Stat 107: Introduction to Business and Financial Statistics

Class 19: Multiple Regression and Logistic Regression

Important Dates: Final Exam

- December 13th, 2-5pm
- 3 hour take home final similar to midterm

The Class Project-2 to 5 people

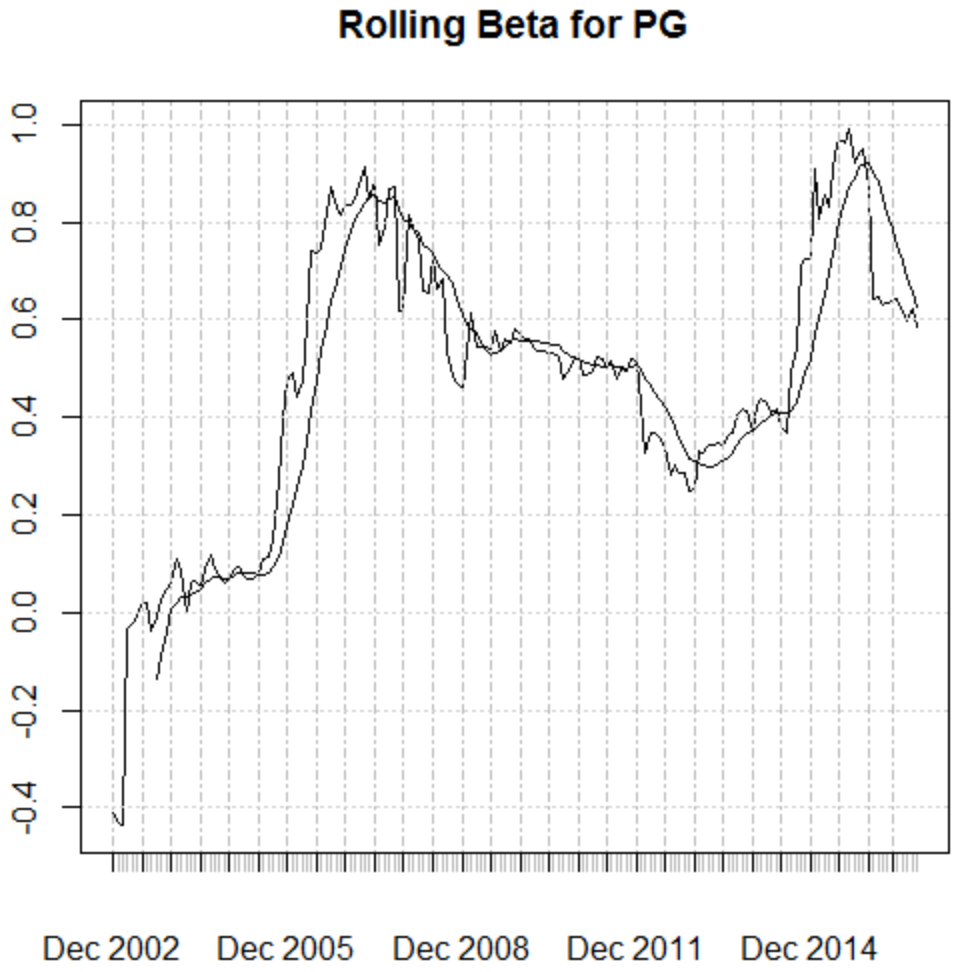
- Important Dates
  - November 16th: project proposal due (ungraded)
  - December 6th: project due (midnight)
- There are least two directions one can take for the final project:
  - Reproducing an existing study in the literature
  - Simulating or Reporting on a particular stock trading strategy
  - (Your own idea at discretion of Instructor)

Rolling Betas

- Beta depends on the time period uses; there is a packaged called `roll` that allows one to easily calculate Beta over different time periods.

```
library(quantmod)
library(roll)
getSymbols("SPY",from="2000-01-01")
getSymbols("IBM",from="2000-01-01")
getSymbols("PG",from="2000-01-01")
spyret=monthlyReturn(Ad(SPY))
ibmret=monthlyReturn(Ad(IBM))
pgret=monthlyReturn(Ad(PG))
betas=roll_lm(spyret,pgret,width=36)
```

SMA as a timing indicator?



Style Analysis

- From Investopedia

Style Analysis

Like

What Does *Style Analysis* Mean?

The process of determining what type of investment behavior an investor or money manager employs when making investment decisions. Virtually all investors subscribe to a form of investment philosophy, and a prudent analysis of a money manager's style needs to be performed before an investor can determine whether the manager will be good fit for his or her personal investment goals and preferences.

Investopedia explains *Style Analysis*

There is virtually an unlimited number of investment styles; however, some of the most common types of investment styles are categorized as growth investing, value investing, large cap investing, small cap investing and active trading.

Some money managers change their investment styles over time, opting to go with one approach while it is working well and then switching to another when the old approach seems to be losing its luster.

As Dan writes…

- **An inaccurate classification system produces wrong signals, puts investor funds in the hands of those who are not necessarily best qualified to manage them, and ultimately allocates assets into projects that are not optimal.**
- **On a micro level, misclassification has implications for individual investors. For example, a fund that labels itself "income" but invests a large portion of its assets in small, growth-oriented stocks may have risk and return parameters that are inappropriate for a retired couple.**
- **The reliability of any classification system that the investing public relies on is of utmost importance.**

- **Further,** Mutual fund data vendors rely heavily on the funds themselves for classification information.

Returns Based Classification Method

- Style analysis is a more recent returns-based approach to measuring the performance of an investment fund.
- Sharpe (1992) pioneered it in the early nineties by developing an ‘asset class factor model’ to distinguish the performance of different funds with respect to ‘style’ and ‘selection’.
- Style analysis can also be viewed as reverse engineering the asset mix in a portfolio.

Returns Based Classification Method

- The objective of style analysis is to construct a benchmark portfolio, from a set of known indices (for which returns are available), against which to compare the performance of an investment fund's actively-managed portfolio.
- Ideally, the indices should reflect activity in different asset classes, they should be mutually exclusive and exhaustive, and their assets publicly quoted so that they can be tracked 'passively'.
- For example, Sharpe used 12 indices to cover the range of investment options available to US funds. The indices were chosen to have as little overlap as possible.

Style Analysis

- Consider the 6 most widely used categories of equity mutual funds:
  - aggressive growth (f_1)
  - growth (f_2)
  - growth-income (f_3)
  - income (f_4)
  - International (f_5)
  - small capitalization (f_6)

The Model

- We fit the model

$$r_t = b_1 f_{1t} + b_2 f_{2t} + b_3 f_{3t} + b_4 f_{4t} + b_5 f_{5t} + b_6 f_{6t} + e_t$$

- Sharpe suggests finding the weights b_i's to minimize the tracking error:

$$\sum_t \left| r_t - b_1 f_{1t} + b_2 f_{2t} + b_3 f_{3t} + b_4 f_{4t} + b_5 f_{5t} + b_6 f_{6t} \right|$$

- Since they make up a portfolio, the weights are constrained to sum to 100% with the individual weights normally lying between 0% and 100%. (However this latter condition can be modified for funds that are allowed to hold assets short.)

Style Weights

- The weights determined by optimization are called the style weights and, when combined with the indices, form the benchmark portfolio.
- We say that the fund with optimized style weights for the different indices is of the same style as the investment fund.
- The weights can be found using SOLVER in Excel.

Example

- Consider the following indices
  - JKD : Large Core Index Fund
  - JKG : Mid Cap Core Index Fund
  - JKJ : Small Cap Core Index Fund
  - EFA : International Index Fund
  - AGG : Bond Fund

FMAGX Style Analysis

- What we did in R

```
getSymbols("JKD")
getSymbols("JKG")
getSymbols("JKJ")
getSymbols("EFA")
getSymbols("AGG")
getSymbols("FMAGX")
jkd=monthlyReturn(Ad(JKD))
jkg=monthlyReturn(Ad(JKG))
jkj=monthlyReturn(Ad(JKJ))
efa=monthlyReturn(Ad(EFA))
agg=monthlyReturn(Ad(AGG))
fmagx=monthlyReturn(Ad(FMAGX))

fit=lm(fmgax~-1+jkd+jkg+jkj+efa+agg)  (no intercept)
```

The FMAGX Output

- summary(fit)

```
Coefficients:
    Estimate Std. Error t value Pr(>|t|)
jkd  0.14573    0.21408   0.681  0.49947
jkg  0.76318    0.26326   2.899  0.00572 **
jkj -0.13730    0.20966  -0.655  0.51580
efa  0.27465    0.12828   2.141  0.03761 *
agg -0.03724    0.23619  -0.158  0.87542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘
’ 1

Residual standard error: 0.02168 on 46 degrees of freedom
Multiple R-squared: 0.9124,     Adjusted R-squared: 0.9029
```

Rolling the Estimates

- To ensure that the style of the fund doesn't change over time, analysts will roll the data.
- That means they will calculate the style weights over rolling time periods (12-36 month windows), to see if the style weights are drastically changing over time.

Constraining the Weights

- Very easy to read paper on style analysis (on web site)

Style Analysis with Solver

- See this week's homework
- Spreadsheet styleanalysis.xls

Style Weights	
JKD	0
JKG	0.027419
JKJ	0
EFA	0
AGG	0.175302
SHY	0.797279
constraint	1
sum of deviations	0.096746

More than 1 factor?

- The market model is a one factor model: The only determinant of expected returns is the systematic risk of the market. This is the only factor.
- What if there are multiple factors that determine returns?
- Multifactor Models: Allow for multiple sources of risk, that is multiple risk factors.

Multifactor Models

- Use other factors in addition to market returns:
  - Examples include industrial production, expected inflation etc.
  - Estimate a beta or *factor loading* for each factor using multiple regression

Example: Multifactor Model Equation

$$R_i = \beta_0 + \beta_1 R_M + \beta_2 R_{GDP} + \beta_3 R_{IR} + \varepsilon_i$$

R_i = Return for security i

β_2= Factor sensitivity for GDP

β_3 = Factor sensitivity for Interest Rate

ε_i = Firm specific events

Multifactor Models

- The Market Model say that a single factor, **B**eta, determines the return between a portfolio and the market as a whole.
- Suppose however there are other factors that are important for determining portfolio returns.
- The inclusion of additional factors would allow the model to improve the model`s fit of the data.
- The best known approach is the three factor model developed by Gene Fama and Ken French.

Fama-French 3-Factor Model

- They added these two factors to a standard market model

$$R_{i,t} = \alpha_i + \beta_{i1}(R_{m,t}) + \beta_{i2}SMB_t + \beta_{i3}HML_t + \varepsilon_{i,t}$$

SMB = “small [market capitalization] minus big”

"Size" This is the return of small stocks minus that of large stocks. When small stocks do well relative to large stocks this will be positive, and when they do worse than large stocks, this will be negative.

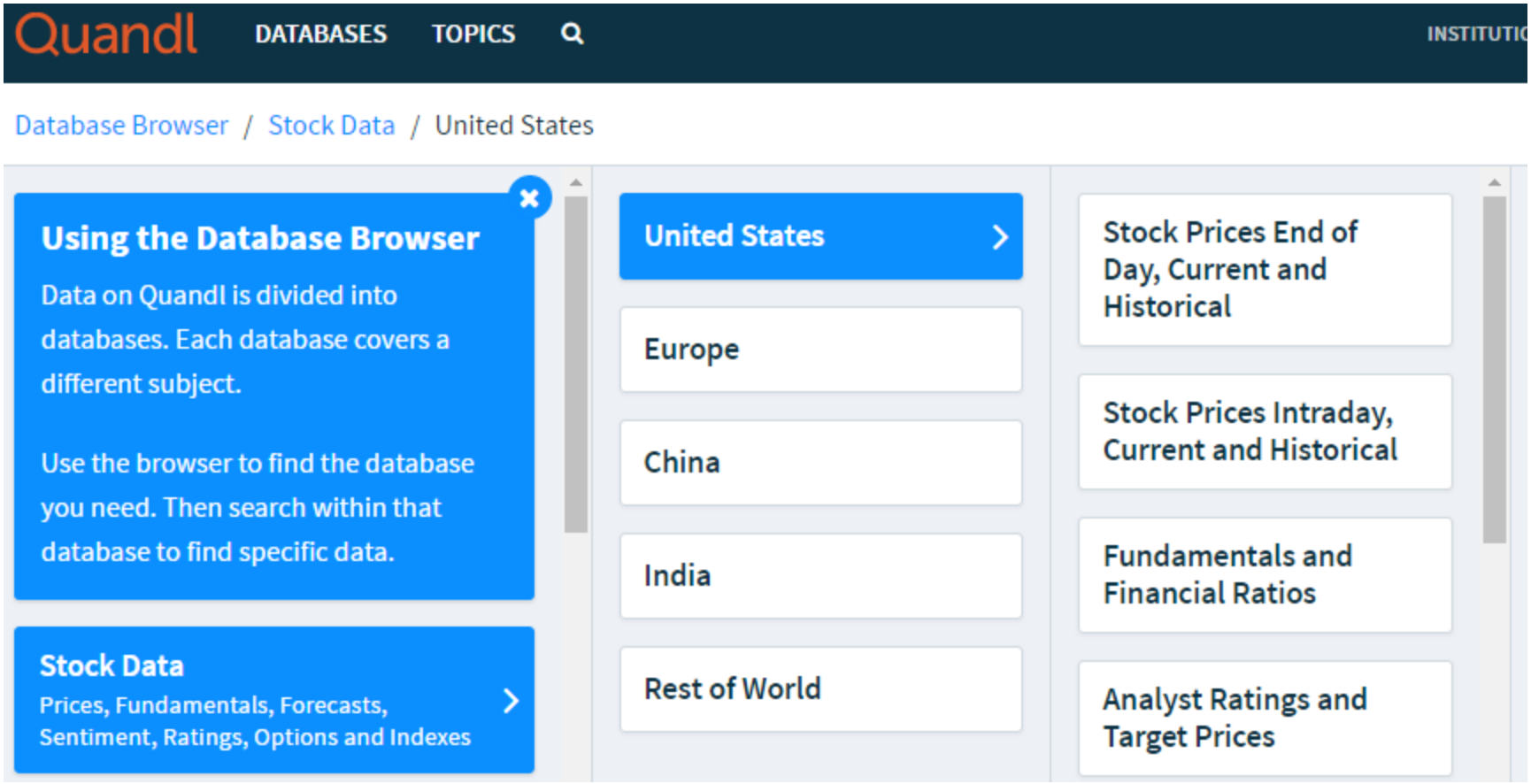
HML = “high [book/price] minus low”

"Value" This is the return of value stocks minus growth stocks, which can likewise be positive or negative.

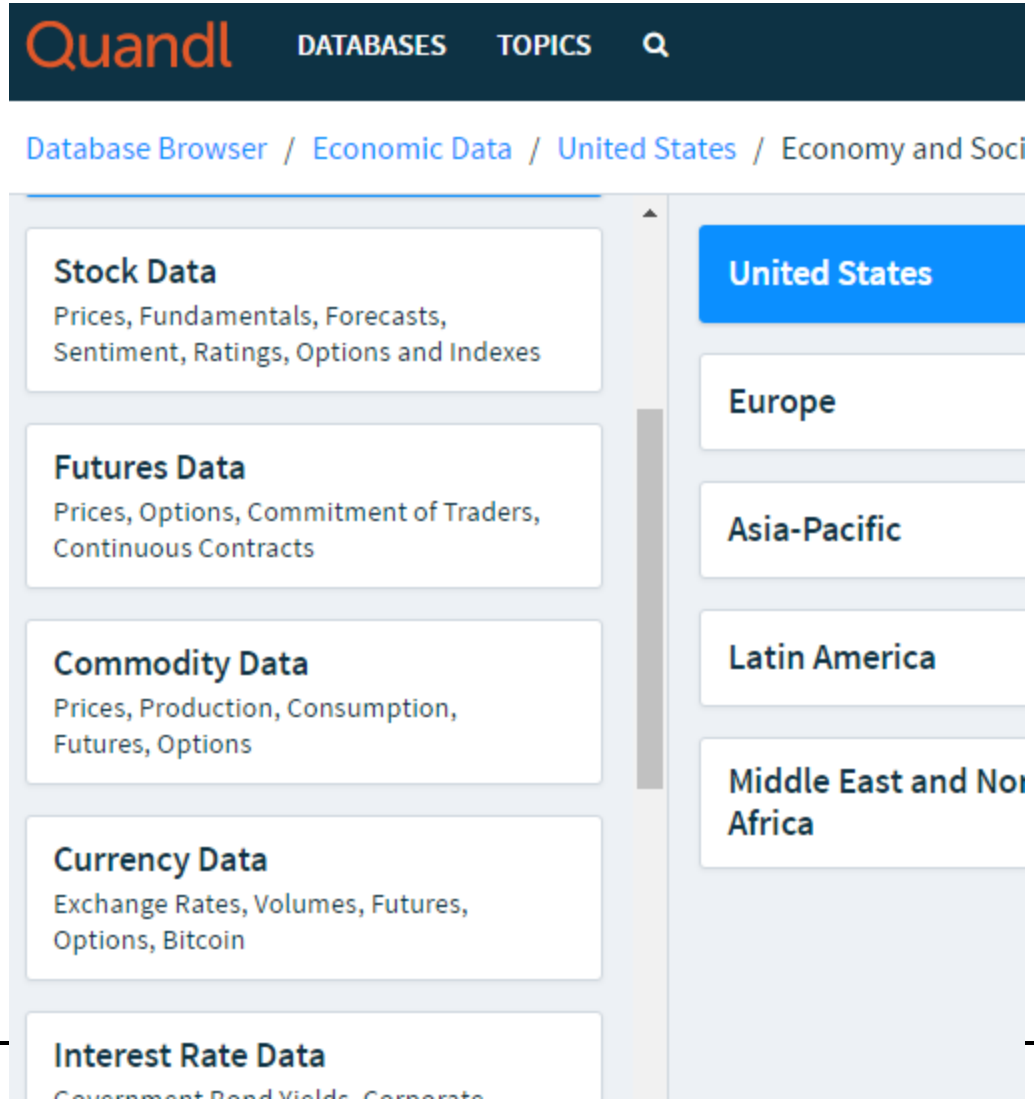
http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/Data_Library/f-f_factors.html

The Fama-French Three Factor model explains over 90% of stock returns.

Getting the Data: Package Quandle



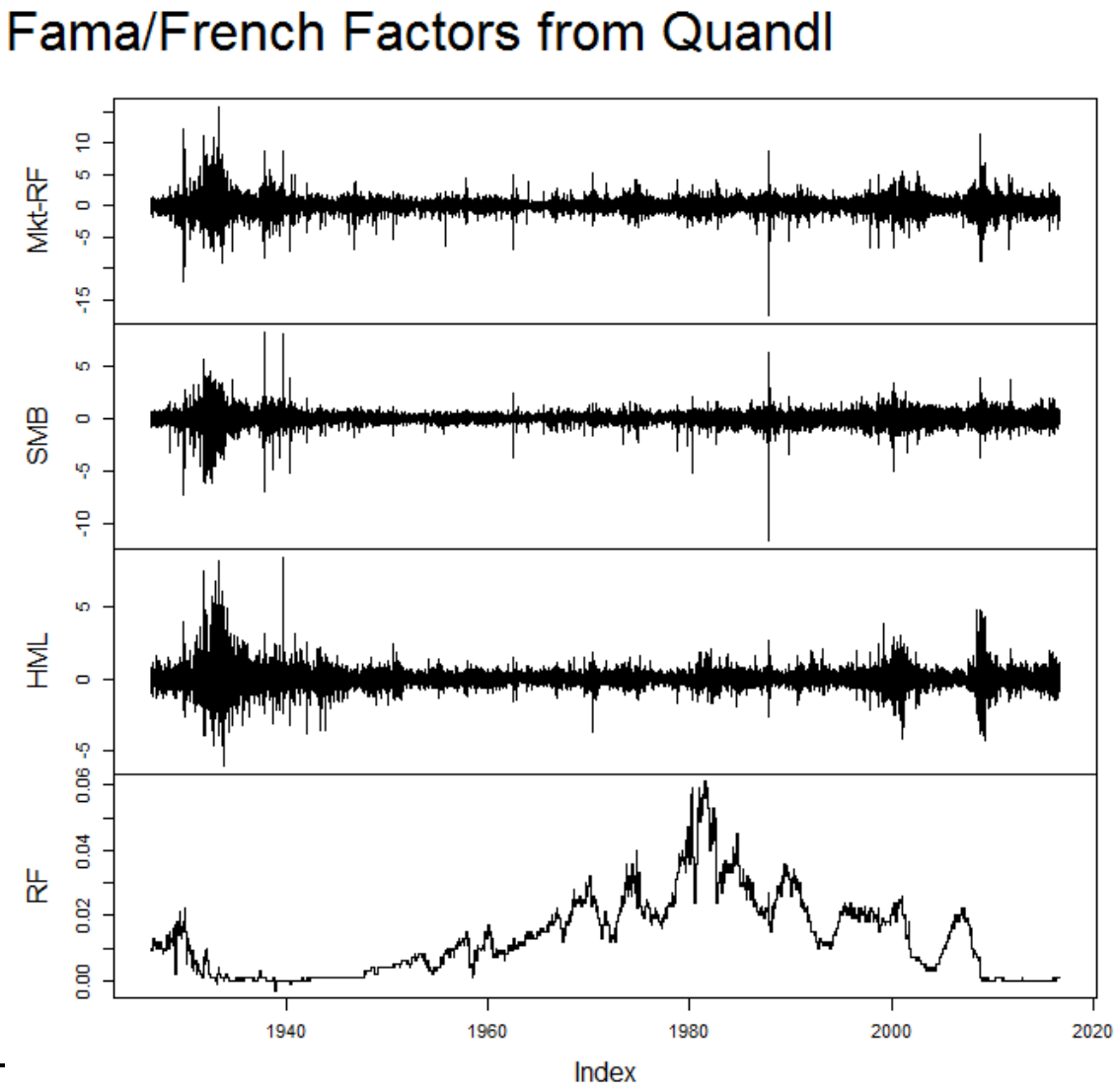
Quandl: Free and Premium Data



Can get Data from French's Site

```
# use Quandl Kenneth French Fama/French factors
# http://www.quandl.com/KFRENCH/FACTORS_D

library(Quandl)
library(quantmod)

f <- Quandl(
  "KFRENCH/FACTORS_D"
)

f <- as.xts(f[,-1],order.by=f[,1])

plot.zoo( f, main = NA )
mtext(
  text = "Fama/French Factors from Quandl"
  , adj = 0
  , outer = T
  , line = -2
  , cex = 2
)
```

Fama French Data



Back to Regression-Why Least Squares?

- As a quick aside, (least squares) regression fits a line by solving the equation

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

`lm(y~x)`

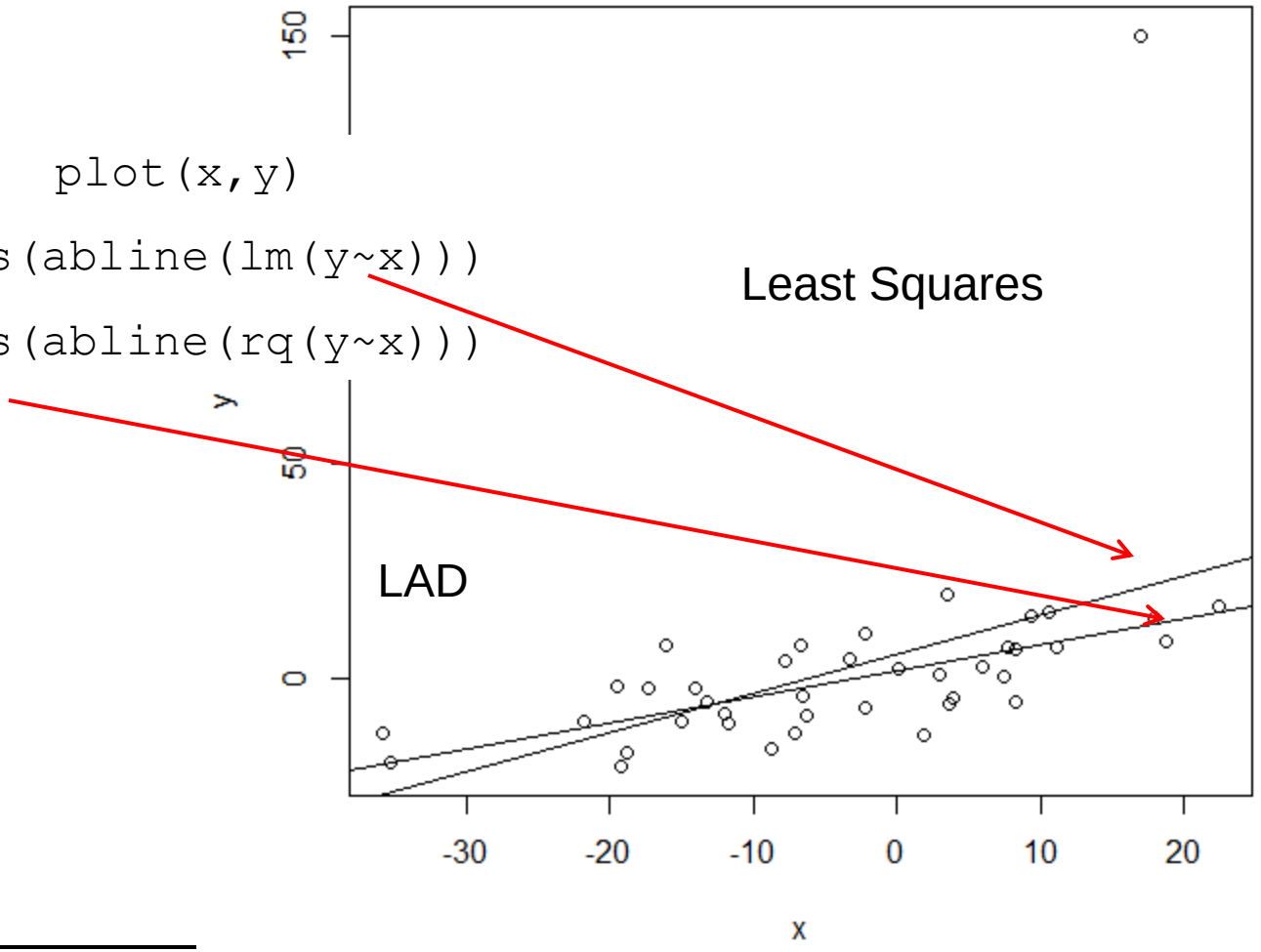
- Why not solve the following equation? This is called LAD (least absolute deviation) regression.

$$\min_{b_0,b_1} \sum_{i=1}^{n} | Y_i - b_0 - b_1 X_i |$$

`rq(y~x)`
`[in package quantreg]`

Least Squares versus MAD

```
plot(x,y)
lines(abline(lm(y~x)))
lines(abline(rq(y~x)))
```

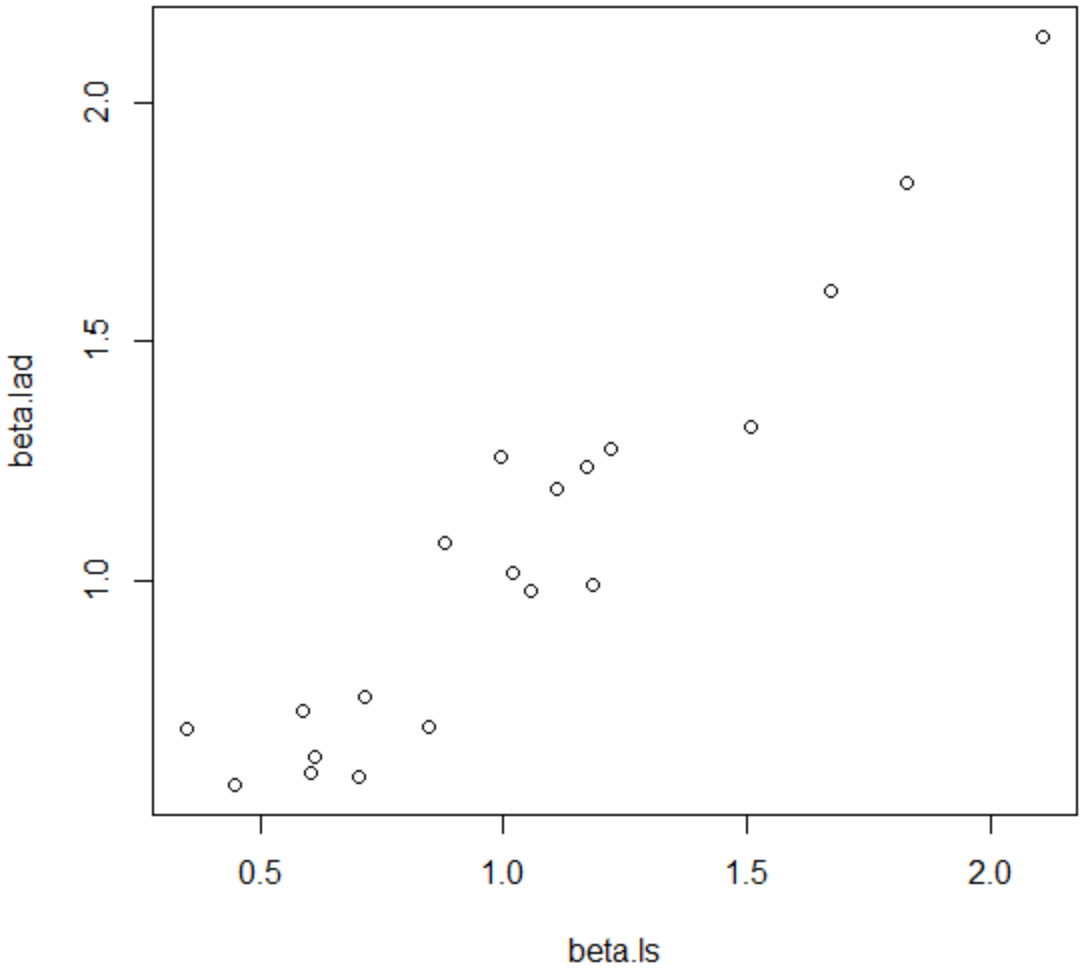


LAD Betas

```
asset.names=c("ATVI", "ADBE", "AKAM", "ALTR", "AMZN",
"AMGN", "APOL", "AAPL", "AMAT", "ADSK", "ADP" , "BIDU",
"BBBY", "BIIB","BMC",  "BRCM" ,"CHRW" ,"CA" ,  "CELG",
                        "ANF")

                    n=length(asset.names)
                      getSymbols("SPY")
                  spy.ret=monthlyReturn(Ad(SPY))
                        beta.ls=1:n
                       beta.lad=1:n

                      for(i in 1:n){
          x=getSymbols(asset.names[i],auto.assign=FALSE)
                    x.ret=monthlyReturn(Ad(x))
               beta.ls[i]=coef(lm(x.ret~spy.ret))[2]
              beta.lad[i]=coef(rq(x.ret~spy.ret))[2]
                              }
```

Which ones are correct?

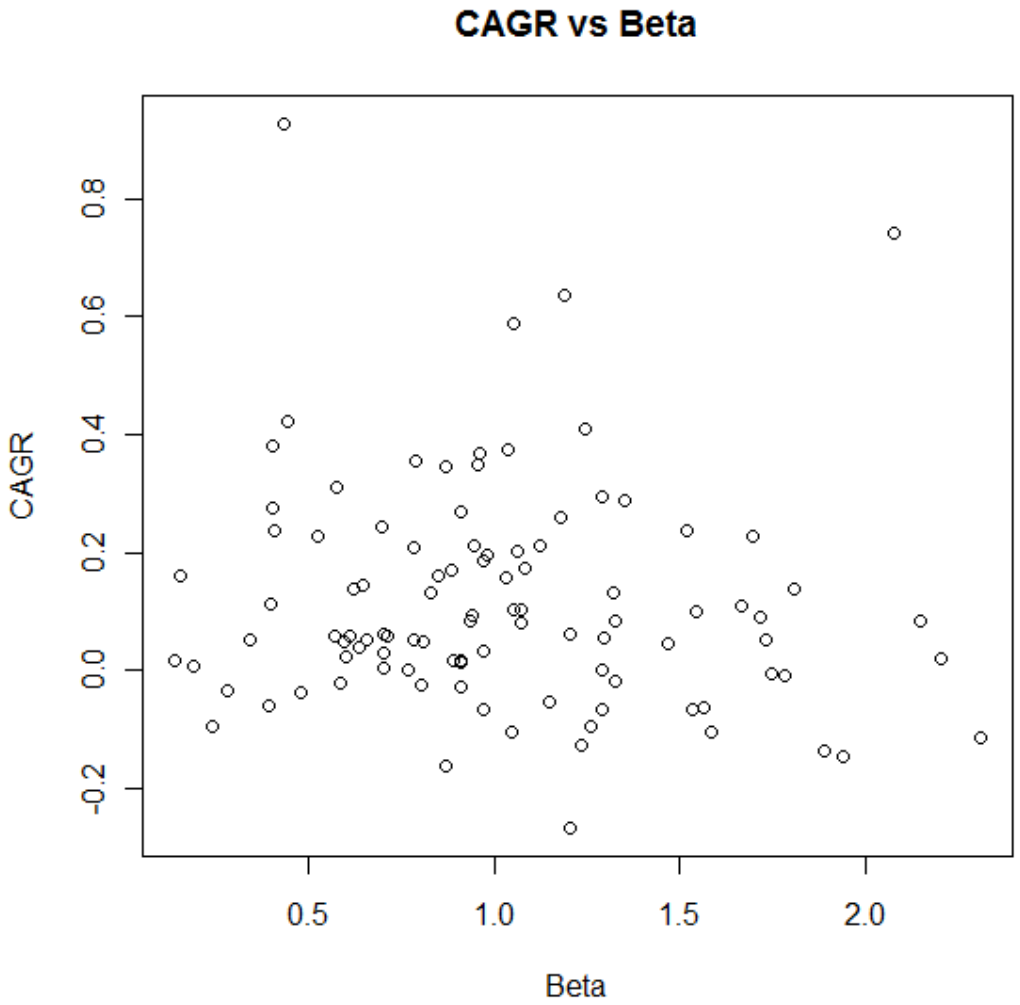
```
> cor(beta.ls,beta.lad)
[1] 0.9531647
```



CAGR and Beta

- According to the CAPM, a bigger Beta implies a bigger expected stock return.
- So what should a graph of CAGR (compounded annual return) versus Beta look like?
- Probably not what you expect.

Beta vs CAGR for Nasdaq 100 stocks



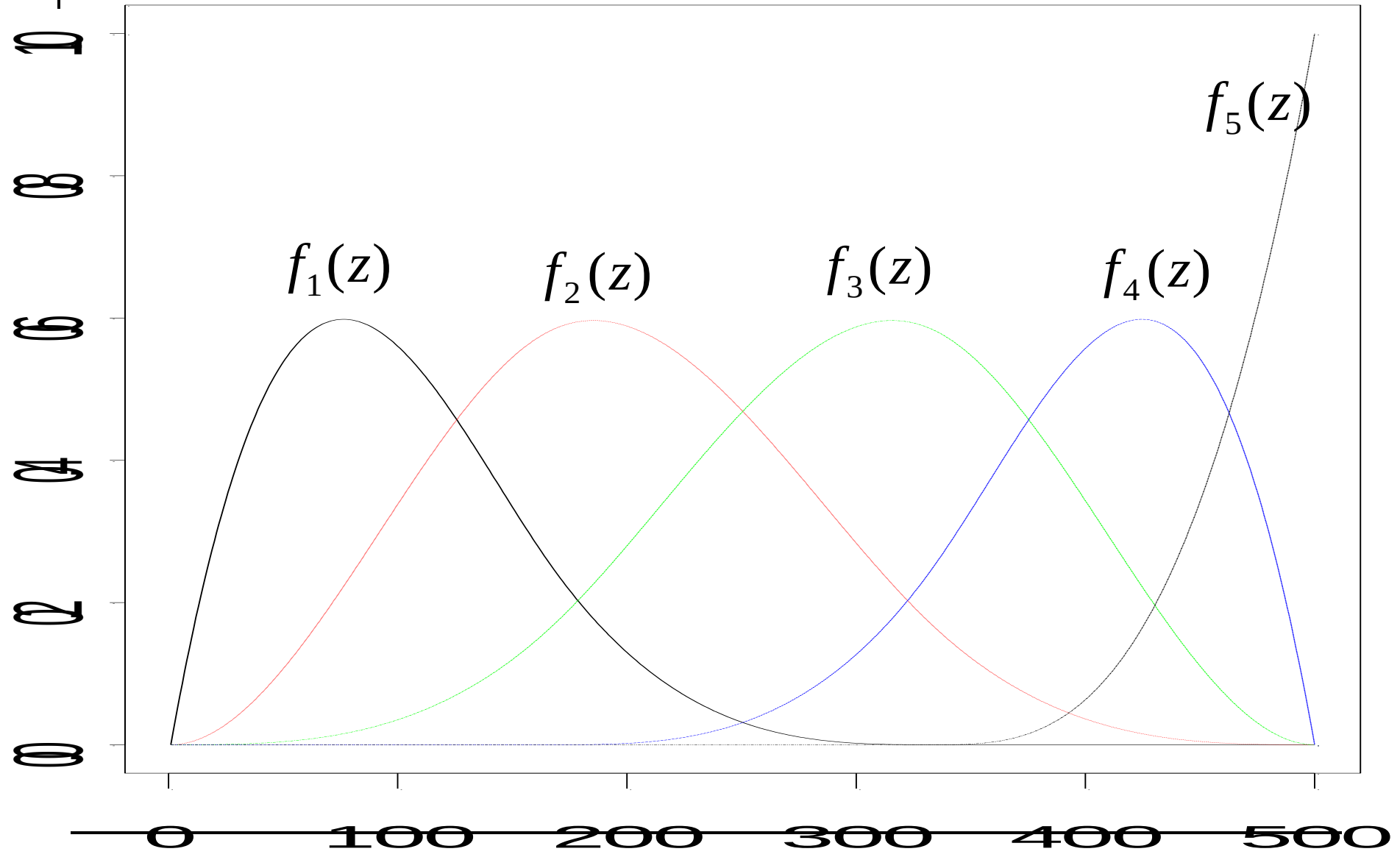
To be fair, we should see what this looks like during a bull market like 1996-2000, but it will look even worse.

Semi-Parametric Modeling

- Before we move to logistic regression, a quick look at some interesting functions built into R.
- We have already seen the idea of splines with density estimation; it can also be used to model non-linear relationships between x and y.

$$Y = f(z) + \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

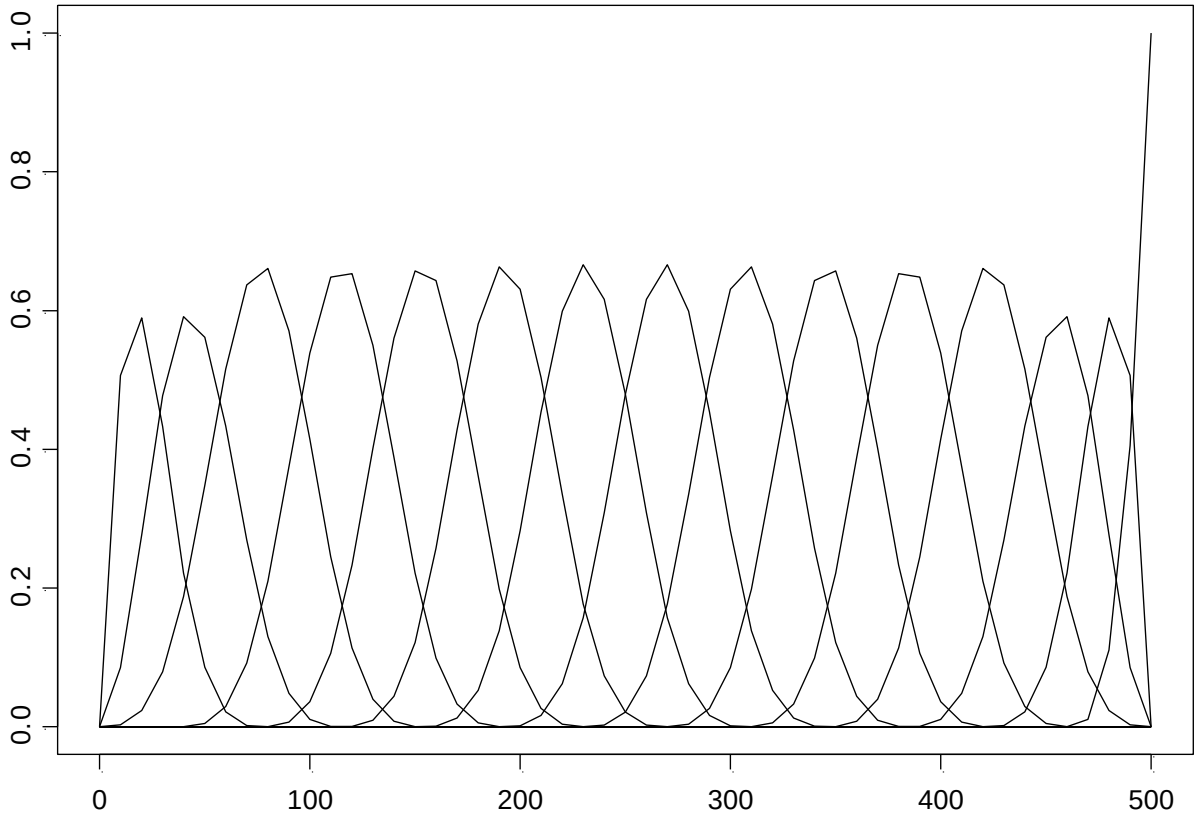
Cubic Spline Basis Functions



The plot on the last slide was made using the following code (library splines was loaded):

```
x=seq(0,500,10)

foo=bs(x,5)

matplot(x,foo,type="l")
```

There is almost no limit as to how many basis functions you can compute:



$$f(z) \approx \sum_{i=1}^{5} \alpha_i f_i(z)$$

$$Y = f(z) + \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

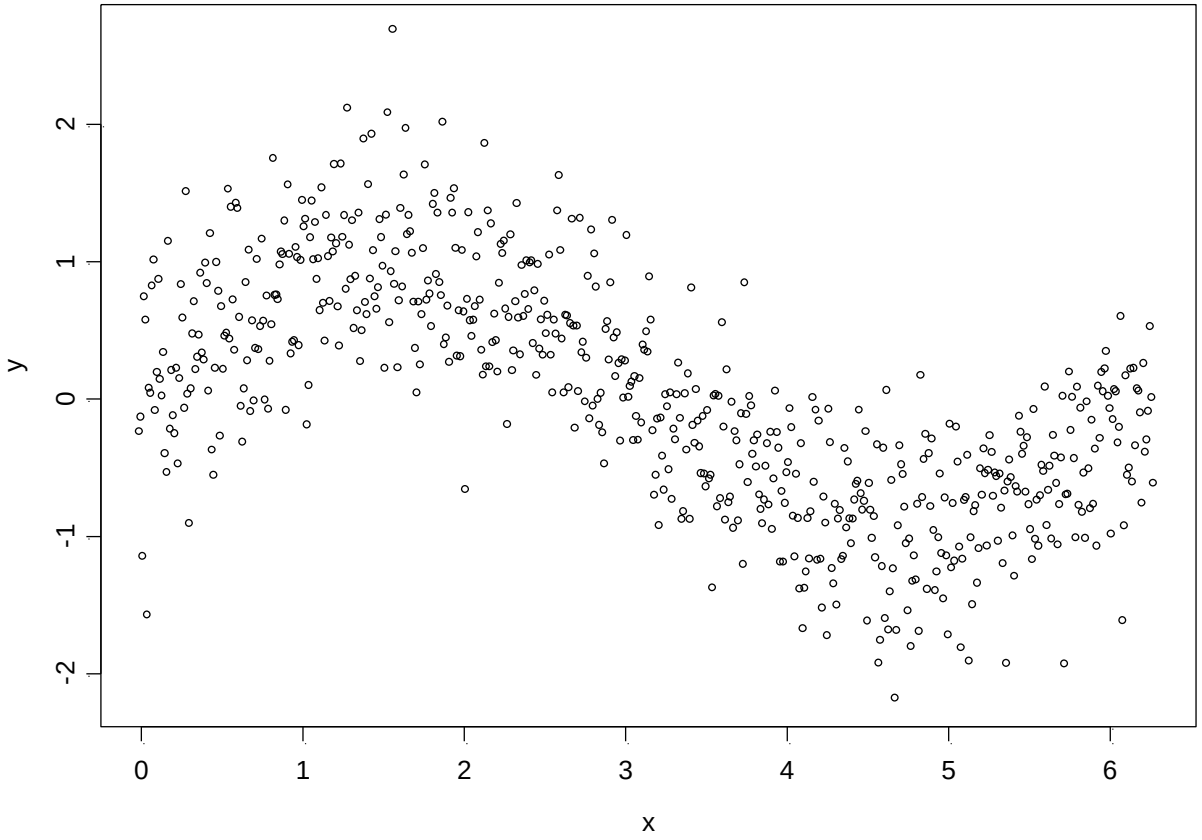
$$\Rightarrow$$

$$Y \approx \sum_{i=1}^{5} \alpha_i f_i(z) + \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

We use ordinary linear regression to maximize over all the unknown parameters.

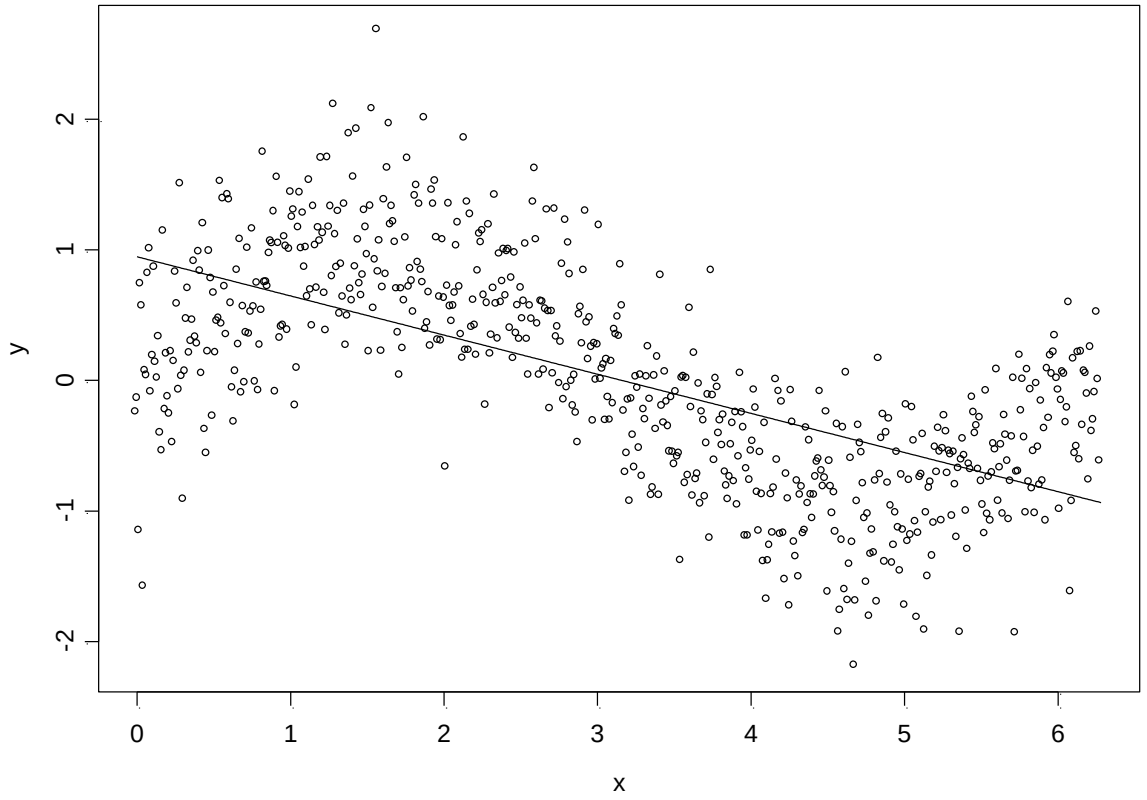
Example

```
> x=seq(0,2*pi,.01)
> length(x)
[1] 629
> y=sin(x)+rnorm(629,s=.5)
```



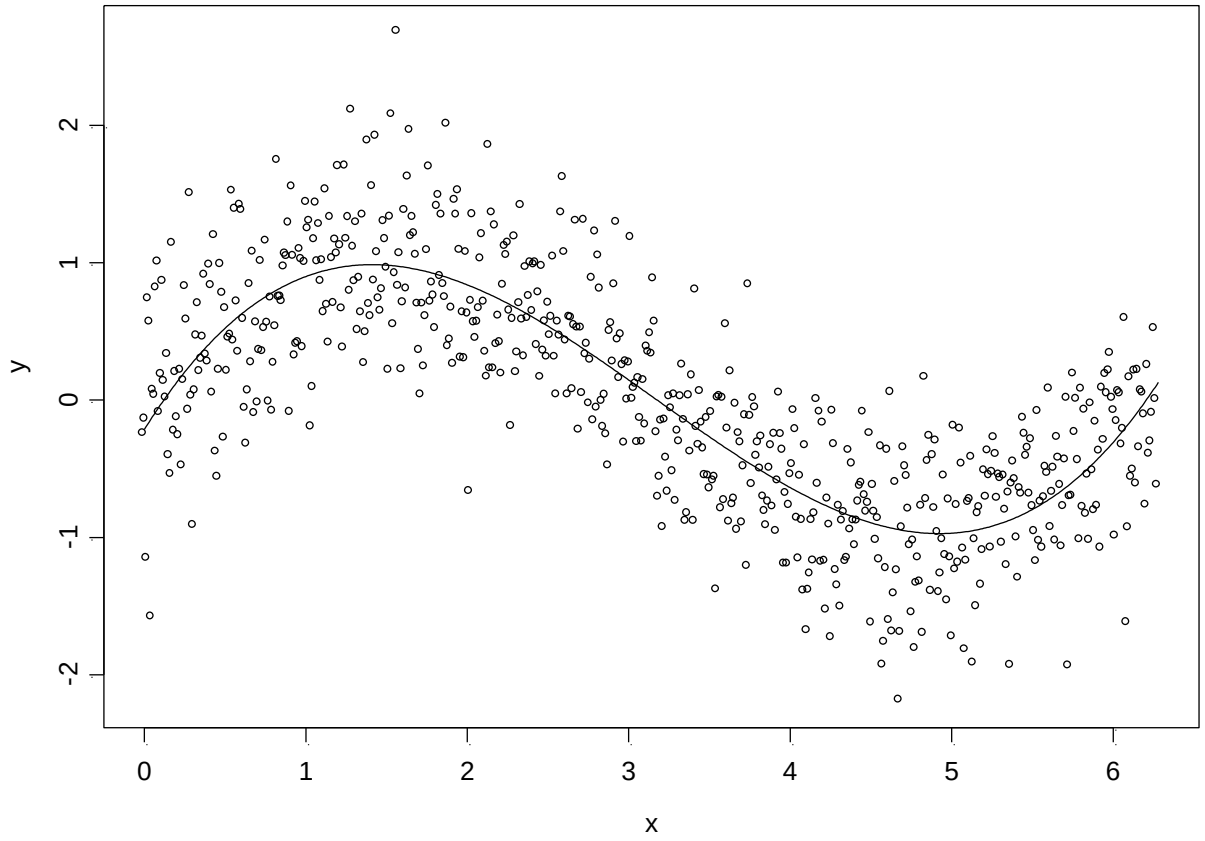
Regression ? (No way)

```
> fit=lm(y~x)
> plot(x,y)
> lines(x,fit$fitted.values,type="l")
```



Spline Fit

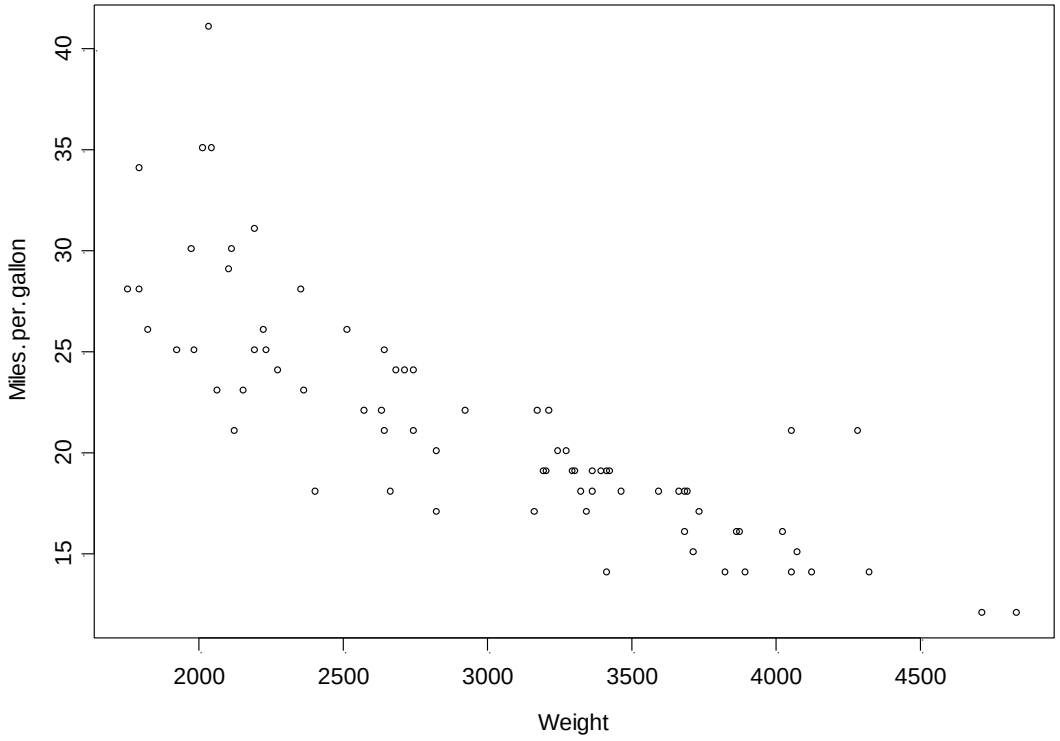
```
> fit=lm(y~bs(x))
> plot(x,y)
> lines(x,fit$fitted.values,type="l")
```



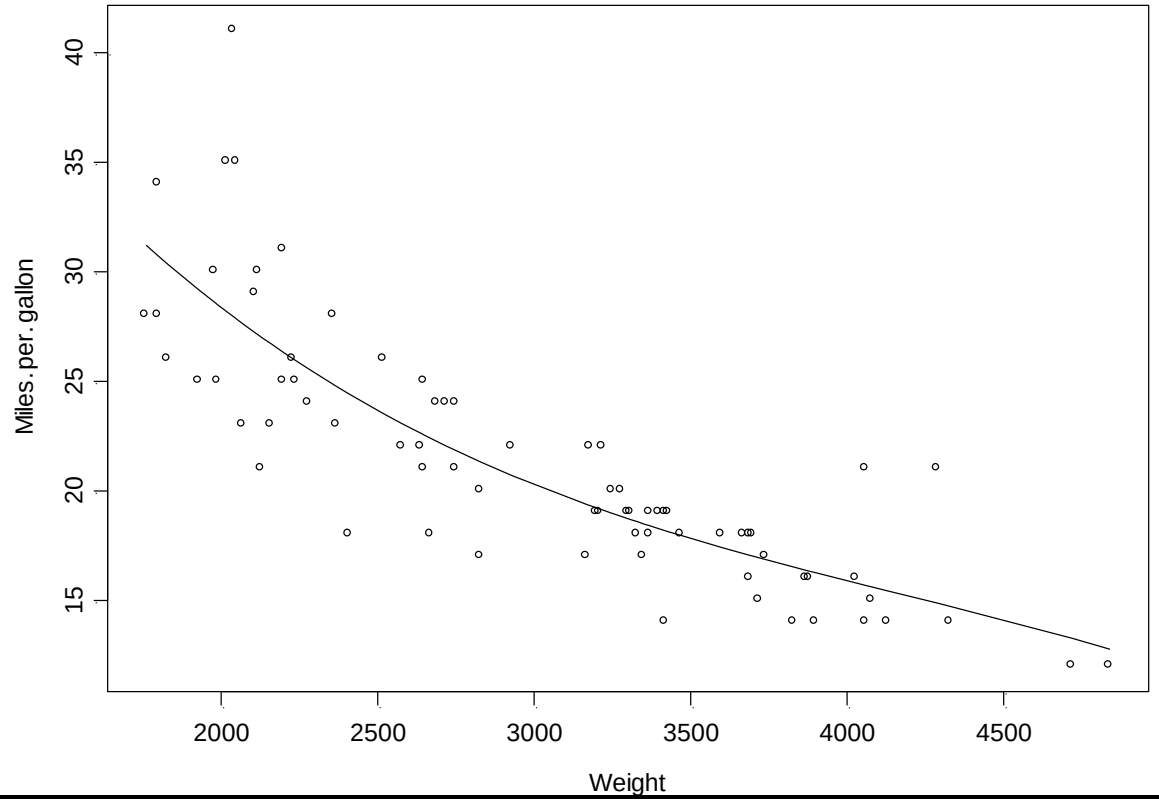
Example:

Car data : model mpg as a function of weight

$$mpg = f(weight) + \varepsilon$$



```
> fit=lm(Miles.per.gallon~bs(Weight,3))
    > plot(Weight,Miles.per.gallon)
          > ord=order(Weight)
> lines(Weight[ord],(fitted(fit))[ord])
                    >
```



```
> summary(fit)

Call: lm(formula = Miles.per.gallon ~ bs(Weight, 3))
Residuals:
    Min      1Q  Median    3Q   Max
 -6.415 -1.556 -0.2815 1.265 13.06

Coefficients:
                  Value Std. Error  t value Pr(>|t|)
   (Intercept)  31.2145   1.3047    23.9243   0.0000
bs(Weight, 3)1 -13.0441   4.1146    -3.1702   0.0023
bs(Weight, 3)2 -14.0931   2.8697    -4.9110   0.0000
bs(Weight, 3)3 -18.4401   2.8696    -6.4260   0.0000

Residual standard error: 3.209 on 70 degrees of freedom
Multiple R-Squared: 0.705
F-statistic: 55.76 on 3 and 70 degrees of freedom, the p-value is 0
```

Notes:

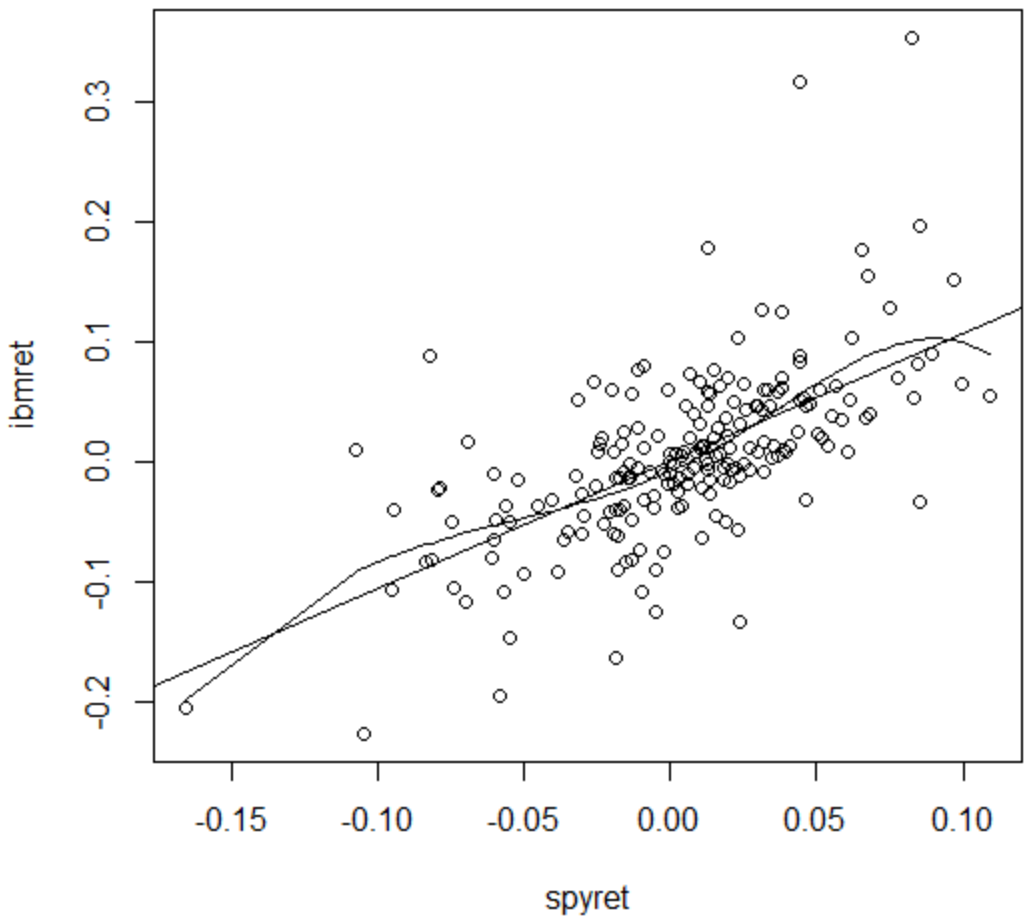
Too few knots give a rough fit, too many knots give too smooth a fit.

One can model each variable in the regression using a spline:

```
Fit=lm(Price~bs(Miles.per.gallon,3)+bs(Weight,3)+bs(Headroom,3))
```

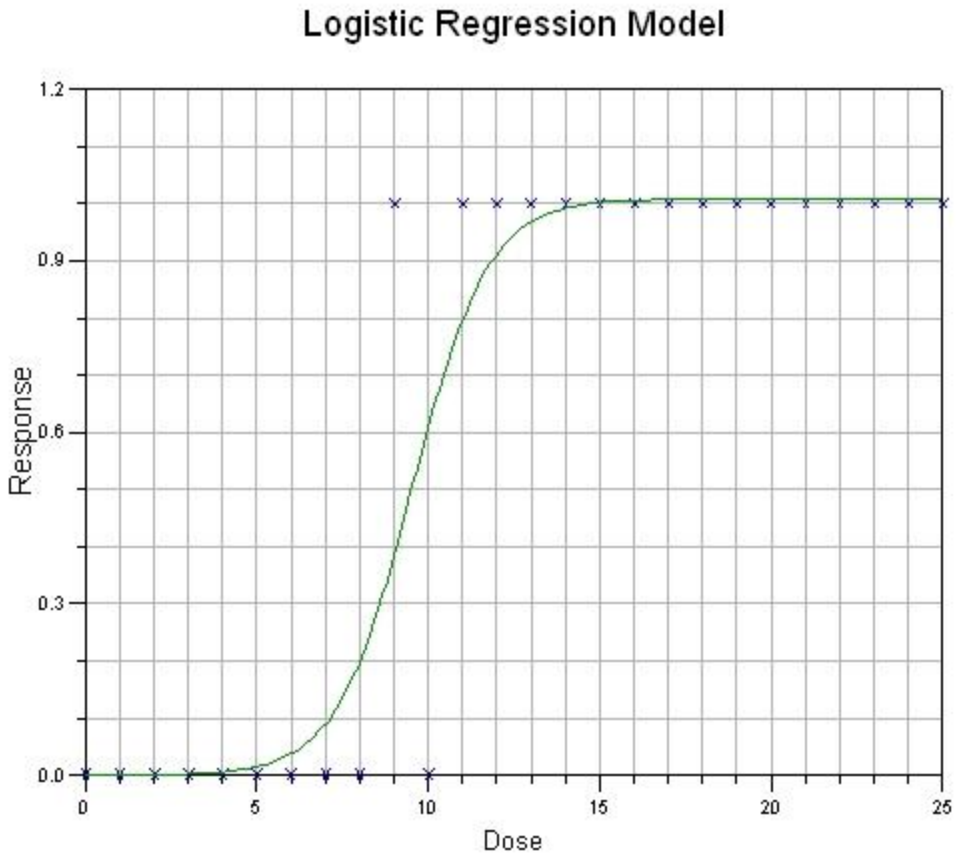
One drawback is that this uses up a lot of degrees of freedom (needs a lot of data).

Example




```
 plot(spyret,predict(fit1))
> plot(spyret,ibmret)
> fit=lm(ibmret~spyret)
> fit1=lm(ibmret~bs(spyret,5))
> abline(fit)
> lines(spyret,predict(fit1))
```

Logistic Regression



Overview of categorical regression models

Binary: Two categories (stata command: logistic y x1 x2 x3…)

- vote in last election?
- patient cured after treatment?
- startup failed?

Ordinal: More than two categories and assume ordered (mlogit)

- not favorable, favorable, extremely favorable
- not likely, likely, very likely
- no coverage, partial coverage, maximum coverage in insurance

Nominal: More than two categories but not ordered (ologit)

- mode of transportation – bus, car or train?
- employment status – employed, unemployed, out of work force

Count: Number of times something has happened

- number of patents a company has attained
- number of times a country goes to war

Logistic Models Probabilities

- Suppose
  - Y=0 if SP500 return neg tomorrow
  - Y=1 if SP500 return pos tomorrow
  - Want to model P(Y=1|today's data)

- Suppose
- Y=0 if you should go to bonds, 1 if stocks, 2 if cash…model P(Y=0), P(Y=1), etc…

Predictions can go wrong!

- Now regression works fine and well if Y is continuous (or nearly so), but things can go wrong if Y is discrete.
- When we mean wrong, we mean that predictions can be silly.

Example: Failing or Passing an Exam

- Let us define a variable ‘Outcome’
  - Outcome = 0 if the individual fails the exam
    = 1 if the individual passes the exam
- We can reasonably assume that Failing or Passing an exam depends on the quantity of hours students studied.
- Note than in this case, the dependent variable takes only two possible values.

The Data Set

- Our dataset contains information about 14 students

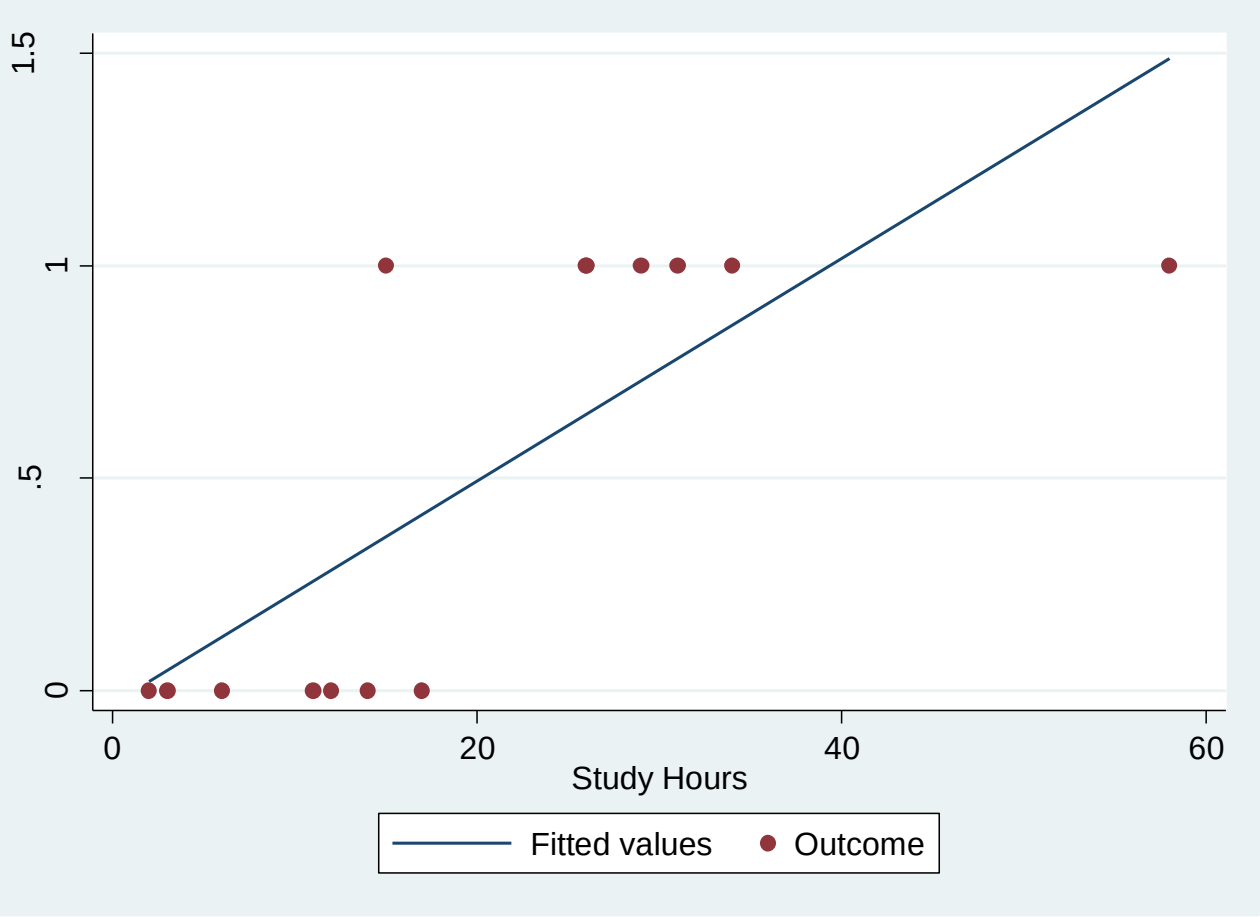
Student id	Outcome	Quantity of Study Hours
1	0	3
2	1	34
3	0	17
4	0	6
5	0	12
6	1	15
7	1	26
8	1	29
9	0	14
10	1	58
11	0	2
12	1	31
13	1	26
14	0	11

The Regression Model

- Nothing too unusual here

```
> fit=lm(outcome~hours,data=mydata)
> summary(fit)

Call:
lm(formula = outcome ~ hours, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48881 -0.27621 -0.03369  0.25842  0.63858

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.031861   0.161591  -0.197  0.84699
hours        0.026219   0.006483   4.044  0.00163 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3513 on 12 degrees of freedom
Multiple R-squared:  0.5768,    Adjusted R-squared:  0.5415
F-statistic: 16.36 on 1 and 12 DF,  p-value: 0.001627
```

The Fitted Line Plot



What is your prediction if you study more than 40 hours?
What are we actually predicting?????

Problems with using regression

- Linear regression is not the appropriate model to use if the Y variable is binary (0/1).
- This is true for several reasons
  - Predictions might be outside the (0,1) range
  - The noise is clearly not normal (is there noise?)

In the usual regression model set-up, we assume that Y is a linear function of k X variables, plus some random noise:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

Another way of stating this is

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Now regression works fine and well if Y is continuous (or nearly so), but things can go wrong if Y is discrete.

A special case is when Y takes on only the values 0 or 1.

For a 0,1 random variable Y, we have that

$$E(Y) = 0P(Y = 0) + 1P(Y = 1)$$

Thus for a <u>0,1 random variable Y</u>

$$E(Y) = P(Y = 1)$$

If we use regression as we know it to model a 0,1, variable, **our regression model is modeling**

$$P(Y=1) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

This is sometimes called a “linear probability model”

Why may this be a bad idea ?

What do we know about probabilities, i.e. what values can they take on ?

How do we fix this?

- We need to come up with some function f(x) so that f(x) is in the interval [0,1] for all x values.
- We can then model P(Y=1|x) = f(x).
- That is, we were using f(x) = b0+b1(x), but this f(x) function doesn't always produce values in the interval [0,1].
- It turns out there are several possible functions to use for f(x).

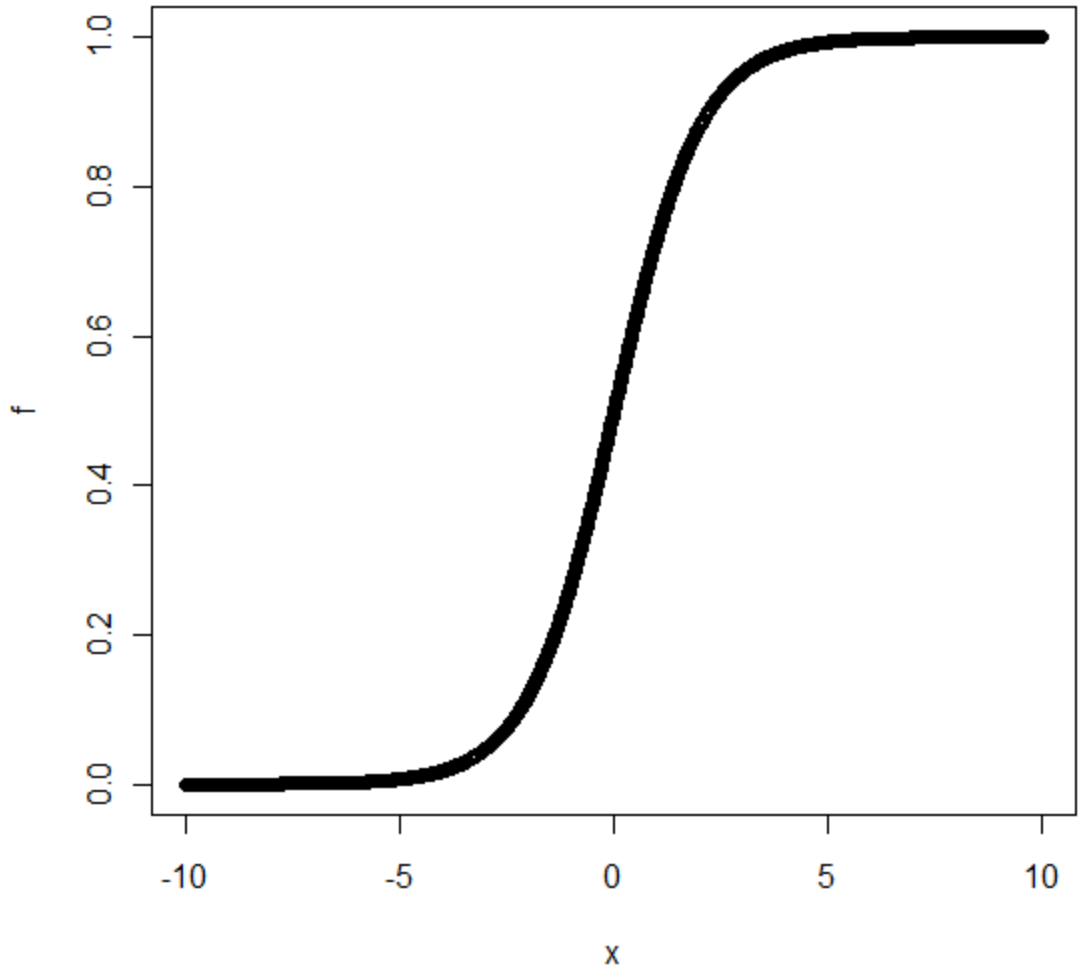
The logistic function

- The logistic function is defined to be

$$f(x) = \frac{e^x}{1+e^x} = \frac{\exp(x)}{1+\exp(x)}$$

- No matter what x value you use as an input, f(x) will always be in the interval [0,1].

Plot of the Logistic Function



The Logistic Regression Model

- The logistic regression model says to model

$$P(Y=1)=\frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}$$

- This is arbitrary, any function f(x) that produces numbers in the interval [0,1] can be used. But this is the most popular way of doing it.

Finding the unknown values

- The logistic regression model says

$$P(Y=1)=\frac{\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}{1+\exp(\beta_0+\beta_1 x_1+\cdots+\beta_k x_k)}$$

- Similar to linear regression, the unknown parameters are $\beta_0, \beta_1, \ldots, \beta_k$.
- But the method of estimation is a bit more complicated than simply least squares. A method called "maximum likelihood" is used which is discussed in more advanced courses (stat 111 or econometrics).

Example

- Data on Age and Signs of Coronary Disease

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

Logistic Regression in R

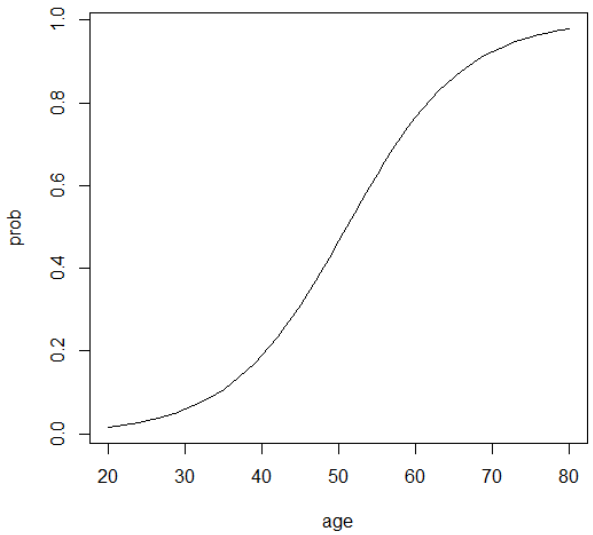
- The output looks like the usual regression output:

```
> fit=glm(cd~age,family=binomial(logit))
> summary(fit)

Call:
glm(formula = cd ~ age, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7025  -0.6497  -0.2377   0.6508   2.1075

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.70846    2.35397  -2.850  0.00437 **
age          0.13150    0.04634   2.838  0.00454 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.987  on 32  degrees of freedom
Residual deviance: 28.672  on 31  degrees of freedom
AIC: 32.672

Number of Fisher Scoring iterations: 5
```

The Fitted Model

- The R Output states that

$$P(outcome = 1) = \frac{\exp(-6.71 + 0.131 * age)}{1 + \exp(-6.71 + 0.131 * age)}$$



Interpreting the Coefficients

- In the linear model, interpreting the coefficients is easy: e.g.

```
> fit=lm(cd~age)
> summary(fit)

Call:
lm(formula = cd ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6833 -0.2800 -0.0261  0.2764  0.8208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.52651    0.21803  -2.415   0.0218 *
age          0.02016    0.00439   4.593 6.87e-05 ***
```

Interpreting the coefficients

- In logistic regression, there is a nonlinear relationship between the probability of success and the covariates:

$$P(outcome = 1) = \frac{\exp(-6.71 + 0.131 * age)}{1 + \exp(-6.71 + 0.131 * age)}$$

- So there is no simple interpretation such as "a unit increase in age increases the probability of cd by blah."

The Marginal Effects

- If you take a discrete data course, or econometrics course, they will discuss how to calculate and interpret the marginal effects for a logistic regression model:

$$\frac{\partial \Pr(Y=1 \mid X_1,\ldots,X_k)}{\partial X_j} = \beta_j \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{\left(1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}\right)^2}$$

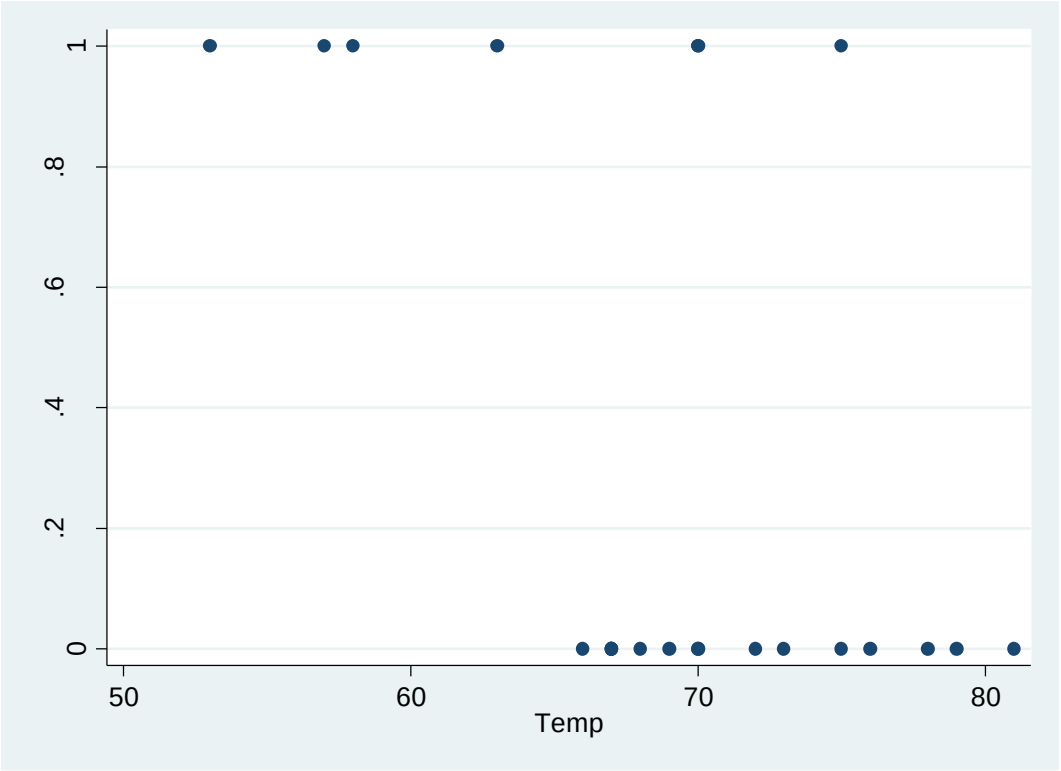
- For now, we'll simply look at plots of predicted probabilities, and understand that positive coefficients increase the probability, and negative coefficients decrease the probability of success.

Example: Space Shuttle Data

- On January 28, 1986, the space shuttle Challenger exploded and seven American astronauts died. Experts agreed that the disaster was caused by two leaky rubber O-rings.
- Prior to the flight, some rocket engineers had theorized that the O-rings would not seal properly due to cold weather.
- The predicted temperature for the launch was 26-29 degrees F.
- The engineers based their theory on data collected on O-ring damage and temperature for all 23 previous launches of the space shuttle.

Graph of the Data

- This is all the data available before the fateful flight.

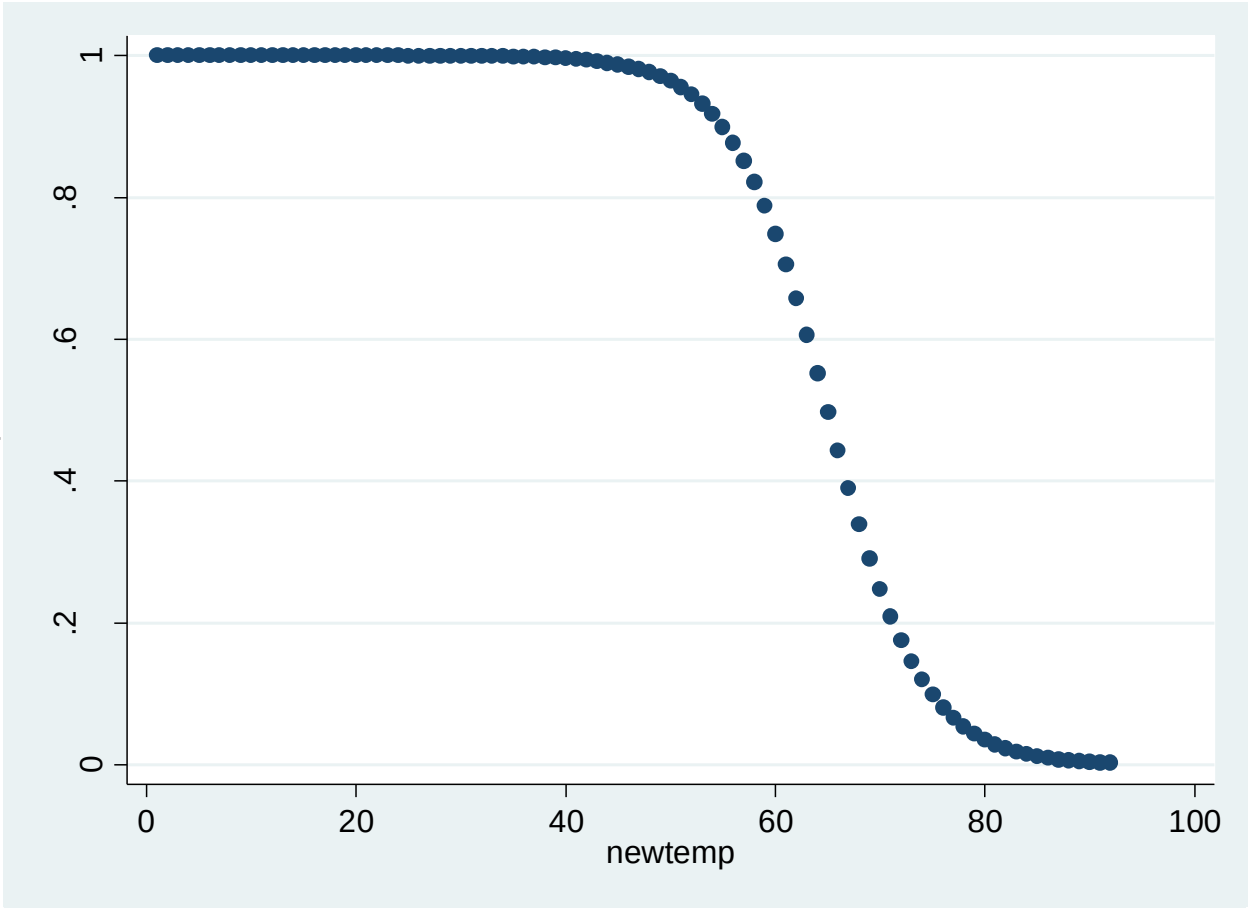


Logistic Regression Output

```
> fit=glm(outcome~temp,family=binomial(logit))
> summary(fit)

Call:
glm(formula = outcome ~ temp, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
temp         -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$P(Failure) = \frac{e^{15.04 - 0.2322(Temp)}}{1 + e^{15.04 - 0.2322(Temp)}}$$

The Output in Graphical Form



Multiple logistic regression
Example: Red Sox Games in 2013

- Study of 162 Red Sox games in 2013
- Response variable
  - *win*: Did the Red Sox win the game?
    win = 1 if they won, 0 if they lost
- Predictor variables
  - *hours*: length of time of game, in hours
  - *attendance*: number of people at game, in thousands
  - *daygame*: 1 if game began before 5pm, 0 if after 5pm
  - *home*: 1 if at home (Fenway Park), 0 if away
  - *previouswin*: 1 if Red Sox won previous game, 0 if they lost

- Source:
http://www.baseball-reference.com/teams/BOS/2013-schedule-scores.shtml

Red Sox 2013 – predictors of wins

Run a logistic regression with all 5 predictors (in R)

```
>summary(glm(win~home+attendance+night+hours+previouswins,
   family="binomial",data=redsox,subset=(year==2013)))

                    Coefficients:
                Estimate Std. Error z value Pr(>|z|)
 (Intercept)  -0.84010    1.54737  -0.543   0.5872
home          0.58612    0.33962   1.726   0.0844 .
 attendance   -0.01760    0.02141  -0.822   0.4112
 night        -0.36585    0.39690  -0.922   0.3566
hours         0.66124    0.35792   1.847   0.0647 .
 previouswins -0.53764    0.34793  -1.545   0.1223
                           ---
    Null deviance: 217.19  on 160  degrees of freedom
Residual deviance: 207.54  on 155  degrees of freedom
      (1 observation deleted due to missingness)
                    AIC: 219.54
```

Use a step-down modeling process

~~(dropping the least important term each step via AIC)~~

Red Sox 2013 – predictors of wins

The final logistic model

```
> summary(step((glm(win~home+attendance+night+hours+previouswins,
    family="binomial",data=redsoxdata,subset=(year==2013)))))

                      Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
     (Intercept)   -1.8156     1.1778  -1.541   0.1232
    home            0.5901     0.3345   1.764   0.0777 .
    hours           0.6850     0.3511   1.951   0.0511 .
     previouswins  -0.4912     0.3429  -1.432   0.1520
                              ---
       Null deviance: 217.19  on 160  degrees of freedom
   Residual deviance: 208.80  on 157  degrees of freedom
       (1 observation deleted due to missingness)
                      AIC: 216.8
            > 1-pchisq(217.79-208.8,df=3)
                      [1] 0.02942414
```

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -1.816 + 0.590(X_1) + 0.685(X_2) - 0.4912(X_3)$$

Scoring Customers-Universal Bank

- We have data on 5000 customers of Universal Bank.
- The data include customer demographic information (age, income,etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan).
- We are interested in what factors contributed to someone accepting the banks personal loan offer.

Scoring Customers-UniversalBank

- The following table describes the various variables that were collected:

ID	Customer ID
Age	Customer's age in completed years
Experience	#years of professional experience
Income	Annual income of the customer ($000)
ZIPCode	Home Address ZIP code.
Family	Family size of the customer
CCAvg	Avg. spending on credit cards per month ($000)
Education	Education Level. 1: Undergrad; 2: Graduate; 3: Advanced/Professional
Mortgage	Value of house mortgage if any. ($000)
Personal Loan	Did this customer accept the personal loan offered in the last campaign?
Securities Account	Does the customer have a securities account with the bank?
CD Account	Does the customer have a certificate of deposit (CD) account with the bank?
Online	Does the customer use internet banking facilities?
CreditCard	Does the customer use a credit card issued by UniversalBank?

Which variables are important?

Logistic regression

Log likelihood = -642.17638

Number of obs	= 5000
LR chi2(11)	= 1877.69
Prob > chi2	= 0.0000
Pseudo R2	= 0.5938

personalloan	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0536101	.061312	-0.87	0.382	-.1737794	.0665592
experience	.0637567	.0609276	1.05	0.295	-.0556592	.1831726
income	.054581	.0026201	20.83	0.000	.0494456	.0597164
family	.6957575	.0743034	9.36	0.000	.5501255	.8413894
ccavg	.1239704	.0396467	3.13	0.002	.0462644	.2016765
education	1.736152	.1150667	15.09	0.000	1.510625	1.961678
mortgage	.0004745	.0005541	0.86	0.392	-.0006117	.0015606
securities~t	-.9368084	.285866	-3.28	0.001	-1.497095	-.3765213
cdaccount	3.822534	.3239396	11.80	0.000	3.187624	4.457444
online	-.6751707	.1570772	-4.30	0.000	-.9830364	-.367305
creditcard	-1.119745	.204992	-5.46	0.000	-1.521522	-.7179679
_cons	-12.19265	1.645167	-7.41	0.000	-15.41712	-8.968185

Back to finance

- Define Y=1 if SP500 return is positive tomorrow, 0 otherwise
- We want to model P(Y=1|known data).
- This would be a timing model.
- Hmmm, what to put in.

Popular

Short-term prediction of exchange traded funds (ETFs) using logistic regression generated client risk profiles

Jerry K. Bilbrey, Jr.
Clemson University

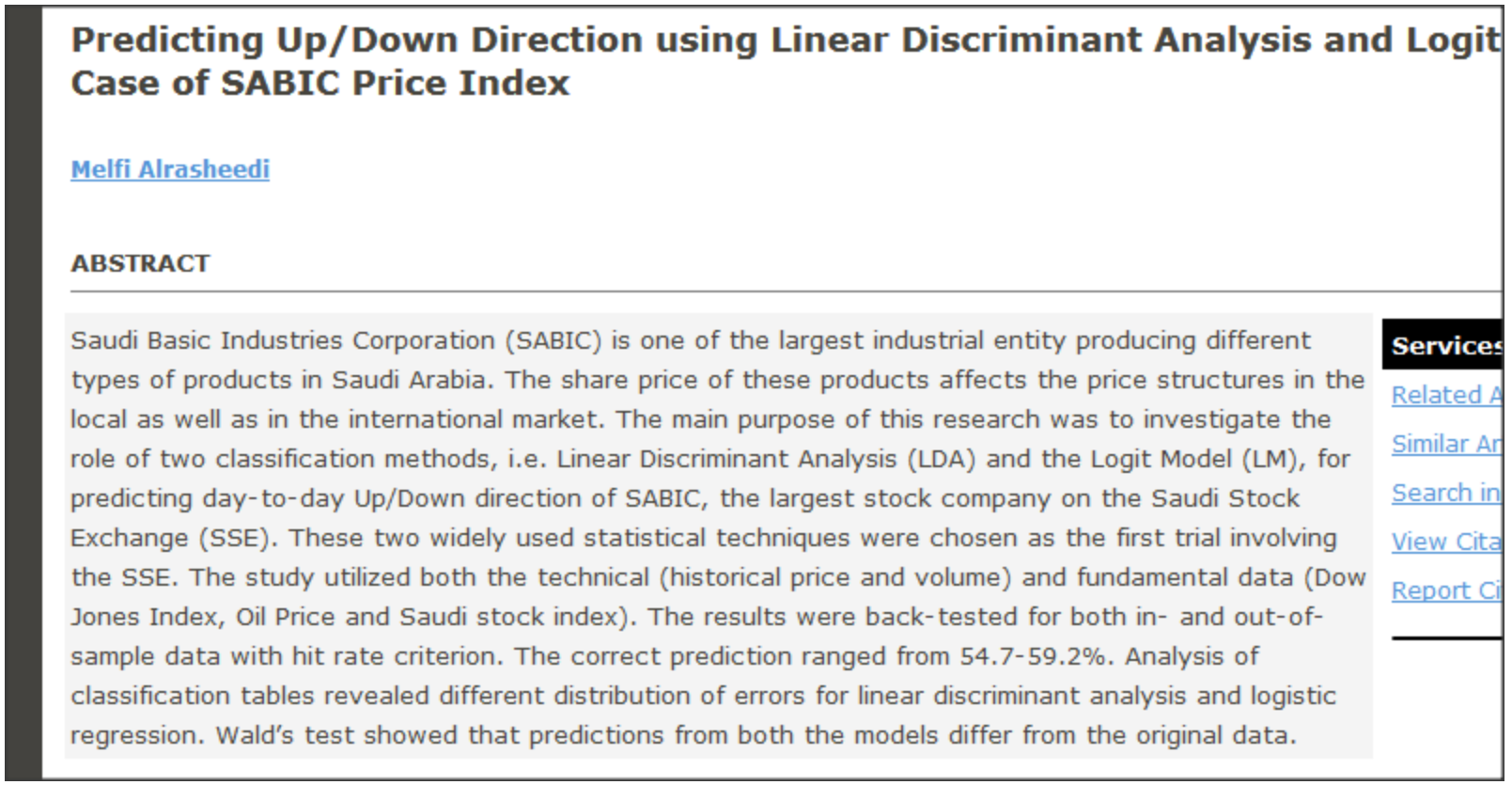
Neil F. Riley
Francis Marion University

Caitlin L. Sams
Anderson University

ABSTRACT

Efficient markets are a major tenet of investment theory. Efficient markets fully reflect all information into the price of a given asset such as stocks, bonds or exchange traded funds (ETFs). This suggests that there are no abnormal profits that can be made based on known public data. This paper presents an approach that goes against the efficient market theory by presenting a method that utilizes price and volume data to predict buy and subsequent sell signals for a list of ETFs. This approach utilizes both linear and logistic regression that develops a method for generating these buy signals. Sell signals are automatically created on the risk profile generated for each ETF by the model. As detailed, the regression based model shows great promise for developing strategies using individual risk based profiles.

Popular Idea



Predicting Up/Down Direction using Linear Discriminant Analysis and Logit Case of SABIC Price Index

Melfi Alrasheedi

ABSTRACT

Saudi Basic Industries Corporation (SABIC) is one of the largest industrial entity producing different types of products in Saudi Arabia. The share price of these products affects the price structures in the local as well as in the international market. The main purpose of this research was to investigate the role of two classification methods, i.e. Linear Discriminant Analysis (LDA) and the Logit Model (LM), for predicting day-to-day Up/Down direction of SABIC, the largest stock company on the Saudi Stock Exchange (SSE). These two widely used statistical techniques were chosen as the first trial involving the SSE. The study utilized both the technical (historical price and volume) and fundamental data (Dow Jones Index, Oil Price and Saudi stock index). The results were back-tested for both in- and out-of-sample data with hit rate criterion. The correct prediction ranged from 54.7-59.2%. Analysis of classification tables revealed different distribution of errors for linear discriminant analysis and logistic regression. Wald's test showed that predictions from both the models differ from the original data.

Services

Related A

Similar Ar

Search in

View Cita

Report Ci

Popular:PhD Thesis

LOGISTIC REGRESSION TO DETERMINE SIGNIFICANT FACTORS

ASSOCIATED WITH SHARE PRICE CHANGE

Variable	Dependent/ Independent	Variable Type
Change in Share Price	Dependent	Binary
Assets/Capital Employed	Independent	Metric
Debt/Assets ratio	Independent	Metric
Debt/Equity ratio	Independent	Metric
Dividend Yield%	Independent	Metric
Earnings/ Share(C)	Independent	Metric
Earnings Yield%	Independent	Metric
Operating Profit Margin%	Independent	Metric
Price/ Earnings	Independent	Metric
Return On Assets%	Independent	Metric
Return on Equity%	Independent	Metric
Return on Capital Employed	Independent	Metric

Popular

Forecasting the Direction and Strength of Stock Market Movement

Jingwei Chen
cjingwei@stanford.edu

Ming Chen
mchen5@stanford.edu

Nan Ye
nanye@stanford.edu

Abstract - Stock market is one of the most complicated systems in the world, and it has connection with almost every part in our life. In this paper, we applied different Machine Learning methods to predict the direction and strength of the market index price movement. Specifically, we looked at Multinomial Logistic Regression, K-nearest neighbors algorithm, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Multiclass Support Vector Machine, and compared their performance results based on test accuracy, robustness, and run-time efficiency.

8. Conclusion

Based on the results from different machine learning methods, we can see that Multinomial Logistic Regression performs the best in all the models in terms of the model robustness, prediction precision, and run-time efficiency. All the models perform better on the weekly data than on the daily data. For Multiclass SVM, scaled predictors can improve the robustness of the model in the prediction.

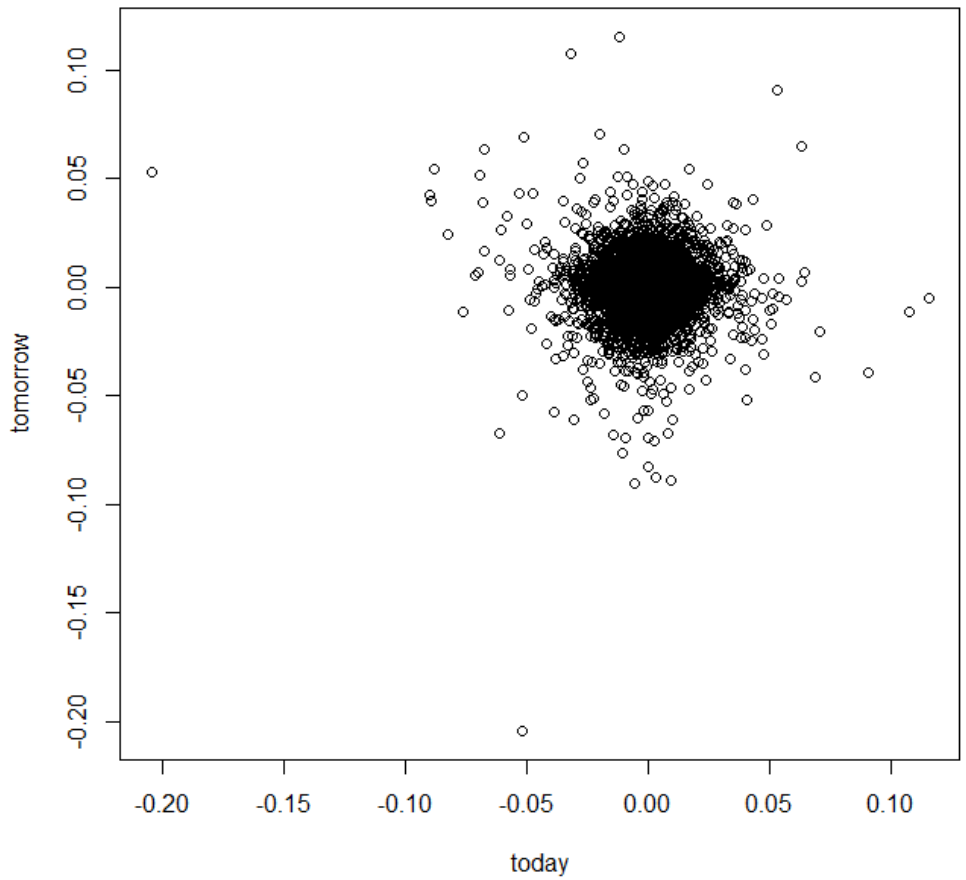
Example: This is hard

- We want to model the probability tomorrow is a positive day, given today's return P(tomorrow +|today)
- We will try this with VFINX (Vanguard SP500 fund) with data from 1980 on.

R Code

```
x=getSymbols(sname,from="2010-01-01",to="2015-01-
01",auto.assign=FALSE)
spyret=dailyReturn(Ad(x))

today=spyret
tomorrow=lag(today,-1)
yesterday=lag(today)
n=length(spyret)

#get rid of missing values
today=as.numeric(today[-c(1,n)])
tomorrow=as.numeric(tomorrow[-c(1,n)])
yesterday=as.numeric(yesterday[-c(1,n)])


ind=1.0*(tomorrow>0)## this is our Y (response) variable
```

Not a lot of structure



Logistic Model

```
 print(summary(glm(ind~today,family="binomial")))

Call:
glm(formula = ind ~ today, family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.351  -1.214   1.108   1.141   1.252

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.08918    0.02265   3.937 8.26e-05 ***
today       -4.06009    1.96986  -2.061   0.0393 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The model

- How to interpret

$$P(Y=1)=\frac{\exp(0.09-4.06today)}{1+\exp(0.09-4.06today)}$$

- If today=0 P(tomorrow +)=52% (slight upward bias)
- If today = 2% P(tomorrow +)=50%
- If today = -2% P(tomorrow +) = 54%