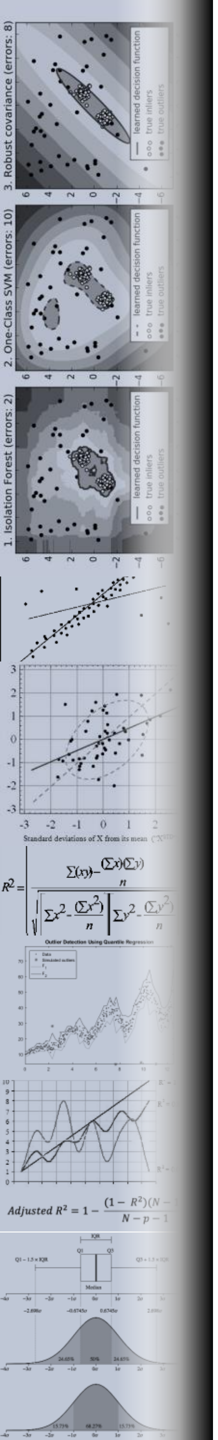
A Novel Robust R-Squared Measure and Its Applications in Linear Regression

Conference Paper *in* Advances in Intelligent Systems and Computing · October 2016

DOI: 10.1007/978-3-319-48517-1_12

Sougata Deb
National University of Singapore



A Novel Robust R-Squared Measure and Its Applications in Linear Regression

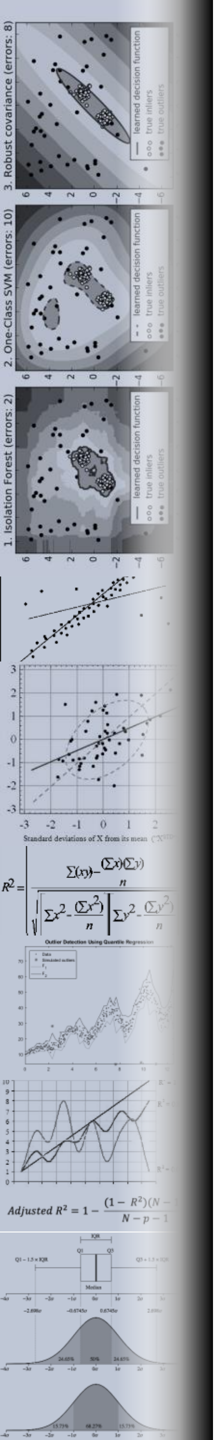
Sougata Deb

This public document is a summary presentation highlighting the key aspects covered in the original paper.

For the peer-reviewed and **published version** of the paper, please access:
https://doi.org/10.1007/978-3-319-48517-1_12

For a raw, unreviewed **pre-print version** of the paper, please let me know of your interest and I will share it privately

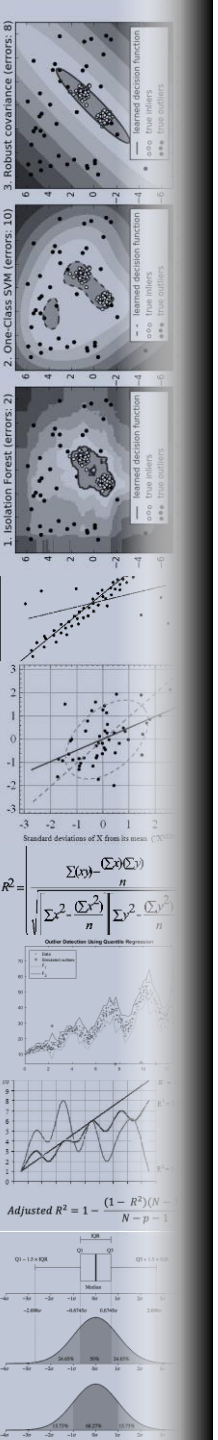
For questions, remarks and feedback, please drop a message to
deb.sougata@gmail.com



A Novel Robust R-Squared Measure

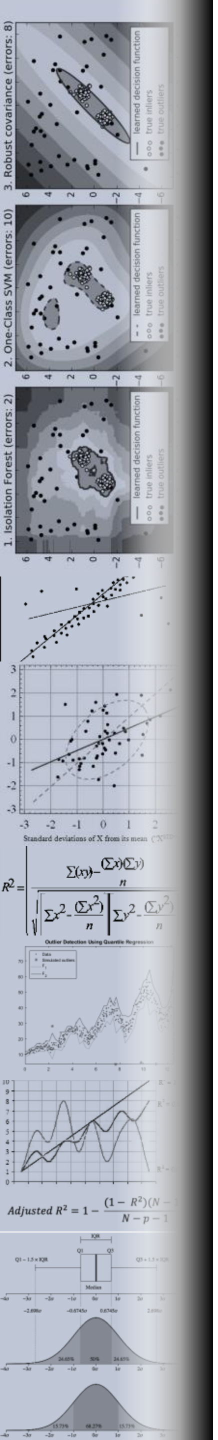
and Its Applications in Linear Regression

Sougata Deb
Analytics Professional and Independent Researcher

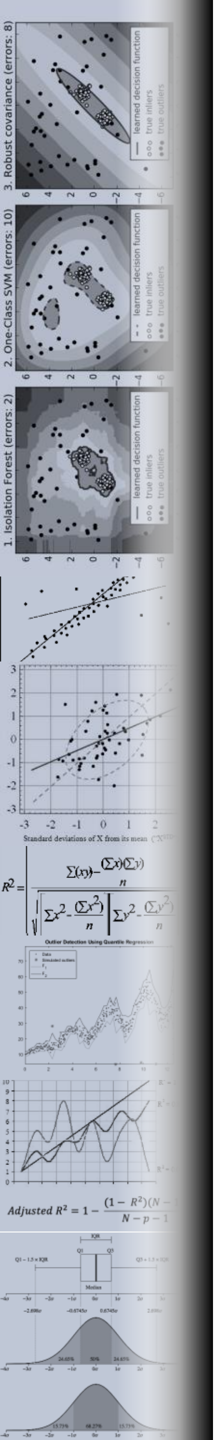


Agenda

- **Problem Overview**
  - R^2, Goodness of Fit, Robust Regression and Contamination Detection
- **Solution Process**
  - ROR^2 Computation
  - Evolution of the Main Algorithm
- **Empirical Results**
  - Results on Synthetic Data
  - Results on Real Datasets
- **Conclusions**



Problem Overview



R-Squared: Why or Why Not?

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

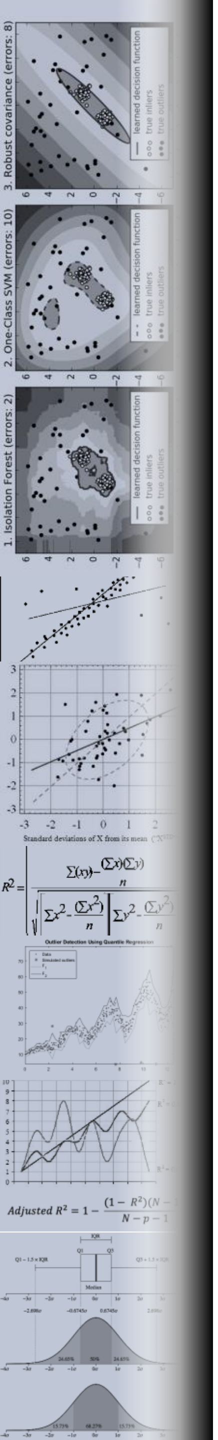
It's difficult to understand model fit using R-squared alone. Research shows that **graphs are essential** to correctly interpret regression analysis results.

In general, the higher the R-squared, the better the model fits your data. However, there are **important conditions** for this guideline.

R-squared ***cannot* determine** whether the coefficient estimates and predictions are biased …

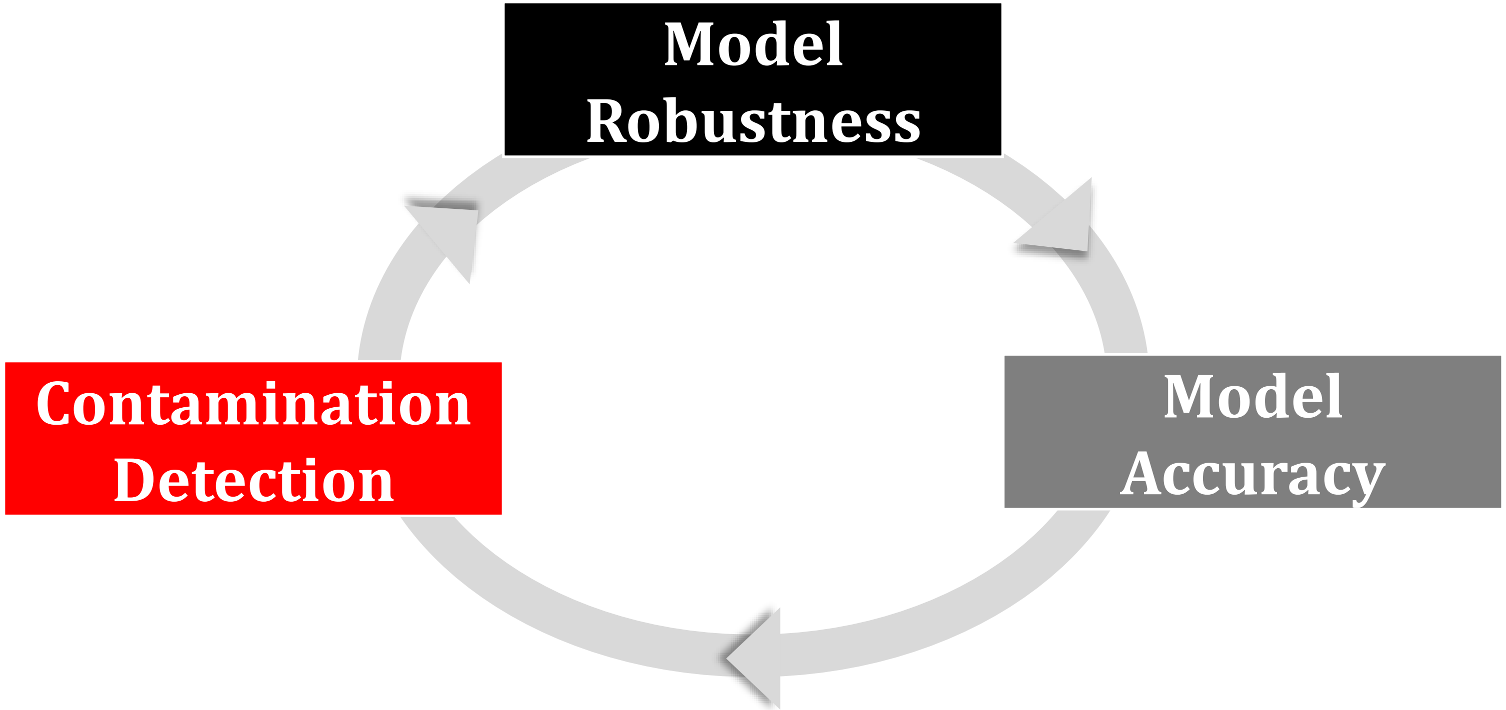
Are Low R-squared Values Inherently Bad?
No! There are **two major reasons** why it can be just fine to have low R-squared values.

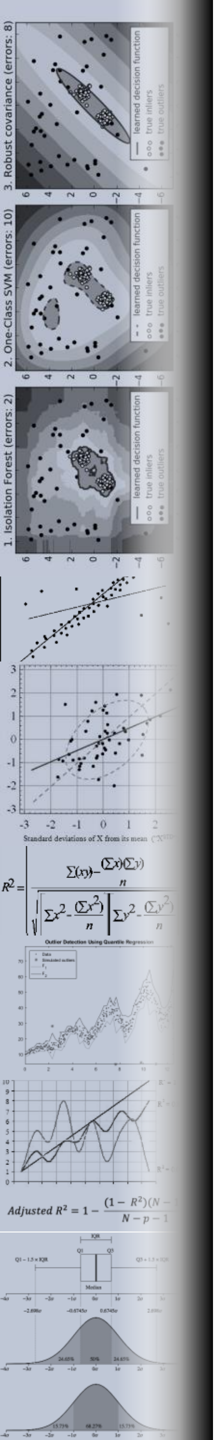
I also showed how it can be a **misleading statistic** because a low R-squared isn't necessarily bad and a high R-squared isn't necessarily good.



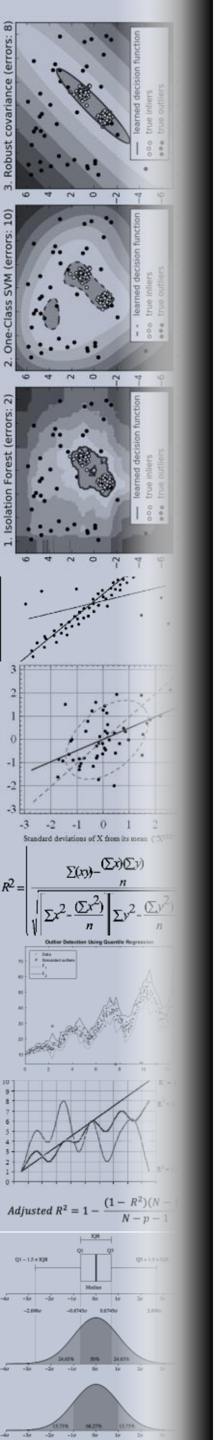
Problem Scope

- Only 1 underlying process generating the data
  - Everything else is contamination: outliers or leverage points





Solution Process



Basic Architecture

- If a model captures this underlying process well
  - **Low error** on *good / regular* observations
  - **High error** on *bad / contaminated* observations

- In such a scenario, we should expect

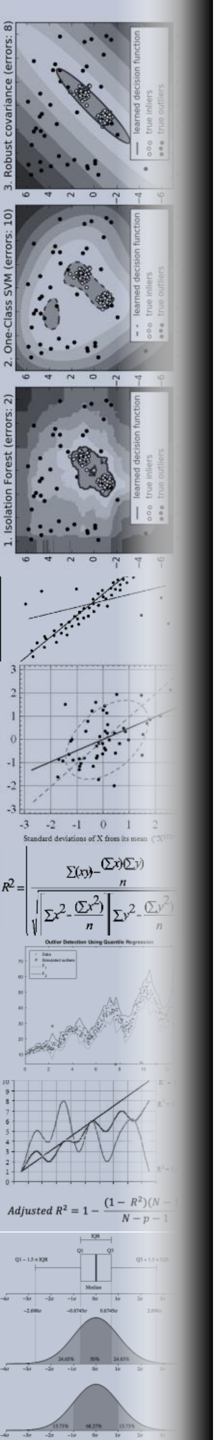
$$\{\, e^2_{(1)}, e^2_{(2)}, \dots, e^2_{(m)}, e^2_{(m+1)}, e^2_{(m+2)}, \dots, e^2_{(n)} \,\}$$

Regular observations ← → **Contaminated Observations**

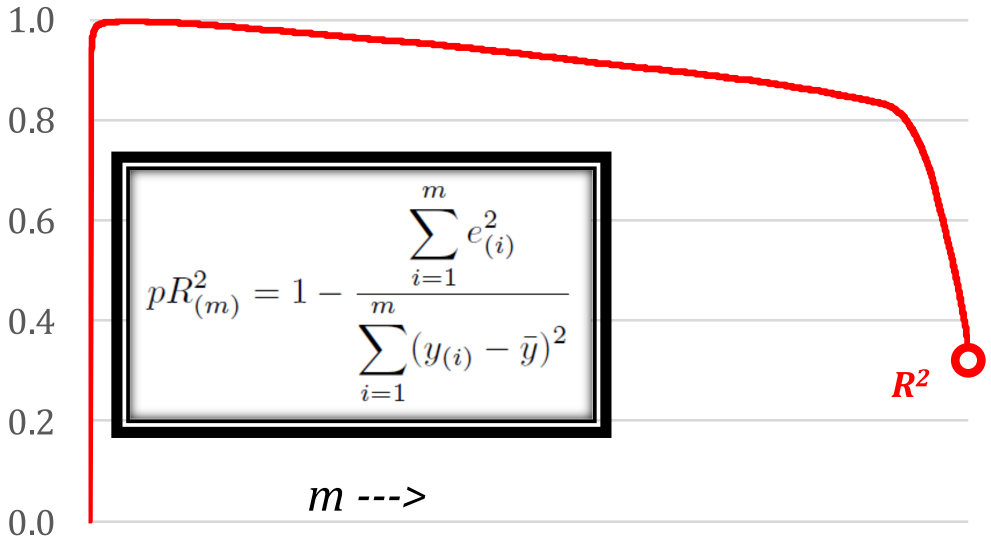
where, $e^2_{(1)} \le e^2_{(2)} \le \cdots \le e^2_{(n)}$

Let $\{\, y_{(1)}, y_{(2)}, \dots, y_{(n)} \,\}$ be the corresponding actuals

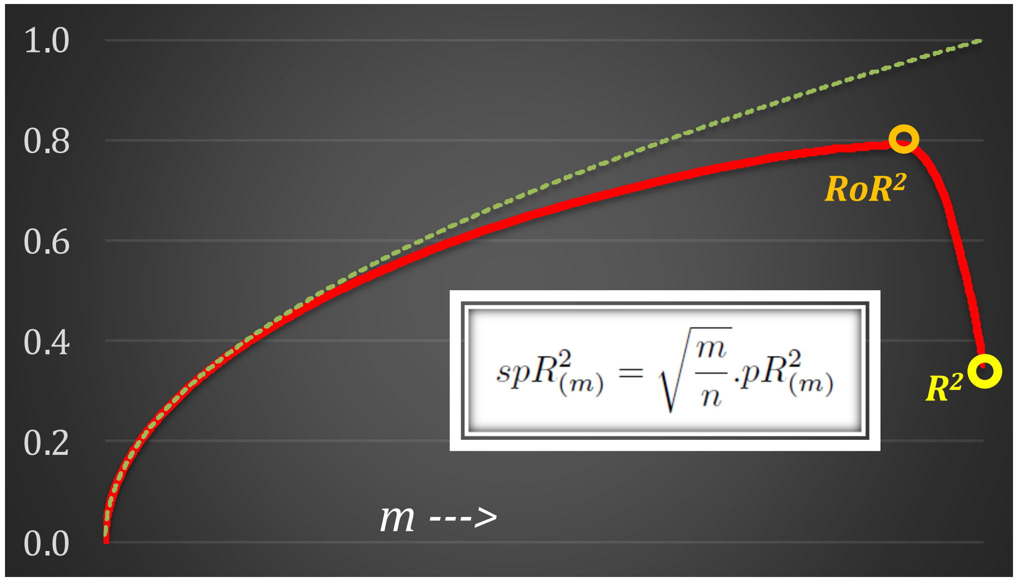
Defining **Progressive** Metrics



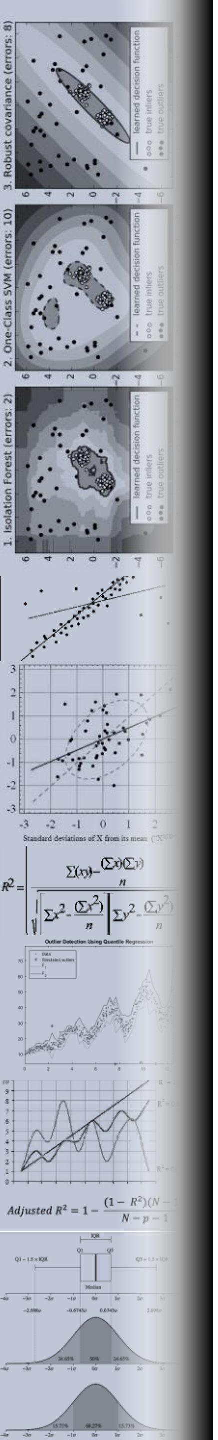
Progressive R^2



Scaled-Progressive R^2

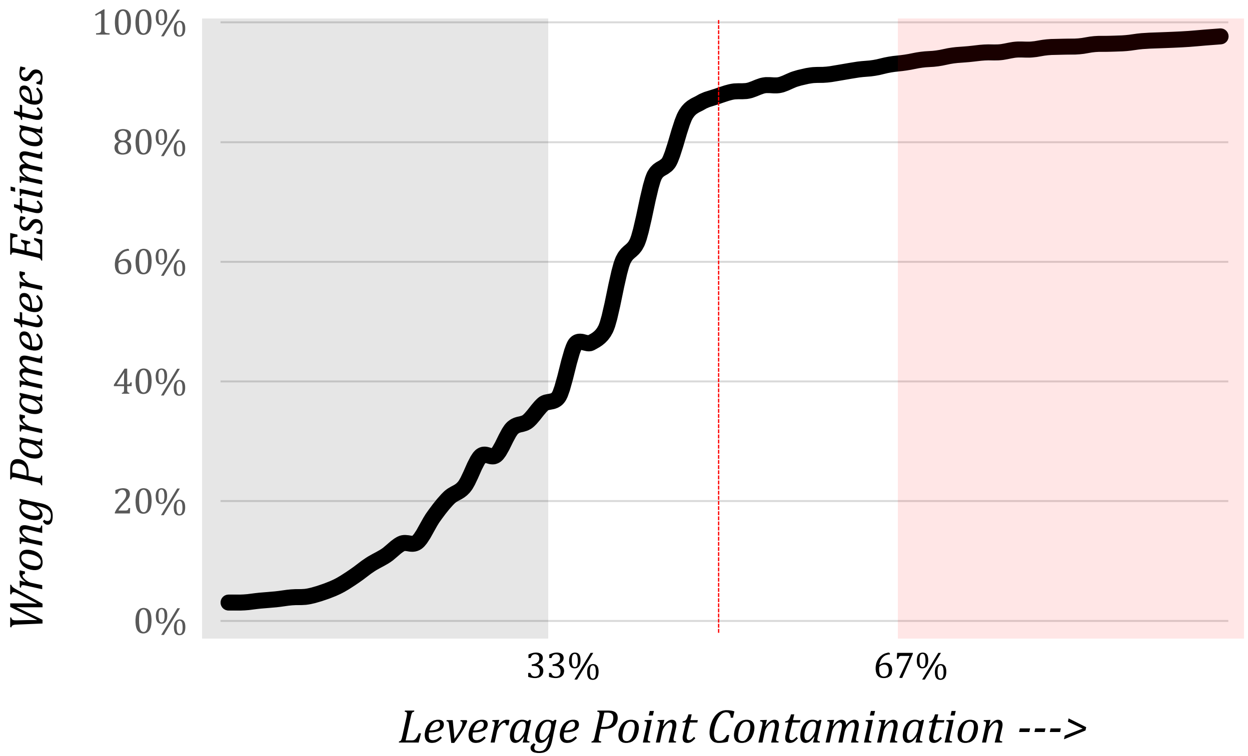


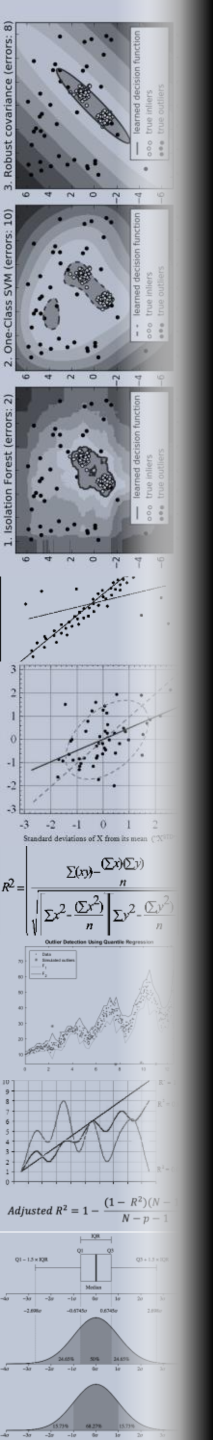
- Take out the outliers, then calculate the R^2
- *"Take out the outliers"* => identify the *true* underlying process / relationship
- *True* relationship is independent of any particular model being evaluated



Trial 1: OLS initialization

- Start with OLS estimates, drop observations beyond *argmax* RoR^2
- Re-estimate regression parameters and repeat the last step
- **Finding:** it works for outliers, but gets trapped by leverage points





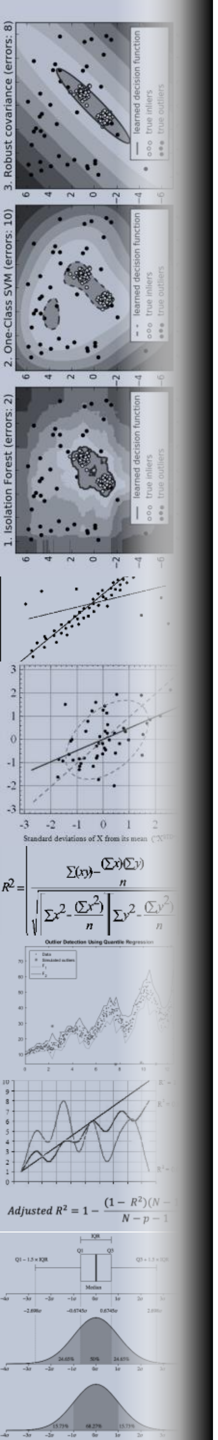
Trial 2: Committee of OLS Initializations

- Apply the same methodology on multiple subsets of the original data

- Check 1: Subset Creation
  - No resampling, partition into k subsets
  - No random partitioning, using distance from median

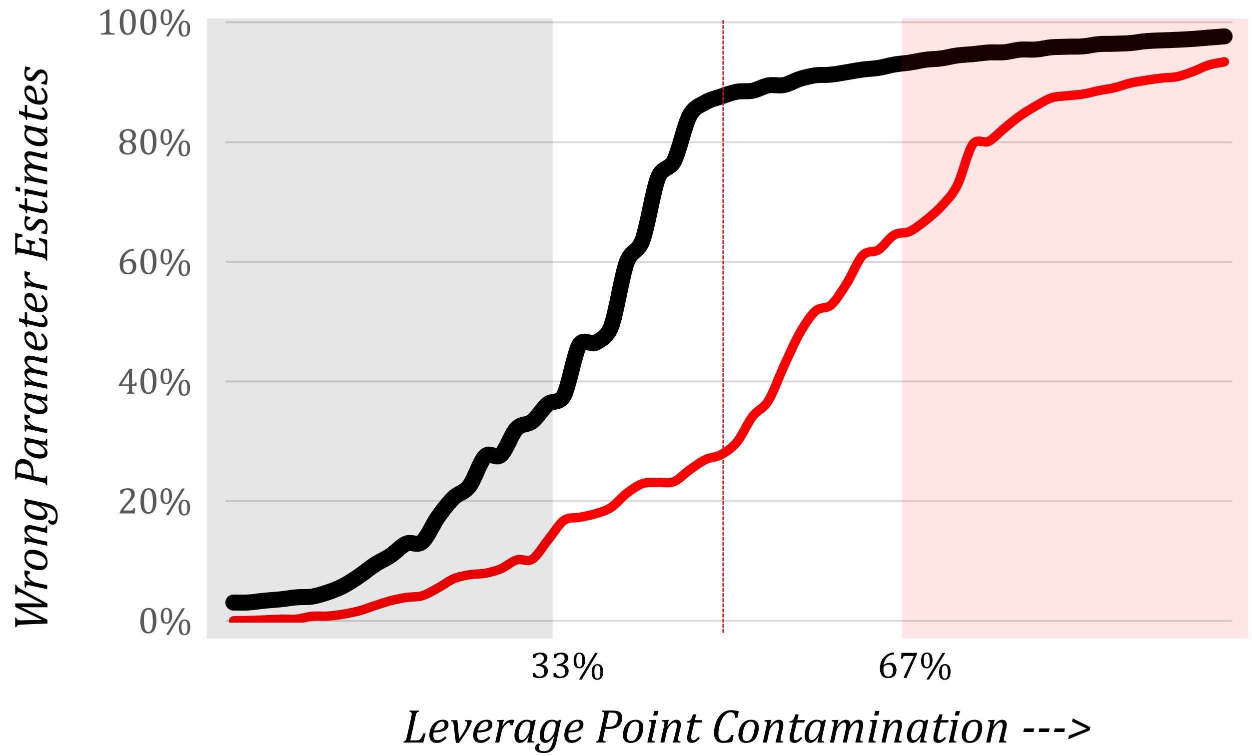
- Check 2: Pruning
  - Discard subsets that end up selecting < 50% of starting sample size
  - Discard subsets whose final parameter estimates are different from majority

  - Define a *distance* metric for different sets of parameter estimates: **Cosine Similarity**
  - Retain only *similar* estimates, final model becomes an average of these estimates

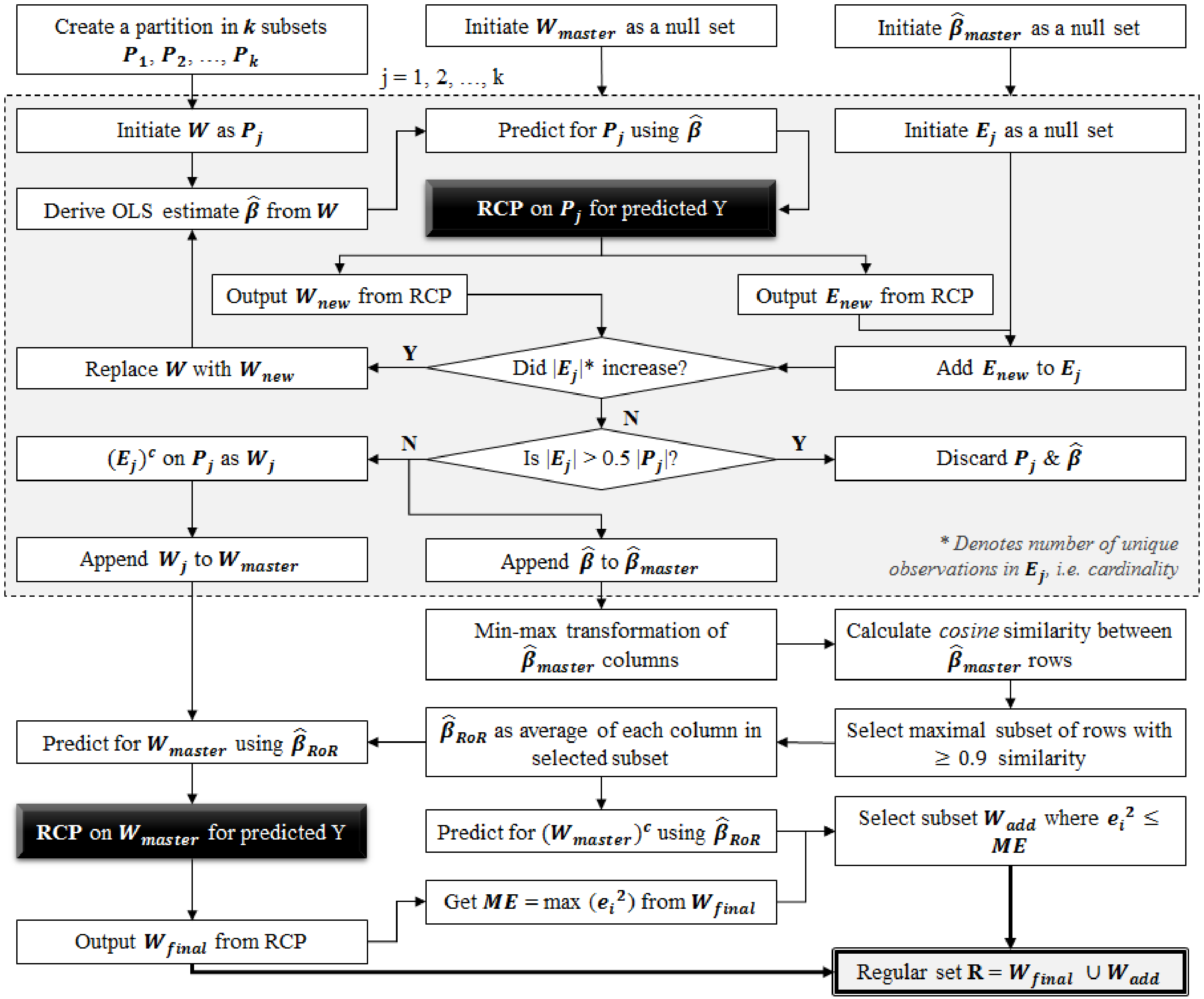


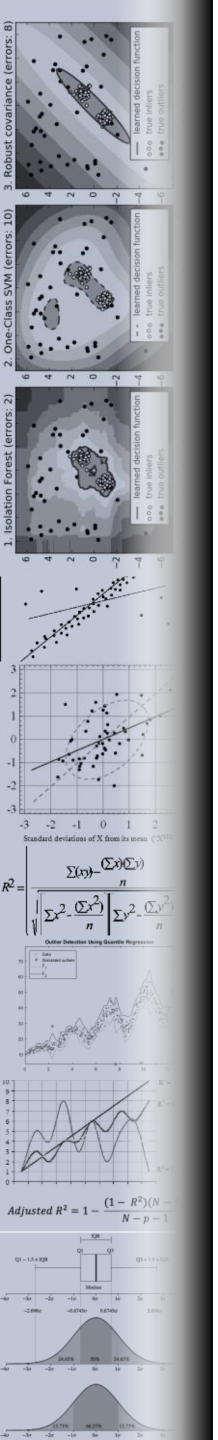
Trial 2: Committee of OLS Initializations

- Check 3: Coverage
  - **Check 2** leads to over-pruning but gives a representative model
  - Use this model on the discarded subsets and select observations with low error (<maximum error for regular observations)

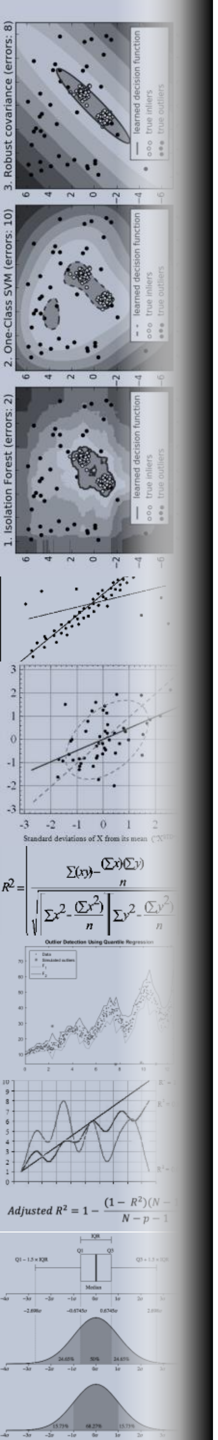


Complete Algorithm at-a-glance





Empirical Results



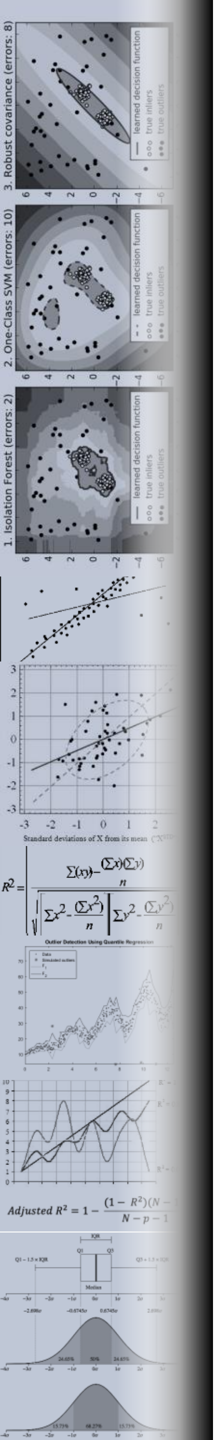
Hypotheses and Simulation Construct

- Two subsets **R** (retained) and **E** (excluded) obtained as output of the algorithm represent *Regular* and *Contaminated* observations respectively
- RoR^2 based on **R** can select the "best" model more effectively
- Simulation Construct
  - Benefit: Ground truth is known

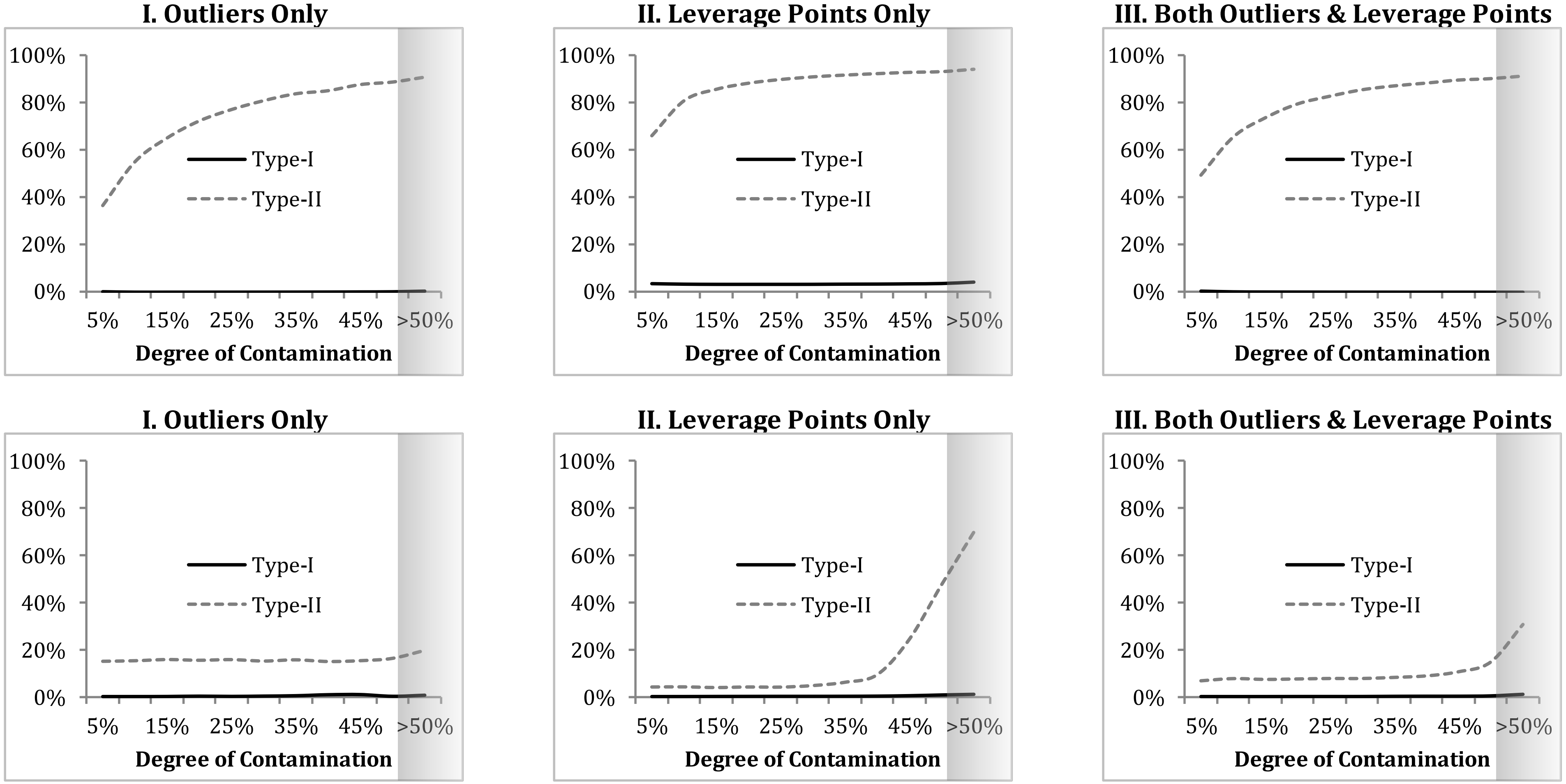
$$\boxed{Y = 20 + X_1 - 2X_2 - 5X_3 + 10X_4 + \mathcal{E}}$$

Outliers Only | **Leverage Points Only** | **Both Outliers and Leverage Points**

  - Elaborate contamination scheme covers extensive varieties of scenarios
  - 35,000 simulated datasets
  - 8 different modeling techniques



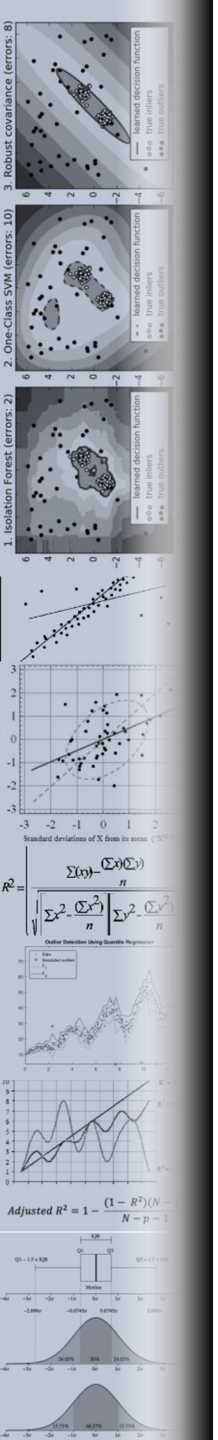
Contamination Detection Performance



Cook's D

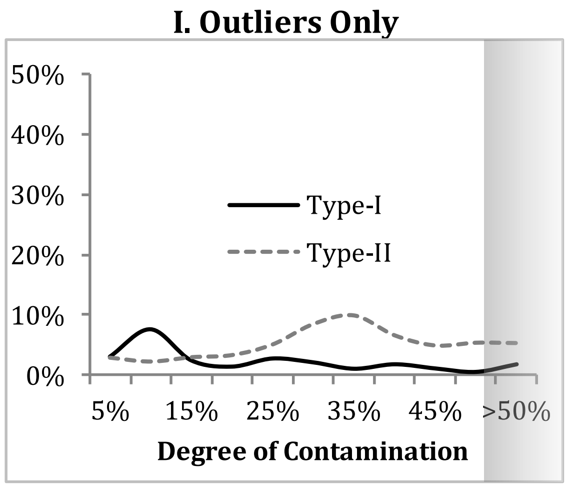
Proposed

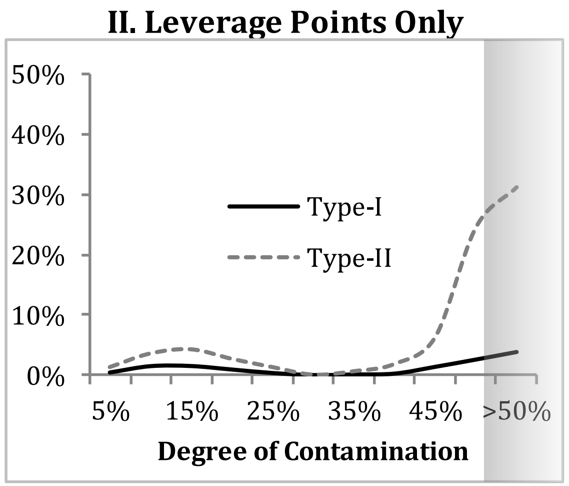
- Type I Error: A *regular* observation **included** in subset **E**
- Type II Error: A *contaminated* observation **included** in subset **R**

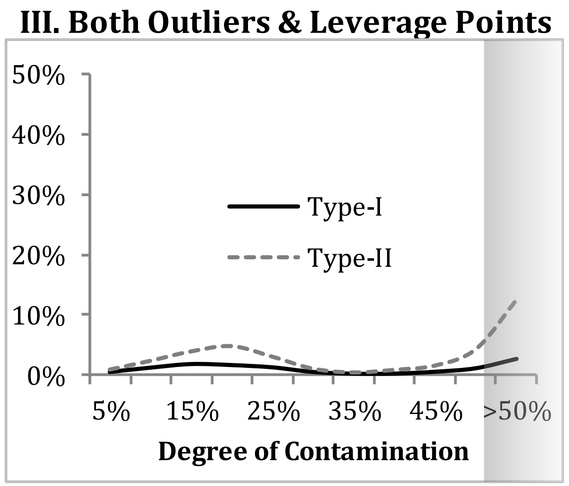


Model Selection Performance

Measure	Type I Error	Type II Error	Measure	Type I Error	Type II Error
R^2	27.0%	28.8%	F-statistic	44.0%	7.1%
Wilcoxon Dispersion	29.1%	22.2%	Median-R^2	23.3%	5.6%
Least Trimmed Square	4.4%	10.3%	RoR^2	**1.4%**	**3.1%**

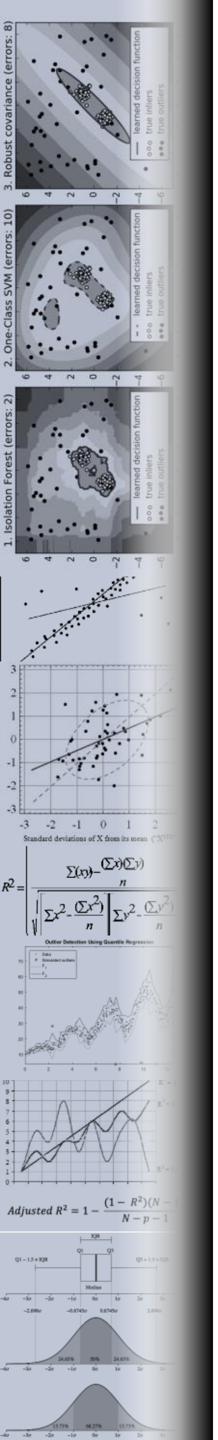






- Type I Error: An “Actual Best” model is **not selected** by the Metric
- Type II Error: Model selected by the Metric is **not** “Actual Best”

Additional Insights on Model Suitability



OLS: *Ordinary Least Square*

QNT: *Quantile Regression using finite smoothing*

HUB: *Huber's M-estimator with Huber weight*

BIS: *Huber's M-estimator with Tukey's bisquare weight*

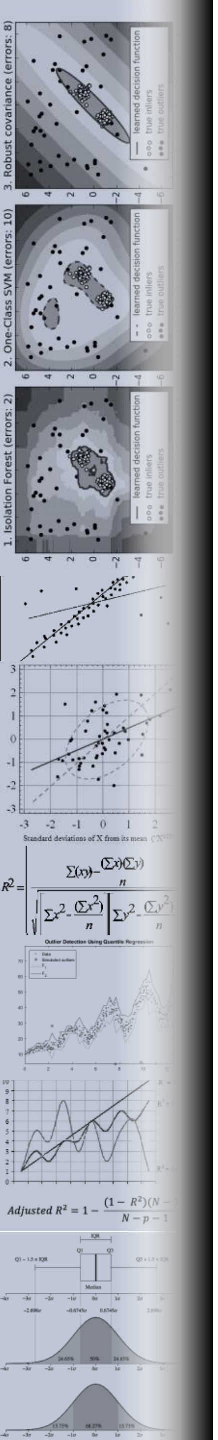
LTS: *Rousseeuw and Leroy's Least Trimmed Square estimator*

MML: *Yohai's MM-estimator with LTS initialization*

SES: *Rousseeuw and Yohai's S-estimator*

ROR: *RoR^2 regression estimator, which is an OLS estimator used on the identified set of regular observations*

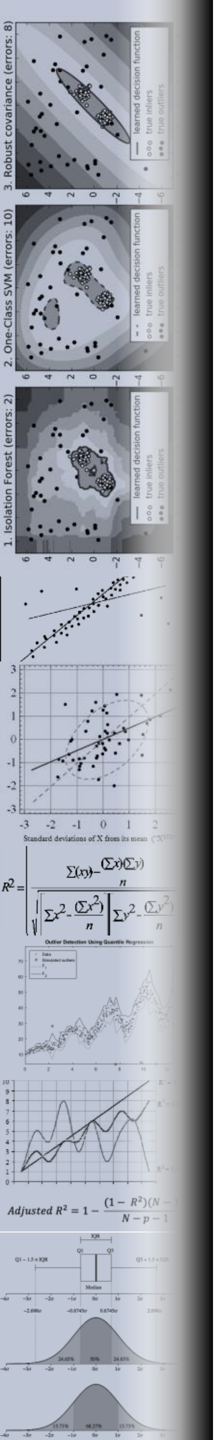
Good	Average	Bad



Performance on Real Datasets

- Real datasets do **not** come with a ground truth
- 3 datasets were chosen from UC Irvine ML repository: Combined Cycle Power Plant (CCPP), Concrete Compressive Strength (CCS) and Boston Housing (BH)
- Root Mean Square Error (RMSE) and correlation coefficient (r) used for validation

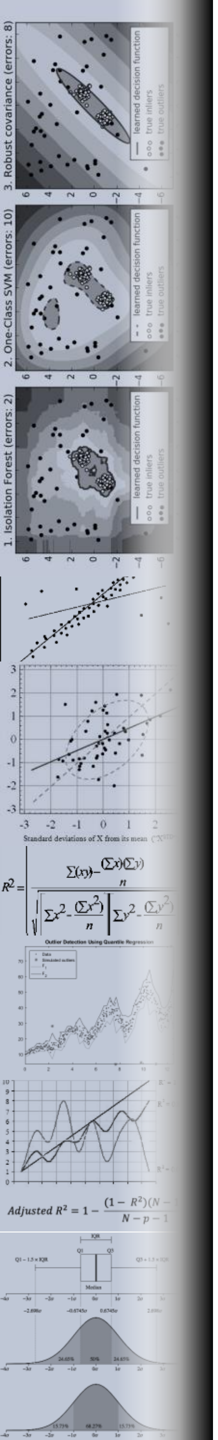
Dataset	CCPP				CCS				BH			
Decision	E	R		Total	E	R		Total	E	R		Total
Obs.	53	**9,515**		9,568	173	**857**		1,030	166	**340**		506
Model	r	r	RMSE	RoR^2	r	r	RMSE	RoR^2	r	r	RMSE	RoR^2
OLS	0.496	**0.967**	**4.314**	**0.931**	0.347	0.850	8.502	0.630	0.778	0.910	0.128	0.563
QNT	0.492	**0.967**	4.326	**0.931**	0.266	0.874	7.793	0.666	0.745	0.932	0.111	0.596
HUB	0.494	**0.967**	4.315	**0.931**	0.319	0.860	8.214	0.645	0.760	0.928	0.114	0.589
BIS	0.494	**0.967**	4.315	**0.931**	0.175	0.906	6.751	0.699	0.746	0.934	0.109	0.598
SES	0.492	**0.967**	4.335	0.930	0.155	**0.913**	6.522	**0.707**	0.722	**0.937**	**0.108**	**0.600**
MML	0.493	**0.967**	4.322	**0.931**	0.155	**0.913**	**6.500**	**0.707**	0.710	0.934	0.110	0.595
LTS	0.489	**0.967**	4.474	0.927	0.152	0.911	6.599	0.705	0.704	0.935	0.113	0.596
ROR	0.495	**0.967**	**4.314**	**0.931**	0.133	0.895	7.211	0.686	0.500	0.921	0.121	0.595



Conclusions and Next Steps

- Contamination detection performance was significantly better than existing model-based methods
- RoR^2 reduced model selection (and interpretation) error to < 5% from 50%+ observed for the traditional R^2 measure
- OLS based on chosen subset of regular observations, performed similar to existing robust regression methods.
  - Outperformed others at extremely high contamination scenarios

- Future Research
  - Evaluate contamination detection performance more extensively
  - Explore if / how the methodology can be customized for non-linear regression
  - Study behavior of RoR^2 in presence of extra or truncated set of covariates (compared to what is used in the underlying model)

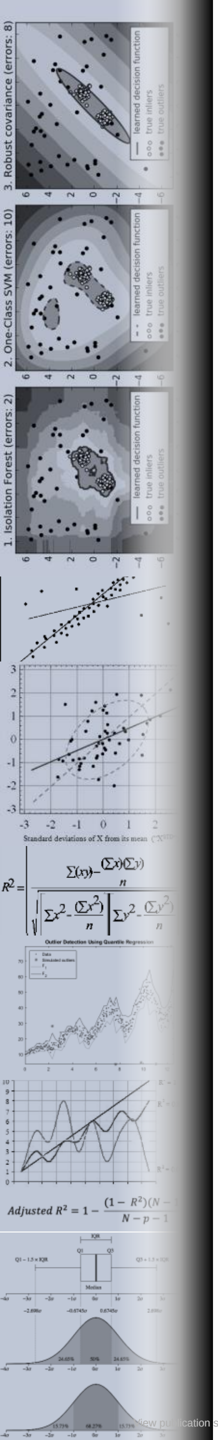


Thank You

For the peer-reviewed and **published version** of the paper, please access:
https://doi.org/10.1007/978-3-319-48517-1_12

For a raw, unreviewed **pre-print version** of the paper, please let me know of your interest and I will share it privately

For questions, remarks and feedback, please drop a message to
deb.sougata@gmail.com



References

1. Peter, J. R.: Least median of squares regression. Journal of the American Statistical Association 79(388), 871-880 (1984)
2. James, P. S.: Outliers and influential data points in regression analysis. Psychological Bulletin 95(2), 334-344 (1984)
3. John, M. S., William, L. S.: Algorithms and complexity for least median of squares regression. Discrete Applied Mathematics 14(1), 93-100 (1986)
4. Santiago, V.: On the behaviour of residual plots in robust regression. Statistics and Econometrics Series 04, Working Paper 93-04 (1993)
5. I-Cheng, Y.: Modeling of strength of high performance concrete using artificial neural networks. Cement and Concrete Research, Vol. 28, No. 12, 1797-1808 (1998)
6. Victoria, J. H., Jim, A.: A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), 85-126 (2004)
7. Peter, J. R., Annick, M. L.: Robust Regression and Outlier Detection (Vol. 589). John Wiley & Sons (2005)
8. Jeff, T. T., Joseph, W. M.: Rank-based analysis of linear models using R. Journal of Statistical Software 14(7), 1-26 (2005)
9. Heysem, K., Pınar, T., Fikret, S. G.: Local and global learning methods for predicting power of a combined gas & steam turbine. In: Proceedings of the International Conference on Emerging Trends in Computer and Electronics Engineering (ICETCEE 2012). Dubai, UAE, pp. 13-18 (2012)
10. S.M.A.Khaleelur, R., M.Mohamed, S., K.Senthamarai, K.: Multiple linear regression models in outlier detection. International Journal of Research in Computer Science 2(2), 23-28 (2012)
11. Moshe, L.: UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science (2013)
12. Pınar, T.: Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. International Journal of Electrical Power & Energy Systems, Volume 60, ISSN 0142-0615, 126-140 (2014)