# A Short Introduction to Curve Fitting and Regression

by Brad Morantz
bradscientist@machine-cognition.com

# Overview

- What can regression do for me?
- Example
- Model building
- Error Metrics
- OLS Regression
- Robust Regression
- Correlation
- Regression Table
- Limitations
- Another method - Maximum Likelihood

# Example

- You have the mass, payload mass, and distance of a number of missiles.
- You need an equation for maximum distance based upon mass of missile & payload
- Regression will let you build a model, based upon these observations that will be of the form
  - Max distance = $B_1$ * total mass + $B_2$ * payload mass + constant
- With this equation, you can answer questions

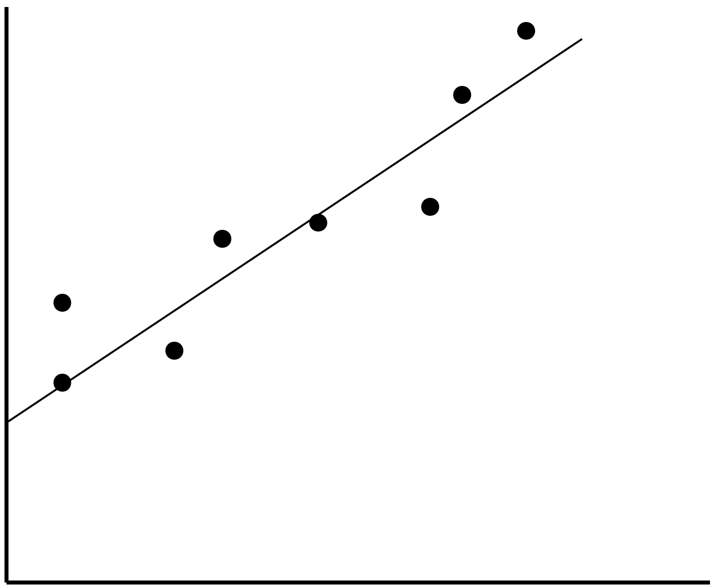
# Model Building

- Why build a model?
  - It can help us to forecast
  - Can assist in design
- Two ways
  - Causal factors
  - Data driven
- Regression is the latter
  - Model based upon observations

# MSE

- Residual
  - Difference between model and actual
  - $e = Y - \hat{Y}$
- If we summed them, the negative and positive would cancel out
- If we took Absolute value, would not be differentiable about origin
- So we take sum of squares
- MSE is mean of sum of squares

- Plot of values
- Line is the model
- Dots are the actual



# RMSE

- RMSE = sqrt(MSE)
- If we accept that MSE is estimator of variance, then RMSE is estimator of standard deviation.
- It is also a metric of how much error there is, or how well a model fits a set of data
- MSE and RMSE are often used as cost functions in many mathematical operations

# OLS Regression

- Ordinary Least Squares
- $Yhat = B_0 + B_1X_1 + \ldots\ldots + B_nX_n + e$
- $B_0$ is the Y intercept
- $B_n$ is the coefficient for each variate
- $X_n$ is the variate
- e is the error
- The program calculates the coefficients to produce a line/model with the least amount of squared error (MSE)

# Robust Regression

- Based on work by Kaufmann & Rousseuw
- Uses median instead of mean
- Not standard practice
- No automatic routines to do it
- Is less affected by outliers

# Simple Test of Model

- Plot out the residuals
- Look at this plot
- Are the residuals *approximately* constant?
- Or do you see a trend that they are growing or attenuating? (heteroscedasticity)
- If it is this latter situation, the causal factors are changing and the model will need to be modified

# Linear Correlation Coefficient

- Shows the relationship between two variables
  - Positive means that they travel in the same direction
  - Negative means that they go in opposite directions
  - Like covariance, but it has no scale
    - Can use for comparisons
- Range is from -1 to +1
  - Zero means no relationship, or independent
- $R_{xx} = X^T X$ (which becomes easy to implement in a matrix language like Fortran or Matlab)

# Correlation and Dependence

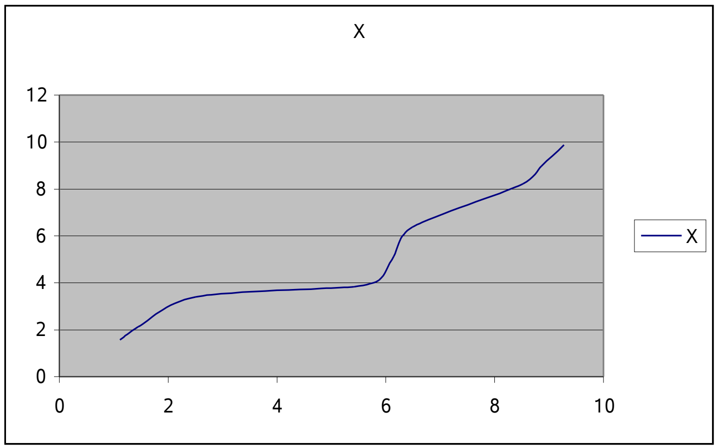
- If two items are correlated, then
  - $r \neq 0$
  - They are not independent
- If they are independent, then
  - $r = 0$
  - $P(a) = P(a|b)$; b is of no effect on a
- Need to have some theory to support above

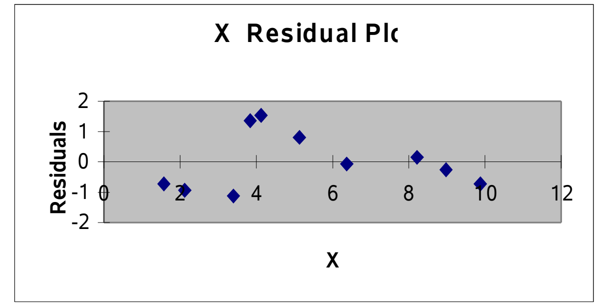
# $R^2$

- Pearsonian correlation coefficient
- Square of r, the correlation coefficient
- Fraction of the variability in the system that is explained by this model
  - Real world is usually 10% to 70%
- Adjusted $R^2$
  - Only use for model building
  - Penalizes for additional causal variables

# Reading a Regression Table

| Y | X |
|---|---|
| 1.113711 | 1.575213 |
| 1.446594 | 2.122573 |
| 2.505273 | 3.399938 |
| 5.416551 | 3.836957 |
| 5.875342 | 4.120897 |
| 6.141051 | 5.128728 |
| 6.484368 | 6.373329 |
| 8.516112 | 8.217644 |
| 8.853574 | 8.980098 |
| 9.272497 | 9.876384 |




This residual plot looks acceptable
About as much on top as on bottom

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.9483 |
| R Square | 0.899273 |
| Adjusted I | 0.886682 |
| Standard | 1.008021 |
| Observati | 10 |

The R Square is very good, so is the adjusted R Square
This model is very explanatory

ANOVA

| | *df* | *SS* | *MS* | *F* | *ignificance F* |
|---|---|---|---|---|---|
| Regressio | 1 | 72.573 | 72.573 | 71.42264 | 2.94E-05 |
| Residual | 8 | 8.128851 | 1.016106 | | |
| Total | 9 | 80.70185 | | | |

The model is good as the F value is >> 3 and the significance is << than 0.05

| | *Coefficient* | *tandard Er* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *ower 95.0* | *Jpper 95.0%* |
|---|---|---|---|---|---|---|---|---|
| Intercept | 0.295648 | 0.7 | 0.422355 | 0.683888 | -1.318555 | 1.909852 | -1.318555 | 1.909852 |
| X | 0.982041 | 0.116201 | 8.451192 | 2.94E-05 | 0.71408 | 1.250002 | 0.71408 | 1.250002 |

X is a good coefficient, as it is statistically significant, but the intercept is not with a P of 0.68 and a t <3

# Limitations of OLS

- Built on the assumption of linear relationship
  - Error grows when non-linear
  - Can use variable transformation
    - Harder to interpret
- Limited to one dependent variable
- Limited to range of data, can not accurately interpolate out of it

# Maximum Likelihood

- Can calculate regression coefficients
  - If probability distribution of error terms is available
- Calculates the minimum variance unbiased estimators
- Calculates the best way to fit a mathematical model to the data
- Choose the estimate for unknown population parameter which would maximize the probability of getting what we have

# References


- *Applied Linear Statistical Models* by Neter, Kutner, Nachtsheim, & Wasserman
- *A Handbook for Linear Regression* by Younger
- *Introduction to Linear Regression Analysis* by Montgomery & Peck
- *The Application of Regression Analysis* by Wittink
- *A Second Course in Business Statistics: Regression Analysis* by Mendenhall & McClave

- Most good statistics books have some information on regression analysis