Taming Regression: A Case Study in How To

- Market your Dyalog APL application:
- Make APL Syntax user friendly to non-APL users.
- Create Things that Competitors can’t do in other languages.
- Design user functions and operators to be consistent with APL primitives.
- Provide APL Programmers with Useful Source Code for other Apps.

ASA AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

SDSS
SYMPOSIUM ON
DATA SCIENCE & STATISTICS
BEYOND BIG DATA: BRIDGING THE GAP
BETWEEN THEORY AND PRACTICE
RICHMOND, VA • JUNE 4–7, 2024

- Over 300 attendees
- 3 or 4 parallel sessions
- 35 attended my presentation "Taming Regression"
- Comments:
  - I assume your program does least-squares regression. Can you also do maximum likelihood?
  - When will the web version be out? My students are reluctant to install a program, but they would be more likely to use a web app.

Jeremy Flood

From: jrflood@aggies.ncat.edu

To: stevemansour@yahoo.com

Thu, Jul 18 at 12:52 PM

Hi Dr. Mansour; I hope this email finds you well today!

I spoke to you briefly during the SDSS conference about your TamStat software package, and was wondering if you'd be free in the coming weeks to discuss it further. As a stat tutor, I absolutely love how Tamstat makes data analysis less intimidating; and as a stat student, I love how the probability wizard makes probability calculations simple and intuitive. The illustrations alone could reduce several lectures into one image!

Planning a Wedding

Key:
known
unknown

- Costs are:
  - $500 to rent the hall
  - $100 per guest

1. What is the final cost for 35 guests?
2. How many guests can you invite with a budget of $8000?
3. You are in the catering business. How much do you charge for the venue? Per person?

Model: $f(x) = \beta_0 + \beta_1 x$

```
f←500∘+100∘×
```

1. Arithmetic: $y = f(x)$

```
f 35 ↔ 4000
```

2. Algebra: $y = f(x)$

```
f⍣¯1⊢8000 ↔ 75
```

3. Regression: $y = f(x)$

Check sales records from other caterers for total costs and number of guests.

```
COST⌹1,⍪GUESTS ↔ 500 100
```

Some issues with regression

- Not just coefficients, but an executable function.
- Confidence or prediction intervals.
- Non-linear relationships between X and Y.
- Non-constant variance over the range of f.
- Qualitative variables.
- Meaningful variable names in multiple regression.

The `regress` Operator*

- The `regress` operator in TamStat can perform:
  - Simple linear regression
  - Multiple linear regression
  - Regression with indicator variables
  - Polynomial regression
  - Variance Stabilizing Transformations
  - Multiplicative Regression
  - Logistic Regression
  - General non-linear regression

*If not specified, operand assumed to be linear

MODEL: NameSpace Result of regress

B: Coefficients – Numeric Vector $(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)$

RSq: R-Squared - Positive Scalar

S: Standard Error – Positive Scalar

YHAT: Fitted Values – Numeric Vector $(\hat{y}_1, \cdots \hat{y}_n)$

E: Residuals – Numeric Vector $(e_1, \cdots e_n)$

AnovaTable: Matrix showing Analysis of Variance

f: Linear Function relating X to Y: $y = f(x)$

Simple Regression Example

- A car dealer runs television ads for five weeks and records the number of cars sold that week.
- He would like to predict sales from the number of ads run.

Week	Television Ads Run	Cars Sold
1	1	14
2	3	24
3	2	18
4	1	17
5	3	27

Simple Linear Regression in TamStat

```
⍝ Predictor Variable
  ADS←1 3 2 1 3
⍝ Response Variable
  SALES←14 24 18 17 27
  MODEL←SALES regress ADS
  MODEL.B ⍝ Coefficents
10 5
  MODEL.Equation
Y←10+(5×X1)+E
  MODEL.f 2 ⍝Sales=f(Ads)
20
```

```
⍝ Estimate mean sales
   MODEL.f confInt 2
16.925 23.075
   .99 MODEL.f confInt 2
14.357 25.643
   MODEL.f 1 2 3
15 20 25
  ⍝ Predict dealer sales
   MODEL.f predInt 1 2 3
 6.7216 23.278
12.469  27.531
16.722  33.278
```

Two Regression Wizards

- Univariate Regression
  - One predictor variable
  - Simple or non-linear regression
  - Output graphic: scatterplot with regression line and confidence bounds
- Multivariate Regression
  - One or more predictor variables
  - Linear regression
  - Expressions are allowed
  - Separate screen for confidence and prediction intervals

User Meetings - Dyalog | Programme - Dyalog | TamStat - Taming Statistics | TamStat / Univariate Regression

https://tamstat-app.dyalog.com/RegWiz_uv

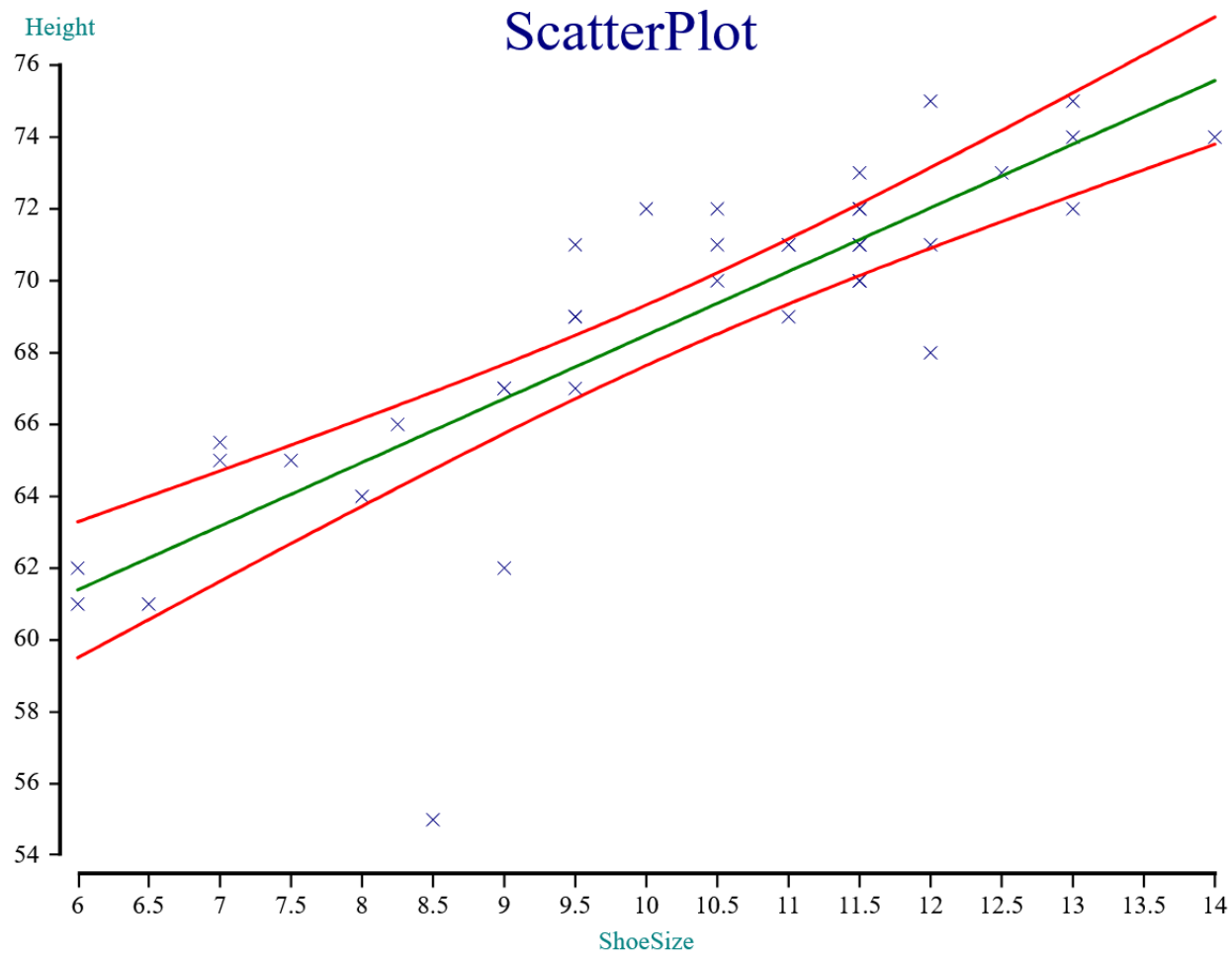
Search

Getting Started | Getting Started | Welcome to Dyalog - ... | Most Visited | Graphics | Release pre-release fo... | Sign out | Other Bookm

TamStat Dataset Descriptive Probability Inference Advanced Help

Expression `0.05 report ⊢ MODEL← Height regress ShoeSize`

Significance	0.050
Model Name	MODEL
Response Variable	Height
Transformation	[None]
Operation	regress
Predictor Variable	ShoeSize



report MODEL

The regression equation is:

Y←10+(5×X1)+E

ANOVA Table

SOURCE	SS	DF	MS	F	P
Regression	100.00	1	100.00	21.43	0.01899
Error	14.00	3	4.67		
Total	114.00	4			

S = 2.16025 R-Sq = 87.72% R-Sq(adj) = 83.63%

Solution

Variable	Coeff	SE	T	P
Intercept	10.0000	2.3664	4.22577	0.02424
X1	5.0000	1.0801	4.62910	0.01899

Multiple Regression

- Multiple regression in TamStat requires the right argument to take on one of three forms:
  - Variable List
  - Matrix
  - Namespace
- For both variable list and matrix examples, the left argument represents the response variable, and the right argument represents the predictor variables

```
⍝ Variable List
MODEL←Weight regress Height ShoeSize
report MODEL
```

```
The regression equation is:

 Y←32.021+(0.29419×X1)+(10.328×X2)+E

 ANOVA Table

 SOURCE                    SS     DF                MS          F
 ------     ---------------- ----- ---------------- ---------
 Regression          43,341      2           21,670      73.36
 Error               29,244     99              295
 ---------- ---------------- ----- ---------------- ---------
 Total               72,585    101

 S =   17.18717  R-Sq =  59.71%  R-Sq(adj) =   58.90%

 Solution

 Variable        Coeff         SE          T          P
 Intercept       32.02      39.87    0.80319    0.42379
 B1               0.29       0.76    0.38781    0.69899
 B2              10.33       1.62    6.39080   <0.00001
```

Multiple Regression (Continued)

```
⍝ Using a Namespace
V←'Weight Height ShoeSize'
 DB←V selectFrom SD
 MODEL←'Weight' regress DB
 MODEL.f 68 9.5
158.84
 .9 MODEL.f predInt 68 9.5
114.95 202.72
```

```
report MODEL
```

```
The regression equation is:

Weight←32.021+(0.29419×Height)+(10.328×ShoeSize)+E

 ANOVA Table

 SOURCE                    SS     DF              MS          F
 ------     ---------------- ----- --------------- --------- --
 Regression          43,341      2          21,670     73.36
 Error               29,244     99             295
 ---------- ---------------- ----- --------------- --------- --
 Total               72,585    101

 S =   17.18717  R-Sq =  59.71%  R-Sq(adj) =   58.90%

Solution

Variable       Coeff         SE          T          P
Intercept      32.02      39.87    0.80319    0.42379
Height          0.29       0.76    0.38781    0.69899
ShoeSize       10.33       1.62    6.39080   <0.00001
```

Indicator Variables

- Character fields are treated as indicator variables.
- Two categories: Creates a Boolean variable whose name is the value with the highest mean response value . 1=the value, 0=not the value.

- More than two categories: TamStat creates multiple indicator variables, one less than number of categories.
- Base Case:

  Category with minimum average response.

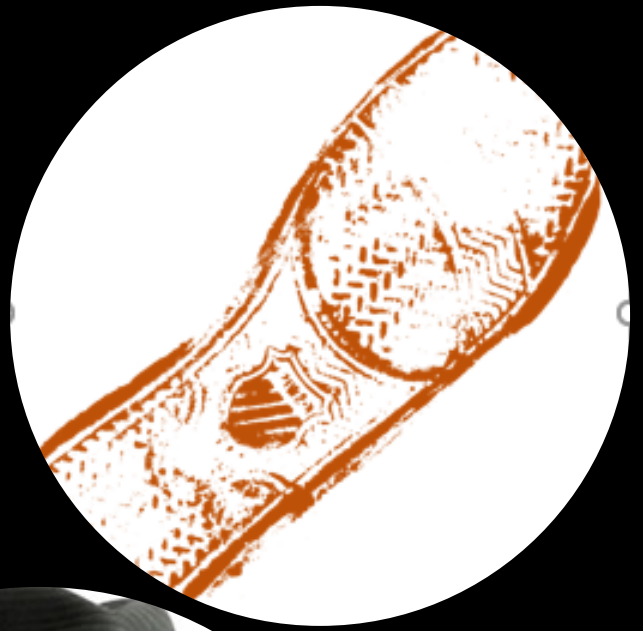
CSI Scranton

You are investigating a murder.

You find a bloody footprint near the body.

It is of a man’s shoe, size 9-1/2.

What is the height of the suspect?



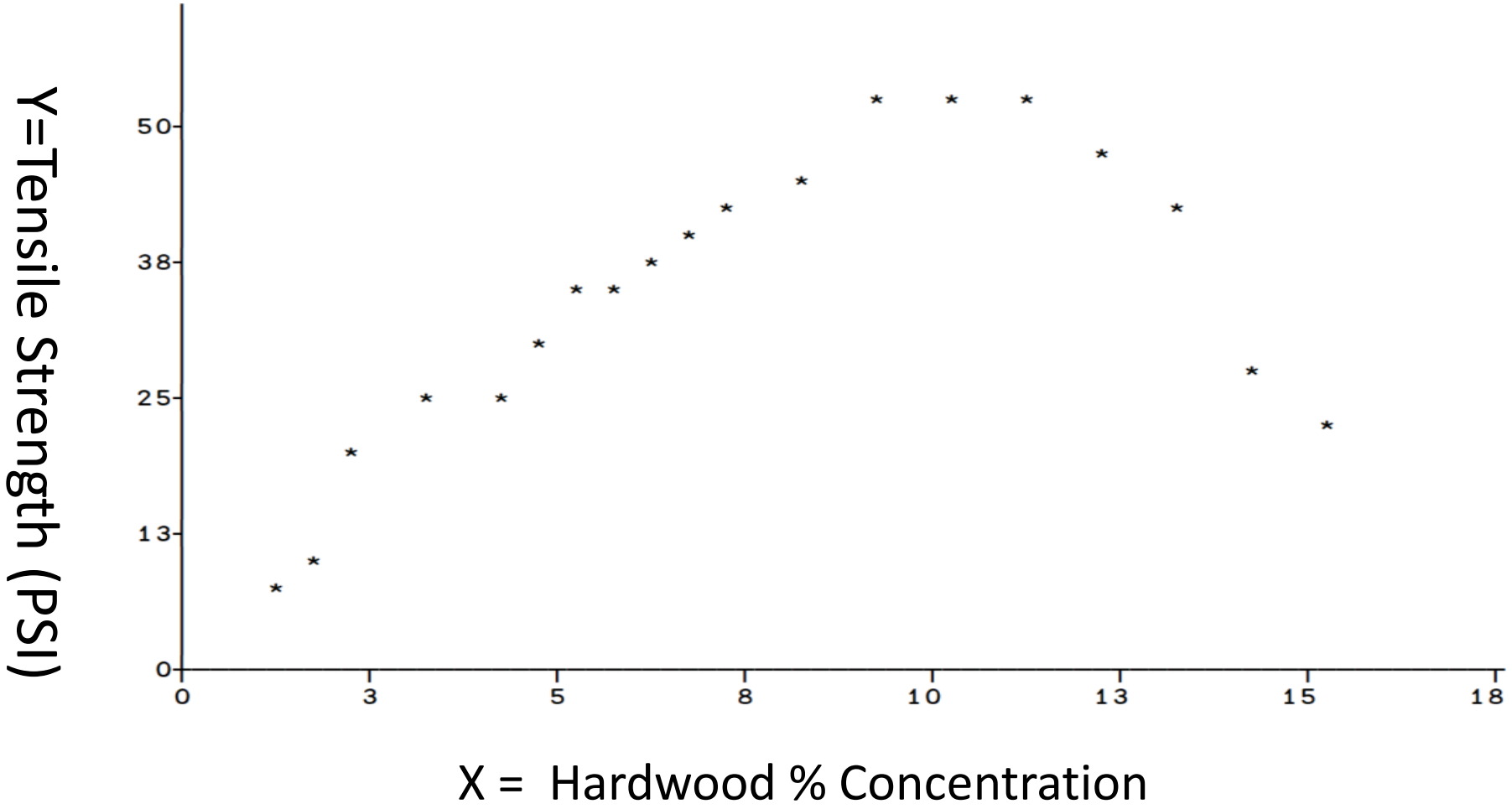
Indicator Variables (Continued)


```
      V←'Height Sex ShoeSize'
      DB←V selectFrom D
      MODEL←'Height' regress DB
      MODEL.Equation
Height←52.242+(2.9777×M)+(1.4031×ShoeSize)+E
      MODEL.f 1 9.5 ⍝ Point Estimate
68.549
      ⍝ 90% prediction interval
      0.9 MODEL.f predInt 1 9.5
64.454 72.645
```

Conclusion:
Suspect is between 5’5” and 6’1”.

Non-Linear Regression – Polynomial

```
scatterPlot show Y X
```

- 

Quadratic Regression (Operand: `quad`)

```
   MODEL←Y quad regress X
   MODEL.B
45.295 2.5463 ¯0.63455
⍝  ↑       ↑        ↑
⍝ Int   Linear   Square
   MODEL.g 7 10 15
44.581 47.511 27.012
   MODEL.g confInt 10
44.402 50.619
   MODEL.g predInt 15
15.752 38.273
```

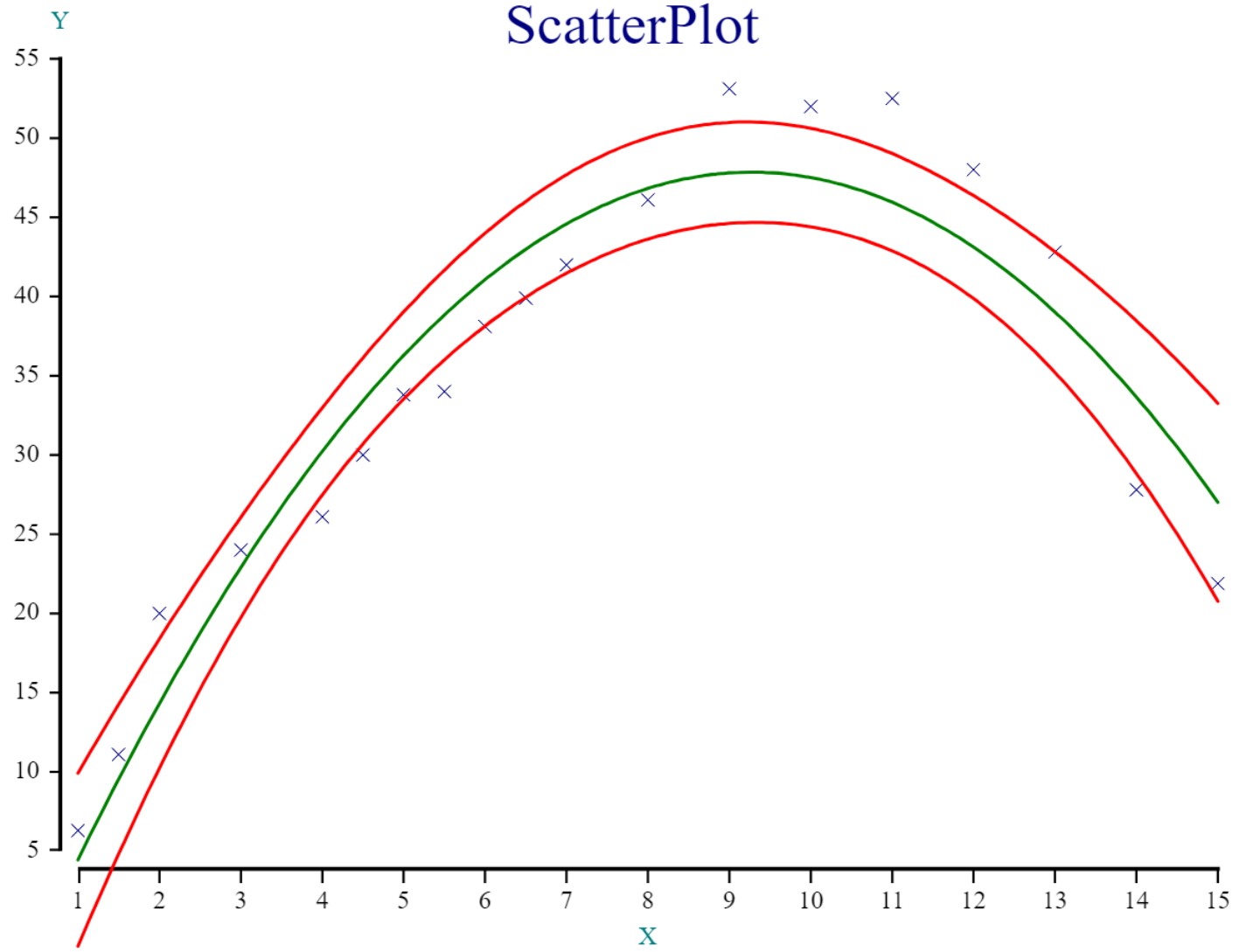
```
   MODEL.Equation
Y←45+(2.6×X-7.3)+(¯0.6×(X-7.3)*2)+E
   MODEL.Coeff
Variable    Coeff     SE         T
Intercept   45.29   1.48    30.55
X1           2.55   0.25    10.03
X2          ¯0.63   0.06   ¯10.27
```

TamStat / Univariate Regression Wizard

TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression `0.05 report ⊢ MODEL← Y quadratic regress X`

Significance	0.050
Model Name	MODEL
Response Variable	Y
Transformation	quadratic
Operation	regress
Predictor Variable	X



The regression equation is:

Y←45.295+(2.5463×X-7.2632)+(¯0.63455×(X-7.2632)*2)+E

ANOVA Table

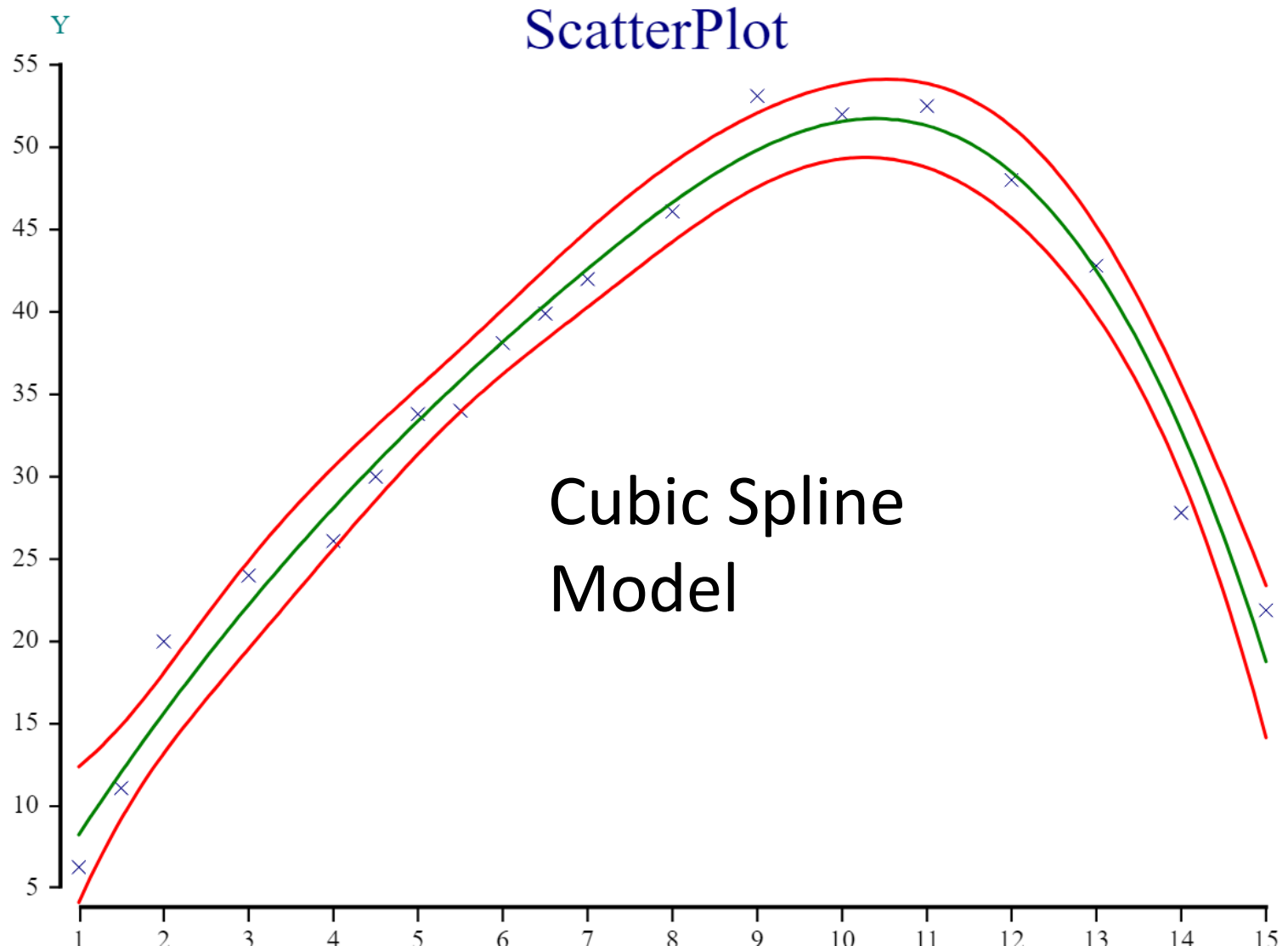
SOURCE	SS	DF	MS	F
Regression	3,104.2	2	1,552.1	79.43
Error	312.6	16	19.5	

TamStat / Univariate Regression Wizard

TamStat File Dataset Data Descriptive Probability Inference Advanced Help

Expression `0.05 report ⊢ MODEL← Y cubic regress X 7`

Significance	0.050
Model Name	MODEL
Response Variable	Y
Transformation	cubic
Operation	regress
Predictor Variable	X
Knots (Optional)	7



Logistic Regression – Boolean Response Variable

- Can you predict a person's sex based on height?
- Let `Y←Sex eq 'M'` and `X←Height`
- Logistic regression uses the following formula:

```
YHAT←÷1+*-B+.×1,X
```

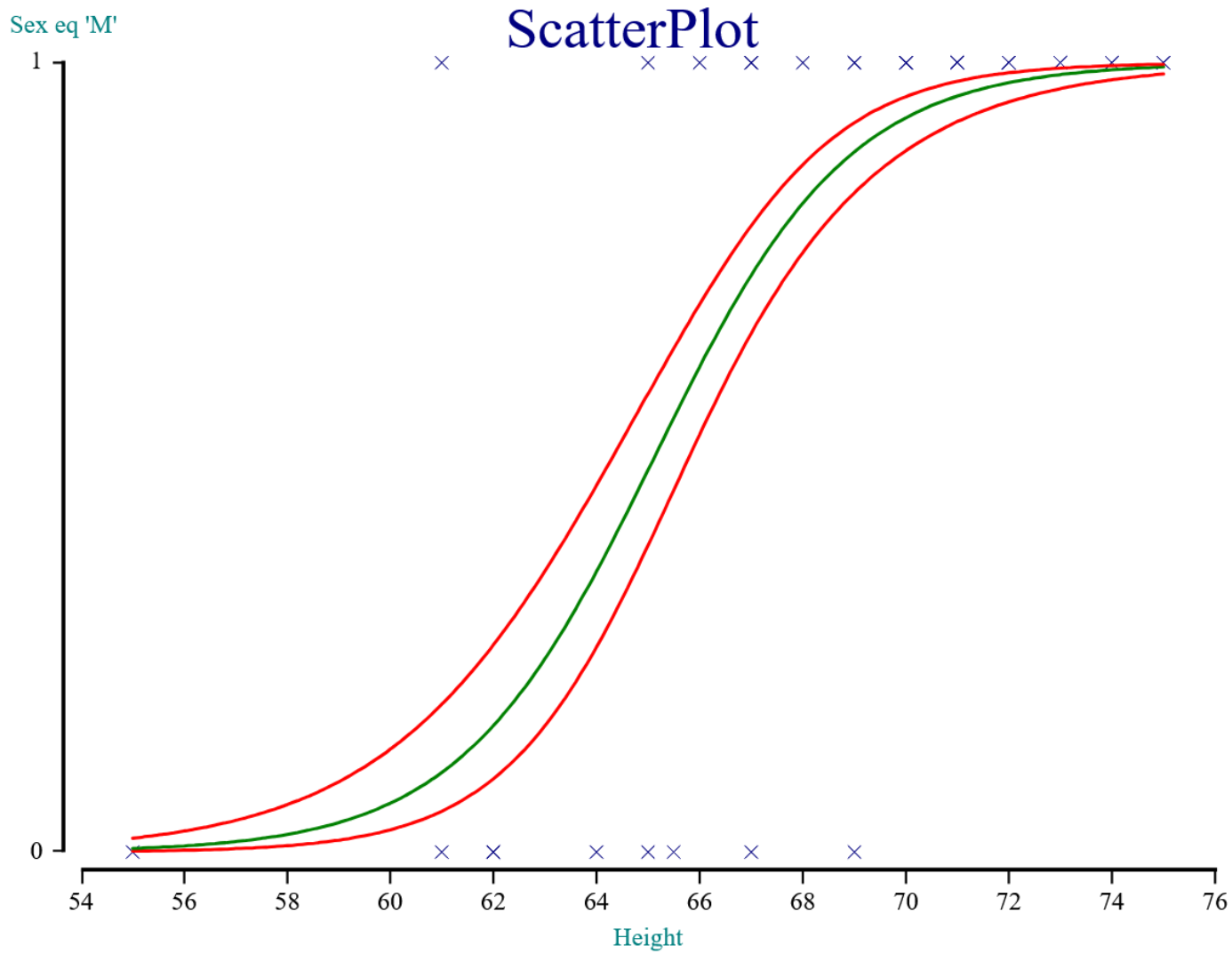
- `YHAT` is between 0 and 1 and represents the probability that `Y=1`.

- The slope and intercept in B must be solved numerically by providing a starting point.

```
B←0.2 0.001   ⍝ Guess
ML←Y logit regress X B
ML.B ⍝ Intercept,Slope
¯34.568 0.53083
ML.g 68 ⍝ P(Male|68")
0.8218
```

Expression `0.05 report ⊢ MODEL← (Sex eq 'M') logit regress Height ( 0.2 0.001 )`

Significance	0.050
Model Name	MODEL
Response Variable	Sex eq 'M'
Transformation	logit
Operation	regress
Predictor Variable	Height
Guess (Int,Slope)	0.2 0.001



How to Download and Install TamStat

TamStat runs on Windows and the Mac, is free to use, and can be downloaded from the following website:

https://tamstat.dyalog.com

A Web Version is also available on the site.

Conclusion

- Market your Dyalog APL application
  - Go to conferences where there are domain experts.
  - Bring a laptop and demo your software during refreshment breaks
  - Create Graphics that will “wow” the customer.
- Make APL Syntax user friendly to non-APL users.
  - Make functions flexible and use English-like syntax
- Create Things that Competitors can’t do in other languages.
  - Use operators and create functions dynamically
- Design user functions and operators to be consistent with APL primitives.
  - Where appropriate, many defined functions can be made to behave just like scalar functions.
  - Create ambi-valent functions which supply default values.
- Provide APL Programmers with Useful Source Code for other Apps.
  - TamStatCore functions are available on GitHub