Thank you to ABET for sponsoring this webinar…

AMERICAN ACADEMY OF ENVIRONMENTAL ENGINEERS & SCIENTISTS®

ABET

BE CONFIDENT

With ABET accreditation, your environmental engineering and science programs will equip students with the knowledge, skills and confidence to tackle the world's greatest challenges.

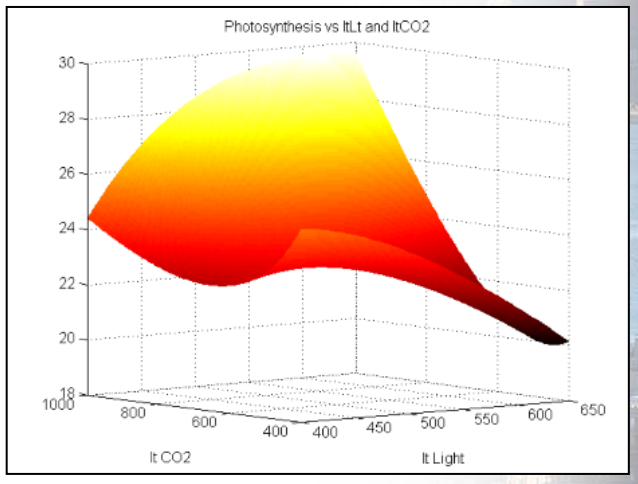
We will begin our presentation in a few minutes…

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Modeling Complex Environmental Data Without the Black Box

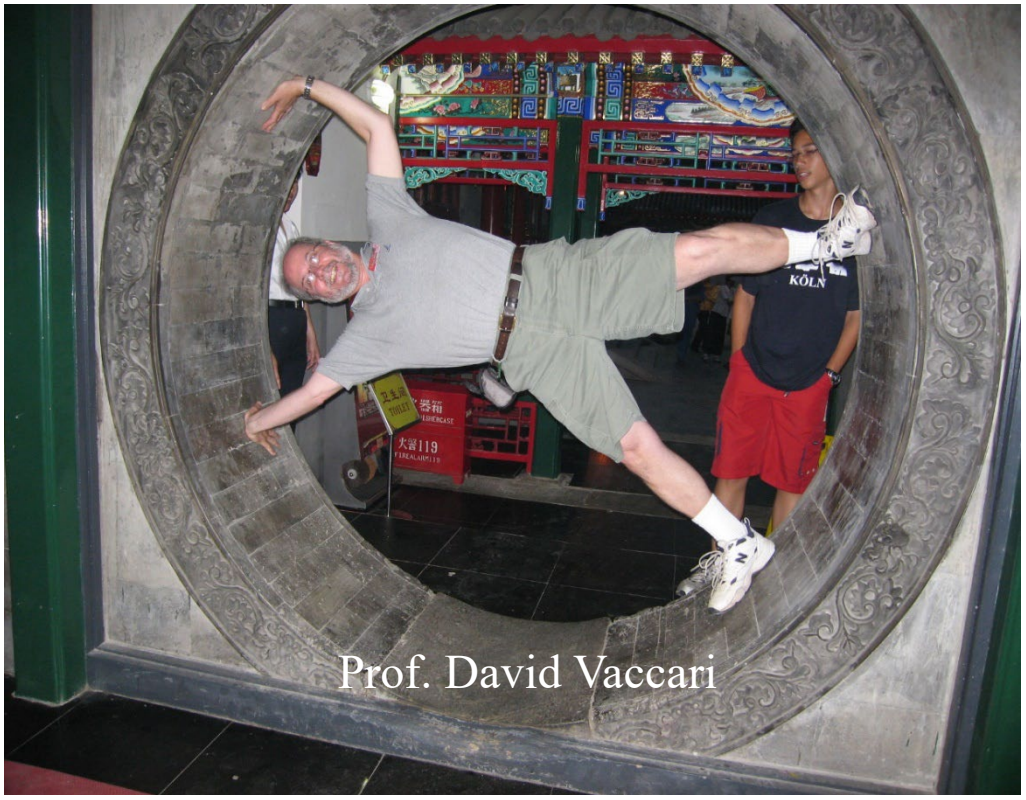
Multivariate Polynomial Regression

AAEES Webinar
February 16, 2022



David A. Vaccari
dvaccari@stevens.edu

AMERICAN ACADEMY OF ENVIRONMENTAL ENGINEERS & SCIENTISTS®



Prof. David Vaccari

Environmental Engineer specializing in:

- Wastewater treatment
- Statistical modeling
- Phosphorus resources

Rutgers University:

- Environmental science
- Chemical engineering

Stevens Institute of Technology:

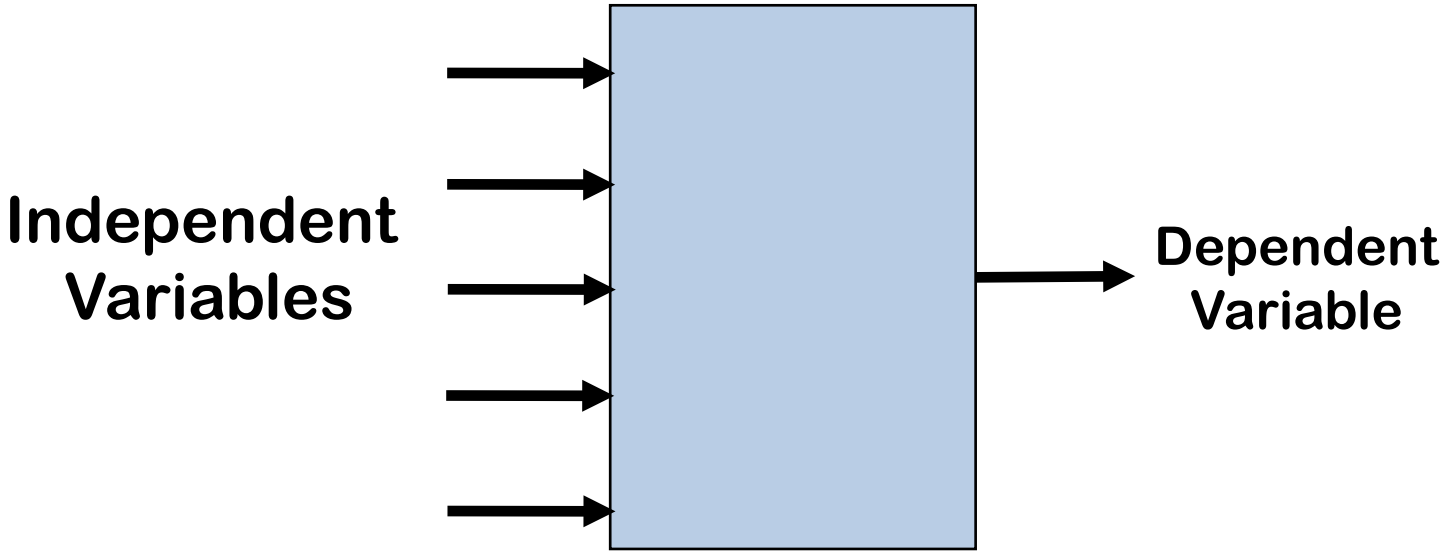
- Professor of environmental engineering

- Licensed Professional Engineer
- Board-Certified Environmental Engineer
- President-elect, AAEES
- Fellow of ASCE

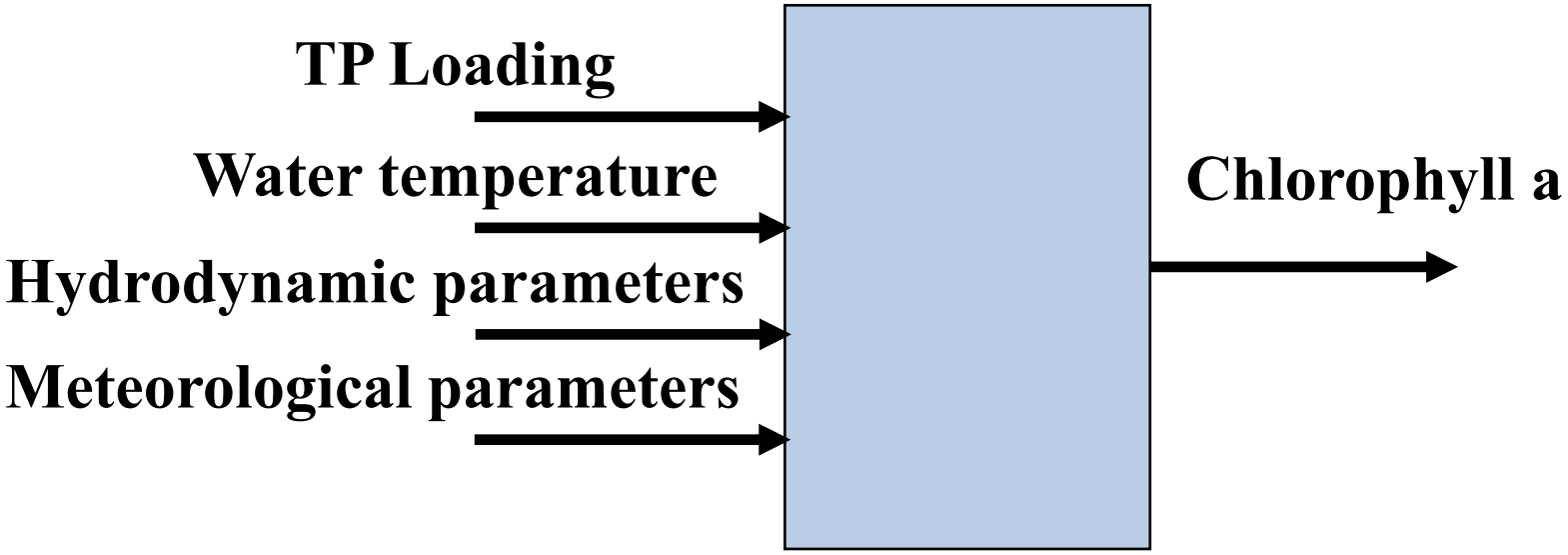
Outline

1. Introduction – modeling approaches
    - Artificial Neural Networks
    - Simple Linear Regression (review) and Goodness-of-Fit
    - Multilinear Regression
2. Multivariate Polynomial Regression
3. *TaylorFit* – Overview and features
    - Cross-validation and Final validation
    - Model analysis – residuals and sensitivity analysis
    - Confidence intervals and Prediction Intervals
4. Example applications
    - Macrophyte Indices
    - Photosynthetic productivity for NASA
    - Phytoplankton occurrence in the Hudson and East Rivers
    - Retail Industrial Sales

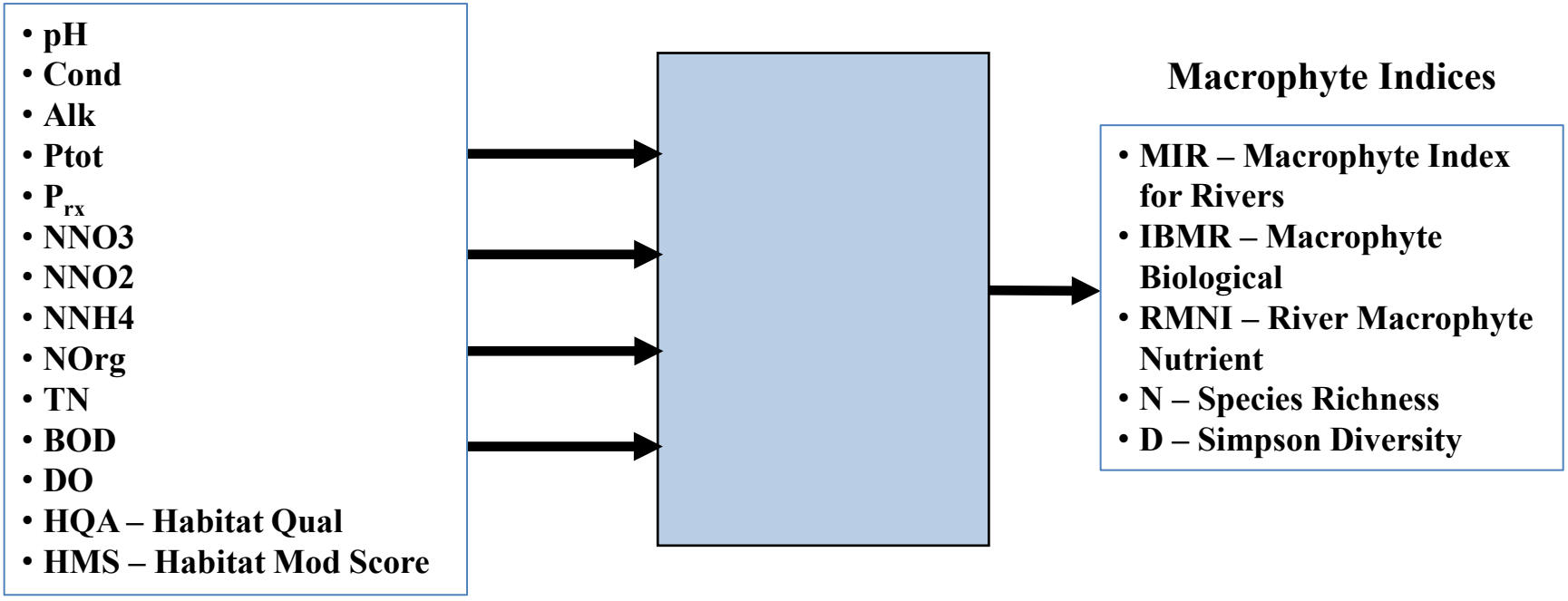
The Problem of Multivariable Correlation and Complexity



Water Quality

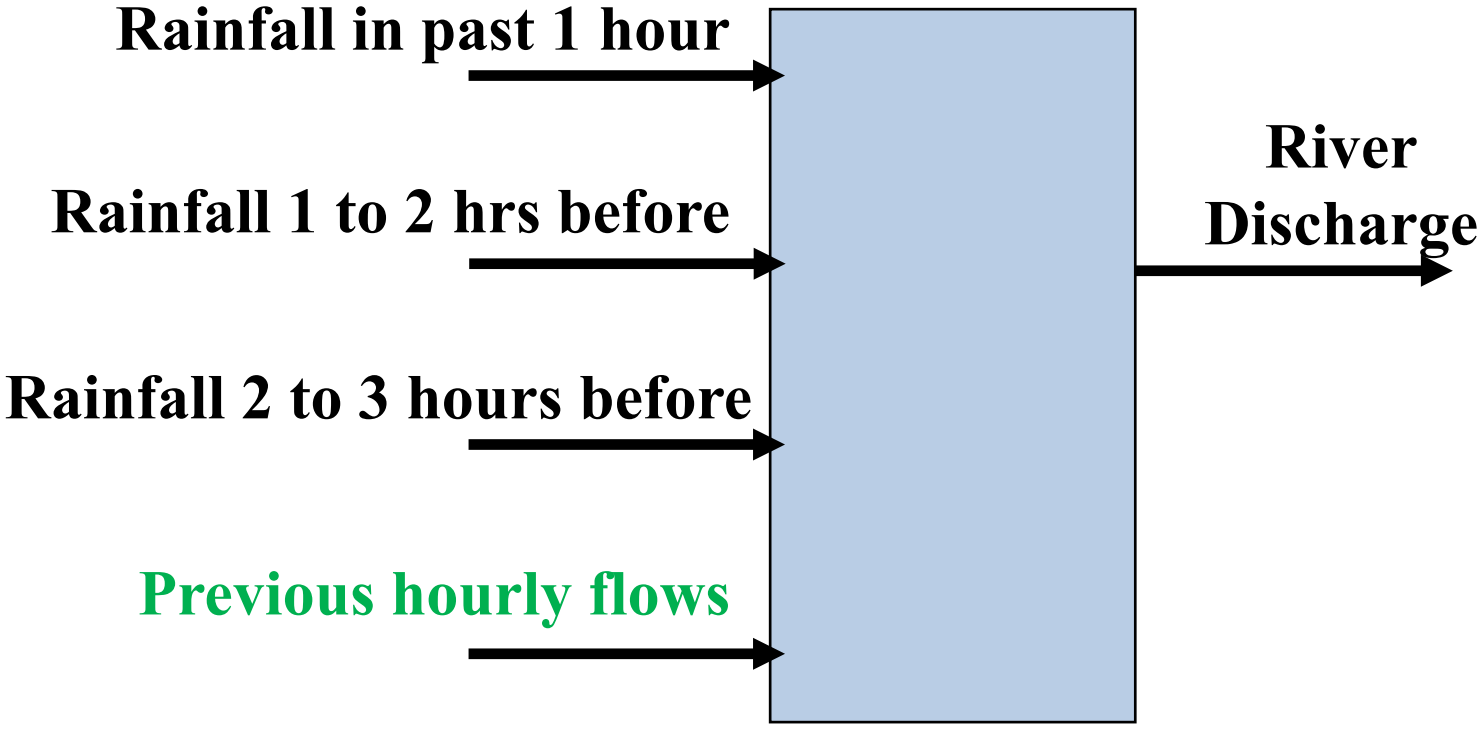


Water Quality



Vishwa Shah, Sarath Chandra K. Jagupilla *, David A. Vaccari, Daniel Gebler (2021). Non-Linear Visualization and Importance Ratio Analysis of Multivariate Polynomial Regression Ecological Models based on River Hydromorphology and Water Quality. Section: Ecohydrology, *Water* 2021, 13, 2708. https://doi.org/10.3390/w13192708.

Time-Series Example



Predictive Modeling Approaches

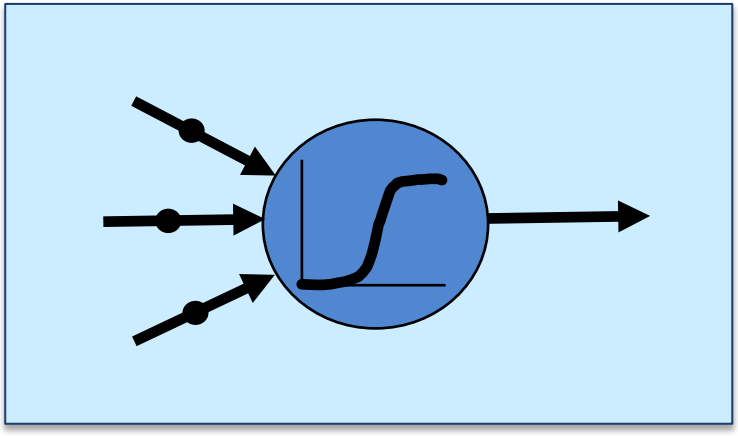
Linear

- Multilinear Regression (MLR)

Nonlinear

- Artificial Neural Networks (ANNs)
- Multivariate Polynomial Regression (MPR)

Artificial Neural Networks



Artificial Neural Networks

Pros

- Accurate
- Describes complex behavior
- Handles high dimensional problems

Cons

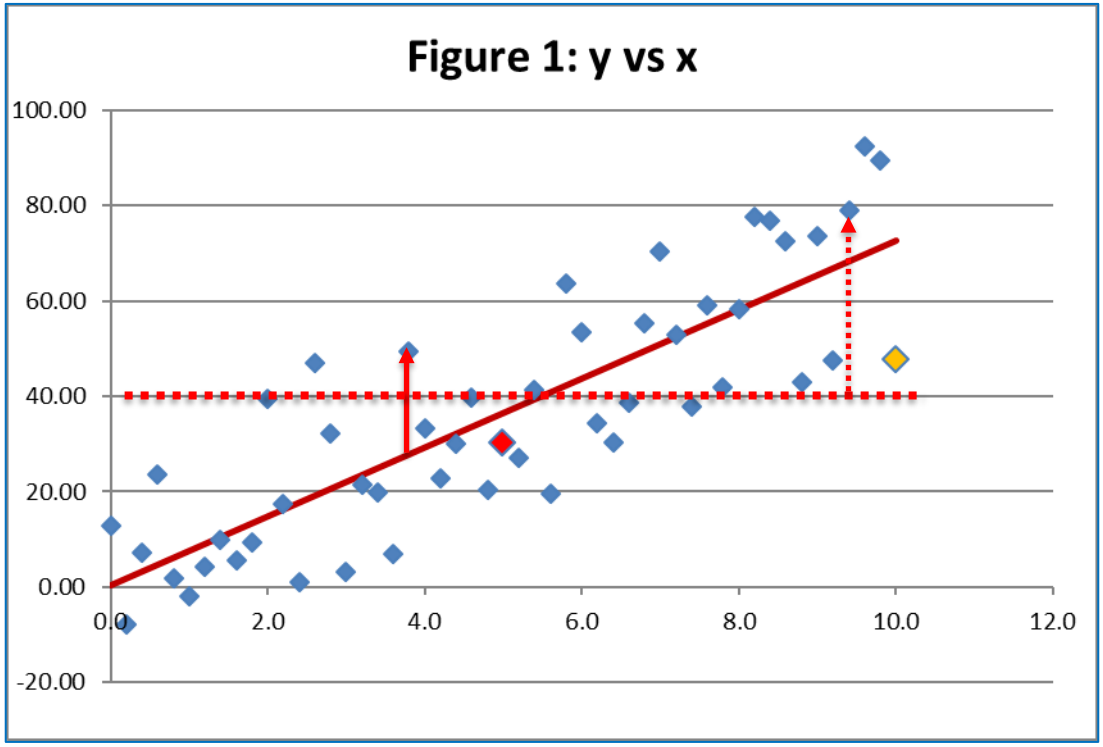
- Complex
- Difficult to analyze
- Difficult to communicate
- Susceptible to overfitting
- Hidden behavior: **Black Box!**

Simple Linear Regression

Optimization problem:
Find slope and intercept that minimize the Sum of Squares of the Error (SSE)

TSS is sum of squares of deviation from the mean

$$0 < SSE < TSS$$



Worst case: SSE = TSS

I.e., if x provides no predictive power

And, of course, the "best" outcome would be SSE = 0

Our first Global Goodness-of-Fit Statistic

$$R^2 = 1 - \frac{SSE}{TSS}$$

$$0.0 \leq R^2 \leq 1.0 \quad \textbf{(}\textbf{\textit{for }}\textbf{\textit{\underline{fit}}}\textbf{ only)}$$

Interpretation:
Approximately the fraction of the variance in y that is explained by the model

Multilinear Regression (MLR)

- MLR model

$$y = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + \ldots a_n \cdot x_n$$

- Problem of always increasing R^2

How do you determine how good is good enough?

- GOF statistics that penalize complexity
- Parsimonious selection of terms based on *t*-statistic

Global GoF Statistics that Penalize Complexity

- All are based on ***TSS, SSE, n_d, n_p***

$$R^2 = 1 - \frac{SSE}{TSS}$$

$$df = n_d - n_p$$

$$SSR = TSS - SSE$$

$$TMS = \frac{TSS}{n_d - 1}$$

$$MSE = \frac{SSE}{df}$$

$$MSR = \frac{SSR}{n_p}$$

$$R^2_{adj} = 1 - \frac{MSE}{TMS}$$

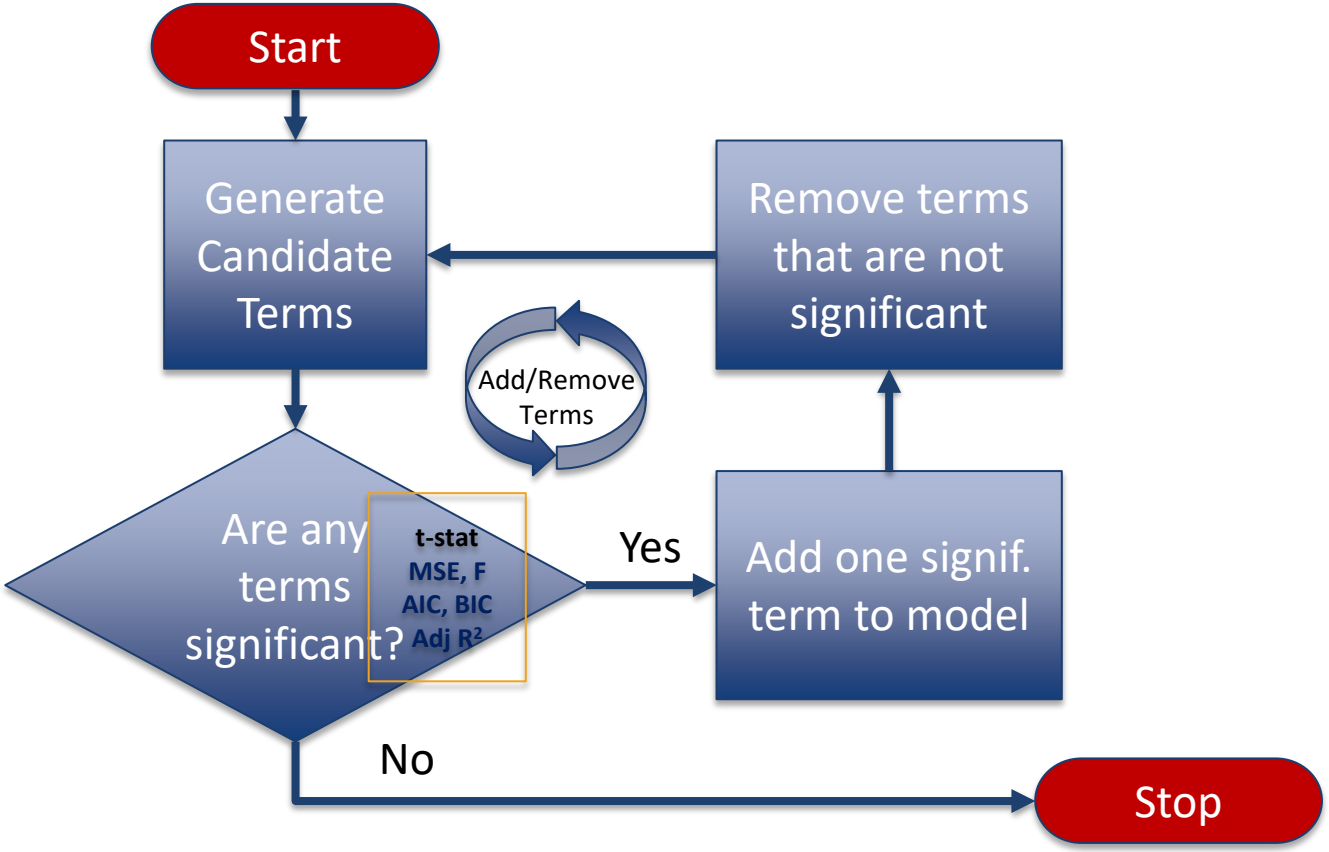
$$F = \frac{MSR}{MSE}$$

"Signal-to-noise ratio"

$$Probability\ of\ a\ larger\ value\ of\ F\ occurring\ by\ chance: = p(F, n_p, df)$$

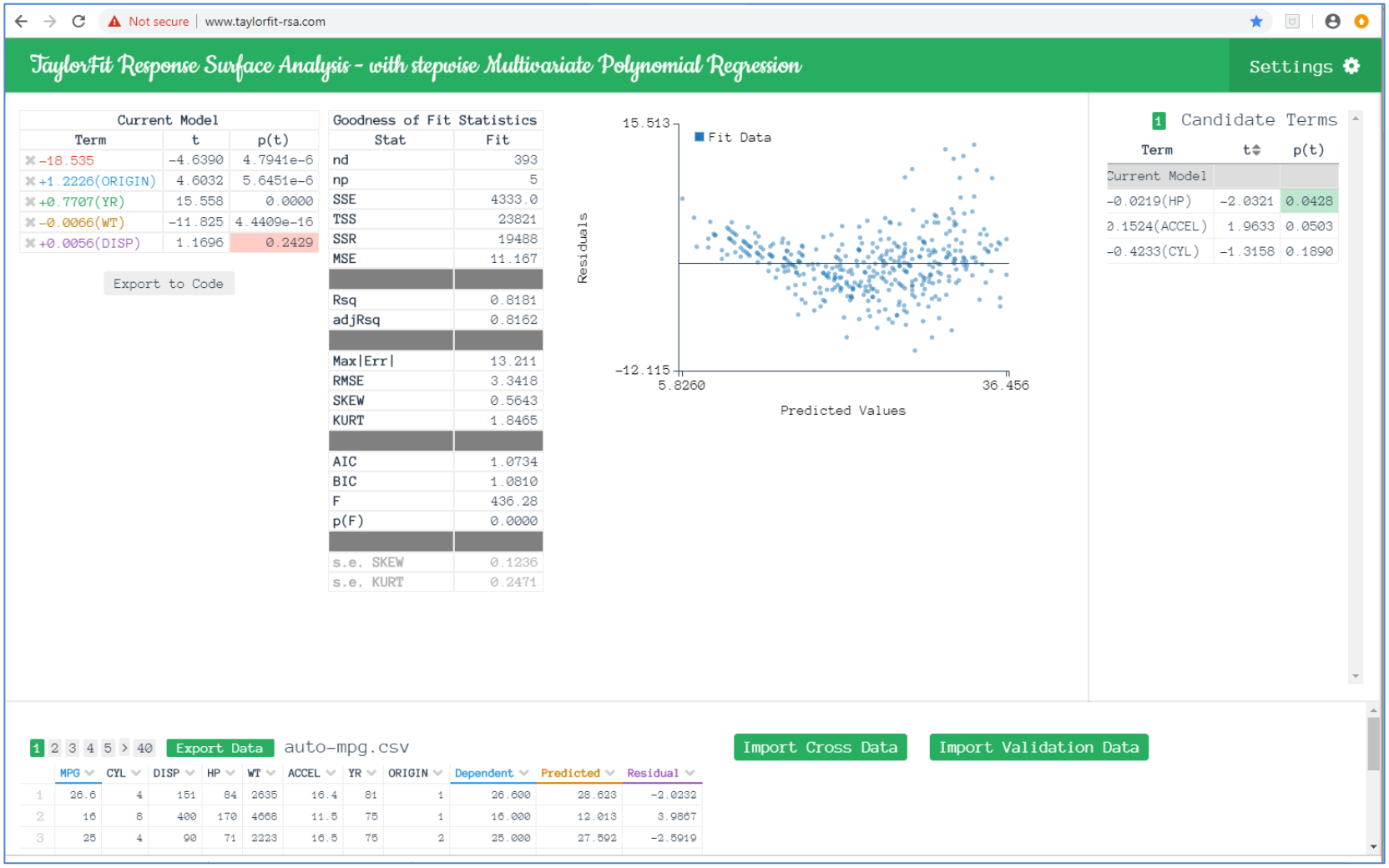
$$F = \frac{R^2_{adj}}{1 - R^2_{adj}} \cdot \frac{df}{n_p}$$

How to make models concise:
Stepwise Regression for MLR

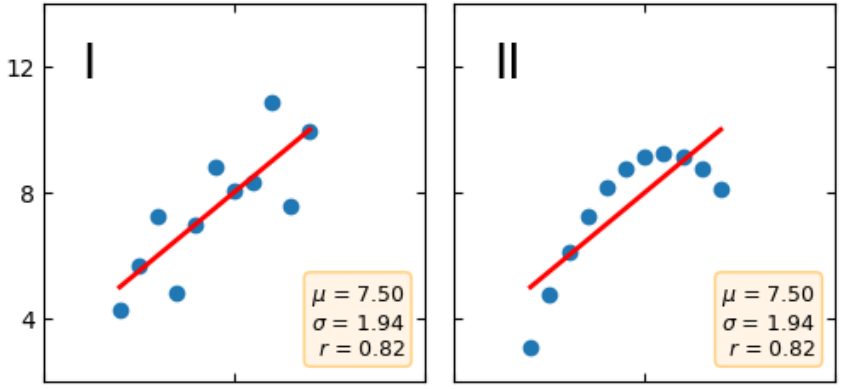


Demonstration

MLR with TaylorFit



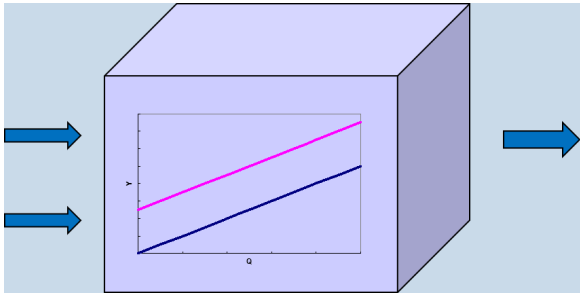
Model Specification Bias



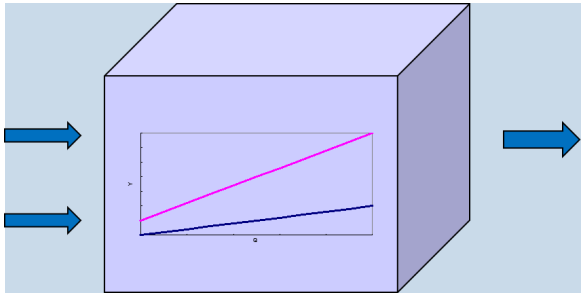
Both have the same mean, standard deviation and R^2

https://matplotlib.org/3.2.1/gallery/specialty_plots/anscombe.html

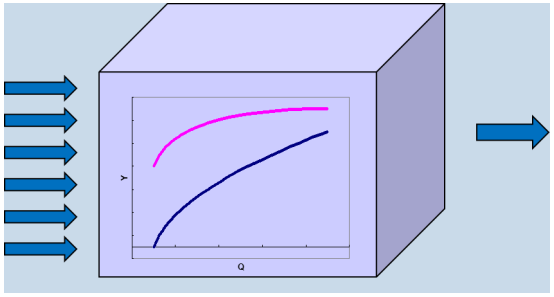
Multivariate Polynomial Regression Response Surface Analysis



Multilinear Regression



Interactions



Multivariate Polynomial (including ratios)

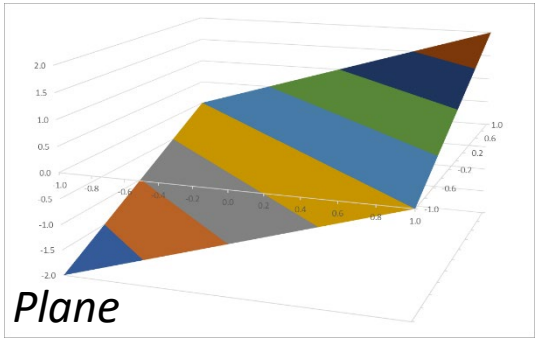
Model: $Z = a_0 + a_1 \cdot X + a_2 \cdot Y + a_3 \cdot X \cdot Y + a_4 \cdot X^2 \cdot Y^3 + a_5 \cdot X \cdot Y^{-1}$

Sensitivity: $\frac{\partial Z}{\partial X} = a_1 + a_3 \cdot Y + 2\, a_4 \cdot X \cdot Y^3 + a_5 \cdot Y^{-1}$

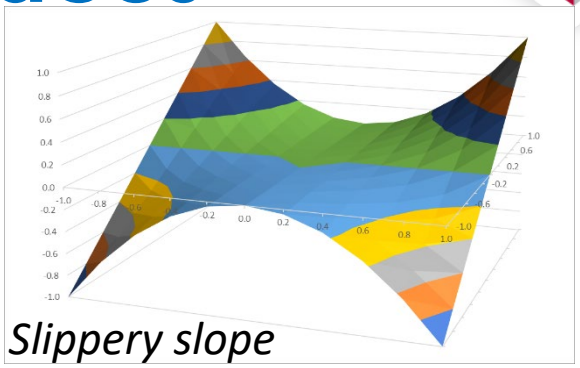
If you do not consider nonlinear relationships such as these, ***and they exist in the data***, <u>then the model will be biased</u>

Polynomial Response Surfaces

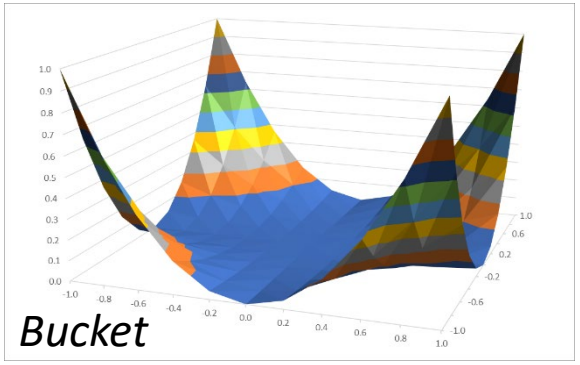
$$Z = X + Y$$



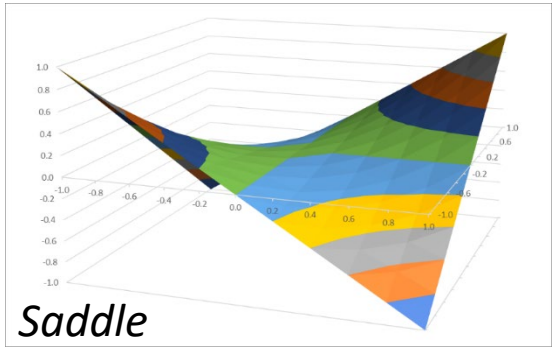
$$Z = X \cdot Y^2$$



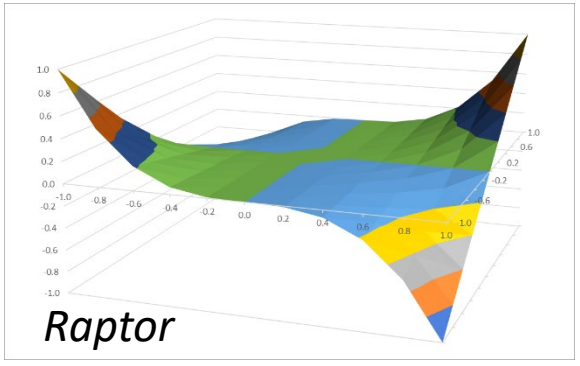
$$Z = X^2 \cdot Y^2$$



$$Z = X \cdot Y$$



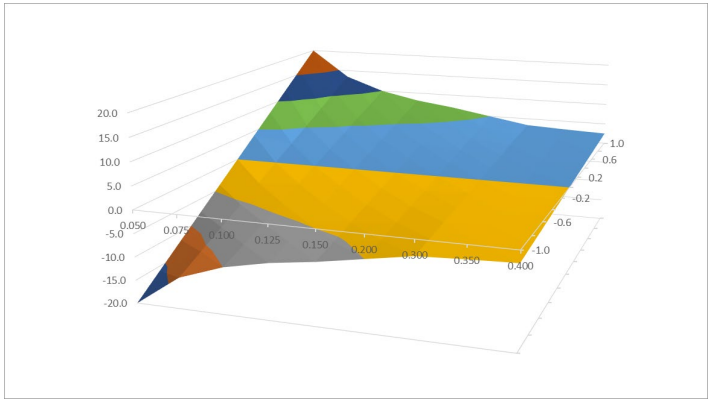
$$Z = X \cdot Y^3$$



Response Surfaces Involving Ratios

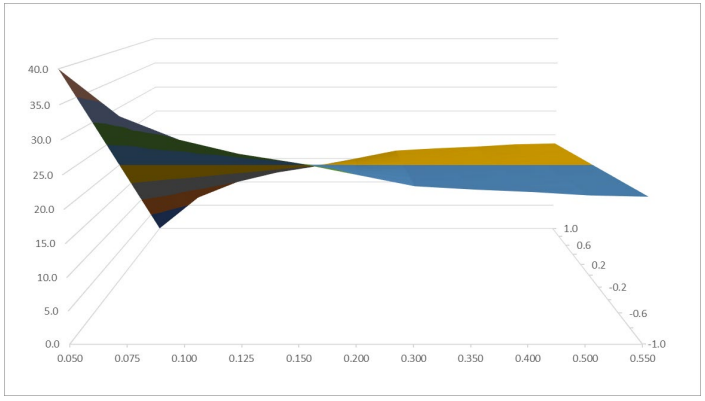
Simple ratio

$$Z = X \cdot Y^{-1}$$



Saturation

$$Z = a - X \cdot Y^{-1}$$



Example Multivariate Polynomial Model

$$Y = a_0 + a_1Q + a_2R + a_3S$$

$$+ a_4QR + a_5QS + a_6RS + a_7QRS$$
$$+ a_8Q^2 + a_9R^2 + a_{10}S^2$$

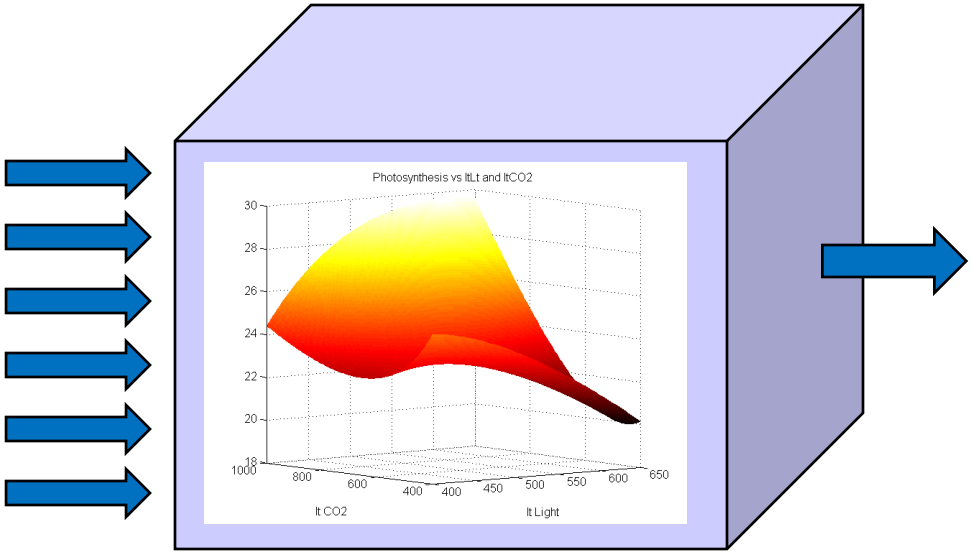
$$+ a_{11}Q^2R + a_{12}Q^2S + a_{13}Q^2R^2 + a_{14}Q^2S^2$$
$$+ a_{15}R^2S + a_{16}R^2S^2 + a_{17}QR^2 + a_{18}QS^2 + a_{19}RS^2$$

$$+ a_{20}Q^2RS + a_{21}Q^2R^2S + a_{22}Q^2RS^2 + a_{23}QR^2S^2$$
$$+ a_{24}QR^2S + a_{25}QRS^2 + a_{26}Q^2R^2S^2$$

We need a way to limit model complexity

- **STEP-WISE REGRESSION**

www.TaylorFit-RSA.com



- Accurate
- Transparent
- Transportable
- Unbiased
- Tractable
- Resistant to overfitting

Another way to ensure conciseness: Cross-validation

- Dual criteria for adding terms:
  - $p(t_{stat}) \leq \alpha$
  - Selected **GoF statistic** *for cross-validation file* associated with the candidate term improved over current model

Final Validation – The Gold Standard for Modeling

- Reserve a third partition of the dataset, called the **validation dataset**

Multivariate Polynomial Regression

TaylorFit Response Surface Analysis - with stepwise Multivariate Polynomial Regression

Settings

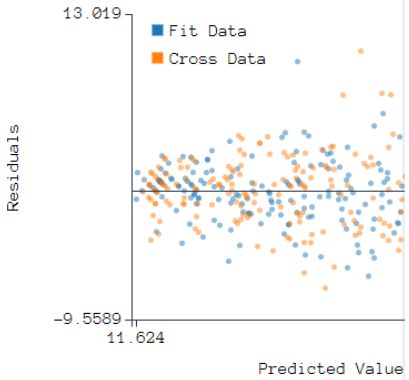
Current Model

Term	t	p(t)
-886.50	-4.2943	2.7922e-5
+6.5598(YR)	4.9124	1.9351e-6
-7.8022e-5(WT)(YR)	-2.9289	0.0038
-0.0668(HP)(ORIGIN)	-5.0127	1.2248e-6
+1.6814e-5(HP)(WT)	3.5751	4.4399e-4
+0.0028(WT)(ORIGIN)	5.6056	7.2234e-8
+30896$(YR)^{-1}$	3.9927	9.3291e-5
+2.9060e+6$(WT)^{-1}(YR)^{-1}$	3.7324	2.5050e-4

Export to Code

Goodness of Fit Statistics

Stat	Fit	Cross
nd	198	195
np	8	8
SSE	1572.9	1405.2
TSS	11960	11861
SSR	10387	10455
MSE	8.2782	7.5143
Rsq	0.8685	0.8815
adjRsq	0.8636	0.8771
Max|Err|	11.137	10.309
RMSE	2.8772	2.7412
SKEW	0.8535	0.3129
XKURT	2.4853	1.5959
AIC	0.9987	0.9579
BIC	1.0107	0.9698
F	179.25	198.77
p(F)	0.0000	0.0000
s.e. SKEW	0.1741	0.1754
s.e. XKURT	0.3482	0.3508



1 2 3 4 5 › 21 **Candidate Terms**

Term	MSE	t	p(t)
Current Model	7.5143		
-0.0032(HP)(ACCEL)	7.4427	-2.3533	0.0196
-0.1546(ACCEL)(ORIGIN)$^{-1}$	7.4444	-1.3104	0.1916
0.0174(WT)(ACCEL)$^{-1}$	7.4485	1.7873	0.0755
0.5345(ACCEL)$^{-1}$(YR)	7.4522	1.3044	0.1937
24.284(ACCEL)$^{-1}$(ORIGIN)$^{-1}$	7.4529	0.8024	0.4233
0.1944(HP)(ACCEL)$^{-1}$	7.4536	0.8871	0.3761
47.006(ACCEL)$^{-1}$	7.4552	1.5270	0.1284
0.0058(ACCEL)$^{-1}$(YR)^{2}	7.4574	1.0732	0.2846
6.0625e-6(WT)^{2}(ACCEL)$^{-1}$	7.4583	2.4331	0.0159
-0.0049(ACCEL)^{2}(ORIGIN)$^{-1}$	7.4621	-1.2035	0.2303

1 2 3 4 5 › 20 Export Data auto-mpg.csv

	MPG	CYL	DISP	HP	WT	ACCEL	YR	ORIGIN	Dependent	Predicted	Residual
1	26.6	4	151	84	2635	16.4	81	1	26.600	28.701	-2.1014
2	16	8	400	170	4668	11.5	75	1	16.000	13.424	2.5765
3	25	4	90	71	2223	16.5	75	2	25.000	27.423	-2.4228
4	18.1	6	258	120	3410	15.1	78	1	18.100	19.818	-1.7175

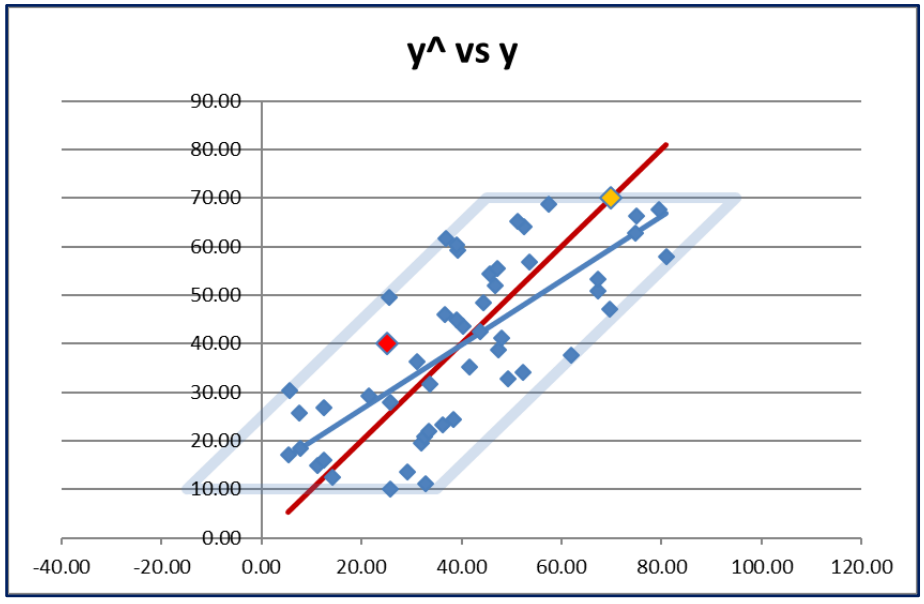
Graphical Examination of Fit

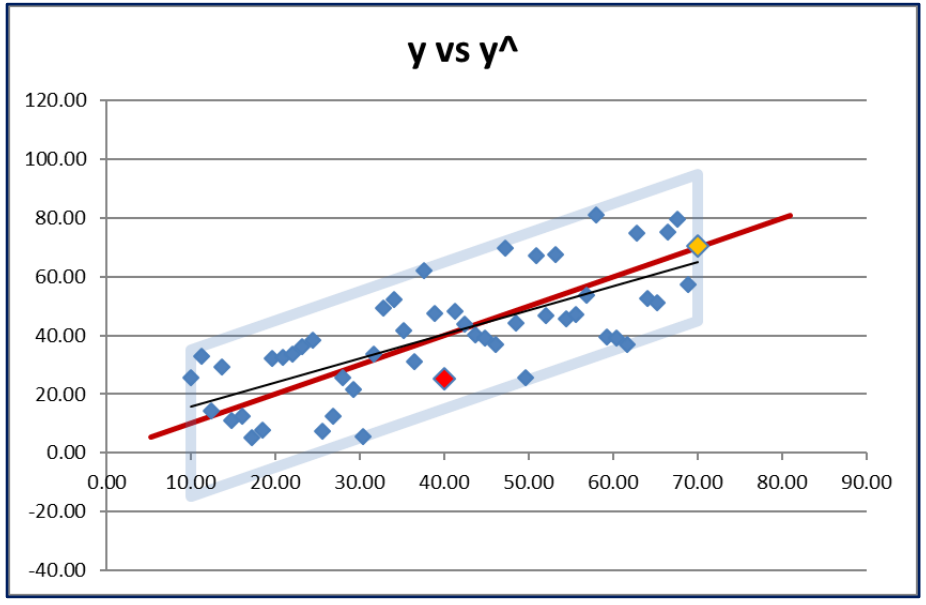
Plot predicted versus observed?

Never!

Always plot:

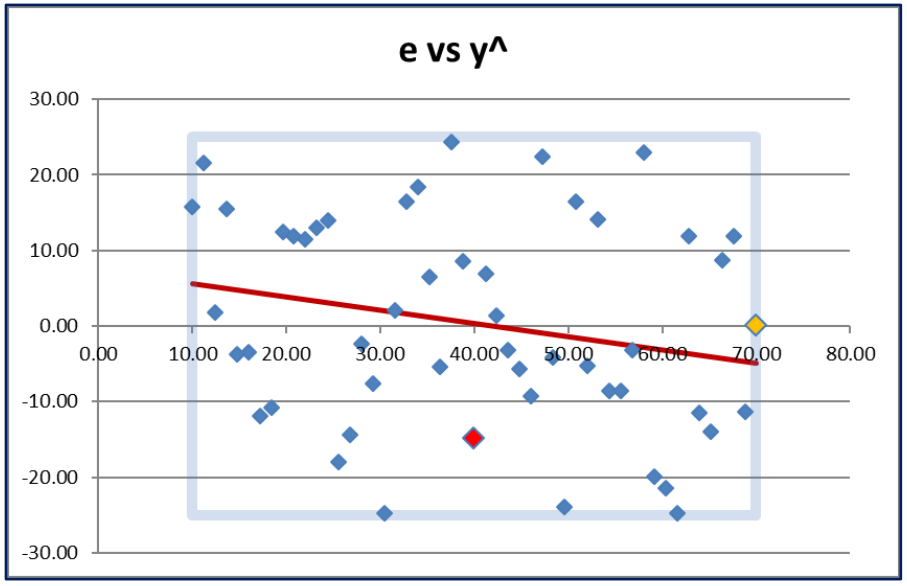
Observed versus **Predicted**

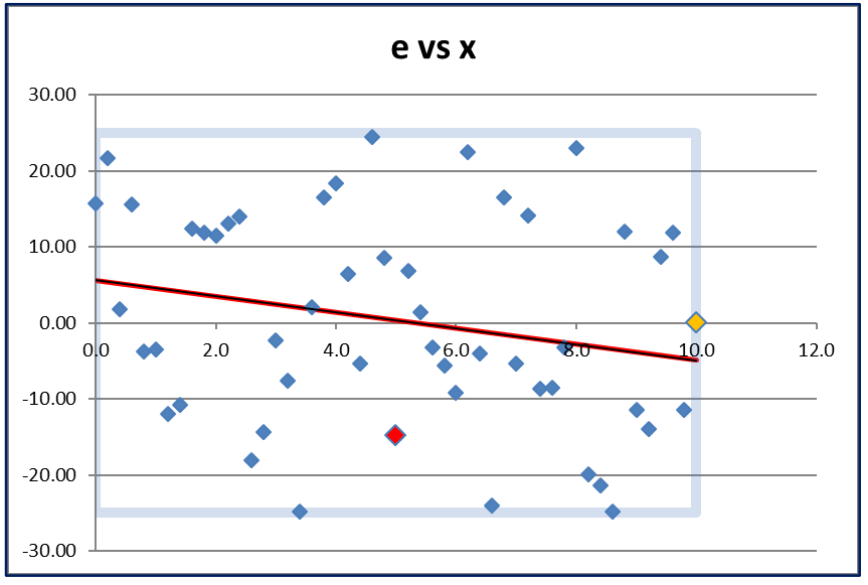




Examining Residuals

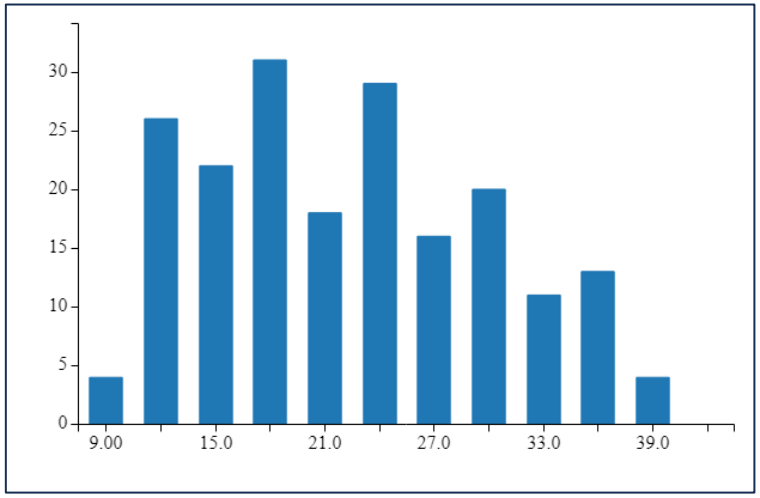
Even better: Plot residuals instead of predictions:
versus (all) data, by sequence, by time, etc.

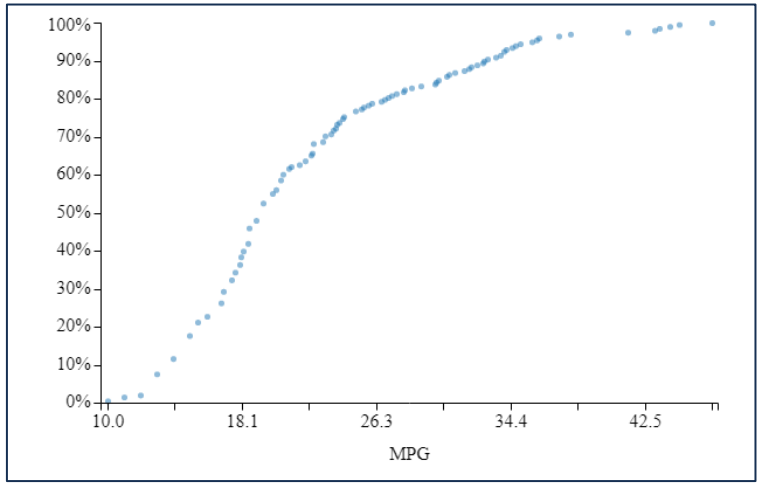


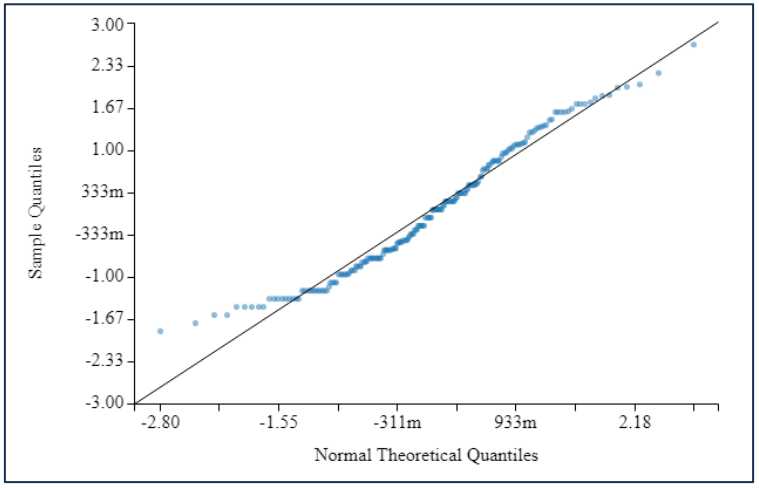


Residuals plot should appear random - no trends

Histogram, CFD, Q-Q Plot







Model Analysis

- Sensitivity
- Sensitivity distribution
- Importance ratio
- Confidence intervals

Nonlinear Local Sensitivity

Sensitivity is not constant.
So how do we represent it?

As a distribution,
based on the data:

Compute sensitivity for each datum (row)

*Mean, std. dev., **rms***
histogram, CFD, etc.

Dependence of sensitivity on units

$$MPG\left(\frac{mi}{gal}\right) = 45.6 - 0.0074 \cdot WT(lbs)$$

$$\psi_{WT} = -0.0074$$

$$MPG\left(\frac{mi}{gal}\right) = 45.6 - 0.0164 \cdot WT(kg)$$

$$\psi_{WT} = -0.0164$$

Sensitivity depends on the measurement units,
so sensitivity of different variables cannot be compared with each other

Importance ratio

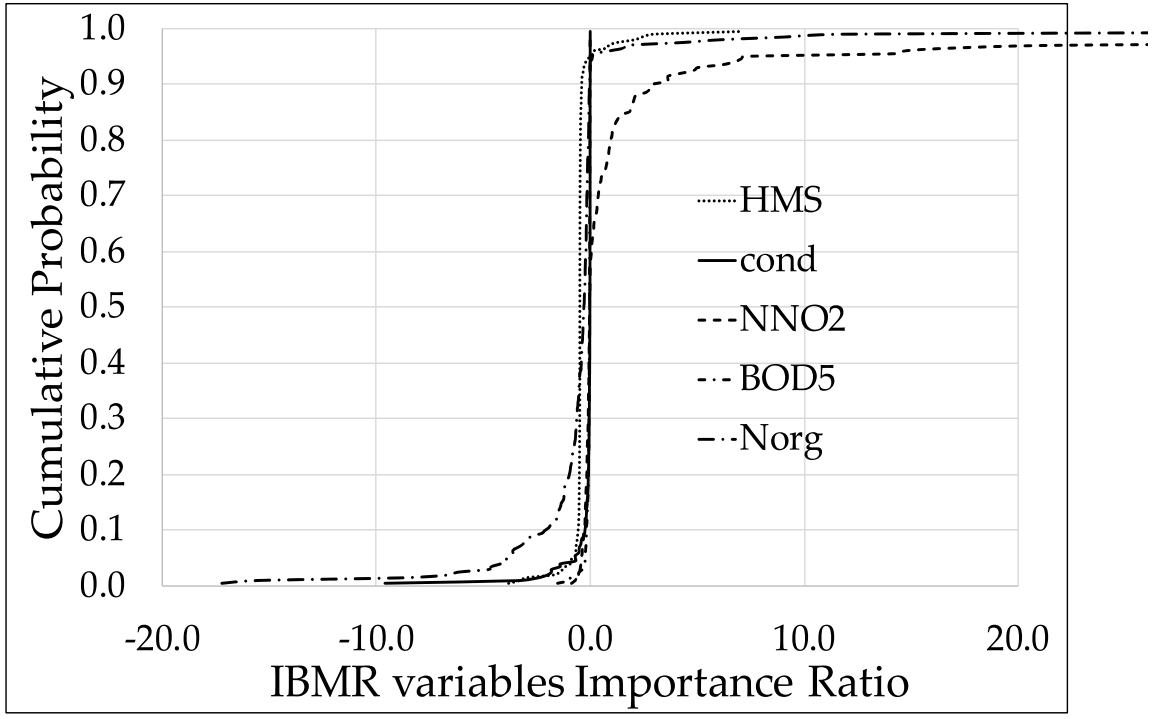
Importance ratio is a normalized sensitivity:

$$IR_{x_i} = \frac{\partial y / s.e._y}{\partial x_i / s.e._{xi}}$$

Importance ratio is dimensionless
so different I.R.'s can be compared.

Interpretation: A measure of
how much a change in x_i relative to its spread
changes y relative to its spread

Importance Ratio - Distribution



Prediction Intervals and Confidence Intervals

For Prediction: Confidence Intervals
Indicates the expected spread of the prediction

Predicted value ± C.I. (at, say, 95% confidence level)

For Observations: Prediction Intervals
Indicates the expected spread of a measured observation

Observed value ± P.I. (at, say, 95% confidence level)

Examples

- Macrophyte Indices
- Phytoplankton occurrence in the Hudson and East Rivers
- Photosynthetic productivity for NASA
- Performance of a Retail Industrial Equipment and Supply outlet

Macrophyte Index Models

Index	Equation
MIR	$3.34780 \times 10^{1} + 1.7305 \times 10^{0}(Ptot)^{-1}$
RMNI	$8.2576 \times 10^{0} - 6.2496 \times 10^{-4}(HQA)^{2}(NNO3)(Ntot)^{-1} - 1.416 \times 10^{-1}(alkal)^{-1}(Ptot)^{-1}(NNO2)^{-1} + 1.1 \times 10^{-3}(cond)(alkal)^{-1}(NNO2)^{-1} - 1.5393 \times 10^{-4}(Ptot)(NNH4)^{-1}O2^{2} + 2.2185 \times 10^{-5}(Ptot)^{-1}(NNO3)^{-1}(NNO2)^{-1}$
IBMR	$9.5335 \times 10^{0} - 2.30 \times 10^{-2}(Ptot)^{-1}(NNO2)(BOD5) + 3.226 \times 10^{-1}(Ptot)^{-1}(Norg)^{-1} - 1.4 \times 10^{-3}(alkal)(Ptot)^{-1}(Ntot)^{-1} + 1.7618 \times 10^{6}(HMS)^{-2}(cond)^{-2}(NNH4)^{-1} - 7.72 \times 10^{-2}(HMS)^{-2}(NNO2)^{-2}(Norg)^{-3}$
D	$6.806 \times 10^{-1} - 3.5578 \times 10^{-6}(HQA)^{-1}(HMS)^{2}(O2)^{2} - 8.8749 \times 10^{-6}(PPO4)^{-2}(NNH4)^{-1}(BOD5)^{-2} - 3.88 \times 10^{-2}(HMS)^{-1}(NNO3)^{-1}(BOD5)^{2} + 1.0894 \times 10^{-5}\ (HQA)^{-2}(cond)^{2}(NNO3)^{-1}$
N	$3.2233 \times 10^{1} - 9.30 \times 10^{-2}(HMS) - 9.4650 \times 10^{1}(pH)(alkal)^{-1} - 3.438 \times 10^{-1}(HQA)^{-1}(alkal) - 3.1979 \times 10^{2}(cond)^{-1}(BOD5) - 1.9307 \times 10^{0}(HMS)^{-1}(NNO3)(NNO2)^{-1} + 1.2670 \times 10^{-4}(NNO2)^{-1}(NNH4)^{-1}(Ntot) - 9.64 \times 10^{-2}(cond)(NNO2)(O2)^{-1}$

Indices of trophic and ecological status of the rivers:

Macrophyte Index for Rivers (MIR)

River Macrophyte Nutrient Index (RMNI)

Macrophyte Biological Index for Rivers (IBMR)

Biological diversity indices:

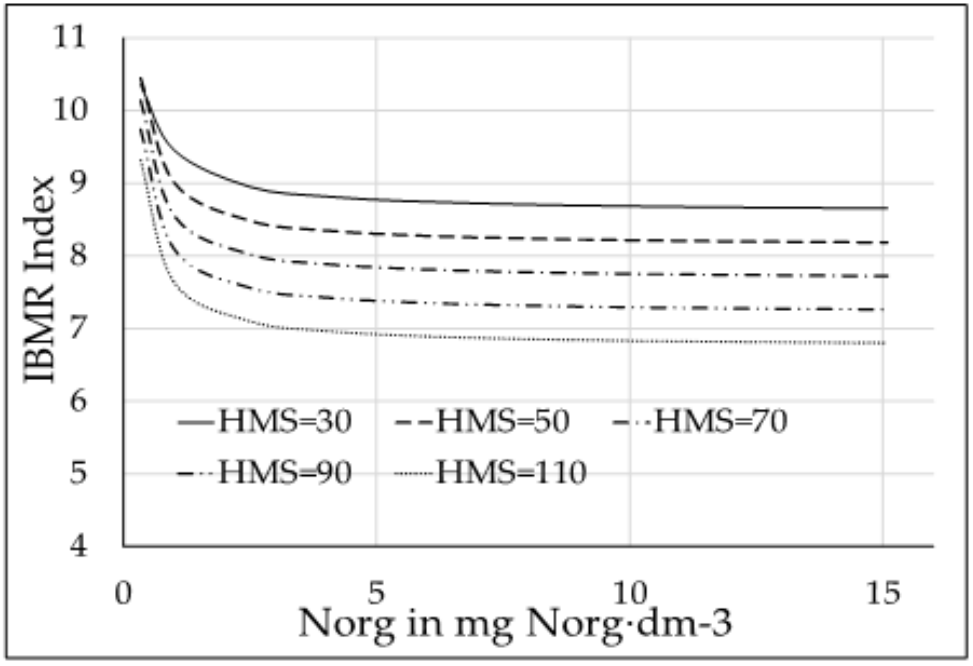
Simpson index (D)

Species richness (N)

Vishwa Shah, Sarath Chandra K. Jagupilla *, David A. Vaccari, Daniel Gebler (2021). Non-Linear Visualization and Importance Ratio Analysis of Multivariate Polynomial Regression Ecological Models based on River Hydromorphology and Water Quality. Section: Ecohydrology, *Water* 2021, 13, 2708. https://doi.org/10.3390/w13192708.

Nonlinear Behavior of IBMR

$$
\begin{aligned}
IBMR = \; & 9.5335 \times 10^{0} \\
& -2.30 \times 10^{-2}(Ptot)^{-1}(NNO2)(BOD5) \\
& +3.226 \times 10^{-1}(Ptot)^{-1}(Norg)^{-1} \\
& -1.4 \times 10^{-3}(alkal)(Ptot)^{-1}(Ntot)^{-1} \\
& +1.7618 \times 10^{6}(HMS)^{-2}(cond)^{-2}(NNH4)^{-1} \\
& -7.72 \times 10^{-2}(HMS)^{-2}(NNO2)^{-2}(Norg)^{-3}
\end{aligned}
$$



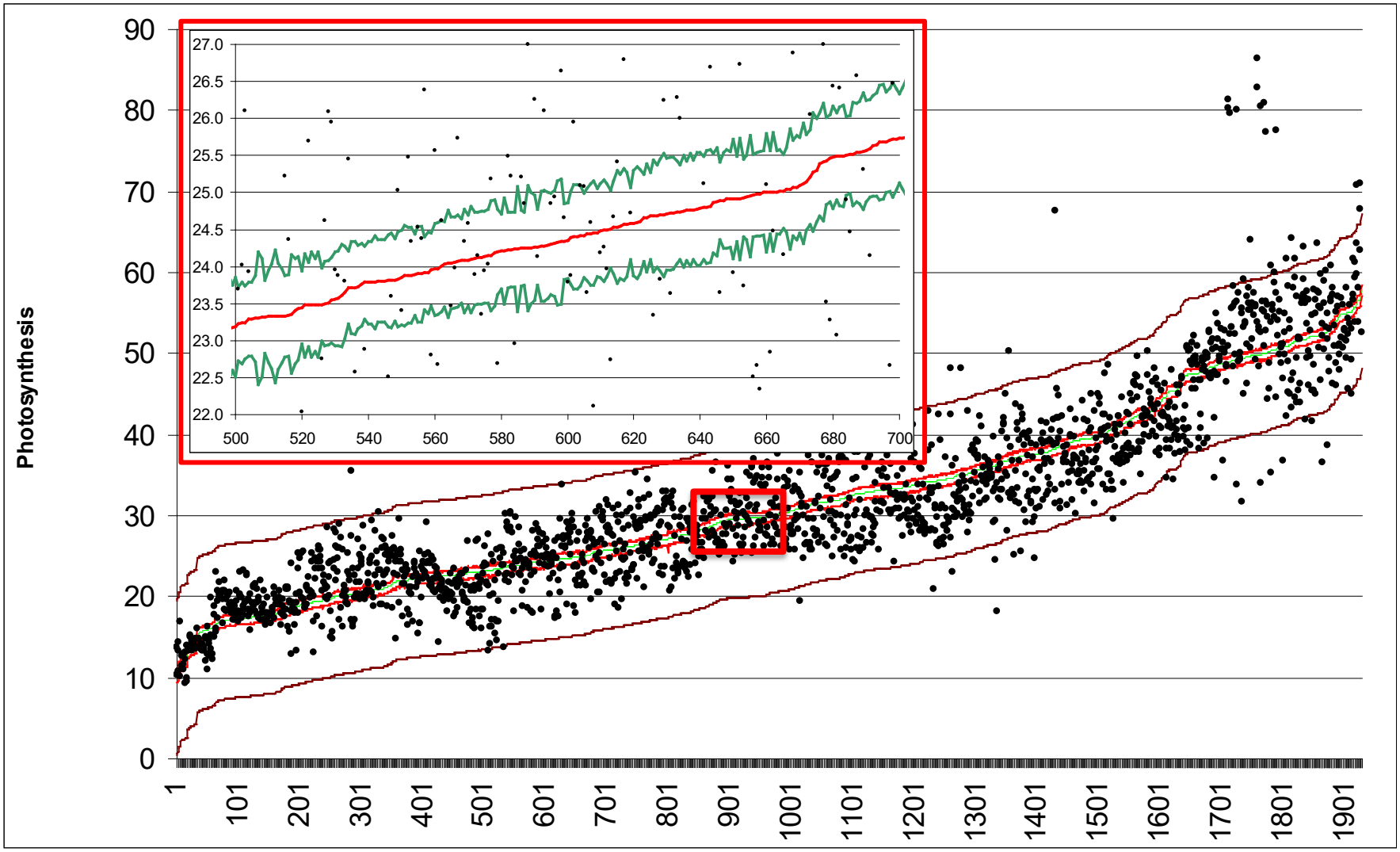
IBMR vs organic nitrogen with river morphology as a parameter.
(HMS – River Hydromorphological Index)

GoF Performance of MPR vs. ANNs

	Type of Model	Validation R^2	Degrees of Freedom (df) for Error	Mean Square Error (MSE)	F-Stat	p(F)
MIR	MPR	0.58	12	10.856	1.11	0.373
	ANN	0.702	52	9.790		
RMNI	MPR	0.65	9	0.048	1.00	0.452
	ANN	0.715	52	0.050		
IBMR	MPR	0.47	9	0.364	0.88	0.551
	ANN	0.532	52	0.411		
D	MPR	0.237	8	0.008	1.00	0.447
	ANN	0.284	52	0.009		
N	MPR	0.33	11	8.392	0.90	0.544
	ANN	0.415	52	9.288		

$p(F) > 0.05$ means the difference between MPR and ANN model accuracy was not statistically significant

Experimental Photosynthesis Data, Prediction, Confidence Intervals for Observations and Predictions

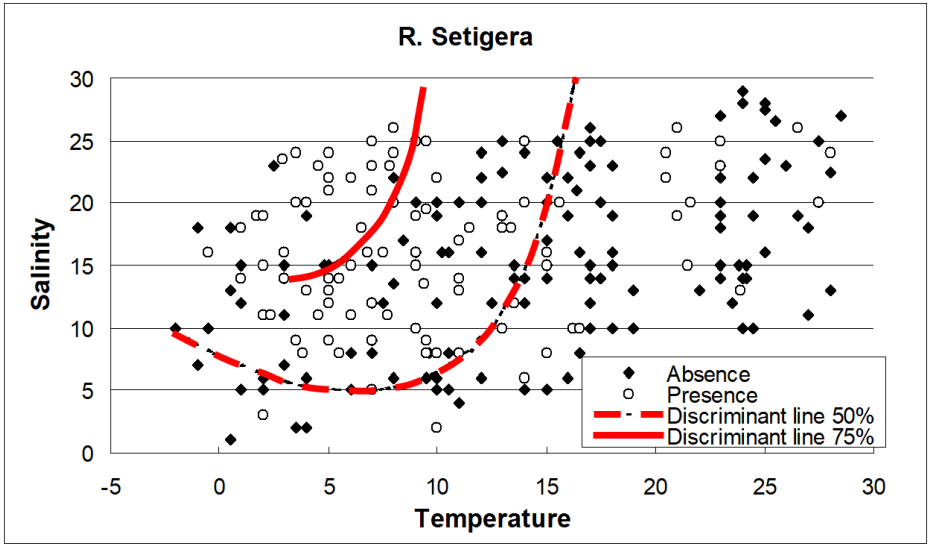


Classification model for presence of phytoplankton

- Presence or absence of 3 phytoplankton species (coded variable)

- Salinity
- Water temperature
- Dissolved oxygen
- pH
- Secchi disk depth
- Percent cloud cover
- Tidal stage
- Wind

$$y = 1.17 \cdot 10^{-2} \cdot pH^2 - 9.74 \cdot 10^{-3} \cdot T \cdot DO + 8.8 \cdot 10^{-4} \cdot T \cdot DO^2 + 2.37 \cdot 10^{-4} \cdot S \cdot DO^2 + 5.3 \cdot 10^{-4} \cdot pH \cdot DO^2 - 0.34$$



Z. Wang, D.A. Vaccari, M. Levandowsky. Multivariate Polynomial Modeling of Phytoplankton in the Lower Hudson River. Platform presentation, American Society of Limnology and Oceanography, summer meeting, Savannah Georgia, 2004.

Performance of Retail Industrial Sales Outlet

$$GM = 3.72 + 0.56 \cdot SE - \frac{1}{15.5 \cdot M \cdot Y} + \frac{93.45}{S \cdot C} + \frac{1}{81.7 \cdot Y \cdot W} + \frac{Q}{69.2}$$

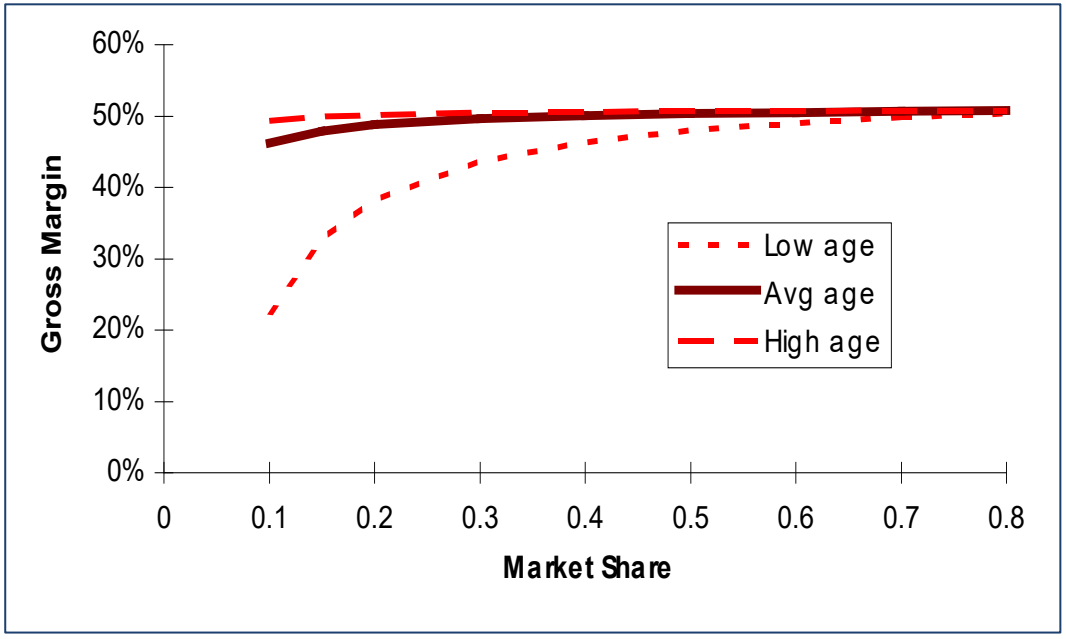
$R^2 = 0.82$ (82%)

Variables used:

- Supply/equipment ratio
- Market share
- Years at location
- Sales
- Competition intensity
- Walk-in/delivery ratio
- Quality

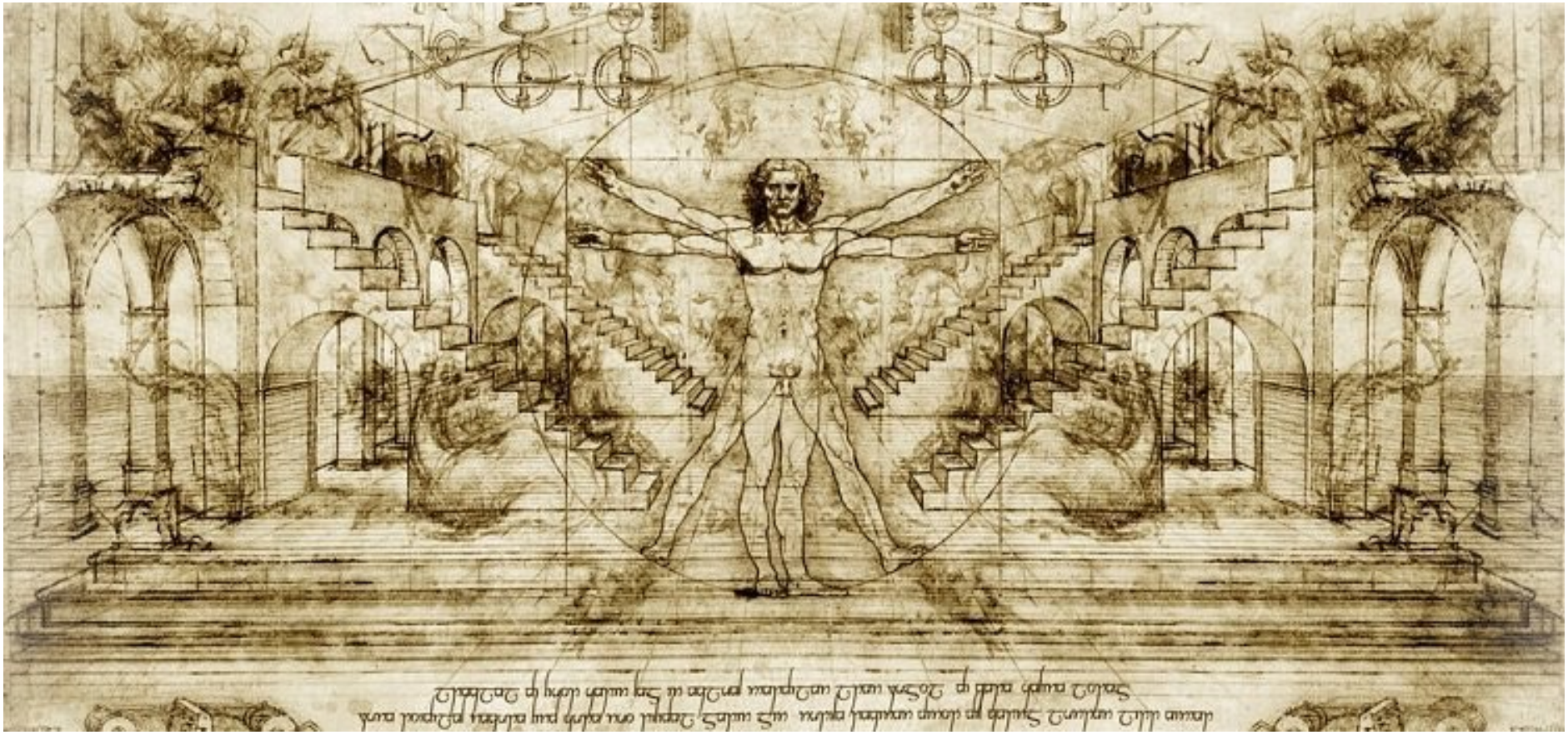
Variables not significant:

- Industrial concentration
- Rural/Urban location
- Highway visibility
- # of staff
- Average years of service
- # of salespersons



Effect of *Market Share* with *Years at Location* as a parameter

Modeling is Art as well as Science



The Vitruvian Man – Leonardo Da Vinci

https://www.leonardodavinci.net/

Thank you

www.TaylorFit-RSA.com

Information:

dvaccari@stevens.edu

Thank you to our Patrons

AMERICAN ACADEMY OF ENVIRONMENTAL ENGINEERS & SCIENTISTS®

LA DWP Los Angeles Department of Water & Power

BLACK & VEATCH

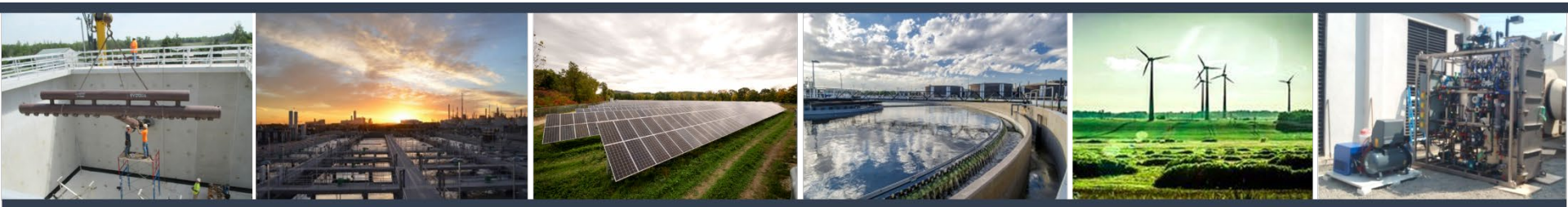
Geosyntec consultants

THE METROPOLITAN WATER DISTRICT OF SOUTHERN CALIFORNIA

DSWA DELAWARE SOLID WASTE AUTHORITY

LOS ANGELES COUNTY SANITATION DISTRICTS
Converting Waste Into Resources

dc water is life®



Leadership and Excellence in Environmental Engineering and Science

Thank you for attending our webinar today.

AMERICAN ACADEMY OF ENVIRONMENTAL ENGINEERS & SCIENTISTS®

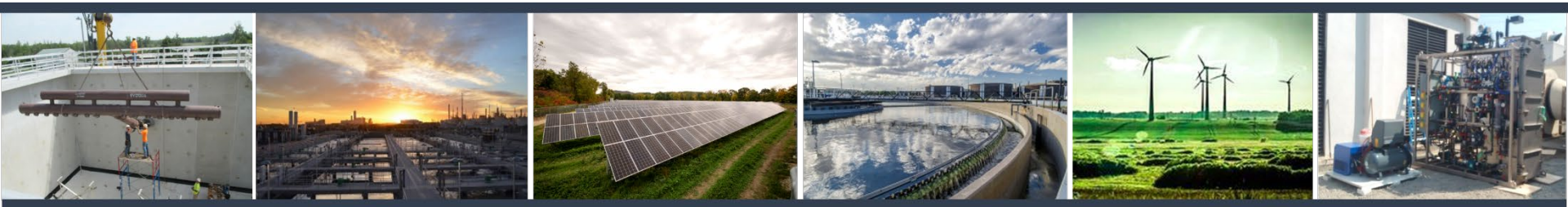
Would you like to attend our next webinar?
Join us March 2nd as Los Angeles County Sanitation District discusses their San Gabriel River Watershed Project, Balancing Water Needs of the Community and the Environment.
Go to AAEES.org to register.

Would you like to watch this webinar again?
A recording of today's event will be available on AAEES.org tomorrow.

Not an AAEES member yet?
To determine which type of AAEES membership is the best fit for you, please go to AAEES.org or email Marisa Waterman at mwaterman@aaees.org

Need a PDH Certificate?
You will be emailed a PDH Certificate for attending this webinar within two weeks.

Questions?
Email Marisa Waterman at mwaterman@aaees.org with any questions you may have.



Bibliography

Users' Manual for *TaylorFit* Response Surface Analysis Using Multivariate Polynomial Regression. (2019) www.TaylorFit-RSA.com.

Non-time-series:

Vaccari, D.A., Nonlinear Analysis of Retail Performance, *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering* (CIFEr), pp252-258, New York, NY, March 24-26, 1996.

Vaccari, D.A. and J. Levri, "Multivariable Empirical Modeling of ALS Systems Using Polynomials," *Life Support and Biosphere Science*, vol. 6 pp. 265-271, (1999).

Jagupilla, S.J.K., D.A. Vaccari; R.I. Hires, "Multivariate Polynomial Time-Series Models and Importance Ratios to Identify Fecal Coliform Sources," *ASCE Journal of Environmental Engineering*, v136, n7, pp657-665 (2010).

V. Shah, S.C. K. Jagupilla *, D. A. Vaccari, D. Gebler (2021). Non-Linear Visualization and Importance Ratio Analysis of Multivariate Polynomial Regression Ecological Models based on River Hydromorphology and Water Quality. Section: Ecohydrology, *Water* 2021, 13, 2708. https://doi.org/10.3390/w13192708.

Z. Wang, D.A. Vaccari, M. Levandowsky. Multivariate Polynomial Modeling of Phytoplankton in the Lower Hudson River. American Society of Limnology and Oceanography, summer meeting, Savannah Georgia, 2004.

Time-series:

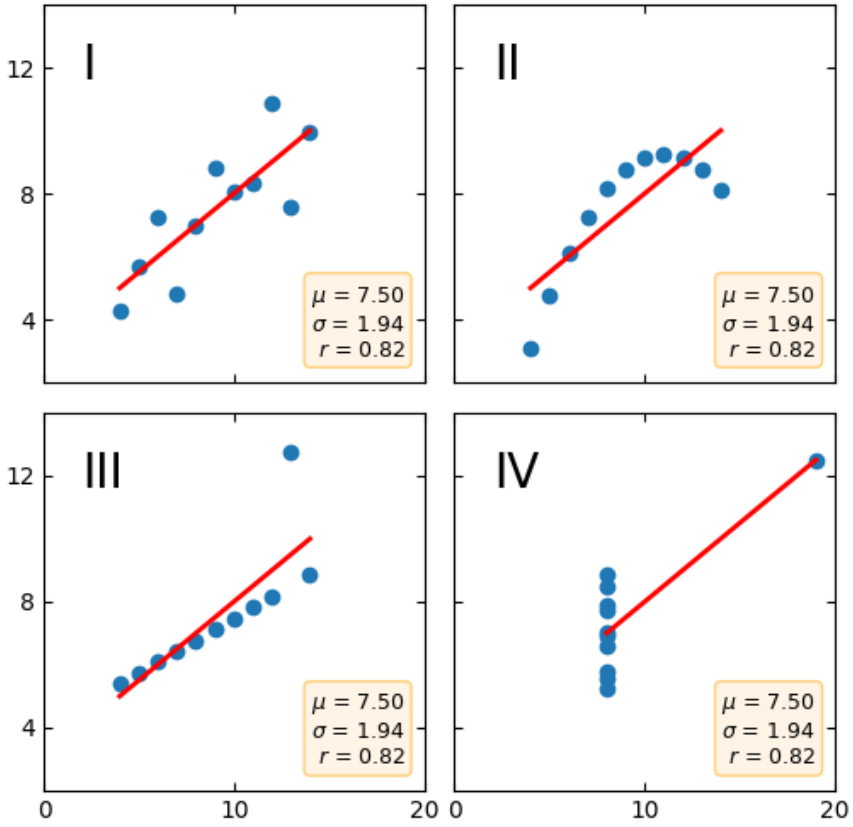
Vaccari, D. A. and H.K. Wang, "Multivariate polynomial regression for identification of chaotic time series," *Mathematical and Computer Modelling of Dynamical Systems*, v13, n4, pp395–412 (2007).

Watson, R., Nieradka, B. and Vaccari, D. A. (2017), Empirical Dynamic Material Flow Model for Tungsten in the USA. *Journal of Industrial Ecology*. doi:10.1111/jiec.12555.

Empirical Modeling Pitfalls

- Overfitting
- Spurious correlation
- Polynomial wiggle
- Outliers
- Influential points
- Multicollinearity
- Nonlinear bias
- Explosive behavior
- Heteroscedasticity

Anscombe's Quartet



All have the same
mean, standard deviation and R^2

I – Typical regression
II – Nonlinear
III – Outlier
IV – Influential point

Shows the importance of graphical examination of the data and not just relying on statistical measures

https://matplotlib.org/3.2.1/gallery/specialty_plots/anscombe.html