Three common statistical mistakes in reservoir characterization

Frank Male and Jerry L Jensen,

The University of Texas at Austin

STARR
State of Texas Advanced Resource Recovery

The University of Texas at Austin
Center for Subsurface Energy and the Environment
Cockrell School of Engineering

Follow along:

Jupyter notebooks at https://github.com/frank1010111/statistical_missteps with binder and Colab badges

Only three? Topics

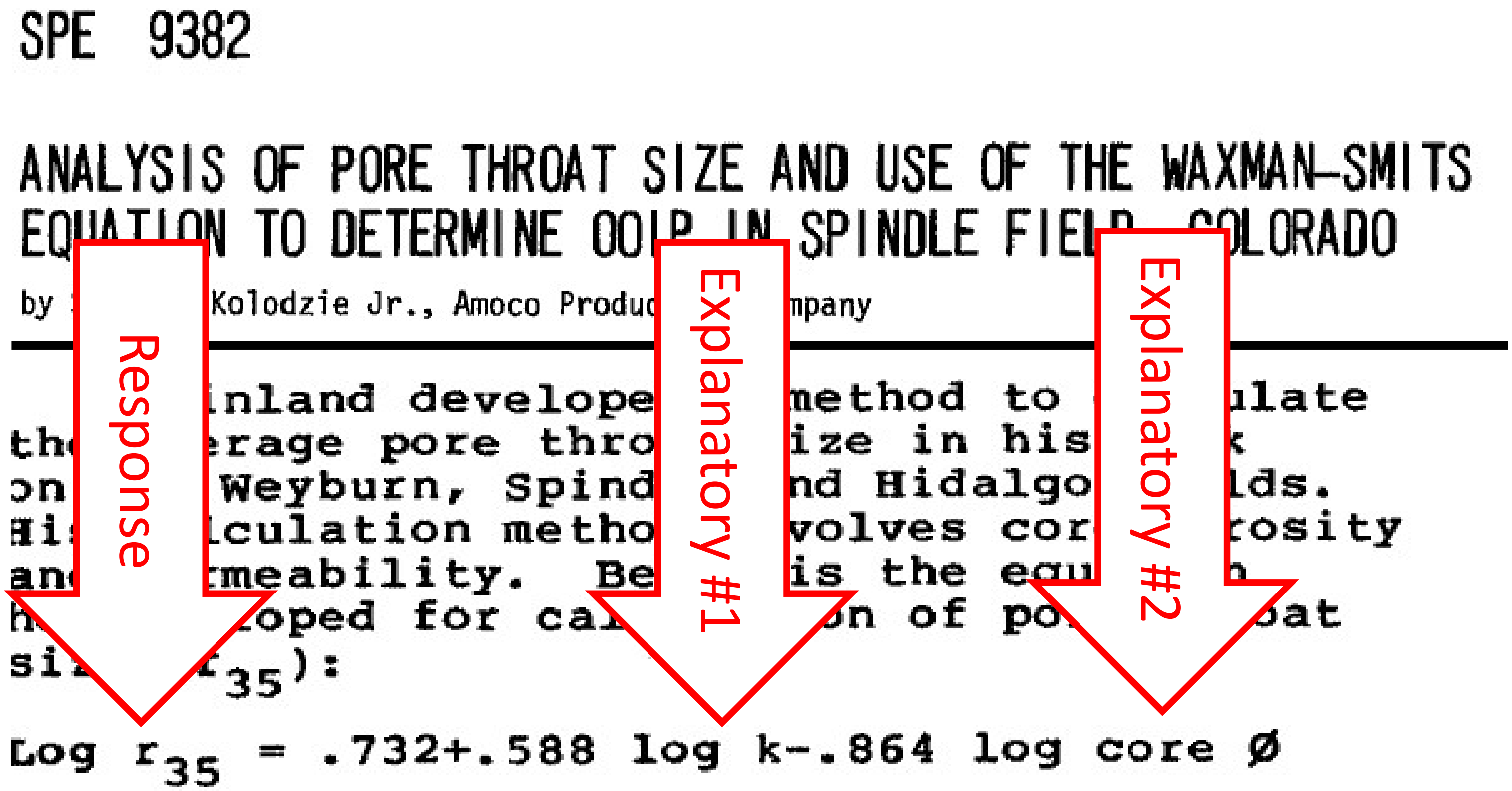
- Mistake 1: Algebraic manipulations with linear regression derived models
  - Brief review of least squares linear regression concepts
- Mistake 2: Working with log-transformed variables
- Mistake 3: Interpreting R^2

A few words about study motivation

- We all make mistakes, including authors, reviewers, and editors
- We cite specific papers for three reasons
    - With best will and effort, mistakes remain
    - To demonstrate these mistakes actually happen
    - As a caution to future investigators
    - As authors and researchers, take time to look at earlier literature
        - As reviewers, pay attention to the literature review
        - As editors, be careful with superficial assessments

Mistake 1
Algebraic manipulations with Linear models

Example: The Winland Equation (as published)



SPE 9382

ANALYSIS OF PORE THROAT SIZE AND USE OF THE WAXMAN-SMITS EQUATION TO DETERMINE OOIP IN SPINDLE FIELD, COLORADO

by Kolodzie Jr., Amoco Produc mpany

inland develope method to ulate the erage pore thro ize in his k on Weyburn, Spind nd Hidalgo lds. Hi lculation metho volves cor rosity an rmeability. Be is the equ n h loped for ca on of po oat si r_{35}):

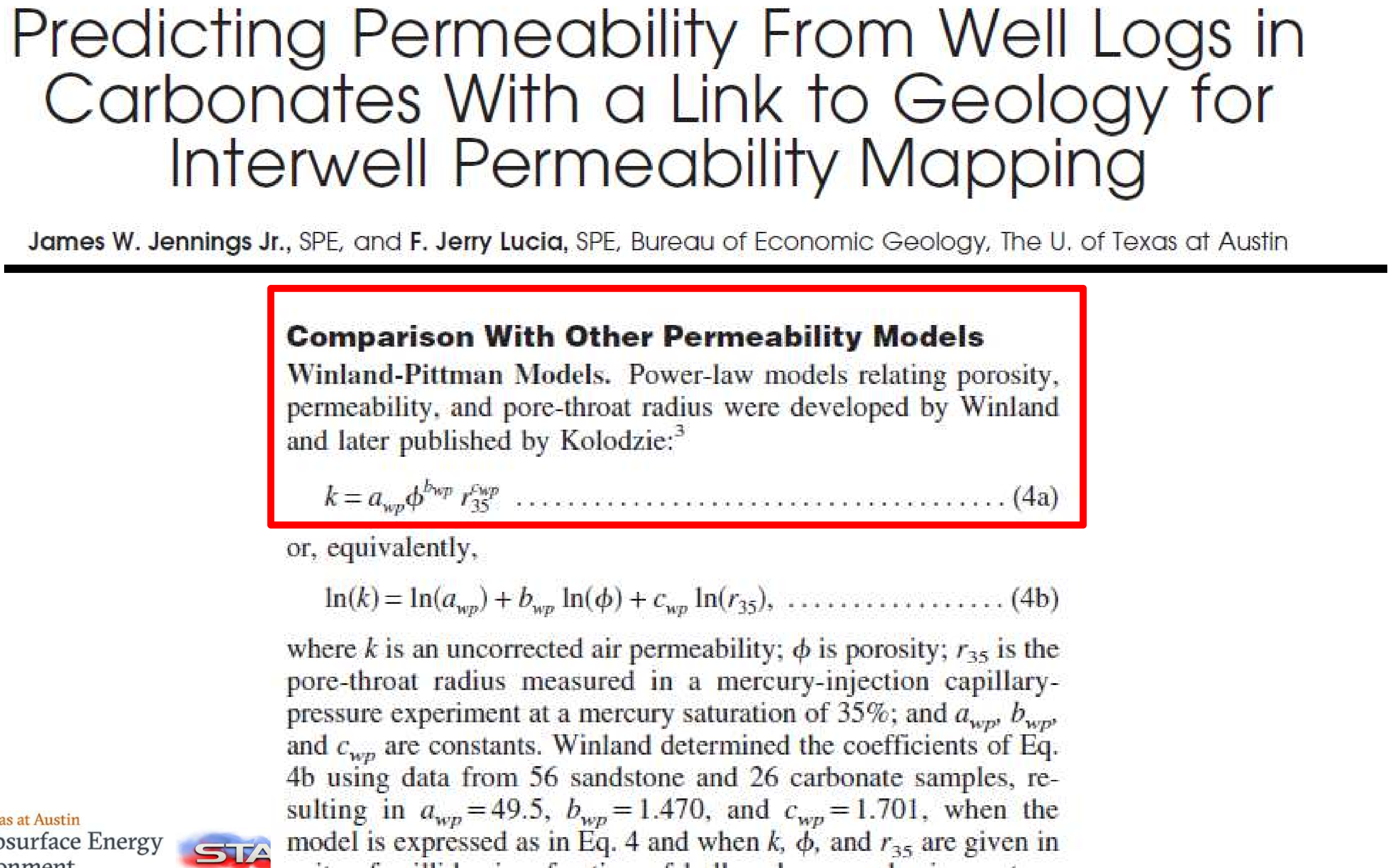
$$\text{Log } r_{35} = .732+.588 \log k-.864 \log \text{core } \emptyset$$

Response

Explanatory #1

Explanatory #2

Winland Equation as described (Jennings and Lucia)



Predicting Permeability From Well Logs in Carbonates With a Link to Geology for Interwell Permeability Mapping

James W. Jennings Jr., SPE, and **F. Jerry Lucia**, SPE, Bureau of Economic Geology, The U. of Texas at Austin

Comparison With Other Permeability Models

Winland-Pittman Models. Power-law models relating porosity, permeability, and pore-throat radius were developed by Winland and later published by Kolodzie:[3]

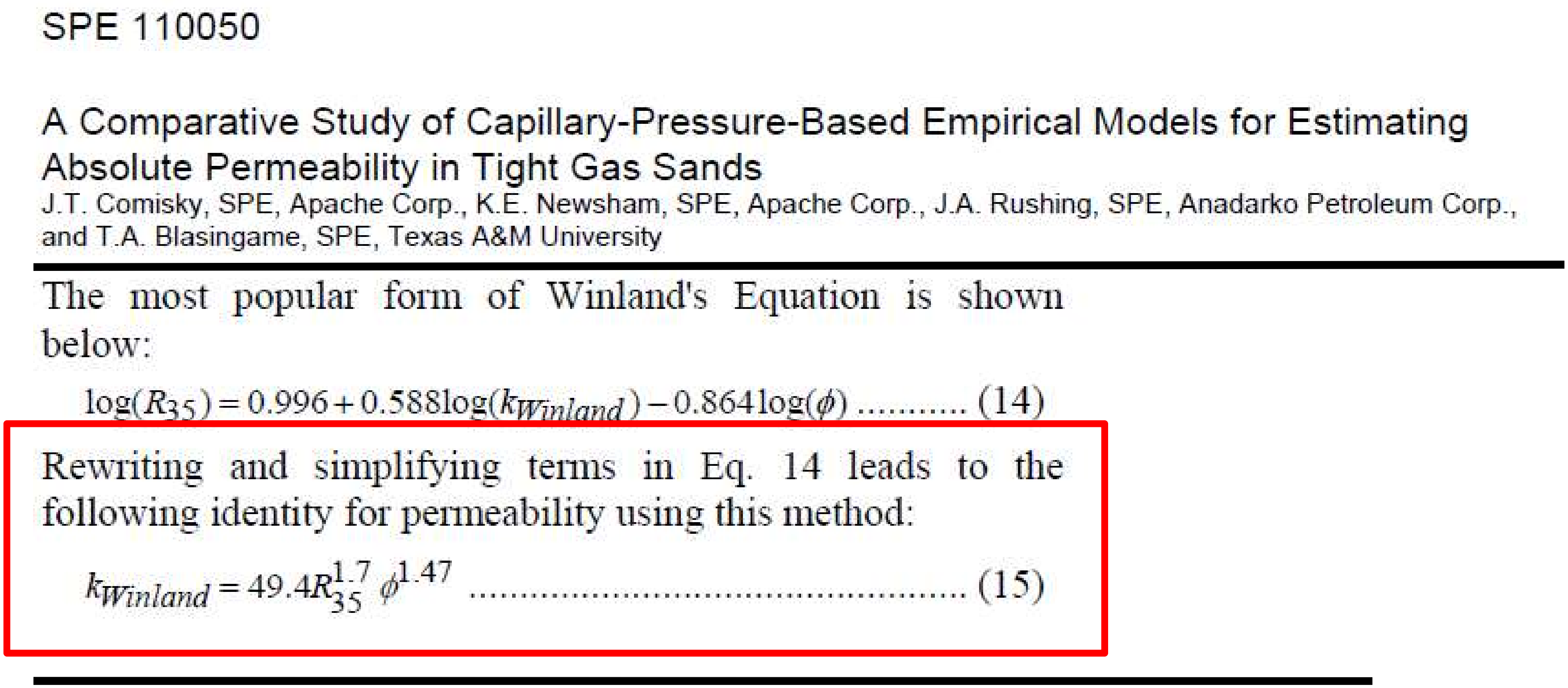
$$k = a_{wp}\phi^{b_{wp}}\, r_{35}^{c_{wp}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots \text{(4a)}$$

or, equivalently,

$$\ln(k) = \ln(a_{wp}) + b_{wp}\ln(\phi) + c_{wp}\ln(r_{35}), \qquad \ldots\ldots\ldots \text{(4b)}$$

where k is an uncorrected air permeability; ϕ is porosity; r_{35} is the pore-throat radius measured in a mercury-injection capillary-pressure experiment at a mercury saturation of 35%; and a_{wp}, b_{wp}, and c_{wp} are constants. Winland determined the coefficients of Eq. 4b using data from 56 sandstone and 26 carbonate samples, resulting in $a_{wp} = 49.5$, $b_{wp} = 1.470$, and $c_{wp} = 1.701$, when the model is expressed as in Eq. 4 and when k, ϕ, and r_{35} are given in

The Winland Equation as described (Comisky et al.)



SPE 110050

A Comparative Study of Capillary-Pressure-Based Empirical Models for Estimating Absolute Permeability in Tight Gas Sands

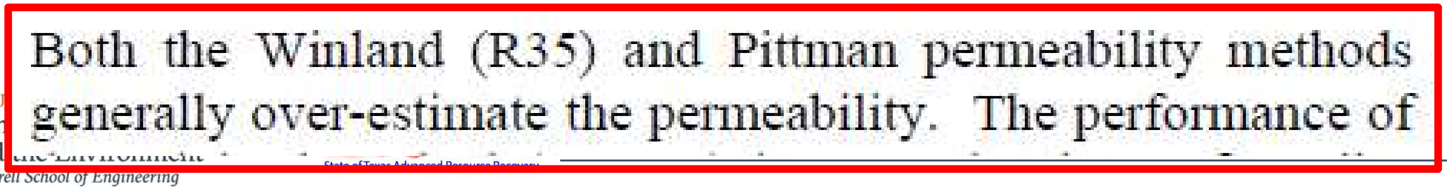
J.T. Comisky, SPE, Apache Corp., K.E. Newsham, SPE, Apache Corp., J.A. Rushing, SPE, Anadarko Petroleum Corp., and T.A. Blasingame, SPE, Texas A&M University

The most popular form of Winland's Equation is shown below:

$$\log(R_{35}) = 0.996 + 0.588\log(k_{Winland}) - 0.864\log(\phi) \;\ldots\ldots\ldots\ldots\; (14)$$

Rewriting and simplifying terms in Eq. 14 leads to the following identity for permeability using this method:

$$k_{Winland} = 49.4 R_{35}^{1.7}\,\phi^{1.47} \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\; (15)$$



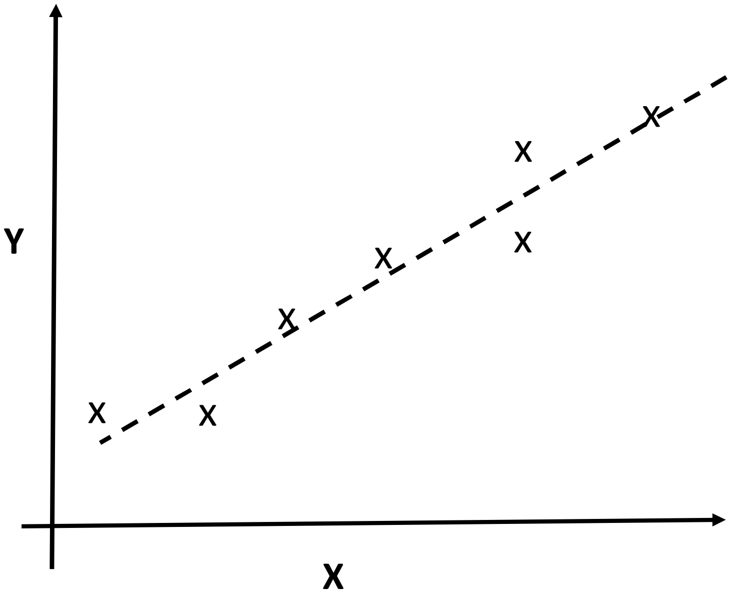
Both the Winland (R35) and Pittman permeability methods generally over-estimate the permeability. The performance of

Brief review of linear regression principles (1)

- Have N data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$
- Roles of variables
  - X is explanatory
  - Y is response

Model: $Y = aX + b + \varepsilon$

- Use LSLR to estimate a and b
- Called “Y-on-X” regression
- e is an RV and very important



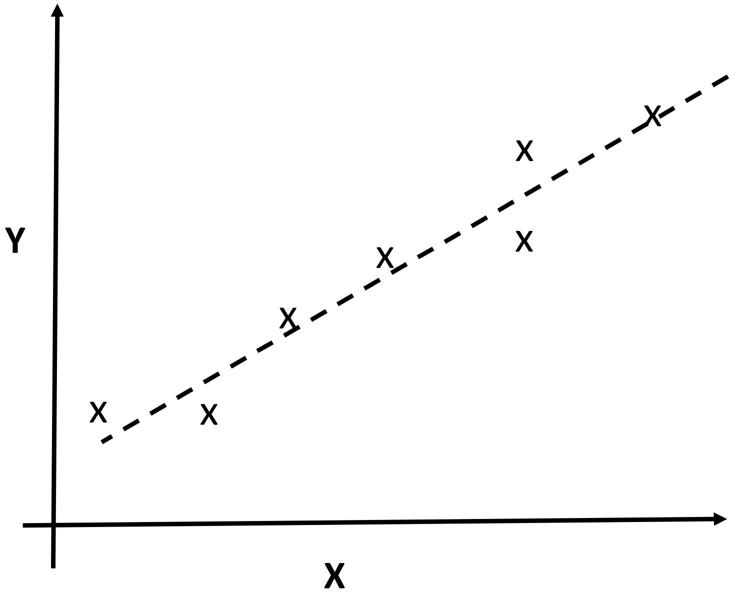
Brief review of linear regression principles (2)

ε has several roles in $Y = aX + b + \varepsilon$

- Relationship not exactly correct
- Y may contain random errors
- Model assumes X measured without error
- Error is independent of X, Y, a, and b

Residuals p

- Difference $Y_i - (aX_i + b) = p_i$
- Show how well model explains Y variability
- $R^2 = 1 - \mathrm{Var}(p)/\mathrm{Var}(Y)$



Brief review of linear regression principles (3)

Alternative line $X = cY + d + \varepsilon^*$

- X-on-Y line
- Residuals q
- X has errors

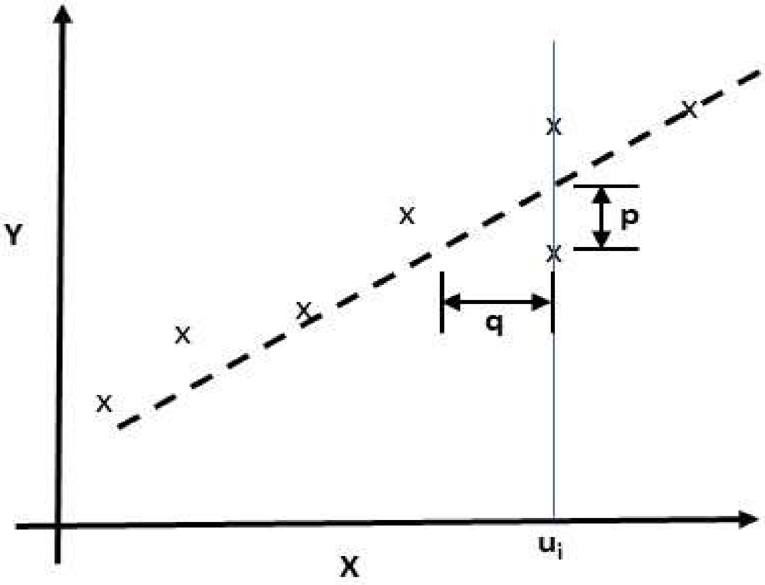
- How are a & b and c & d related?
- Simple algebra takes
  - $Y = aX + b$ and gives

$$X = Y/a + b/a$$

- Suggesting $c = 1/a$ and $d = b/a$

This ignores role of ε: forcing $\varepsilon^* = \varepsilon/a$

In general, $c \neq 1/a$ and $d \neq b/a$



Example data set from Pittman (1992)

- For 25% Hg saturation, his data and linear regression give

$$\log(r_p) = 0.531 \log(k) - 0.350 \log(\phi) + 0.204 \qquad (1)$$

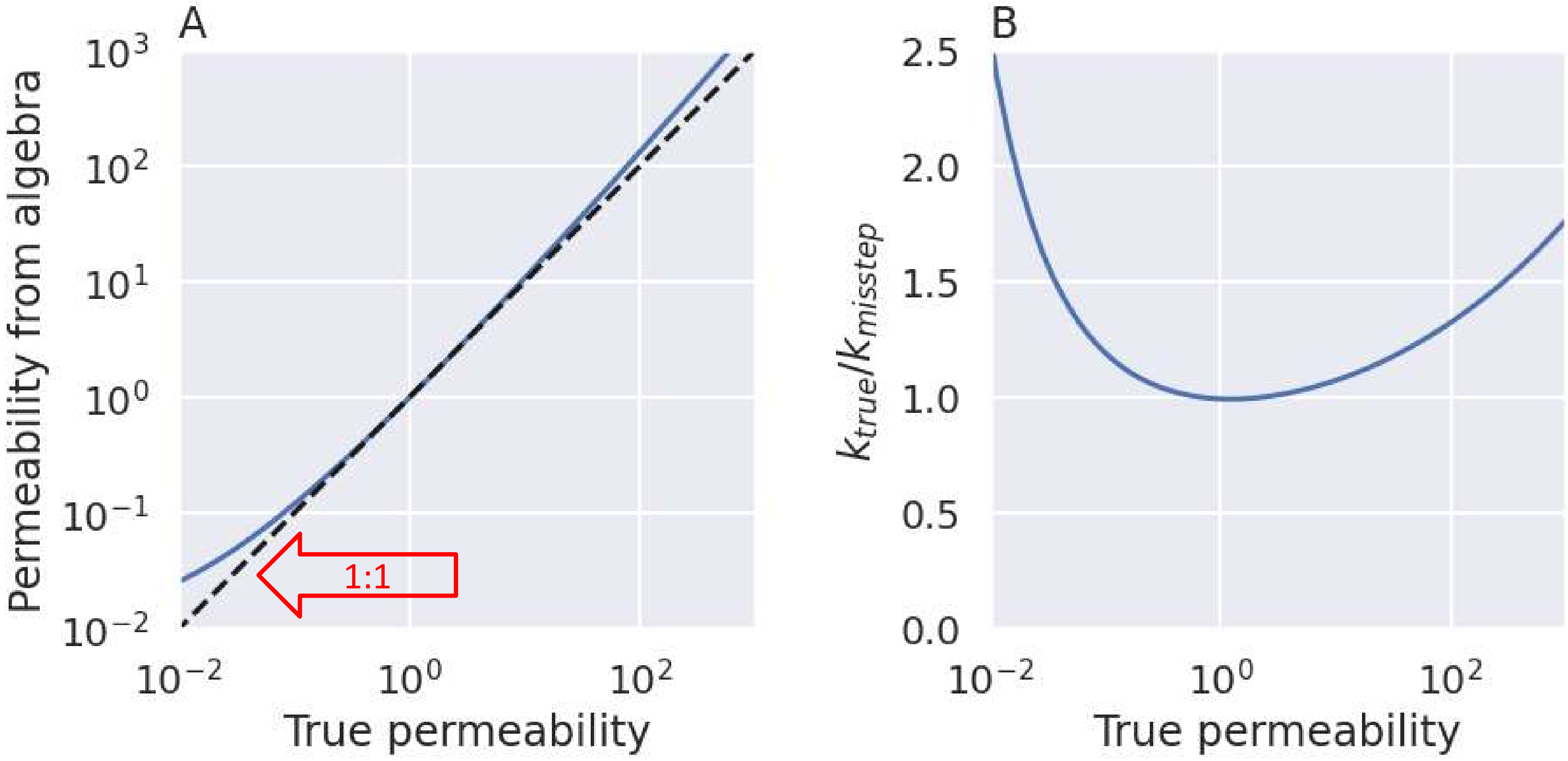
- Also, same data and linear regression give
- $$\log(k) = 1.512 \log(r_p) + 1.415 \log(\phi) - 1.221 \qquad (2)$$

- If we took Eq (1) and used algebra—”naïve” method—we get

$$\log(k) = 1.88 \log(r_p) + 0.659 \log(\phi) - 0.384 \qquad (3)$$

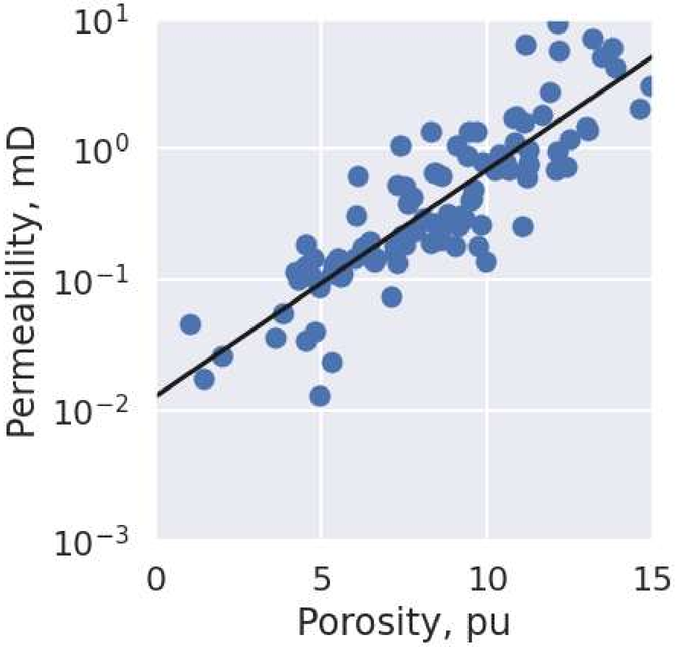
What’s the difference?

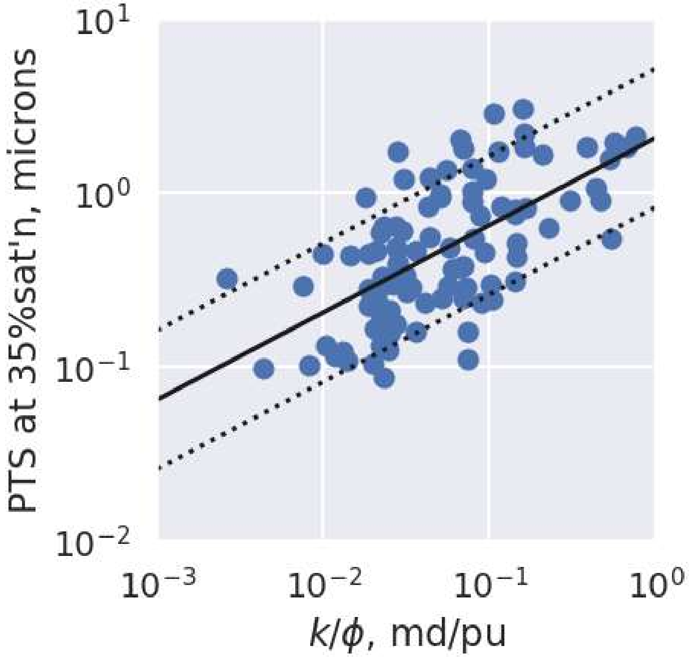
- Incorrect perm larger or equal to correct perm
- Incorrect perm particularly poor on low perms



Another example

- Synthetic dataset N = 100 with Y on X honoring Kwon and Pickett (1975)
- Avg predicted perm ~ 7 x actual perm avg
- 30 to 40% of predicted values > 2 x actual

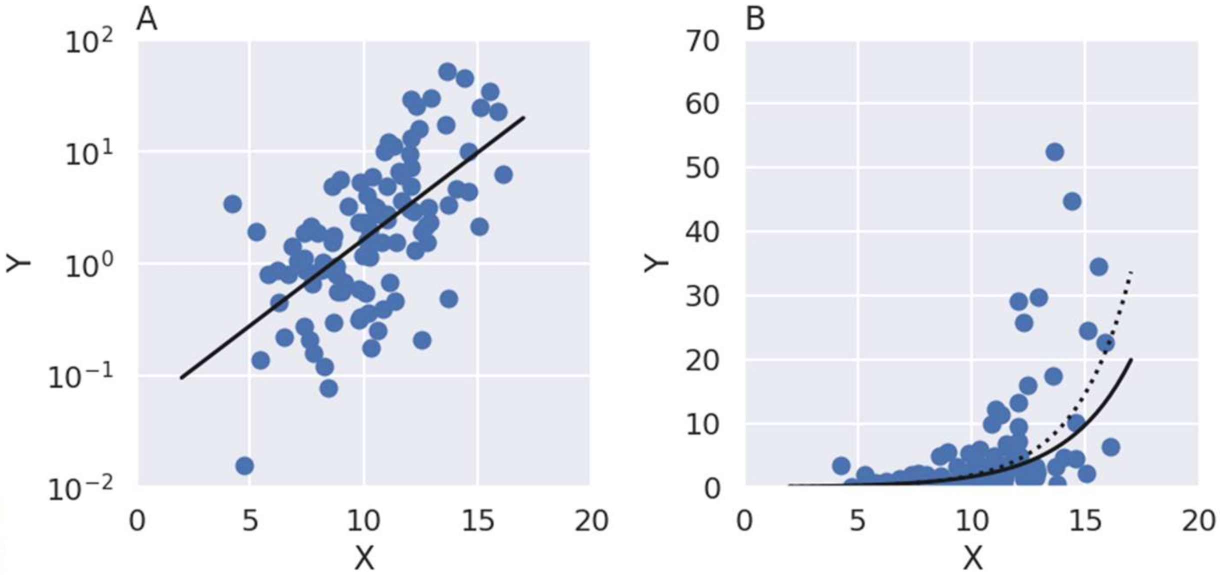




Mistake 2
De-transformation with linear regression model

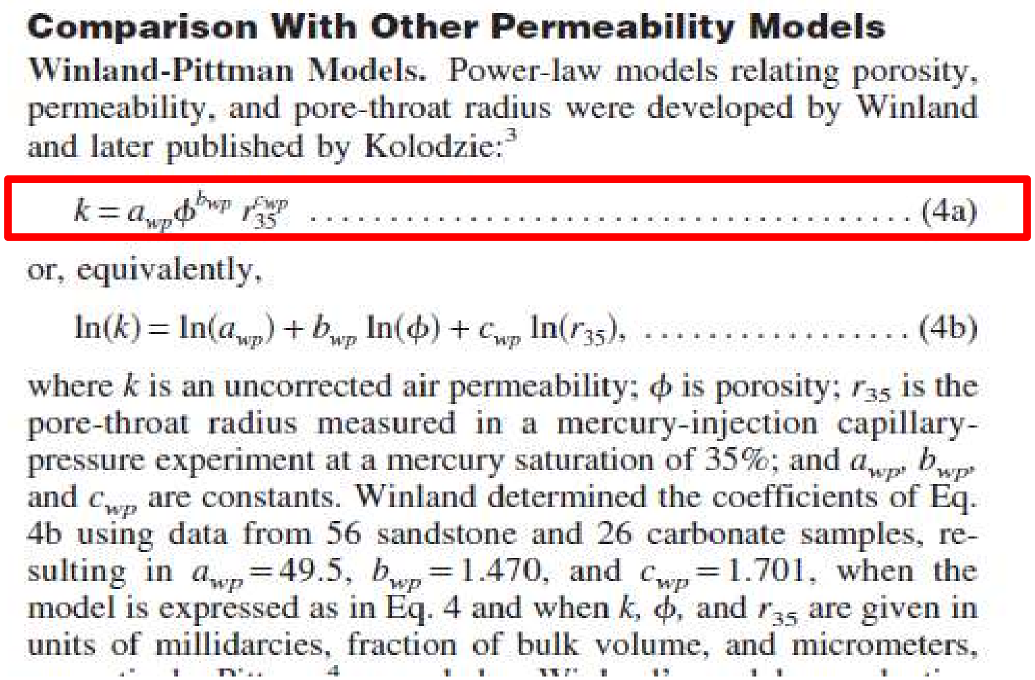
A problem with least-squares processes

- Y-on-X process minimizes $\Sigma[Y_i - (aX_i + b)]^2$
- Equal weight to deviations above and below line
- At X = 7
  - line gives log(Y) = 0
  - difference of 10 – 1 = 9 above line same as 1 – 0.1 = 0.9 below line



Literature examples

- Winland equation
  - Jennings and Lucia (2003)
  - Comisky et al. (2007)
  - Lucia (2007, p. 16)
- Porosity-permeability relationships
  - Worthington (2004)
  - Lucia (1999)
  - Craig (1991)
- Core vs well log upscaling
  - Lucia (2007, p. 89)
  - Hearn et al. (1986)



Comparison With Other Permeability Models

Winland-Pittman Models. Power-law models relating porosity, permeability, and pore-throat radius were developed by Winland and later published by Kolodzie:[3]

$$k = a_{wp}\phi^{b_{wp}}\, r_{35}^{c_{wp}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(4a)}$$

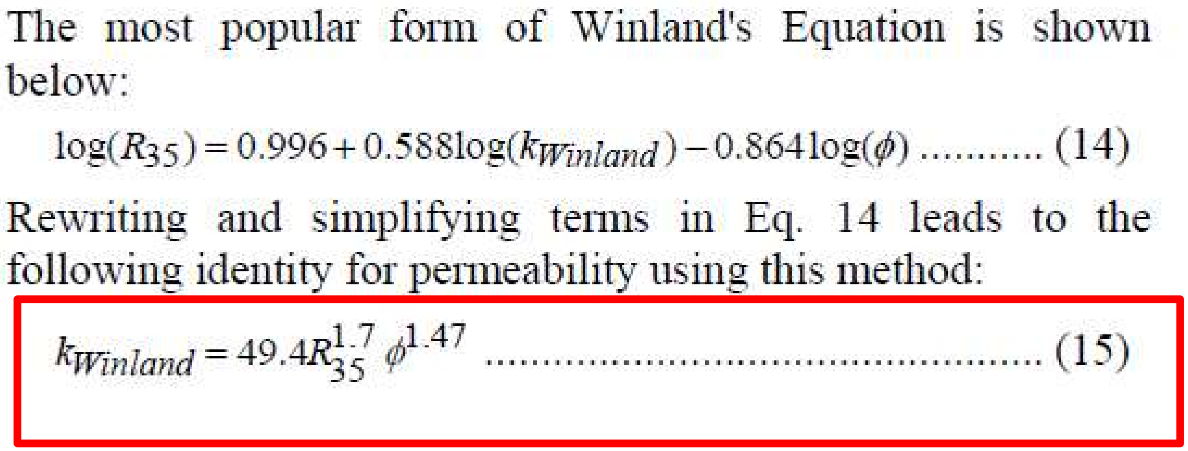
or, equivalently,

$$\ln(k) = \ln(a_{wp}) + b_{wp}\ln(\phi) + c_{wp}\ln(r_{35}), \quad \ldots\ldots\ldots\ldots\ldots \text{(4b)}$$

where k is an uncorrected air permeability; ϕ is porosity; r_{35} is the pore-throat radius measured in a mercury-injection capillary-pressure experiment at a mercury saturation of 35%; and a_{wp}, b_{wp}, and c_{wp} are constants. Winland determined the coefficients of Eq. 4b using data from 56 sandstone and 26 carbonate samples, resulting in $a_{wp} = 49.5$, $b_{wp} = 1.470$, and $c_{wp} = 1.701$, when the model is expressed as in Eq. 4 and when k, ϕ, and r_{35} are given in units of millidarcies, fraction of bulk volume, and micrometers,

Literature examples

- Winland equation
  - Jennings and Lucia (2003)
  - Comisky et al. (2007)
  - Lucia (2007, p. 16)
- Porosity-permeability relationships
  - Worthington (2004)
  - Lucia (1999)
  - Craig (1991)
- Core vs well log upscaling
  - Lucia (2007, p. 89)
  - Hearn et al. (1986)



The most popular form of Winland's Equation is shown below:

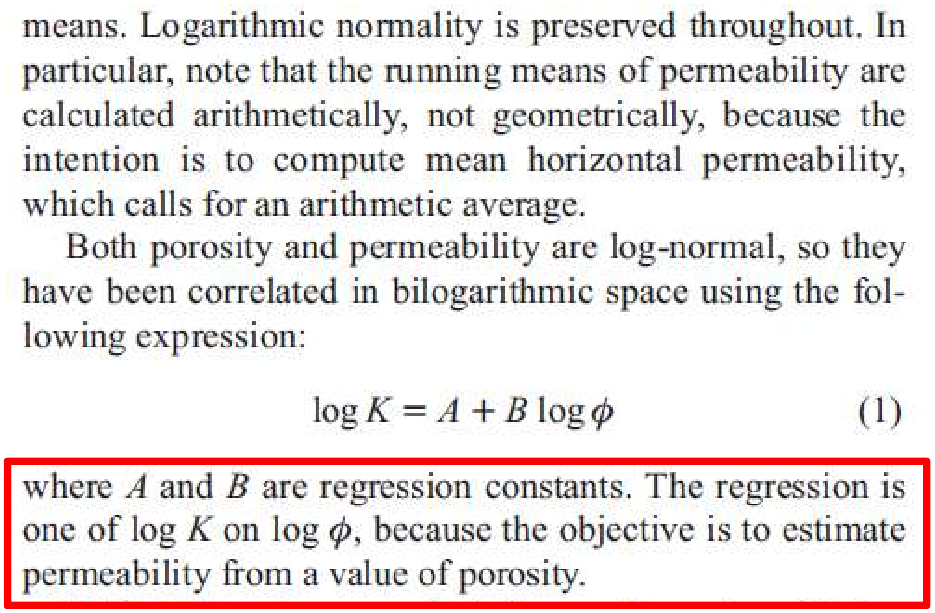
$$\log(R_{35}) = 0.996 + 0.588\log(k_{Winland}) - 0.864\log(\phi) \;........... (14)$$

Rewriting and simplifying terms in Eq. 14 leads to the following identity for permeability using this method:

$$k_{Winland} = 49.4R_{35}^{1.7}\,\phi^{1.47} \;... (15)$$

Literature examples

- Winland equation
  - Jennings and Lucia (2003)
  - Comisky et al. (2007)
  - Lucia (2007, p. 16)
- Porosity-permeability relationships
  - Worthington (2004)
  - Lucia (1999)
  - Craig (1991)
- Core vs well log upscaling
  - Lucia (2007, p. 89)
  - Hearn et al. (1986)



means. Logarithmic normality is preserved throughout. In particular, note that the running means of permeability are calculated arithmetically, not geometrically, because the intention is to compute mean horizontal permeability, which calls for an arithmetic average.

Both porosity and permeability are log-normal, so they have been correlated in bilogarithmic space using the following expression:

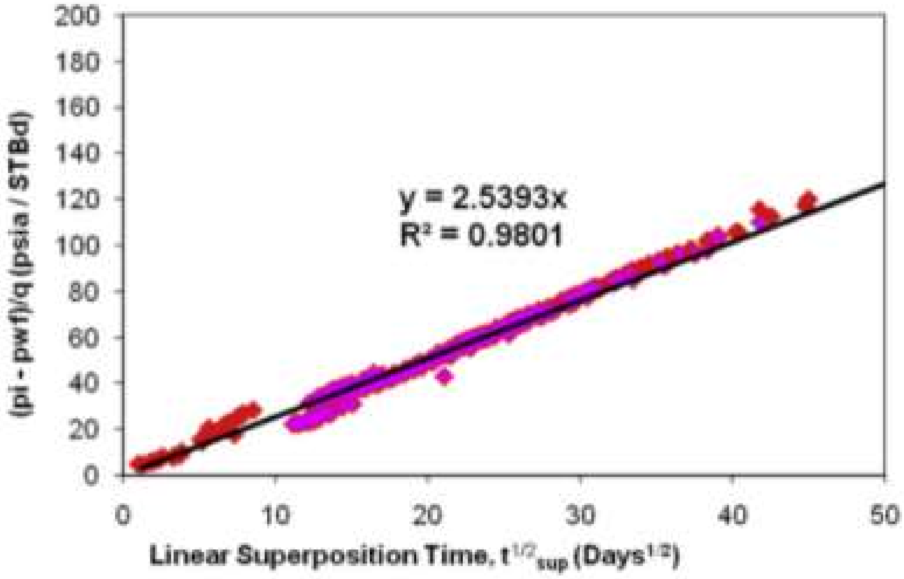
$$\log K = A + B \log \phi \qquad (1)$$

where A and B are regression constants. The regression is one of $\log K$ on $\log \phi$, because the objective is to estimate permeability from a value of porosity.

Mistake 3
R^2 interpretation

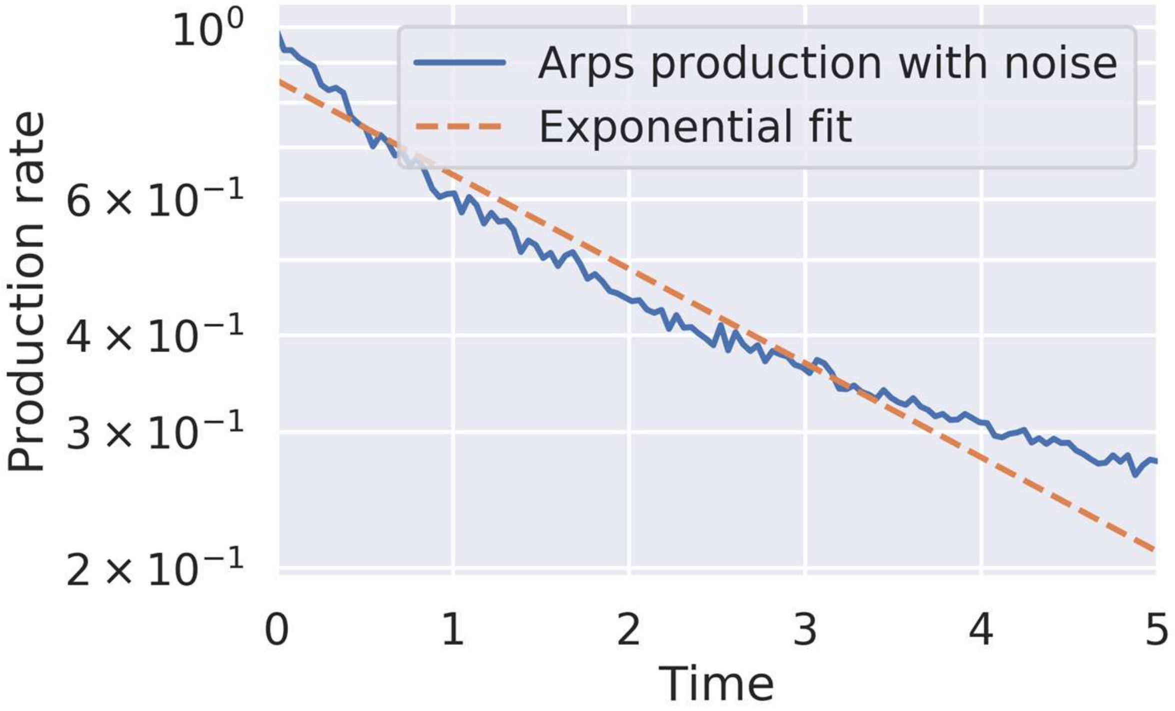
R^2 interpretation issue 1: autocorrelation

- When data is strongly autocorrelated, a model that just predicts the last value has a high R^2, without any actual predictive power
- Often seen in: well log analysis, production analysis
- Examples from literature: Can and Kabir (2014) (see figure), Gupta et al. (2018), Ren et al. (2019)



Synthetic example with autocorrelation

- Input: Arps + noise proportional to rate
- Fitting model: straight exponential
- R^2: 0.95



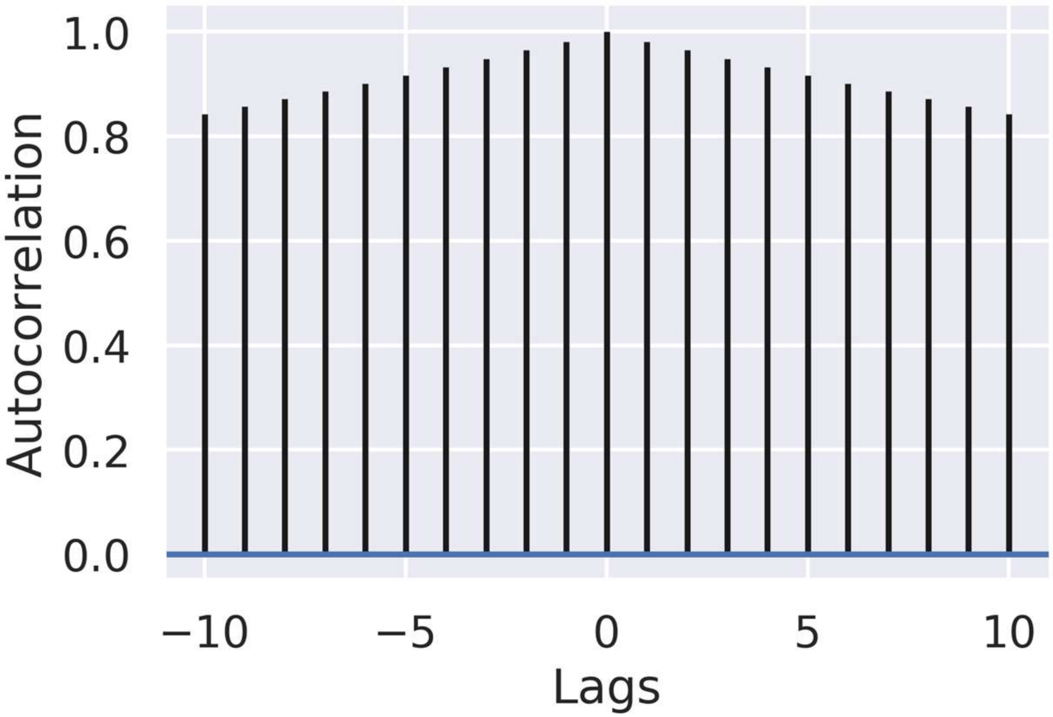
Identifying and correcting for autocorrelation

Diagnosis:

- Calculate autocorrelation
- Do semi-variogram analysis

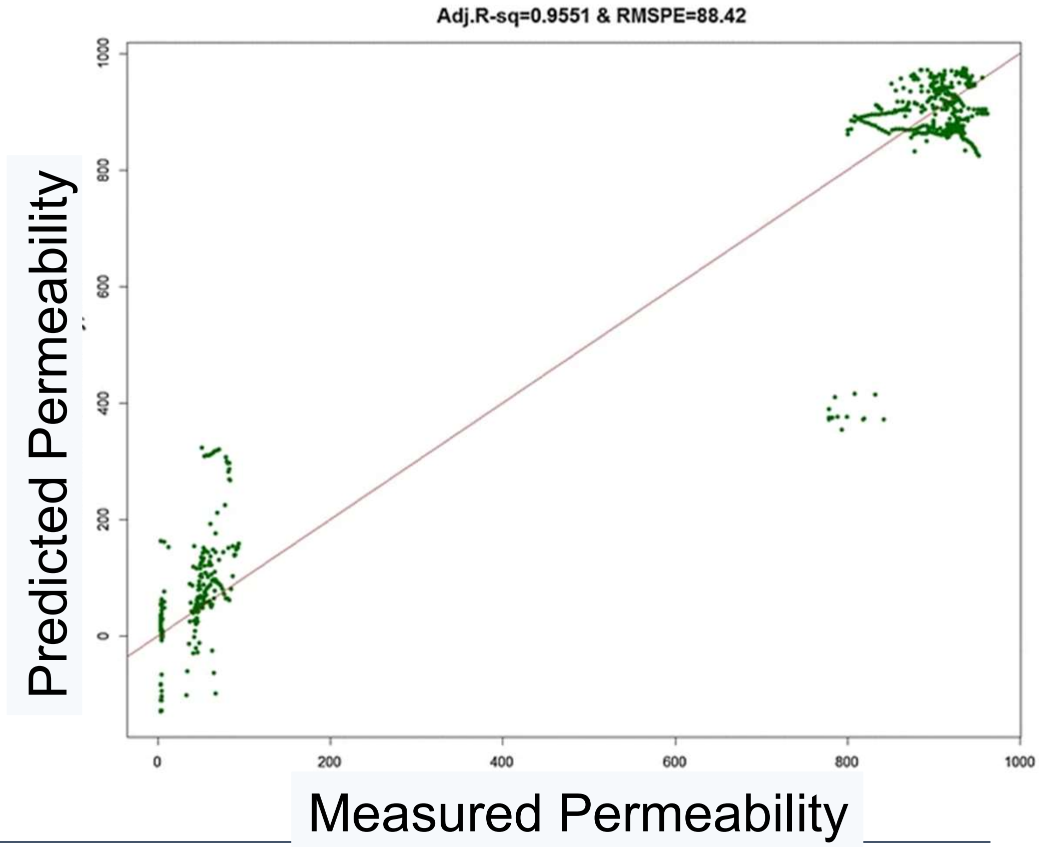
Solutions:

- Compare errors to a naïve model's baseline
- Only predict outside of autocorrelation length
- Use MAE/RMSE rather than R^2
- Newey-West Estimator



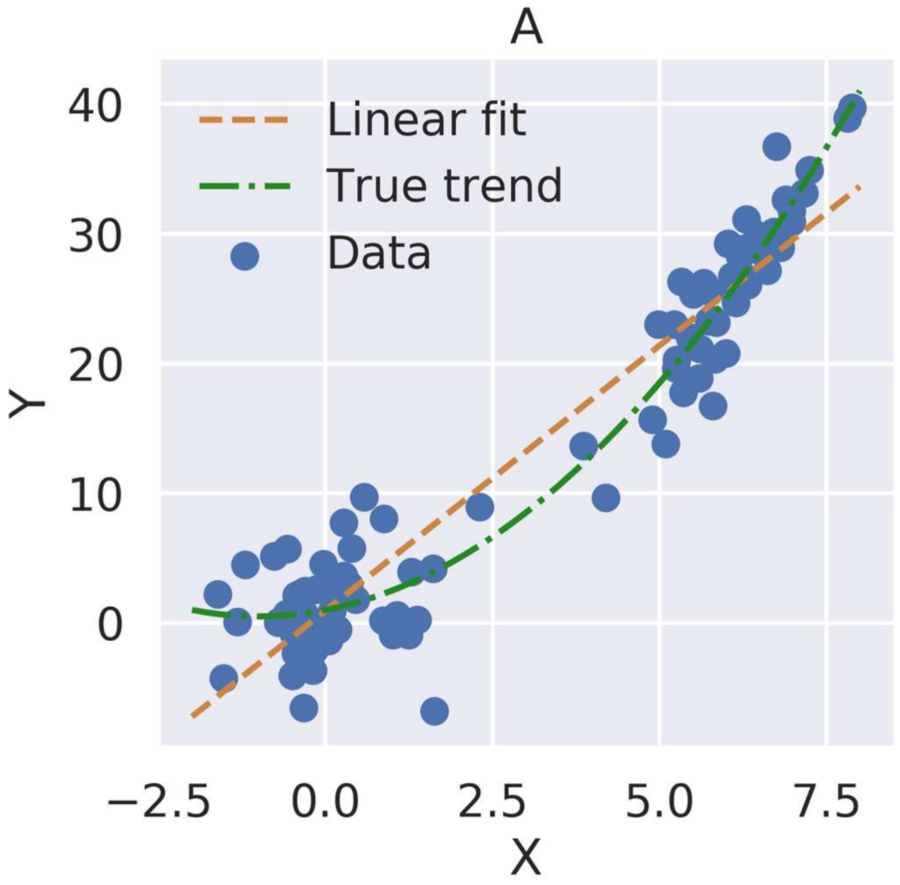
R^2 interpretation issue 2: bimodal inputs

- Inputs: bimodal
  - Cause could be facies
- Model fit within modes: poor
  - Not predicting intra-facies variation
- R^2 looks great
- Examples: Al-Mudhafar (2017) (shown), Ali Ahmadi et al (2012)



Synthetic example of bimodal inputs

- Input: Y ~ X^2, but X has two modes
- Model for fitting: Y ~ X
- R^2: 0.91



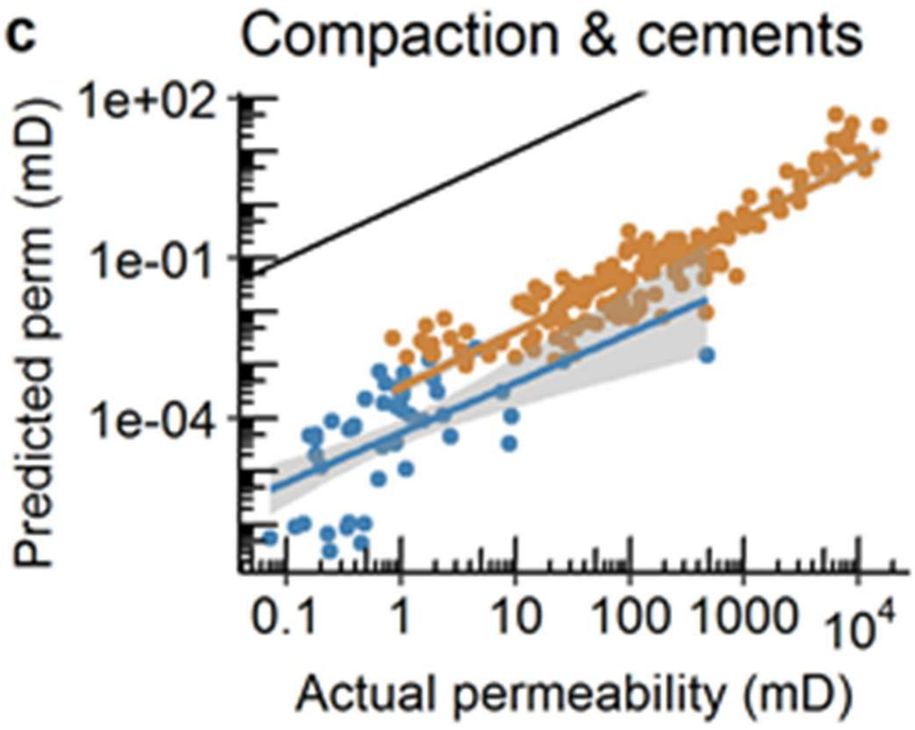
Identifying and correcting bimodal inputs

Diagnosis:

- Histogram your inputs

Solution:

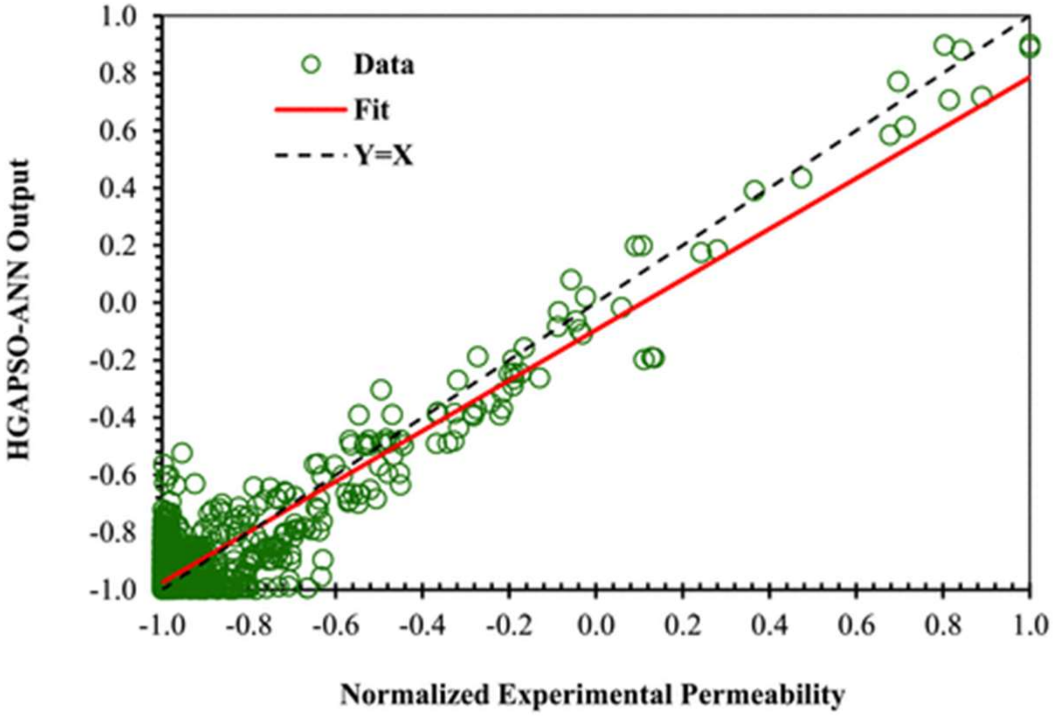
- Split the modes
- Analyze each mode separately



Male, Jensen, and Lake, 2020

R^2 interpretation issue 3: very skewed inputs

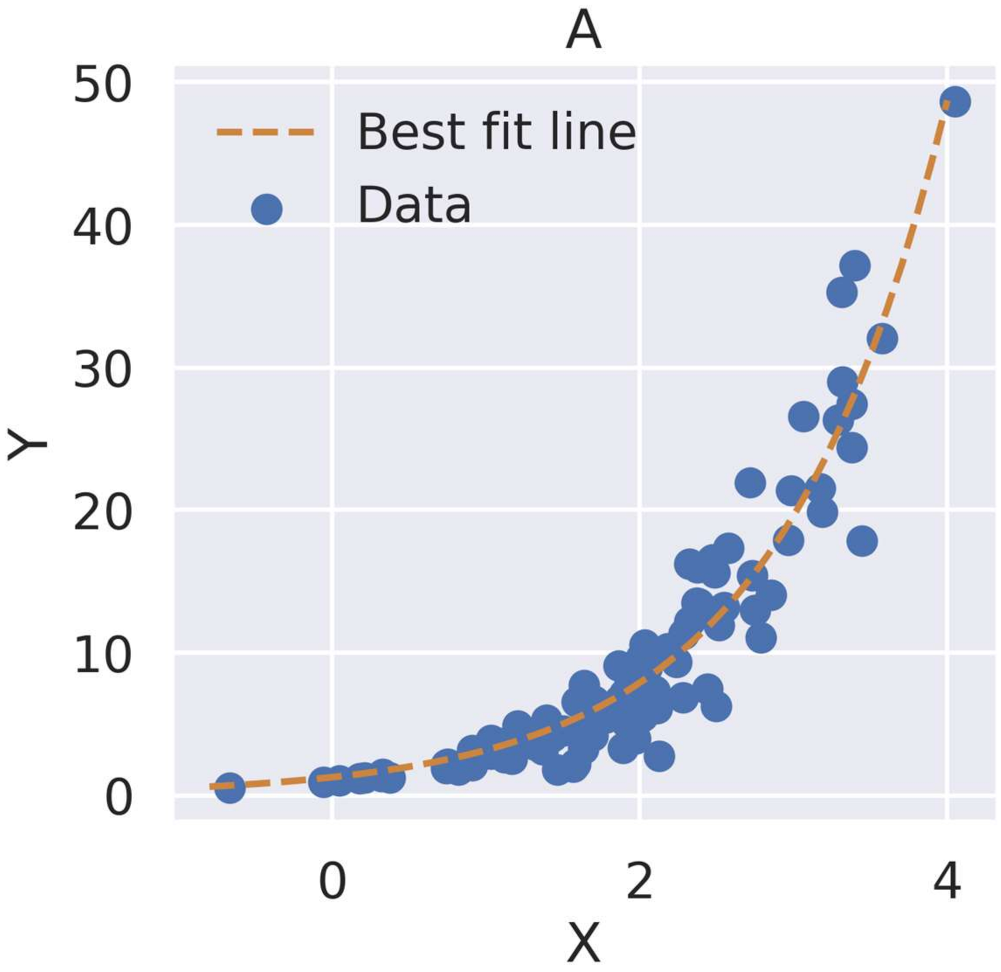
- Skewed response variables can cause errors
- LSLR can handle heteroscedasticity, but the R^2 will be wrong
- Examples: Ahmadi et al (2013) (shown), Rezaee et al (2006)



(a)Training Phase, R^2=0.93885

Synthetic example of skewed inputs

- Input: Y ~ exp(X) + e,
  - e ~ N(0, X^2)
- Model for fit: Y ~ exp(X)
- R^2: 0.72



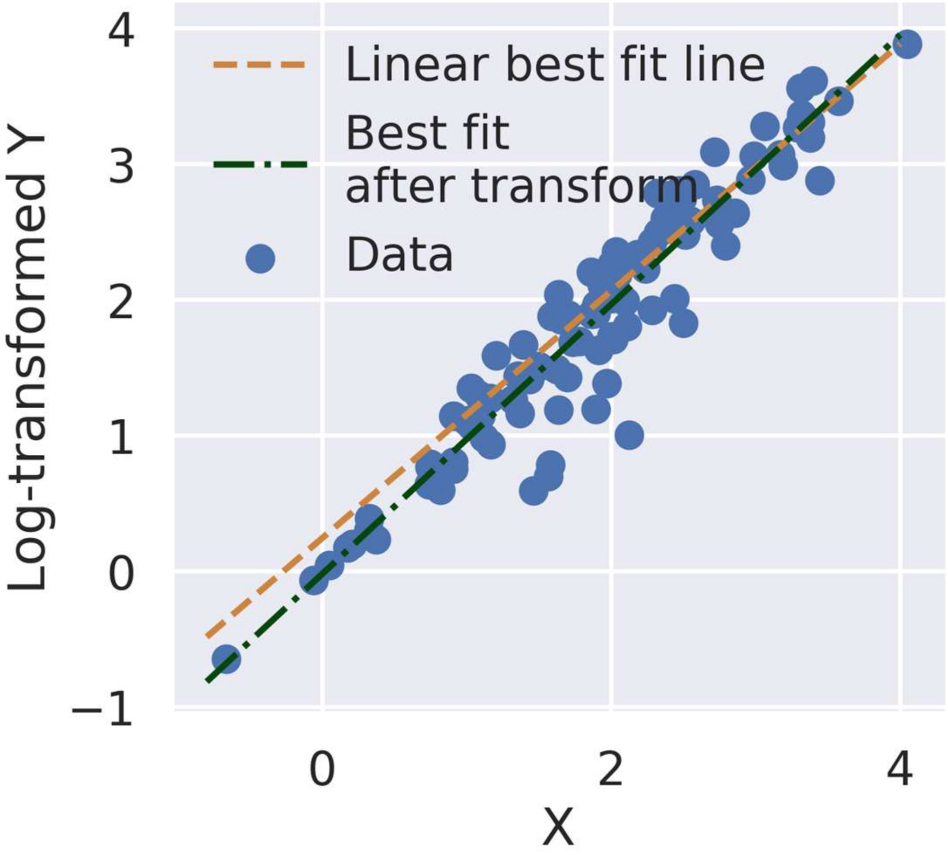
Identifying and correcting skewed inputs

Diagnosis:

- Histogram your inputs
- Check residuals for variance

Solutions:

- Transform variables toward Gaussian
- Use heteroscedasticity-consistent standard errors
- If heteroscedasticity is hurting your regression
  - Run robust regression
  - Trim inputs (very lightly)



Conclusions

- Mistake 1
  - Role switching of linear regression lines gives biased results
  - Winland relation often mis-characterized
- Mistake 2
  - Linear regression with log-transformed response gives biased results
  - De-biasing requires care
- Mistake 3
  - R^2 is can mislead
    - it expects low auto-correlation
    - no change in variance for errors
  - Check your input distributions, plot your residuals
  - If your R^2 is too good to believe, don't believe it

Acknowledgments

- Discussions with colleagues
  - Ian Duncans
  - Larry Lake
  - Behzad Ghanbarian
  - Michael Marder
  - Chris Clarkson
- STARR funding
- Former students (of Jerry's)
  - Jianwei Di
  - Danial Kaviani

STARR State of Texas Advanced Resource Recovery

Need more? Read the preprint at https://eartharxiv.org/repository/view/253/

State of Texas Advanced Resource Recovery

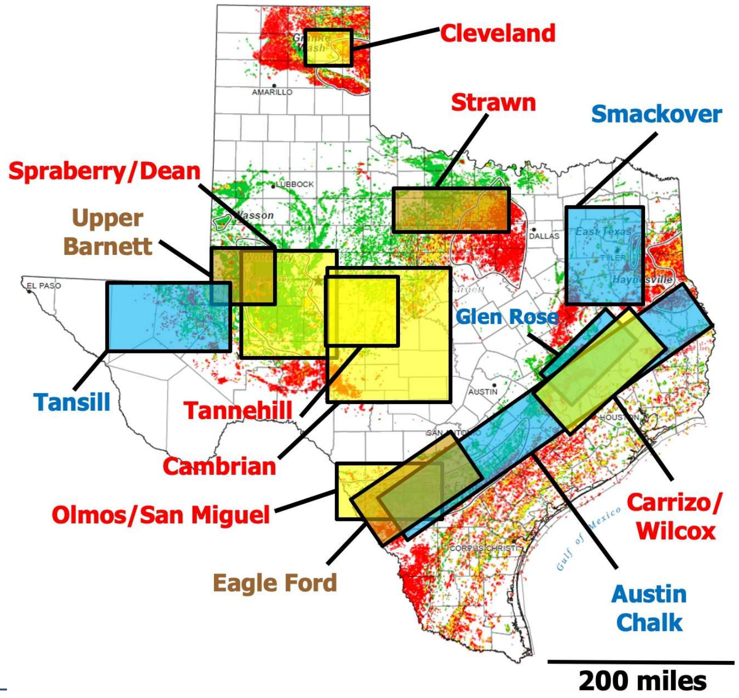
The STARR Mission is to offer research support to help companies in Texas keep energy affordable and plentiful.

To achieve this, they perform regional studies and technology transfer to Texas operators.

The philosophy of STARR is to work with operators in Texas to:

- Maximize recovery efficiency
- Explore in new plays
- Exploit unconventional resources

- STARR PI: Lorena Moscardelli (lorena.moscardelli@beg.utexas.edu)


Cleveland
Strawn
Smackover
Spraberry/Dean
Upper Barnett
Glen Rose
Tansill
Tannehill
Cambrian
Olmos/San Miguel
Carrizo/ Wilcox
Eagle Ford
Austin Chalk
AMARILLO
LUBBOCK
EL PASO
DALLAS
AUSTIN
CORPUS CHRISTI
Gulf of Mexico
200 miles