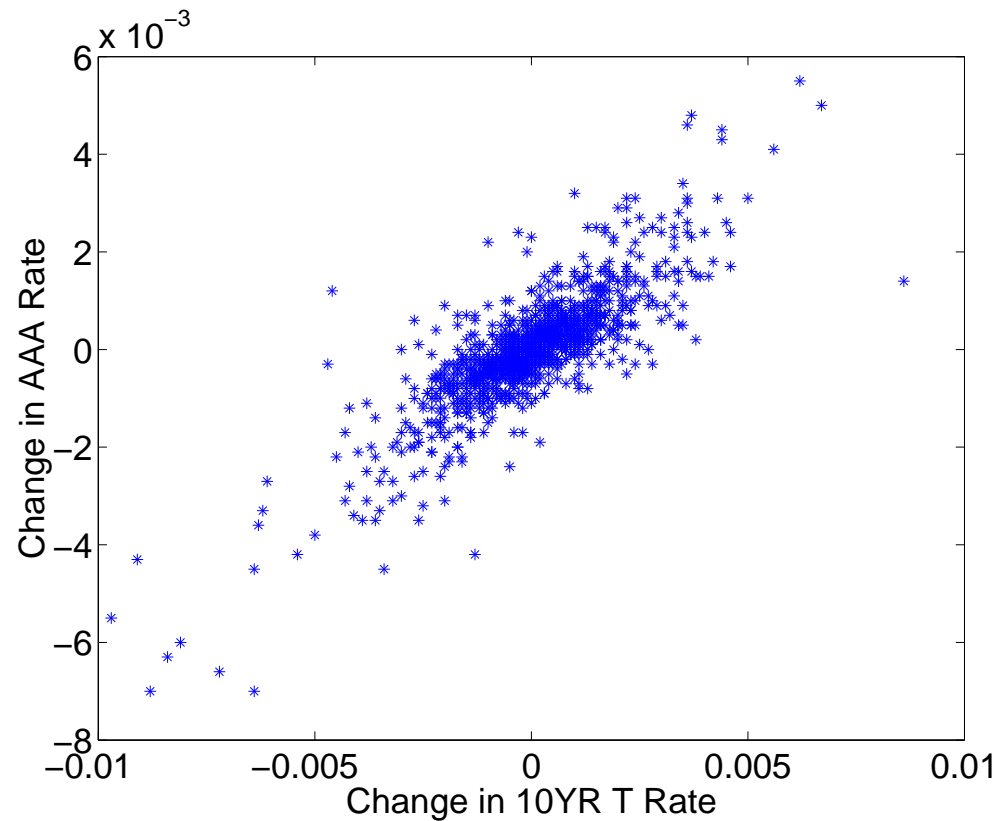
$$s_x^2 = (n - 1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Exercise:

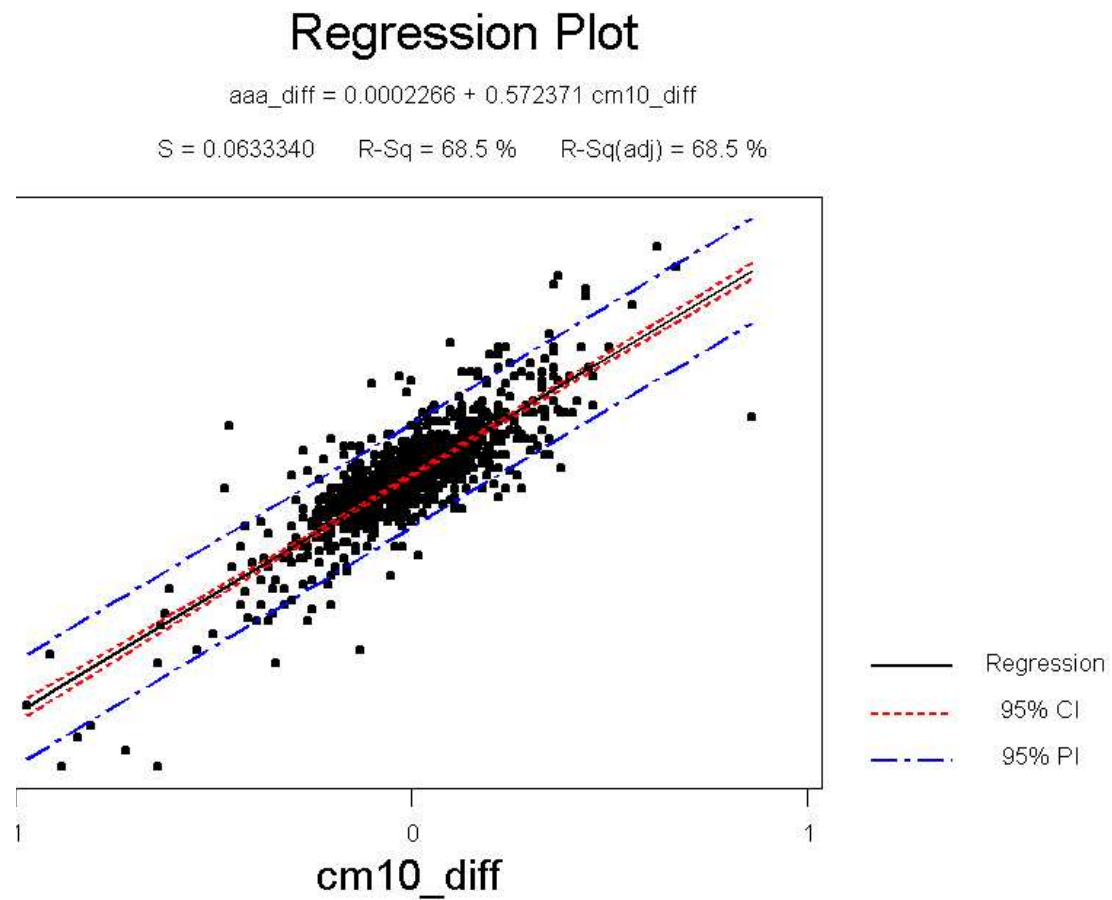
Show that if $\epsilon_1, \dots, \epsilon_n$ are IID $N(0, \sigma^2)$ then the least-squares estimates of β_0 and β_1 are also the maximum likelihood estimates.

Example: Some data on weekly interest rates, from Jan 1, 1970 to Dec 31, 1993, were obtained from the Federal Reserve Bank of Chicago. The URL is:

<http://www.chicagofed.org/economicresearchanddata/data/index.cfm>



Change in “CM10 = 10-YEAR TREASURY CONSTANT MATURITY RATE (AVERAGE, NSA)” plotted against “AAA = MOODYS SEASONED CORPORATE AAA BOND YIELDS”.



Fitted line plot from MINITAB.

Output from fitted line plot in MINITAB

Regression Analysis: aaa_diff versus cm10_diff

The regression equation is

aaa_diff = 0.0002266 + 0.572371 cm10_diff

S = 0.0633340 R-Sq = 68.5 % R-Sq(adj) = 68.5 %

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10.8950	10.8950	2716.14	0.000
Error	1249	5.0100	0.0040		
Total	1250	15.9049			

Fitted Line Plot: aaa_diff versus cm10_dif

Here is the same analysis using “regression” in MINITAB.

The first output is the estimated regression line:

The regression equation is

$\text{aaa_diff} = 0.00023 + 0.572 \text{ cm10_diff}$

1251 cases used 1 cases contain missing values

Next comes the estimates, standard errors, T-statistics, and p-values:

Predictor	Coef	SE Coef	T	P
Constant	0.000227	0.001791	0.13	0.899
cm10_dif	0.57237	0.01098	52.12	0.000

The next output is S = estimate of σ , R^2 and adjusted R^2 :

$S = 0.06333$ $R\text{-Sq} = 68.5\%$ $R\text{-Sq}(\text{adj}) = 68.5\%$

Finally, the analysis of variance table is printed. This table decomposed the variability in Y into several components:

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10.895	10.895	2716.14	0.000
Residual Error	1249	5.010	0.004		
Total	1250	15.905			

- From the output we see that the least-squares estimates of the intercept and slope are 0.000227 and 0.572.
- The missing values created by differencing caused MINITAB to print a warning.

Standard errors, t-values, and p-values

Each of the coefficients in the MINITAB output has three other statistics associated with it:

- SE = standard error
 - This is estimated standard deviation of the least squares estimator and tells us the precision of that estimator.
- t-value
 - This is the t-statistic for testing that the coefficient is 0.

- p-value
 - This is the p-value for the test of the null hypothesis that the coefficient is 0 versus the alternative that it is not 0.
 - The p-value is 0.000 here.
- If the p-value is small as it is here, then this is evidence that the coefficient is **not** 0 which means that the predictor has some effect.

Analysis of variance, R^2 , and F-tests

$$\text{total SS} = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

$$\text{regression SS} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \quad \text{where } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\text{residual error SS} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

$$\text{total SS} = \text{regression SS} + \text{residual error SS}.$$

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = 1 - \frac{\text{residual error SS}}{\text{total SS}}$$

- The **degrees of freedom for regression** is $p =$ number of predictor variables.
 - Note that p is 1 for straight-line regression.
- The **total degrees of freedom** is $n - 1$.
- The **residual error degrees of freedom** is $n - p - 1$.
- The **mean sum of squares (MS)** for any source is its sum of squared divided by its degrees of freedom.
- The **residual MS** is an unbiased estimator of σ^2 .
- The other means sum of squares are used for testing.

$$F = \frac{\text{regression MS}}{\text{residual error MS}}$$

- The **F-statistic** tests null hypothesis that there is no linear relationship between any of the predictors and Y .
- The entry in the column labeled “P” is the **p-value** of this test.
- In our example, the p-value is 0.000 which is very strong evidence against the null hypothesis.
- We conclude that there *is* a relationship between changes in CM10 and changes in AAA.

Regression and best linear prediction

- Note the similarity between the best linear predictor

$$\hat{Y} = E(Y) + \frac{\sigma_{xy}}{\sigma_x^2} \{X - E(X)\},$$

and the least-squares line

$$\hat{Y} = \bar{Y} + \frac{s_{xy}}{s_x^2} (X - \bar{X}),$$

- The least-squares line is a sample version of the best linear predictor.

- ρ_{XY}^2 , the squared correlation between X and Y , is the fraction of variation in Y that can be predicted using the linear predictor.
- The sample version of ρ_{XY}^2 is R^2 .

Multiple Linear Regression

- The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

- β_0 is the intercept.
 - It is the expected value of Y_i when all the X_{ij} 's are zero.
- The regression coefficients β_1, \dots, β_p are the slopes.
 - β_j is the partial derivative of the expected response with respect to the j th predictor:

$$\beta_j = \frac{\partial E(Y_i)}{\partial X_{ij}}.$$

- All coefficients are estimated by least-squares.
- **Example:** weekly interest rates data
 - now with the 30-year Treasury rates as a second predictor.
 - Thus $p = 2$.
- Here is the analysis using SAS.

```
options linesize = 72 ;
data WeeklyInterest ;
infile 'C:\book\SAS\WeeklyInterest.dat' ;
input month day year ff tb03 cm10 cm30 discount prime aaa ;
if lag(cm30) > 0 ;
aaa_dif = dif(aaa) ;
cm10_dif = dif(cm10) ;
cm30_dif = dif(cm30) ;
id = _N_ ;
run ;
title 'Weekly Interest Rates' ;
proc reg ;
model aaa_dif = cm10_dif cm30_dif / ss1 ss2 vif ;
output out=WeeklyInterest predicted=predicted rstudent=rstudent cookd=cookd h=leverage ;
run ;
proc gplot ;
plot rstudent*predicted ;
plot (rstudent cookd leverage cm10_dif cm30_dif)*id ;
plot cm10_dif*cm30_dif ;
run ;
```


The REG Procedure
Dependent Variable: aaa_dif
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	11.35366	5.67683	1357.95	<.0001
Error	876	3.66206	0.00418		
Corrected Total	878	15.01572			

Root MSE	0.06466	R-Square	0.7561
Dependent Mean	-0.00130	Adj R-Sq	0.7556
Coef Var	-4985.33904		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	-0.00010686	0.00218	-0.05	0.9609	0.00148
cm10_dif	1	0.36041	0.04456	8.09	<.0001	11.20585
cm30_dif	1	0.29655	0.04987	5.95	<.0001	0.14781

Parameter Estimates

Variable	DF	Type II SS	Variance Inflation
Intercept	1	0.00001004	0
cm10_dif	1	0.27353	14.03581
cm30_dif	1	0.14781	14.03581

Model Selection

- Model selection means selection of the predictor variables.
- Two principles to balance:
 - Larger models have less bias
 - * they **would** give the best predictions **if** all coefficients could be estimated without error.
 - When unknown coefficients are replaced by estimates, the prediction become less accurate
 - * this effect is worse when there are more coefficients to estimate.

- Thus, larger models have:
 - less bias (**good**)
 - more variability (**bad**)

- Caveat: do not use automatic model selection software blindly.
- R^2 not useful for comparing models of different sizes.
 - It always chooses the largest model
- The adjusted R^2 statistic can be used to select models.

- C_p is a statistic that estimates how well a model will predict.
 - C_p is closely related to the AIC statistic.

- Suppose there are M predictors.
 - $\hat{\sigma}_M^2$ is the estimate of σ^2 using all of them.
 - $SSE(p)$ is the sum of squares for error for a model with only $p \leq M$ predictors.
- n is the sample size
- Then C_p is

$$C_p = \frac{SSE(p)}{\hat{\sigma}_M^2} - n + 2(p + 1)$$

Now let's add more potential predictor variables so there is a total of four. They are the changes in:

"FF = FEDERAL FUNDS RATE"

"CM10 = 10-YEAR TREASURY CONSTANT MATURITY RATE (AVERAGE, NSA)"

"CM30 = 30-YEAR TREASURY CONSTANT MATURITY RATE (AVERAGE, NSA)"

"PRIME = PRIME LENDING RATE CHARGED BY COMMERCIAL BANKS"

Variable Selection Using SAS

```
options linesize = 72 ;
data WeeklyInterest ;
infile 'C:\book\SAS\WeeklyInterest.dat' ;
input month day year ff tb03 cm10 cm30 discount prime aaa ;
if lag(cm30) > 0 ;
aaa_dif = dif(aaa) ;
cm10_dif = dif(cm10) ;
cm30_dif = dif(cm30) ;
ff_dif = dif(ff) ;
prime_dif = dif(prime) ;
run ;
title 'Weekly Interest Rates' ;
proc reg ;
model aaa_dif = cm10_dif cm30_dif ff_dif prime_dif/selection=rsquare adjrsq cp sbc aic ;
run ;
```

- At the beginning of the data set, all values of cm30 are zero.
- These data are actually missing, not zero.
- Treating them as zero, not missing, causes all sorts of bad things to happen. (Earlier results done in MINITAB had this problem.)

Dependent Variable: aaa_dif						
R-Square Selection Method						
Number in		Adjusted				
Model	R-Square	R-Square	C(p)	AIC	SBC	
1	0.7463	0.7460	35.4718	-4778.8055	-4769.24795	
1	0.7379	0.7376	65.5166	-4750.2675	-4740.70994	
1	0.0625	0.0615	2489.033	-3630.0113	-3620.45378	
1	0.0320	0.0309	2598.720	-3601.8083	-3592.25074	

2	0.7561	0.7556	2.1482	-4811.5872	-4797.25086	
2	0.7463	0.7458	37.2491	-4777.0205	-4762.68417	
2	0.7463	0.7457	37.4036	-4776.8714	-4762.53501	
2	0.7391	0.7385	63.2227	-4752.2898	-4737.95341	
2	0.7379	0.7373	67.5166	-4748.2675	-4733.93116	
2	0.0727	0.0706	2454.497	-3637.6104	-3623.27404	

3	0.7563	0.7555	3.5415	-4810.1968	-4791.08170	
3	0.7562	0.7553	3.9224	-4809.8141	-4790.69896	
3	0.7464	0.7455	39.0751	-4775.1885	-4756.07337	
3	0.7392	0.7383	64.8002	-4750.6865	-4731.57137	

4	0.7564	0.7553	5.0000	-4808.7412	-4784.84732	

Number in Model	R-Square	Variables in Model
1	0.7463	cm10_dif
1	0.7379	cm30_dif
1	0.0625	ff_dif
1	0.0320	prime_dif

2	0.7561	cm10_dif cm30_dif
2	0.7463	cm10_dif prime_dif
2	0.7463	cm10_dif ff_dif
2	0.7391	cm30_dif ff_dif
2	0.7379	cm30_dif prime_dif
2	0.0727	ff_dif prime_dif

3	0.7563	cm10_dif cm30_dif ff_dif
3	0.7562	cm10_dif cm30_dif prime_dif
3	0.7464	cm10_dif ff_dif prime_dif
3	0.7392	cm30_dif ff_dif prime_dif

4	0.7564	cm10_dif cm30_dif ff_dif prime_dif

Nonlinear Regression

Often we can derive a theoretical model **but it is not linear.**

Example: the price of par \$1,000 zero-coupon bonds

- The owner of a bond will be paid \$1,000 at maturity.
 - but no payments prior to maturity
- The price of a zero-coupon bond will always be less than par.
- Suppose that there are a variety of bonds with different maturities
 - the i th type of bond has maturity T_i .

- Suppose all market participants agree that the bonds should pay interest at a continuously compounded rate r .
 - We want to estimate r from the price data.
- Under this assumption, the present price of a bond with maturity T_i is

$$P_i = 1,000 \exp(-rT_i).$$

- There will be some random variation in the observed prices.
- Model:

$$P_i = 1,000 \exp(-rT_i) + \epsilon_i.$$

- An estimate of r can be determined by minimizing

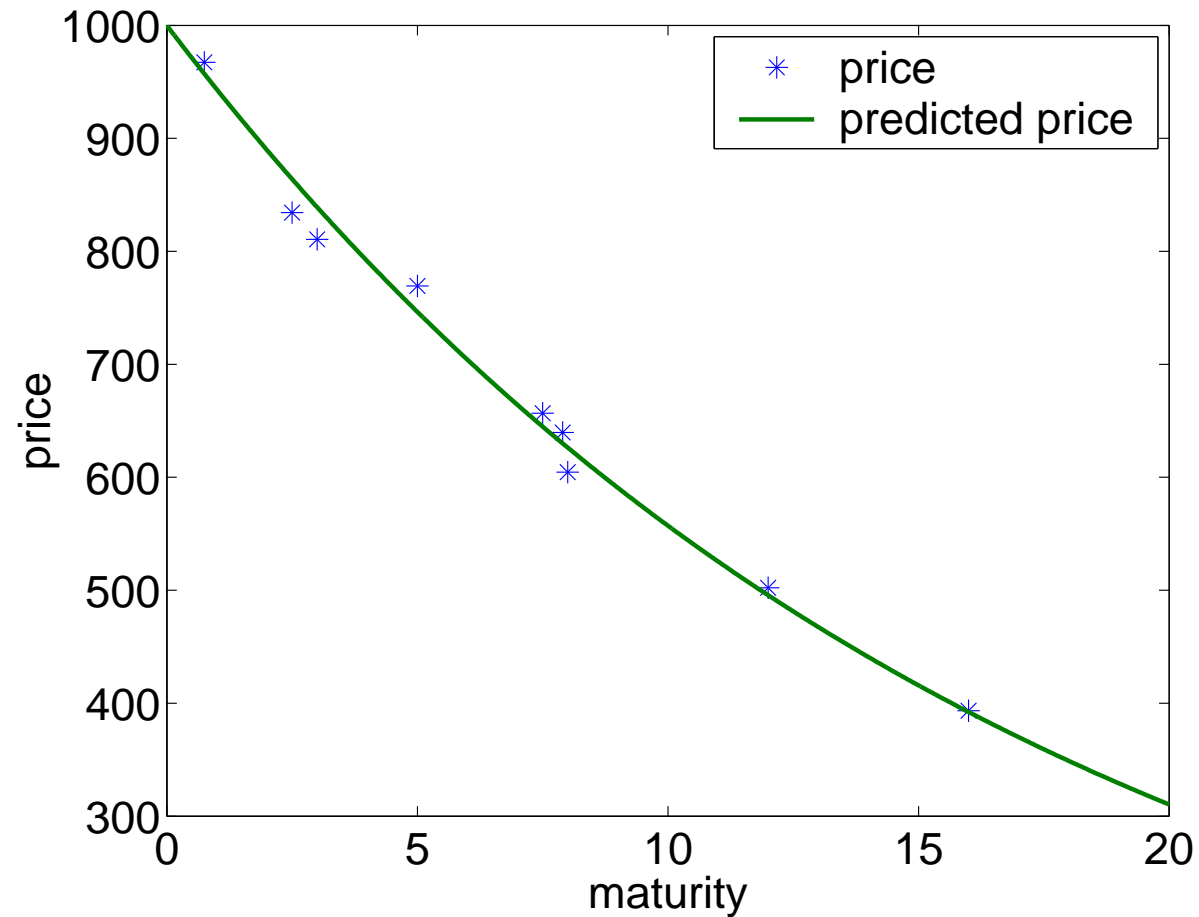
$$\sum_{i=1}^n \left\{ P_i - 1,000 \exp(-rT_i) \right\}^2.$$

- Finding the least-squares estimate requires solving nonlinear equations.

Example:

This example uses simulated data with $r = .06$.

Maturity	Price
0.75	967.26
2.5	834.21
3	810.52
5	769.30
7.5	656.64
7.9	639.71
8	604.61
12	502.11
16	393.38



The data and the predicted price curve using nonlinear regression.

Nonlinear regression in SAS:

```
data bondprices ;  
infile 'c:\courses\or473\data\bondprices.dat' ;  
input maturity price ;  
run ;  
title 'Nonlinear regression using simulated zero-coupon bond data';  
proc nlin ;  
parm r=.02 to .09 by .005 ;  
model price = 1000*exp(-r*maturity) ;  
run ;
```


Here is the SAS output:

Grid Search	
Dependent Variable price	
r	Sum of Squares
0.0200	390066
0.0250	279853
0.0300	192505
0.0350	124990
0.0400	74665.1
0.0450	39230.4
0.0500	16679.7
0.0550	5263.9
0.0600	3456.9
0.0650	9926.6
0.0700	23509.9
0.0750	43191.1

The NLIN Procedure
Iterative Phase
Dependent Variable price
Method: Gauss-Newton

Iter	r	Sum of Squares
0	0.0600	3456.9
1	0.0585	3072.8
2	0.0585	3072.8

NOTE: Convergence criterion met

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	1	4490011	4490011	11689.7	<.0001
Residual	8	3072.8	384.1		
Uncorrected Total	9	4493084			
Corrected Total	8	252587			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
r	0.0585	0.00149	0.0551	0.0619

Residual Plotting

Goals: find problems with model or data

Problems to look for:

- nonnormality
 - outliers can be a problem since they have a large influence on the estimation results
 - a common solution is transformation of the response

- non-constant variance
 - causes inefficient (= too variable) estimates
 - transformation of the response and weighting are common solutions

- nonlinearity (if the model is linear) or, more generally, model misspecification
- **model misspecification** means $E(Y|X_1, \dots, X_p)$ has a functional form different from the model
 - causes biased estimates
 - response transformation, polynomial regression, and nonparametric regression (splines, loess) are common solutions

We assume a general form of the regression model:

$$Y_i = f(\mathbf{X}_i; \boldsymbol{\beta}) + \epsilon_i$$

Predicted values:

$$\hat{Y}_i = f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})$$

Residuals:

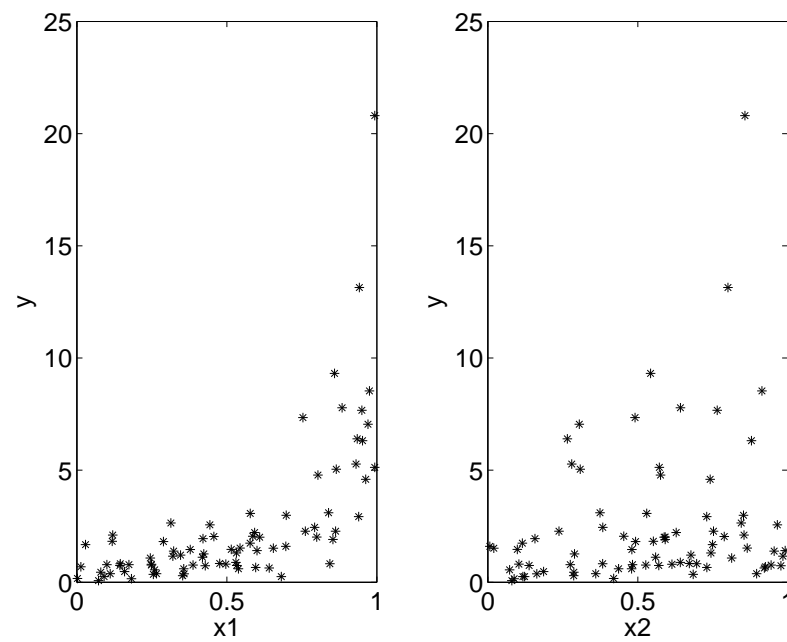
$$e_i = Y_i - \hat{Y}_i$$

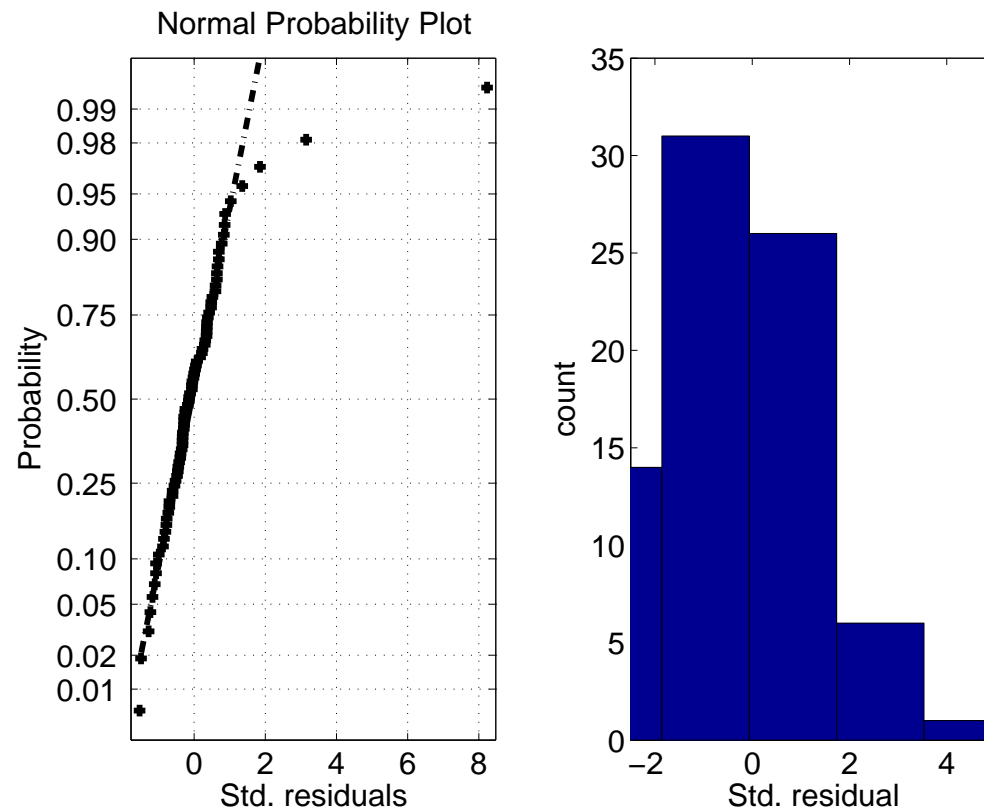
Studentized residuals:

$$\frac{e_i}{\text{se}(e_i)}$$

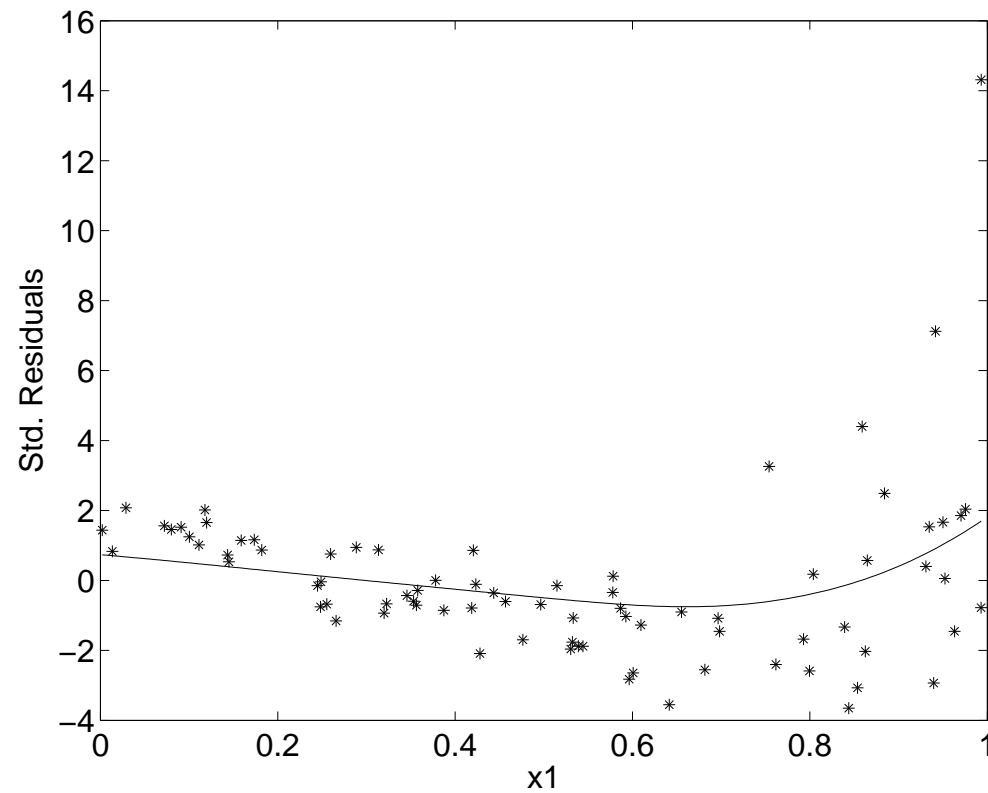
Approximately $N(0, 1)$, if modelling assumptions are true.

Example: Simulated data: Y , X_1 , X_2



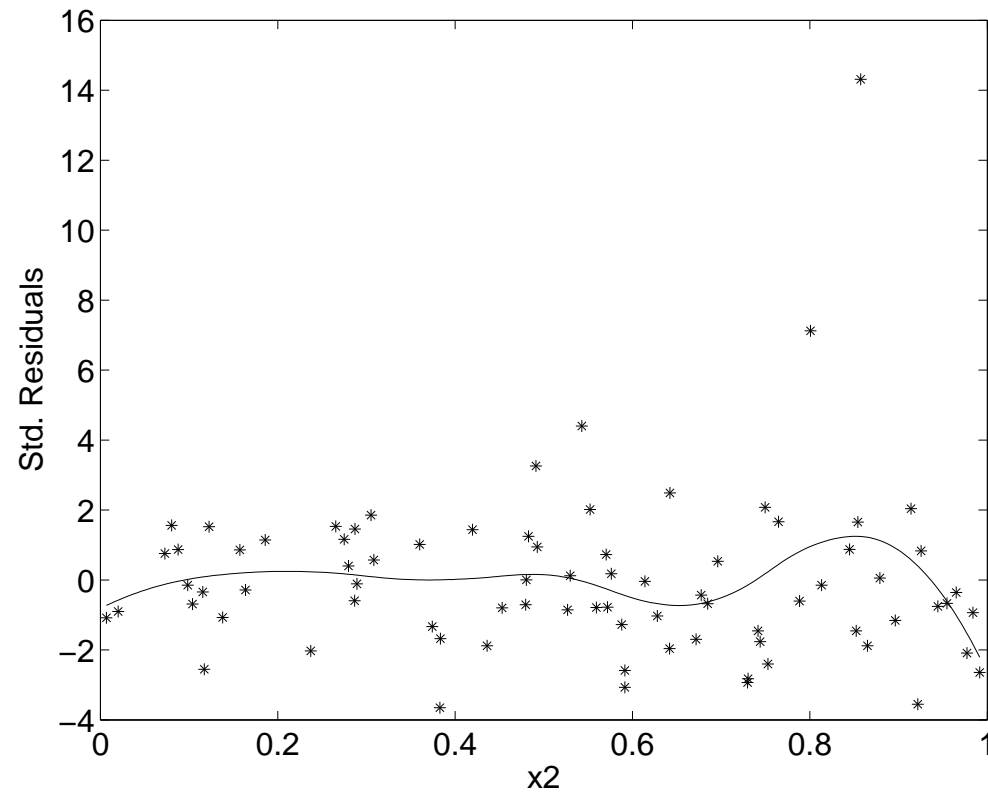


Indication of right skewness, but not severe.

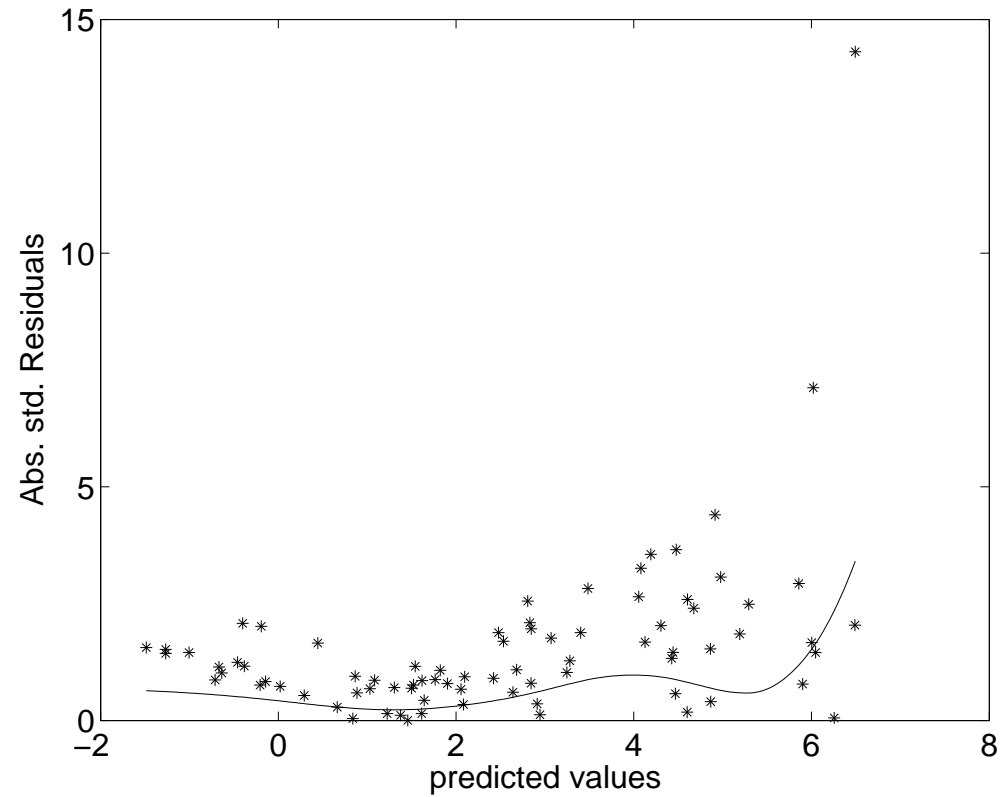


Indication of model misspecification

Model misspecification may cause major problems - fix!



No problems evident



Variance appears nonconstant

Problem of nonconstant variance:

Example: Assume X_1, X_2, X_3 are independent, all with mean μ . $\text{Var}(X_1) = \text{Var}(X_2) = 1$ and $\text{Var}(X_3) = 10$.

$$\text{Var}\{(X_1 + X_2 + X_3)/3\} = (1 + 1 + 10)/9 = 12/9 = \frac{4}{3}.$$

$$\text{Var}\{(X_1 + X_2)/2\} = (1 + 1)/4 = \frac{1}{2}.$$

It seems that dropping X_3 is a good idea.

But this does not seem quite right. Why throw any data?

Most efficient weighted average uses the **inverse variances** as weights.

$$\begin{aligned} & \text{Var}\{(X_1 + X_2 + .1X_3)/2.1\} \\ &= \{1 + 1 + (.1)^2 10\}/(2.1)^2 = \frac{1}{2.1} < \frac{1}{2}. \end{aligned}$$

Example: Estimation of Default Probabilities

Data:

- ratings: 1=Aaa (best),...,16=B3 (worse)
- default frequency (estimate of default probability)

Bluhm, C., Overbeck, L., and Wagner, C. (2003), *An Introduction to Credit Risk Modeling* (denoted **BOW** here)

1. **linear model (poor):**

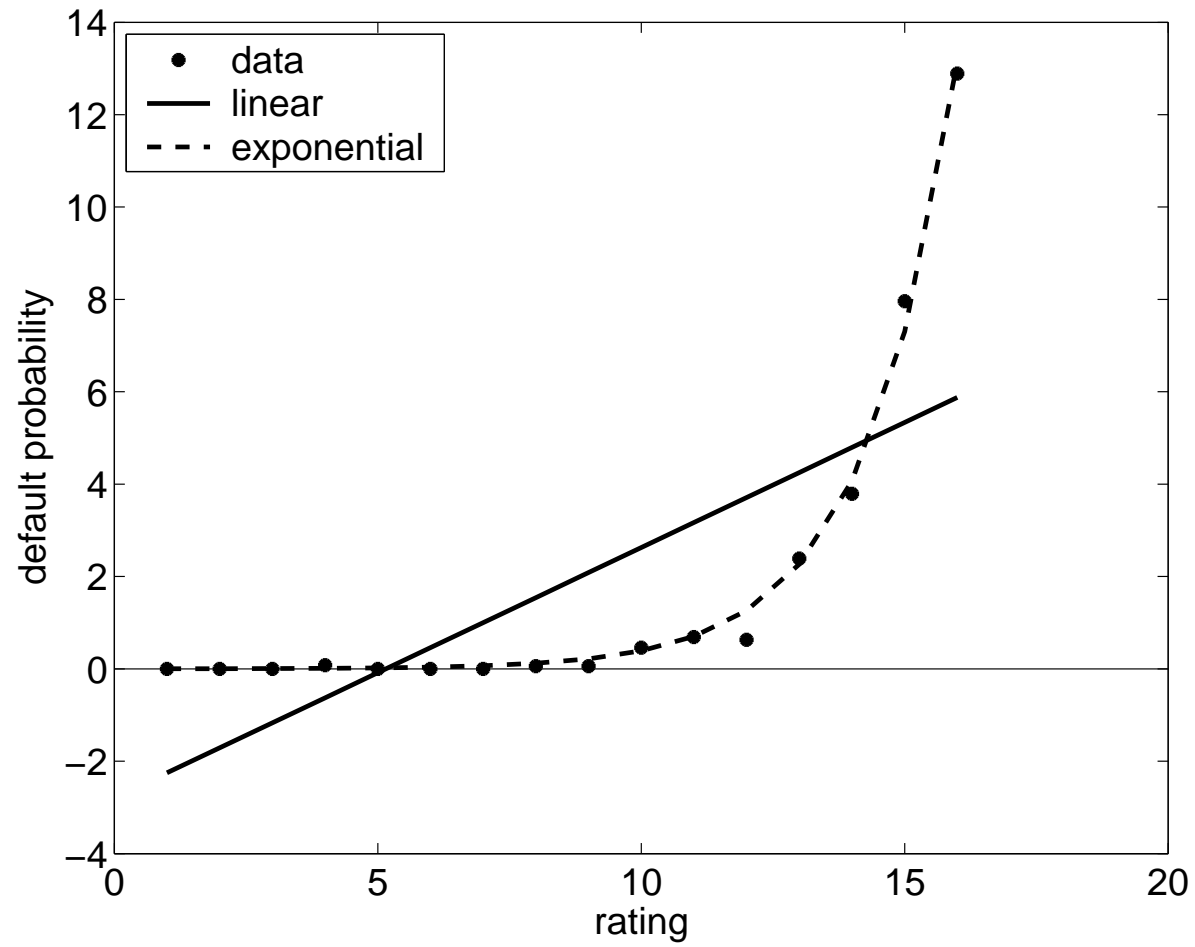
$$\Pr(\text{default}|\text{rating}) = \beta_0 + \beta_1 \text{rating}$$

2. **nonlinear model (good):**

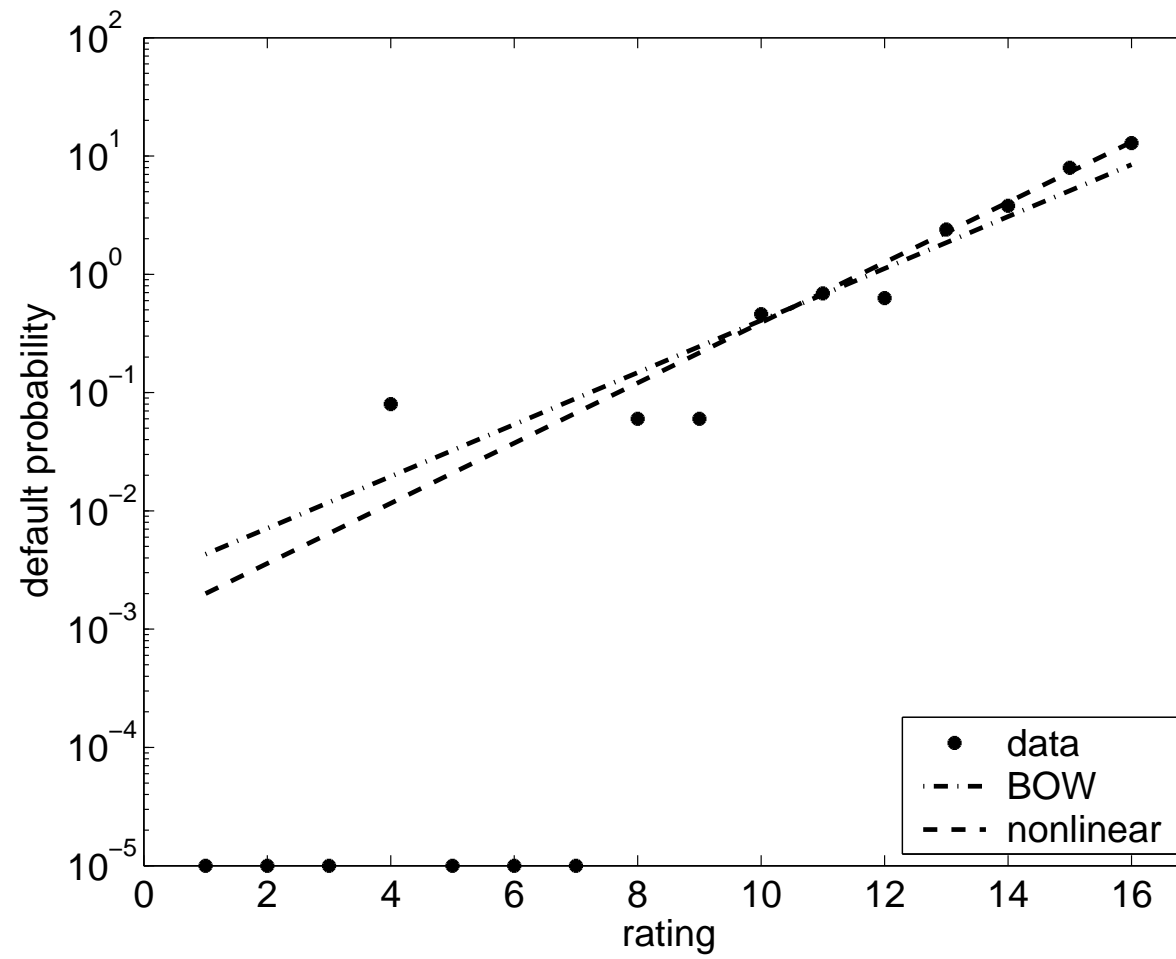
$$\Pr(\text{default}|\text{rating}) = \exp\{\beta_0 + \beta_1 \text{rating}\}$$

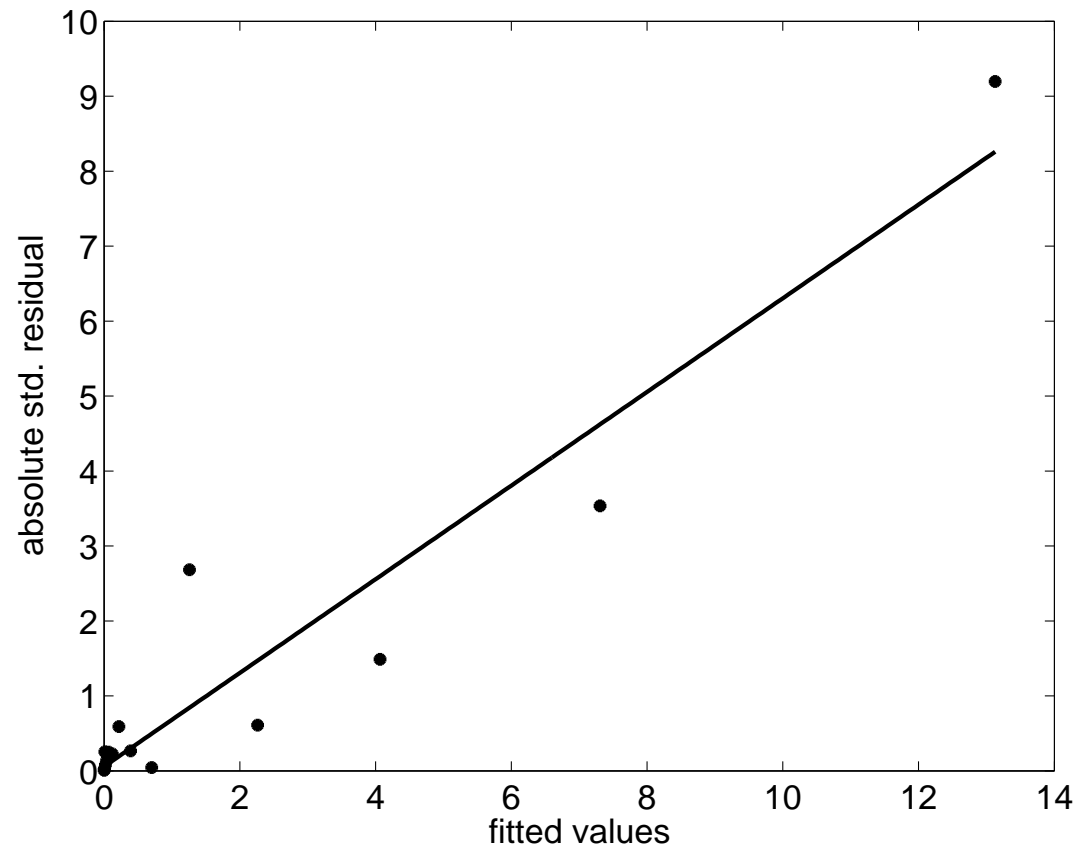
3. **loglinear model (transform of (2) – also good):**

$$\log\{\Pr(\text{default}|\text{rating})\} = \beta_0 + \beta_1 \text{rating} \text{ (used by BOW)}$$

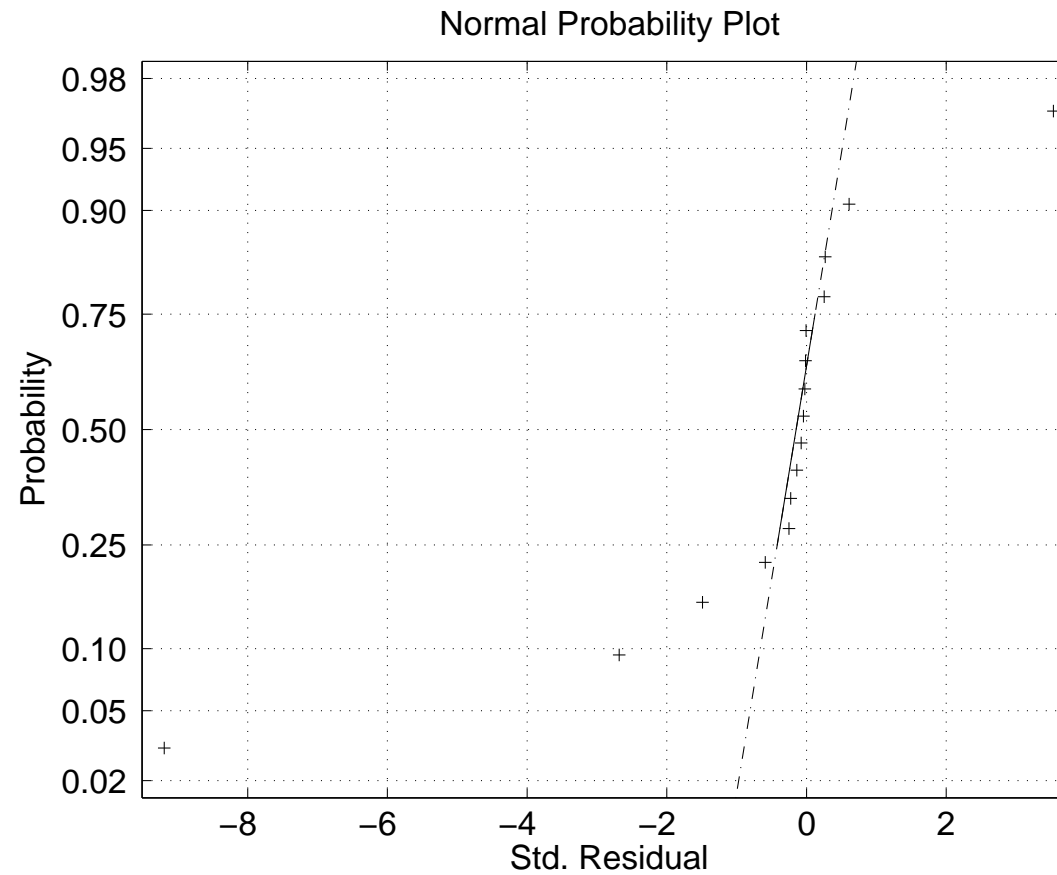


The default probability is given as a percentage.





Nonlinear regression residuals



Nonlinear regression residuals

- **nonlinear model:**

$$\Pr(\text{default}|\text{rating}) = \exp\{\beta_0 + \beta_1\text{rating}\}$$

- **linear/transformation model (BOW's model):**

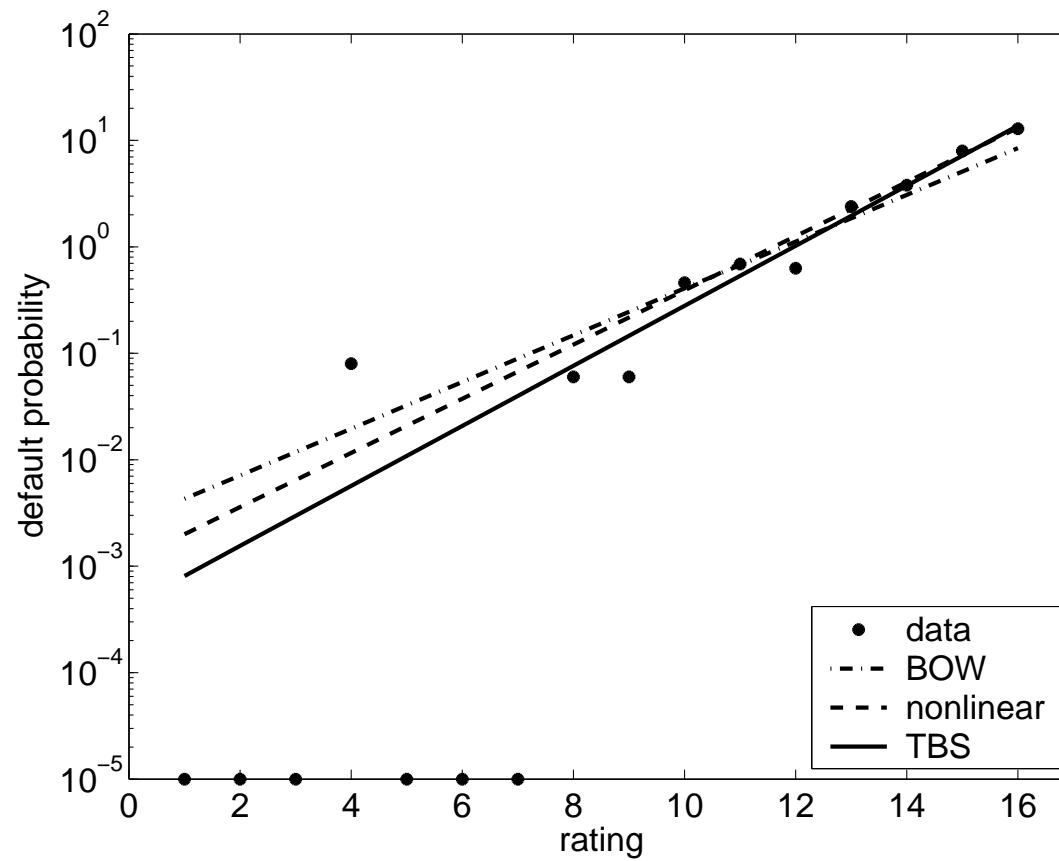
$$\log\{\Pr(\text{default}|\text{rating})\} = \beta_0 + \beta_1\text{rating}$$

- **Problem:** cannot take logs of default frequencies that are 0

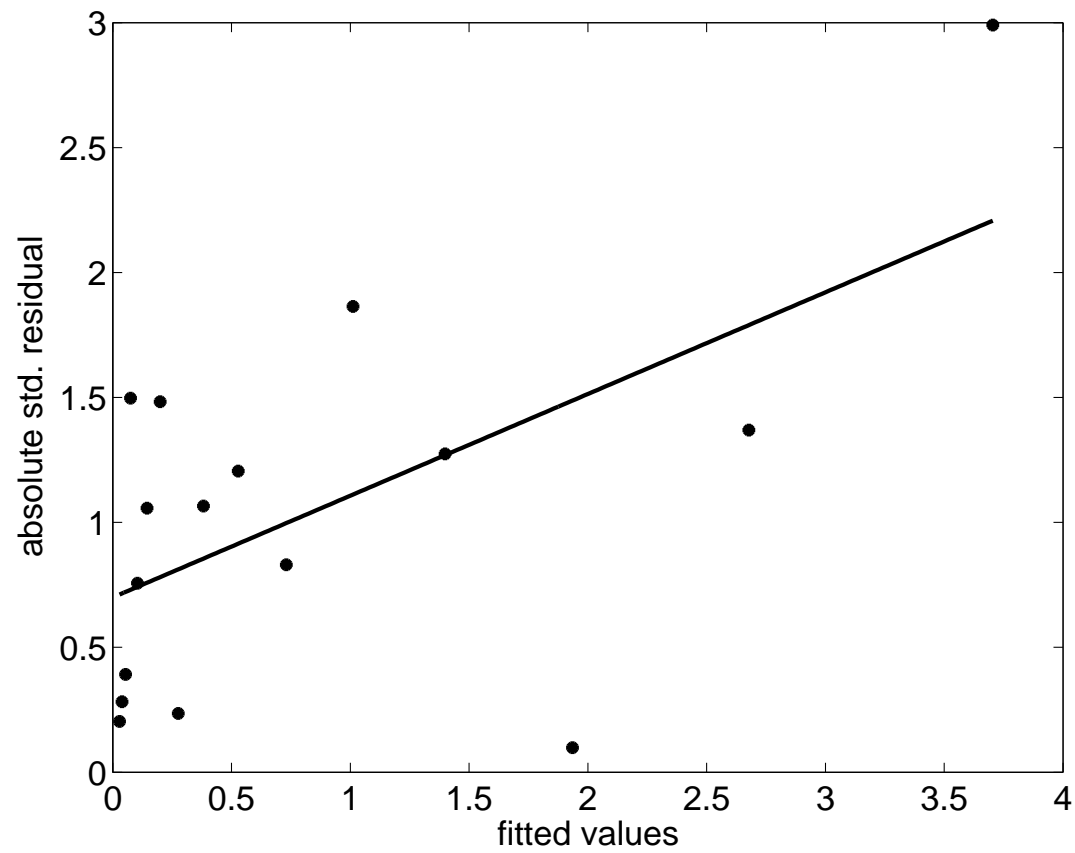
- **Transform-both-sides (TBS) model:**

$$\Pr(\text{default}|\text{rating})^\alpha = \exp[\alpha\{\beta_0 + \beta_1\text{rating}\}]$$

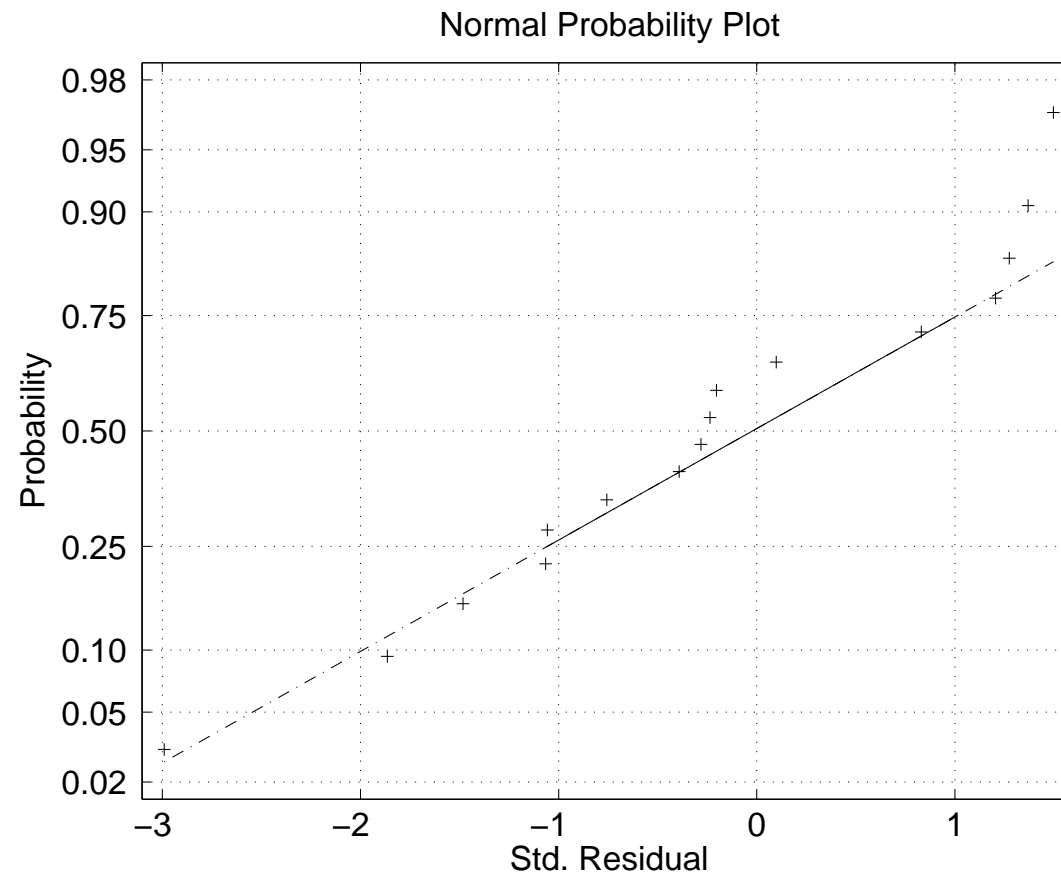
- α chosen by looking for good residual plots
- $\alpha = 1/2$ works well



TBS fit compared to others



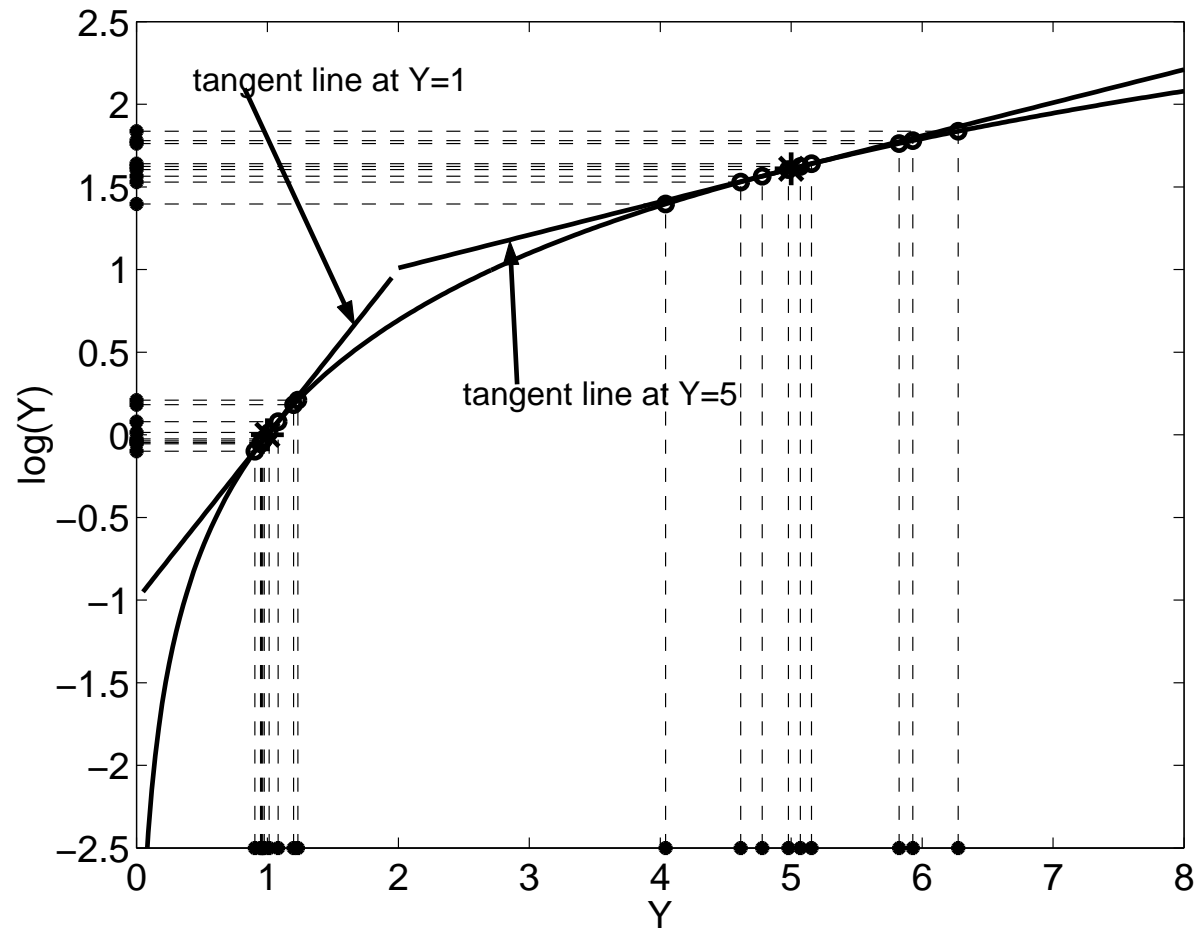
TBS residuals



TBS residuals

method	$\widehat{Pr}\{\text{default} \text{Aaa}\}$	as percent of BOW
BOW	.005%	100%
nonlinear	.002%	40%
TBS	.0008%	16%

How TBS effectively weights data



Taylor series (tangent line) linearization:

$$H(\alpha) \approx H(\alpha_0) + H'(\alpha_0)(\alpha - \alpha_0)$$

for any differentiable H and any fixed α_0 .

Apply to

$$\begin{aligned} & \sum_{i=1}^n [Y_i^\alpha - \{f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\}^\alpha]^2 \\ & \approx (\alpha)^2 \sum_{i=1}^n \left[\frac{Y_i - f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})}{\{f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\}^{-\alpha+1}} \right]^2 \end{aligned}$$

Most appropriate if

$$\text{Var}(Y_i | \mathbf{X}_i) \propto \{f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\}^{2(-\alpha+1)}.$$

From previous slide: most appropriate if

$$\text{Var}(Y_i|\mathbf{X}_i) \propto \{f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\}^{2(-\alpha+1)}.$$

Example: Poisson

- Assume $Y_i|\mathbf{X}_i$ is Poisson with mean $f(\mathbf{X}_i; \hat{\boldsymbol{\beta}})$.
- Variance equals the mean for the Poisson distribution,
so

$$\text{Var}(Y_i|\mathbf{X}_i) = f(\mathbf{X}_i; \hat{\boldsymbol{\beta}}).$$

- Thus, $\alpha = 1/2$ (use square root transformation).

TBS Regression in SAS

```
data DefaultProb ;  
infile 'C:\or473\data\DefaultData.txt' ;  
input rating prob;  
alpha = .5 ;  
transprob = prob**alpha ;  
run ;  
proc nlin ;  
parm beta0=-8 to -1 by 1 beta1=0 to 3 by .5 ;  
model transprob = exp(alpha*(beta0 + beta1*rating)) ;  
output out = outdata p=PredValue r=Residual ;  
run ;
```

```
data outdata;  
set outdata ;  
absresid = abs(Residual) ;  
run ;  
proc gplot;  
plot absresid*PredValue ;  
run ;  
proc univariate normal;  
var Residual ;  
probplot;  
run ;
```

The NLIN Procedure
Dependent Variable SqrtProb
Method: Gauss-Newton
Iterative Phase

Iter	beta0	beta1	Sum of Squares
0	-6.0000	0.5000	1.7369
1	-8.0208	0.6710	0.3230
2	-7.7480	0.6480	0.2748
3	-7.7690	0.6493	0.2747
4	-7.7676	0.6492	0.2747
5	-7.7677	0.6492	0.2747

NOTE: Convergence criterion met.

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Regression	2	28.7353	14.3676	732.21	<.0001
Residual	14	0.2747	0.0196		
Uncorrected Total	16	29.0100			
Corrected Total	15	18.4804			

The NLIN Procedure

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
beta0	-7.7677	0.5182	-8.8791	-6.6563
beta1	0.6492	0.0346	0.5750	0.7234

The UNIVARIATE Procedure

Variable: RESIDUAL

Moments

N	16	Sum Weights	16
Mean	-0.0194237	Sum Observations	-0.3107791
Std Deviation	0.1338344	Variance	0.01791165
Skewness	0.27064364	Kurtosis	-1.1282057
Uncorrected SS	0.27471116	Corrected SS	0.26867468
Coeff Variation	-689.02644	Std Error Mean	0.0334586

Basic Statistical Measures

Location		Variability	
Mean	-0.01942	Std Deviation	0.13383
Median	-0.03524	Variance	0.01791
Mode	.	Range	0.42499
		Interquartile Range	0.24725

Tests for Normality

Test	--Statistic---		-----p Value-----	
Shapiro-Wilk	W	0.937735	Pr < W	0.3222
Kolmogorov-Smirnov	D	0.151914	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.063363	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.402689	Pr > A-Sq	>0.2500

VIF's

```
proc reg ;  
model aaa_dif=cm10_dif cm30_dif ff_dif prime_dif/vif ;  
run ;
```

VIF = Variance Inflation Factor

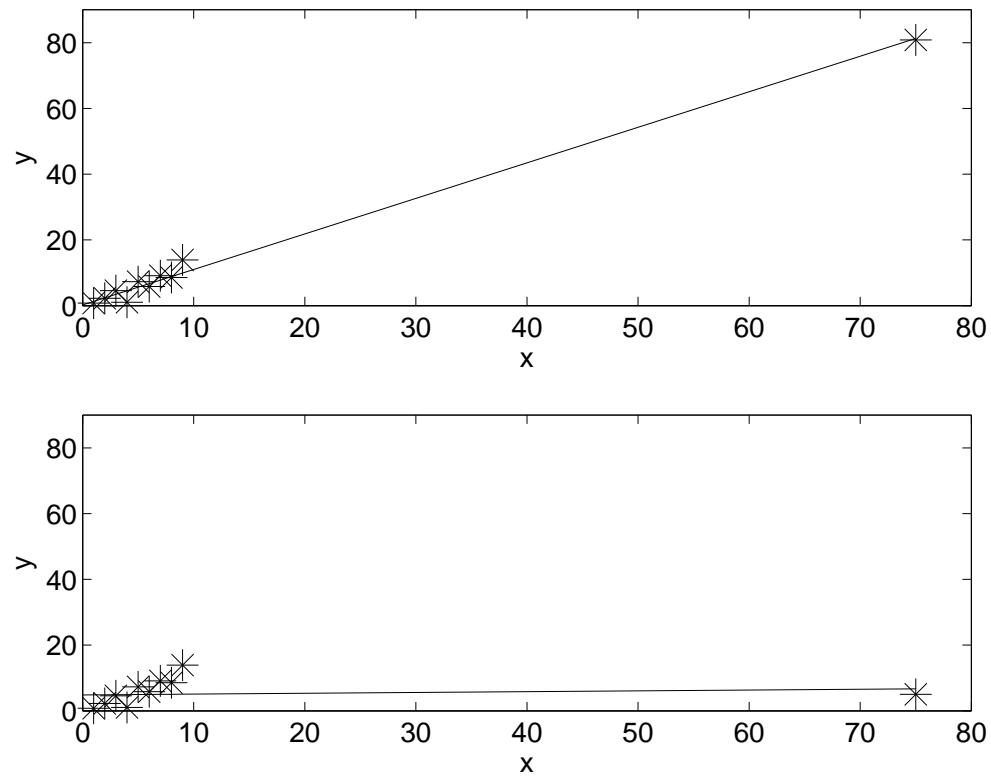
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Type I SS
Intercept	1	-0.00010103	0.00218	-0.05	0.9631	0.00148
cm10_dif	1	0.35510	0.04517	7.86	<.0001	11.20585
cm30_dif	1	0.30093	0.05010	6.01	<.0001	0.14781
ff_dif	1	0.00531	0.00553	0.96	0.3371	0.00254
prime_dif	1	-0.00788	0.01071	-0.74	0.4620	0.00227

Parameter Estimates

Variable	DF	Type II SS	Variance Inflation
Intercept	1	0.00000897	0
cm10_dif	1	0.25860	14.41205
cm30_dif	1	0.15096	14.15236
ff_dif	1	0.00386	1.19941
prime_dif	1	0.00227	1.14743

Influence Diagnostics and Leverage



Linear Regression Example

Three tools for diagnosing problems due to leverage points:

- leverages
- RSTUDENT
- Cook's D

$$\hat{Y}_i = \sum_{j=1}^n H_{ij} Y_j.$$

H_{ij} is a somewhat complex function of the X -variables

- If there is a single X -variable, then

$$H_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{(n-1)s_x^2}$$

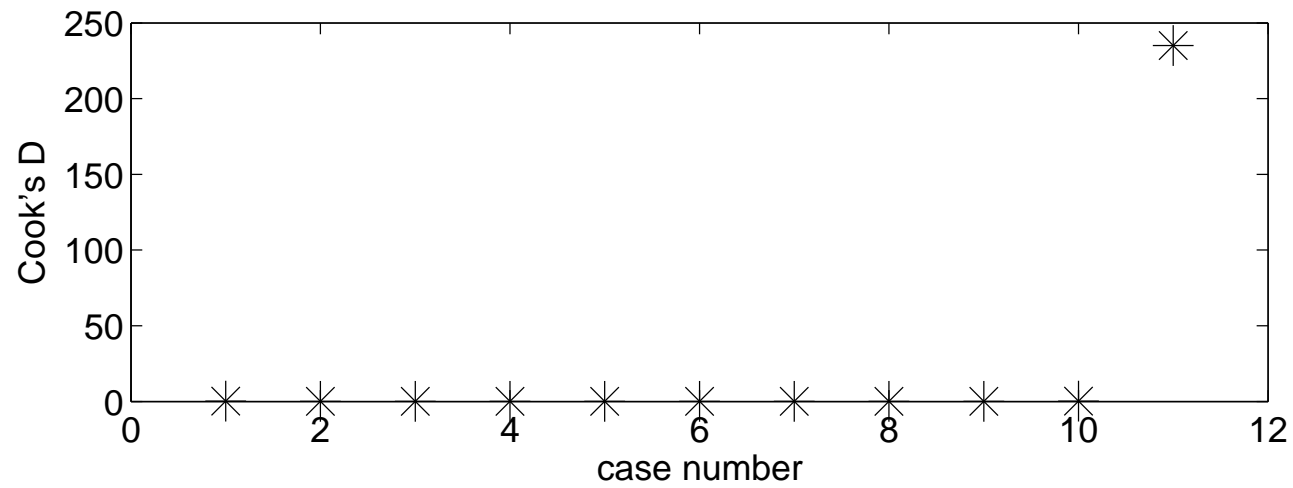
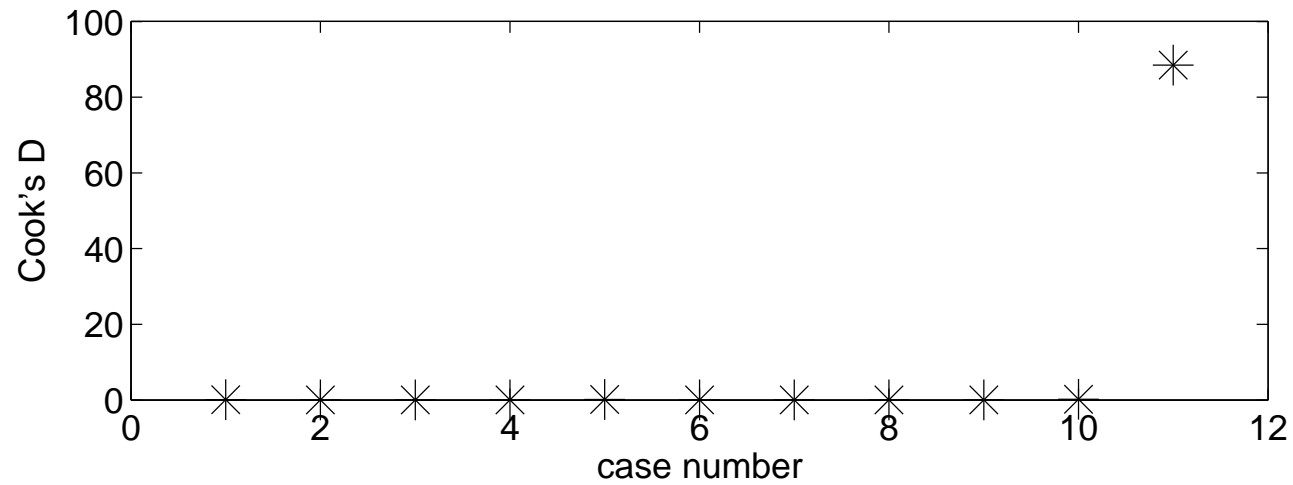
where s_x^2 is the sample variance of the X 's.

H_{11}, \dots, H_{nn} are the **leverages**

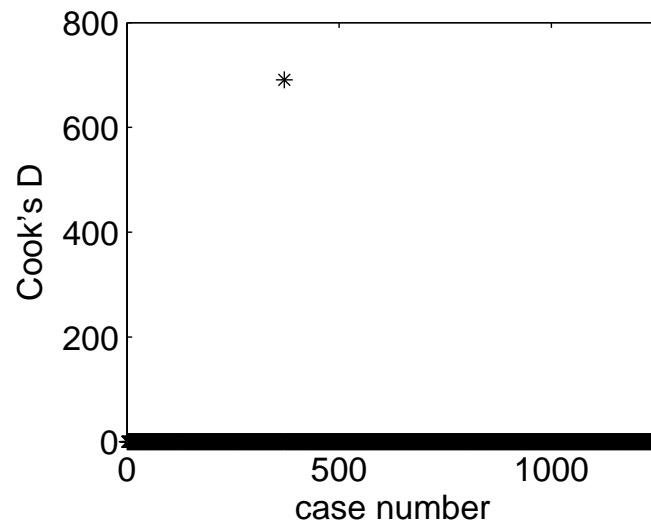
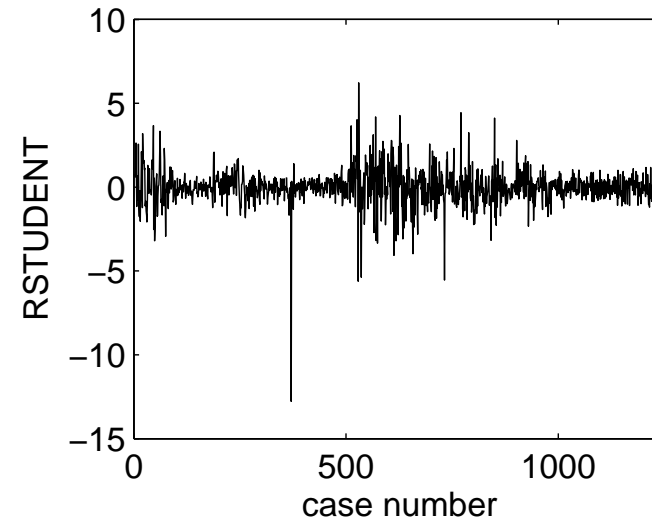
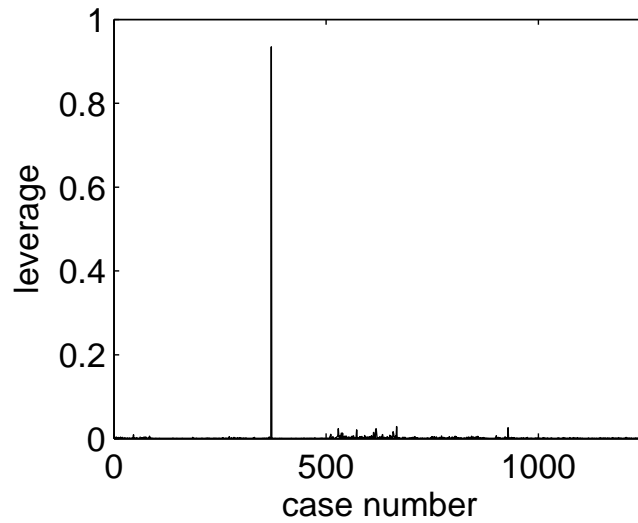
The standard error of $\hat{\epsilon}_i$ is $s\sqrt{1 - H_{ii}}$

$$\text{RSTUDENT} = \hat{\epsilon}_i / \{s_{(-i)} \sqrt{1 - H_{ii}}\}$$

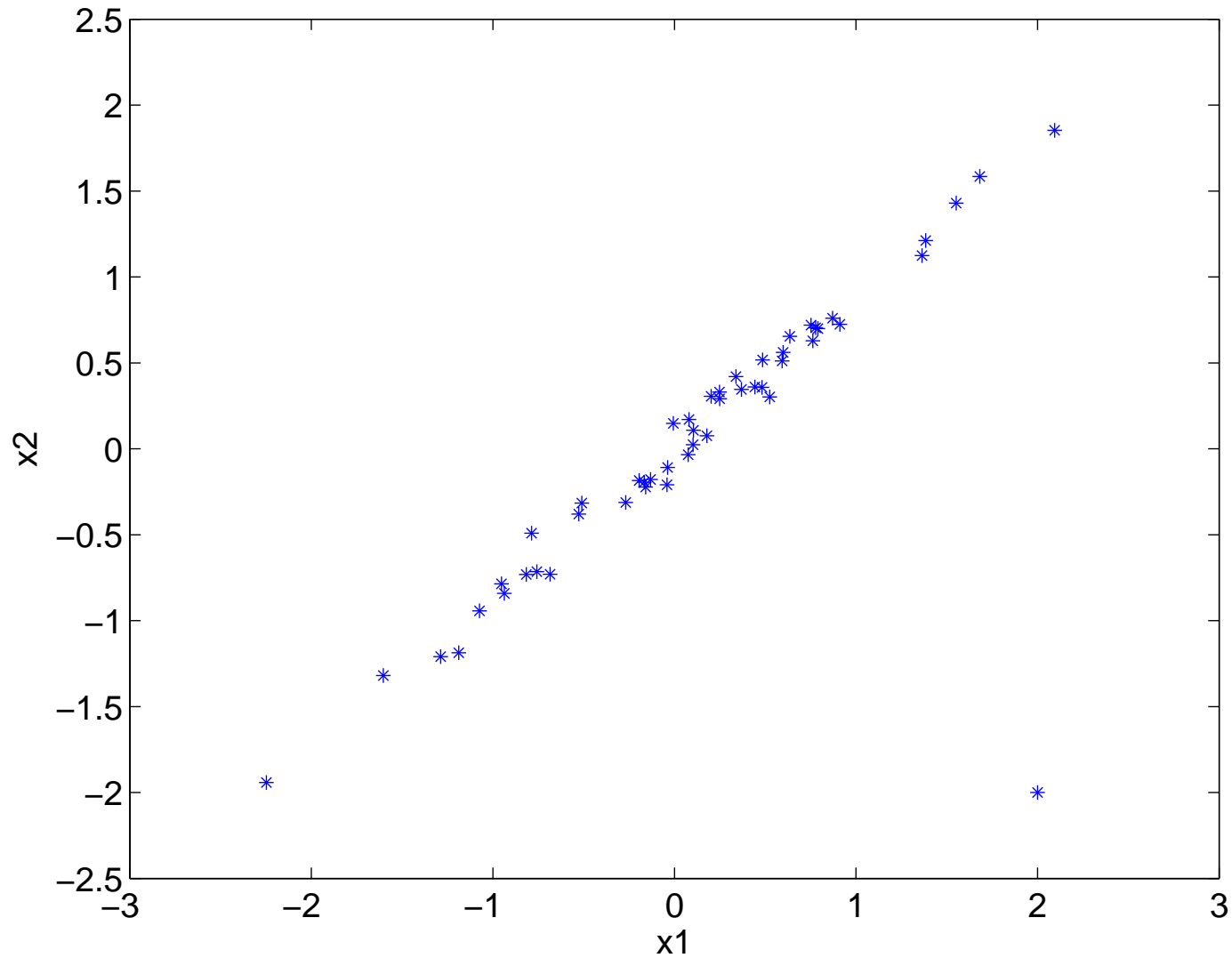
$$\text{Cook's D} = \frac{\sum_{j=1}^n \{\hat{Y}_j - \hat{Y}_j(-i)\}^2}{(p+1)s^2}.$$



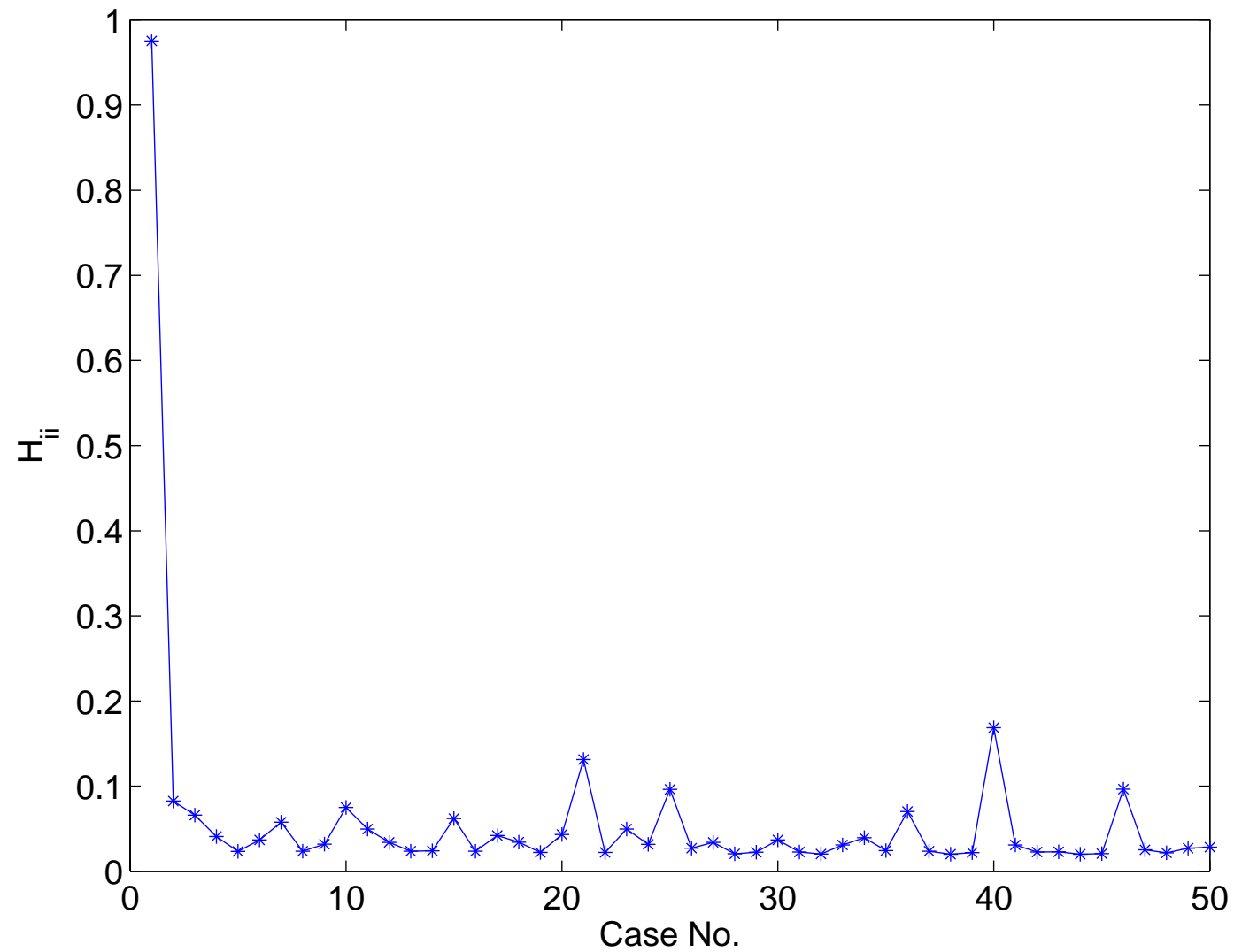
Linear Regression Example



Weekly Interest Rates Example

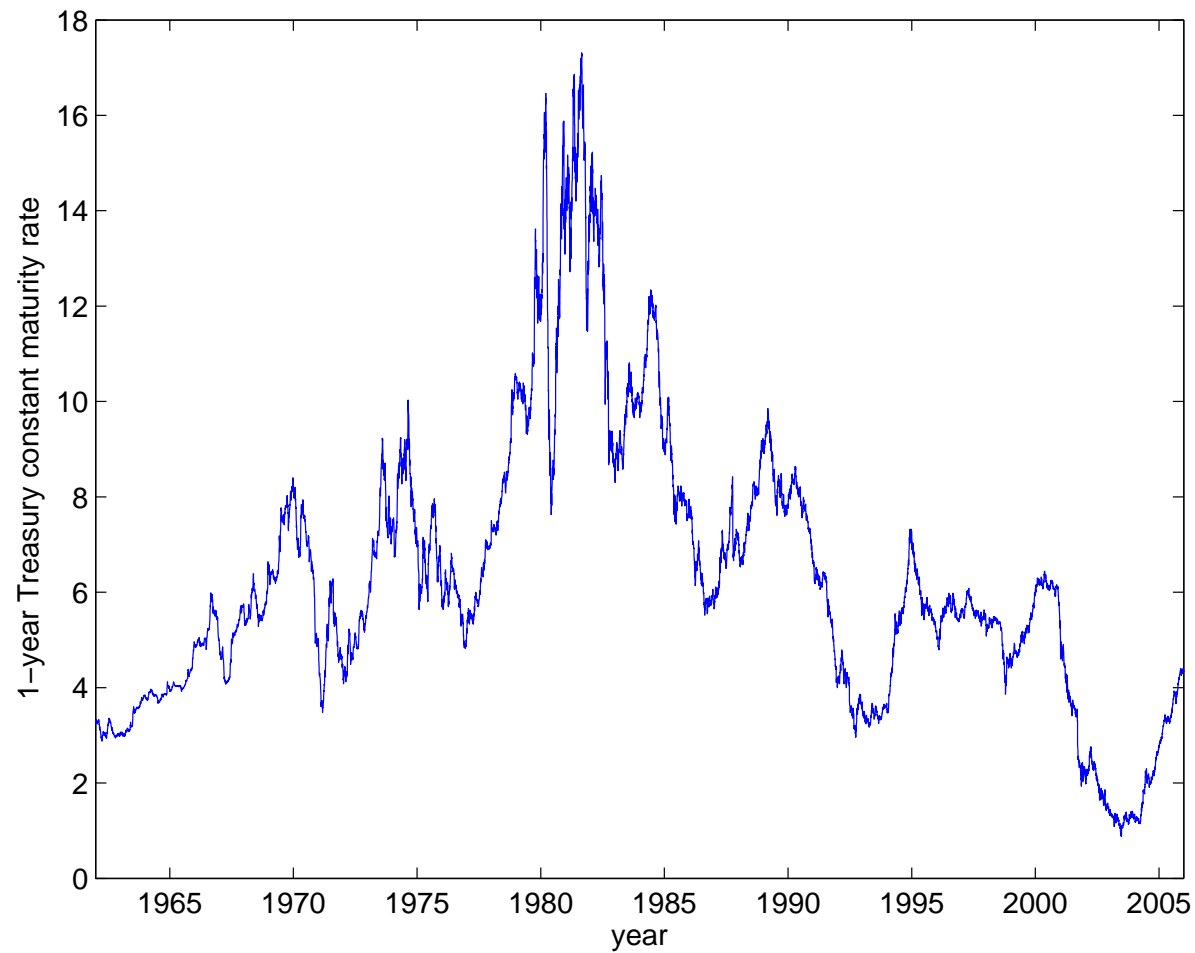


Example of leverage in two dimensions

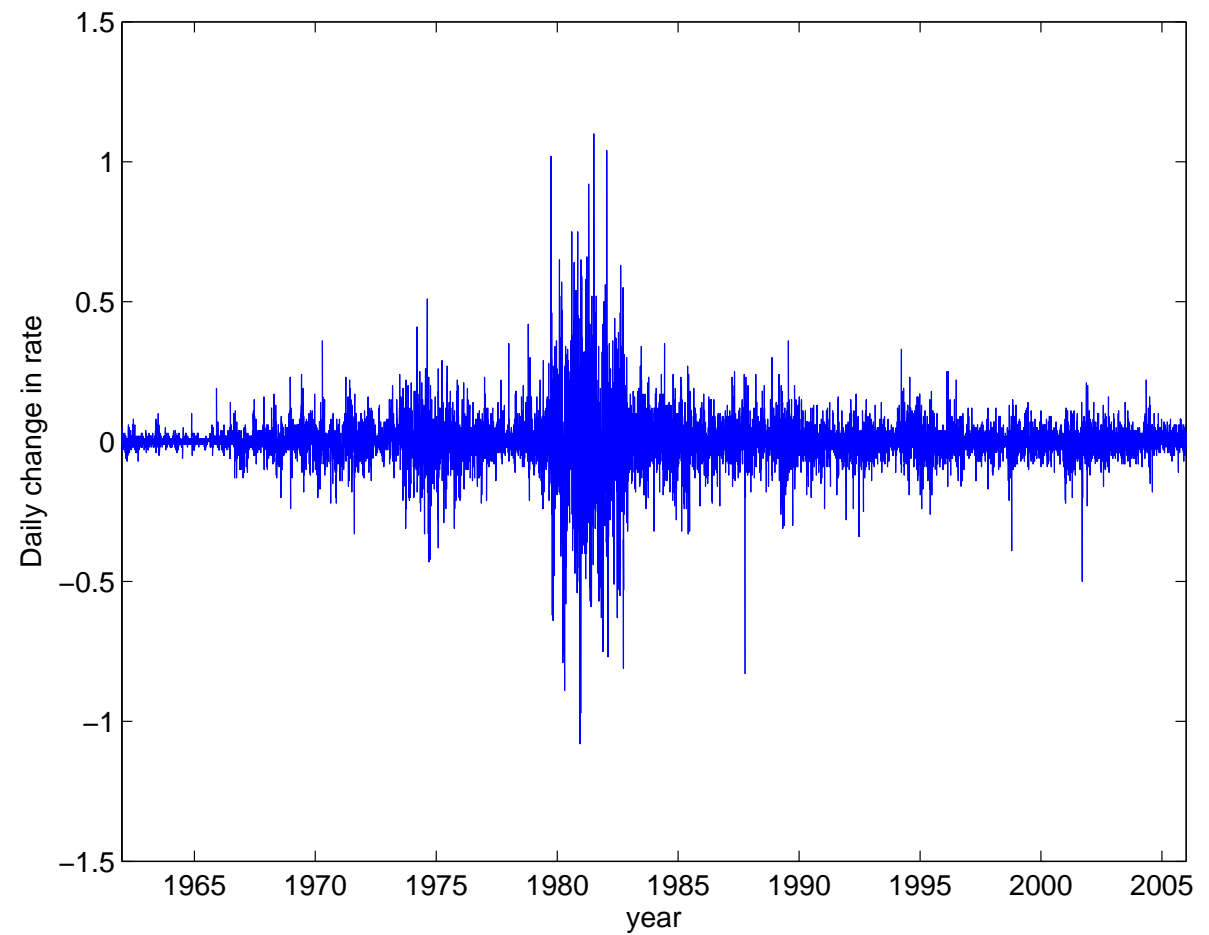


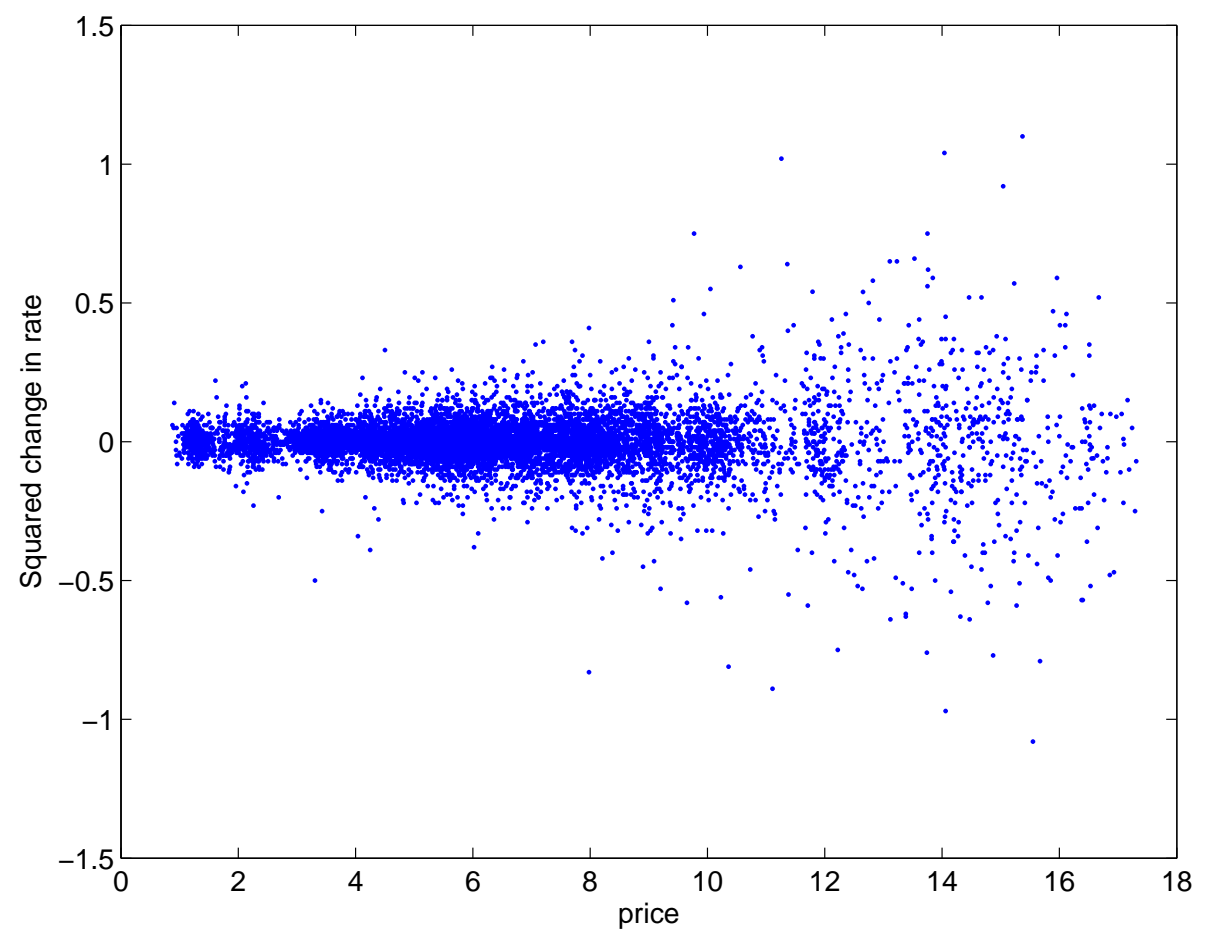
Example of leverage in two dimensions

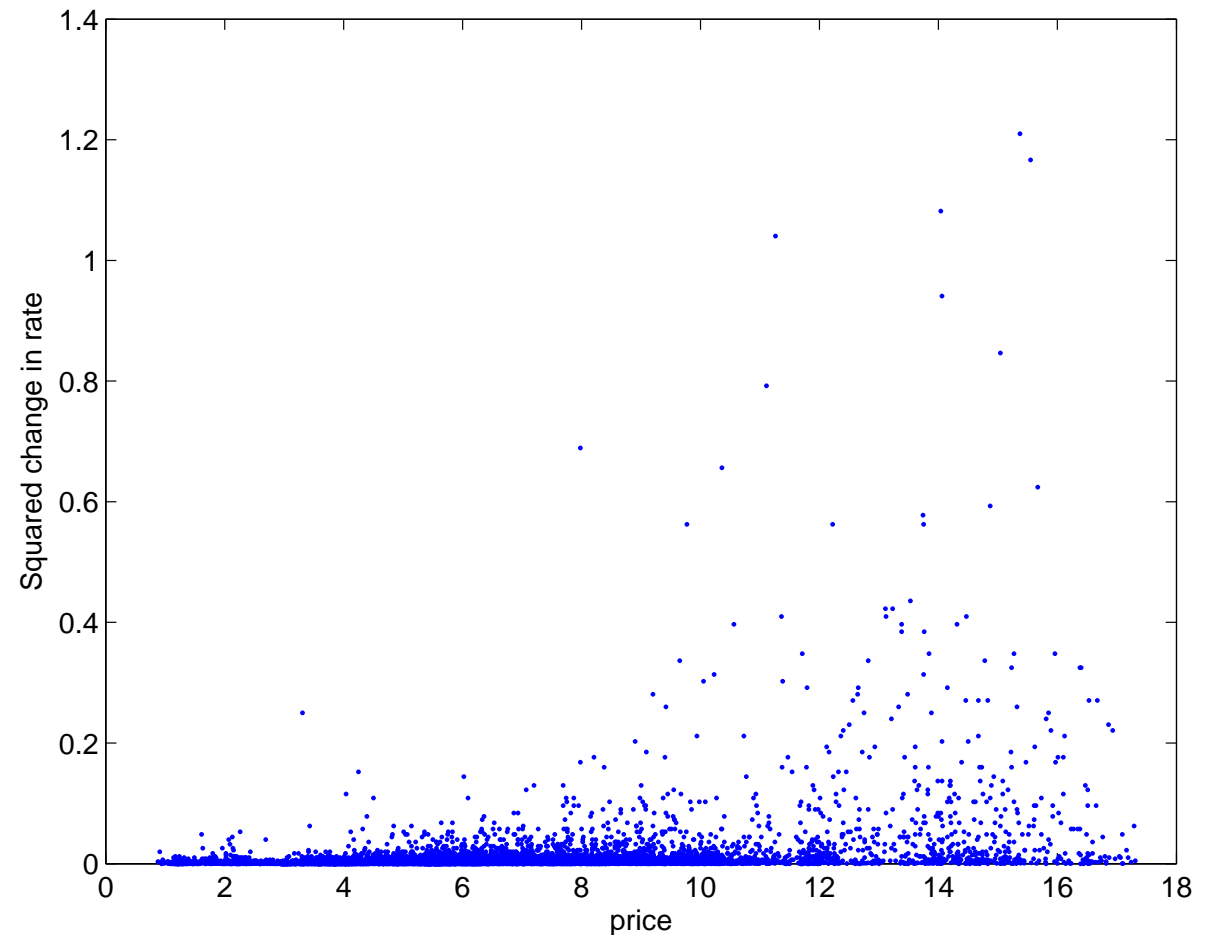
Modeling Interest Rate Volatility



Source: FRED







Power-of-mean model:

- R_t = rate at time t

$$\Delta(R_t) = R_t - R_{t-1}$$

- From plots:

$$E\{\Delta(R_t)\} \approx 0$$

and

$$\text{Var}\{\Delta(R_t)\} \approx \beta_0 R_{t-1}^{\beta_1}$$

- Suggest model

$$\{\Delta(R_t)\}^2 = \beta_0 R_{t-1}^{\beta_1} + \text{noise}$$

```
options linesize 72 ;  
data OneYrTreasury ;  
infile 'C:\Documents and Settings\David Ruppert\My Documents\talks\TempleStatFin\price_data.  
input prices delta_prices ;  
delta_prices2 = delta_prices**2;  
run ;  
proc nlin ;  
parm beta0 = .0001 to .1 by .01 beta1 = 0 to 4 by .25 ;  
model delta_prices2 = beta0*prices**beta1 ;  
run ;
```

Note: $\text{prices} = R_{t-1}$ and $\text{delta_prices} = R_t - R_{t-1}$

The NLIN Procedure
Dependent Variable delta_prices2
Method: Gauss-Newton

Iterative Phase				Sum of
Iter	beta0	beta1		Squares
0	0.000100	2.2500		16.2995
1	0.000069	2.3929		16.2746
2	0.000053	2.4994		16.2373
3	0.000035	2.6647		16.2181
4	0.000022	2.8533		16.1830
5	0.000011	3.1487		16.1165
6	0.000013	3.1657		15.9323
7	0.000014	3.1602		15.9316
8	0.000013	3.1611		15.9316
9	0.000013	3.1610		15.9316

NOTE: Convergence criterion met.

Parameter	Estimate	Approx	Approximate 95% Confidence Limits	
		Std Error		
beta0	0.000013	3.288E-6	7.045E-6	0.000020
beta1	3.1610	0.0936	2.9776	3.3444

Approximate Correlation Matrix

	beta0	beta1
beta0	1.0000000	-0.9960570
beta1	-0.9960570	1.0000000

How do we define a “residual”?

Recall model:

$$\text{Var}\{\Delta(R_t)\} = E\{(\Delta(R_t))^2\} = \beta_0 R_t^{\beta_1}$$

(Assuming $E\{\Delta(R_t)\} = 0$.)

First attempt:

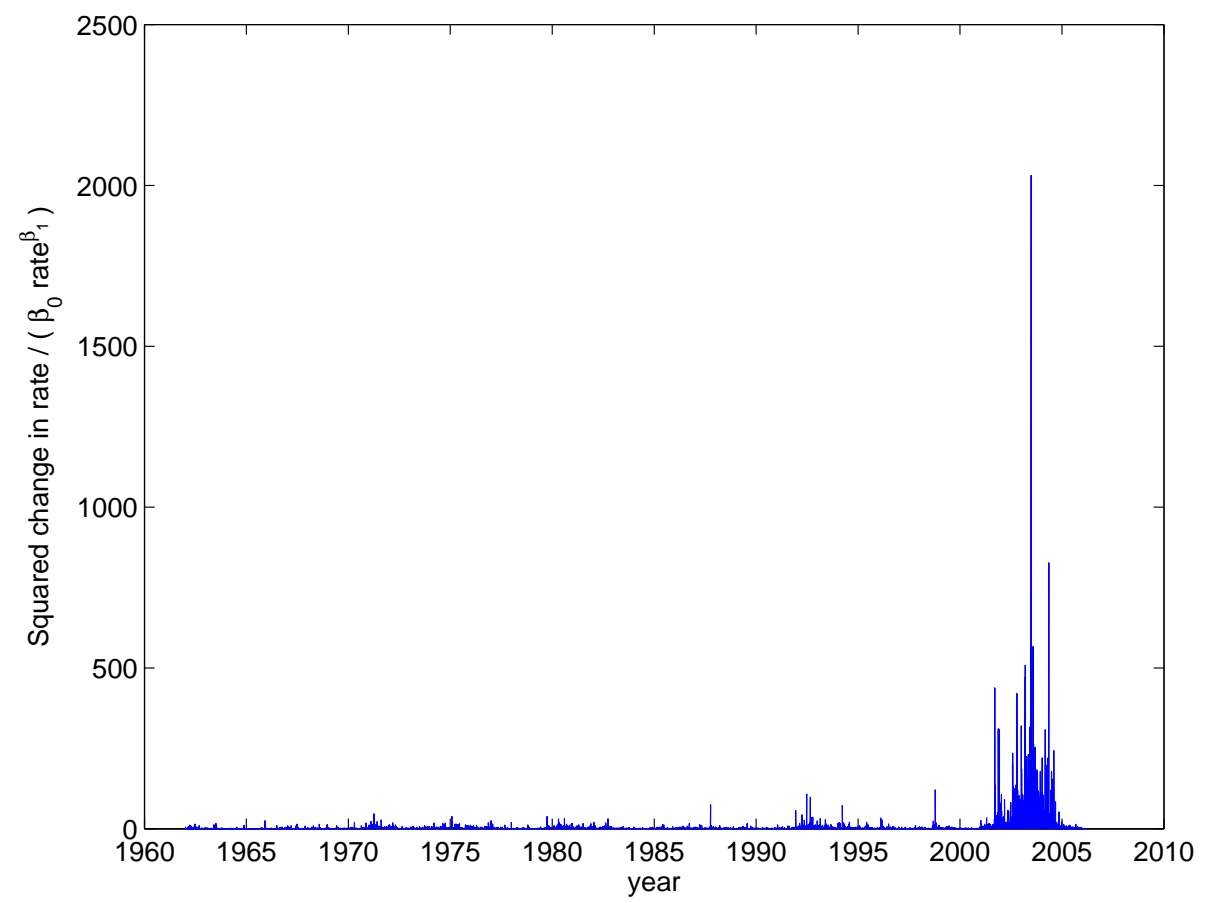
$$\{\Delta(R_t)\}^2 - \beta_0 R_t^{\beta_1}$$

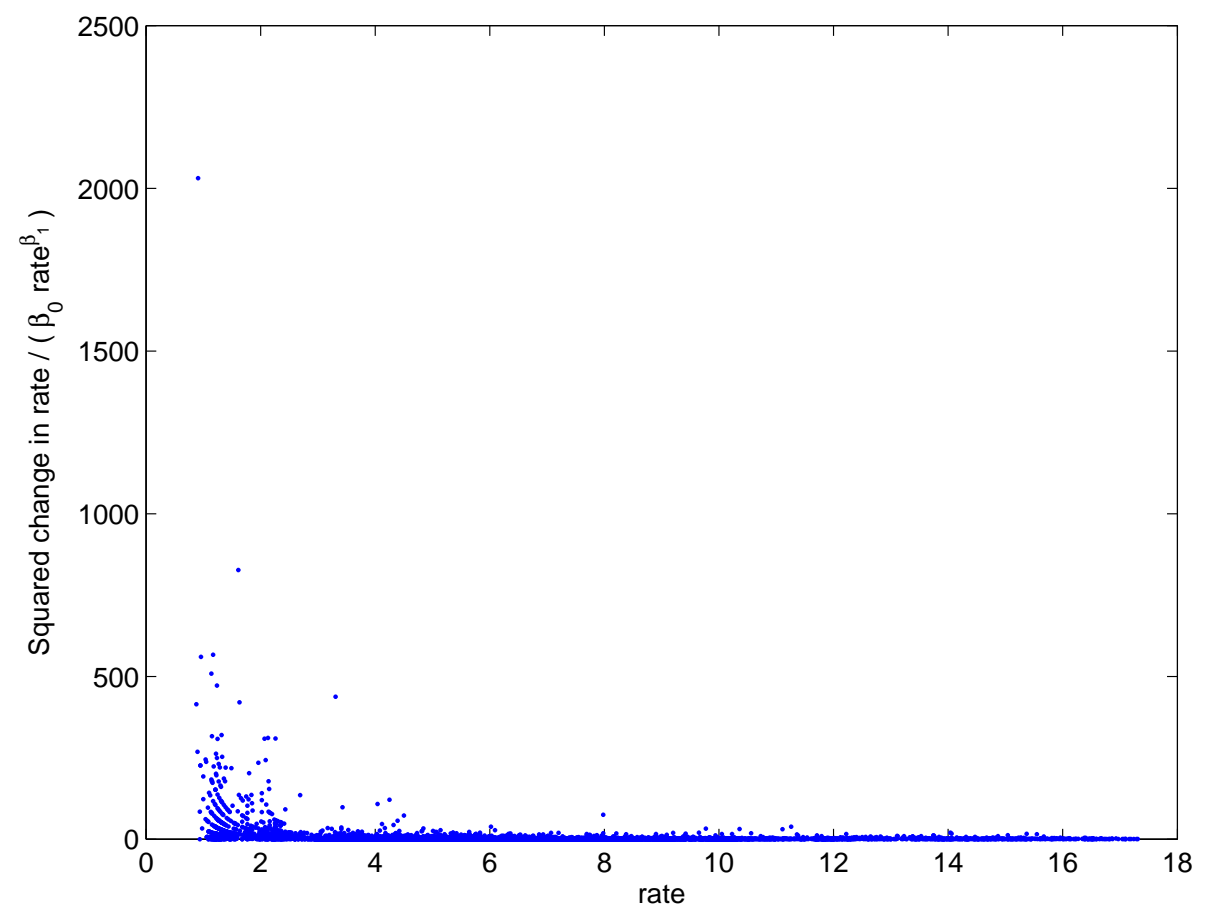
- This does NOT work.
- Has mean zero but not a constant variance

Second attempt:

$$\frac{\{\Delta(R_t)\}^2}{\beta_0 R_t^{\beta_1}}$$

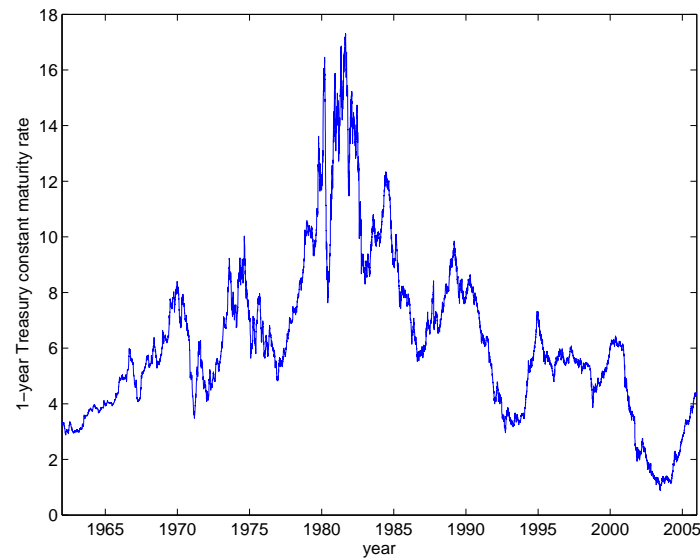
- This DOES work.
- Has constant mean and variance (if model is correctly specified)
- We can check model by plotting this variable against rate and year – we want to see NO pattern





The model has a misspecification problem.

- It does not fit the data well for the period around 2003 when interest rates were at their lows values during the 1962 to 2006 period.





<http://www.springer.com/978-0-387-20270-9>

Statistics and Finance

An Introduction

Ruppert, D.

2004, XXII, 474 p., Hardcover

ISBN: 978-0-387-20270-9