Regression Analysis — What You Should've Been Taught But Weren't, and Were Taught But Shouldn't Have Been

Norm Matloff
University of California at Davis

Bay Area R Users Group
Menlo Park, 19 September, 2017

These slides will be available at
`http://heather.cs.ucdavis.edu/barug0917.pdf`

The Book

The Book

- Was asked to speak about my new book, *Statistical Regression and Classification: From Linear Modelss to Machine Learning*, CRC, 2017

The Book

- Was asked to speak about my new book, *Statistical Regression and Classification: From Linear Modelss to Machine Learning*, CRC, 2017
- I'd wanted to write this book for 30 years, finally did!

What the Book Is NOT

What the Book Is NOT

- It is NOT a computer science book!

What the Book Is NOT

- It is NOT a computer science book!
- Does have a considerable amount of computational material, and uses various CRAN packages, including my **regtools** package.

What the Book Is NOT

- It is NOT a computer science book!
- Does have a considerable amount of computational material, and uses various CRAN packages, including my **regtools** package.
- But if you are looking for a compendium of the ∞-ly many options in **lm()**, or for that matter **caret**, this is not the book for you.

What the Book IS

What the Book IS

- It is STATISTICS book (call it machine learning, if you insist).

What the Book IS

- It is STATISTICS book (call it machine learning, if you insist).
- Tells the REAL TRUTH about regression and classification (as I see it).

What the Book IS

- It is STATISTICS book (call it machine learning, if you insist).
- Tells the REAL TRUTH about regression and classification (as I see it).
- Uses some math (precise formulation of issues, a fair amount of linear algebra, a bit of calculus). but not "math for math's sake."

What the Book IS

- It is STATISTICS book (call it machine learning, if you insist).
- Tells the REAL TRUTH about regression and classification (as I see it).
- Uses some math (precise formulation of issues, a fair amount of linear algebra, a bit of calculus). but not "math for math's sake."
- E.g., Gauss-Markov Theorem is treated as a curiosity, not a central result.

What the Book IS

- It is STATISTICS book (call it machine learning, if you insist).
- Tells the REAL TRUTH about regression and classification (as I see it).
- Uses some math (precise formulation of issues, a fair amount of linear algebra, a bit of calculus). but not "math for math's sake."
- E.g., Gauss-Markov Theorem is treated as a curiosity, not a central result.
- No "Step A, Step B, Step C" formula-plugging!

What the Book IS

- It is STATISTICS book (call it machine learning, if you insist).
- Tells the REAL TRUTH about regression and classification (as I see it).
- Uses some math (precise formulation of issues, a fair amount of linear algebra, a bit of calculus). but not "math for math's sake."
- E.g., Gauss-Markov Theorem is treated as a curiosity, not a central result.
- No "Step A, Step B, Step C" formula-plugging!
- Some sample myth-busting follows.

Myth #I

Myth #I

Myth #1: "Exact" inference in linear models, based on normally distributed Y, homogeneous $Var(Y \mid X)$, etc.

Myth #I

Myth #1: "Exact" inference in linear models, based on normally distributed Y, homogeneous $Var(Y \mid X)$, etc.

- Of course you already knew that is a myth.

Myth #I

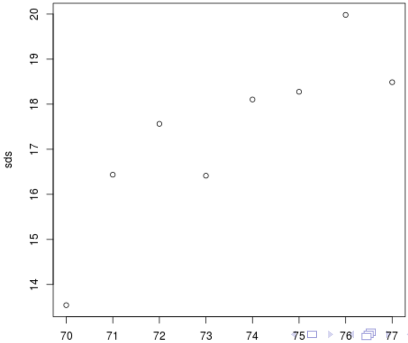
Myth #1: "Exact" inference in linear models, based on normally distributed Y, homogeneous $Var(Y \mid X)$, etc.

- Of course you already knew that is a myth. E.g. no person is 90' tall or has a negative height, thus not normally distributed.

Myth #I

Myth #1: "Exact" inference in linear models, based on normally distributed Y, homogeneous $Var(Y \mid X)$, etc.

- Of course you already knew that is a myth. E.g. no person is 90' tall or has a negative height, thus not normally distributed.
- In typical applications, $s.d.(Y \mid X)$ increases with X, e.g.



Myth #1, cont'd.

Myth #1, cont'd.

- We must accept the fact that the assumptions are only approximate at best.

Myth #1, cont'd.

- We must accept the fact that the assumptions are only approximate at best.
- $\widehat{\beta}$ is approximately MV normal even if the sampled population is not.

Myth #1, cont'd.

- We must accept the fact that the assumptions are only approximate at best.
- $\widehat{\beta}$ is approximately MV normal even if the sampled population is not. So use Z instead of t, χ^2 instead of F etc.

Myth #1, cont'd.

- We must accept the fact that the assumptions are only approximate at best.
- $\widehat{\beta}$ is approximately MV normal even if the sampled population is not. So use Z instead of t, χ^2 instead of F etc.
- To deal with the heteroscedasticity, use the *sandwich estimator*. Widely available, e.g. in CRAN packages **car**, **regtools** (nonlin. reg. case) and **sandwich**.

Myth #2

Myth #2

Myth #2: Transformations of the data (e.g. log , $\sqrt{}$) are usually/often a good idea.

Myth #2

Myth #2: Transformations of the data (e.g. $\log$, $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.

Myth #2

Myth #2: Transformations of the data (e.g. $\log$, $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what?

Norm Matloff
University of California at Davis

Myth #2

Myth #2: Transformations of the data (e.g. $\log$, $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what? To achieve normality?

Myth #2

Myth #2: Transformations of the data (e.g. $\log$, $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what? To achieve normality? See Myth #1!

Myth #2

Myth #2: Transformations of the data (e.g. $\log$, $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what? To achieve normality? See Myth #1!
- FDA actually recommends against transformations.

Myth #2

Myth #2: Transformations of the data (e.g. log , $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what? To achieve normality? See Myth #1!
- FDA actually recommends against transformations.
- Example: Poisson regression.

Myth #2

Myth #2: Transformations of the data (e.g. $\log$, $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what? To achieve normality? See Myth #1!
- FDA actually recommends against transformations.
- Example: Poisson regression.
    - Basically applies a log transformation.

Myth #2

Myth #2: Transformations of the data (e.g. log , $\sqrt{}$) are usually/often a good idea.

- Big distortion, unclear interpretation of coefficients.
- And for what? To achieve normality? See Myth #1!
- FDA actually recommends against transformations.
- Example: Poisson regression.
  - Basically applies a log transformation.
  - But in my book's example (Pima from UCI Machine Learning Data Repository), untransformed Poisson model had a 25% better predictive ability.

Myth #3

Myth #3

Myth #3: R^2 is only for linear models.

Myth #3

Myth #3: R^2 is only for linear models.

- R^2 (on either sample or population level) is the squared correlation between Y and $\widehat{Y}$.

Myth #3

Myth #3: R^2 is only for linear models.

- R^2 (on either sample or population level) is the squared correlation between Y and $\widehat{Y}$.
- Thus is defined for any regression procedure, even nonparametric ones like k-Nearest Neighbor.

Myth #3

Myth #3: R^2 is only for linear models.

- R^2 (on either sample or population level) is the squared correlation between Y and $\widehat{Y}$.
- Thus is defined for any regression procedure, even nonparametric ones like k-Nearest Neighbor.
- Example: Currency data.

Currency data

Norm Matloff
University of California at Davis

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

- Predict *yen* from the rest.

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

- Predict *yen* from the rest.
- Straight linear model yields $R^2 = 0.89$.

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

- Predict *yen* from the rest.
- Straight linear model yields $R^2 = 0.89$. Not bad!

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

- Predict *yen* from the rest.
- Straight linear model yields $R^2 = 0.89$. Not bad!
- But k-NN yields $R^2 = 0.98$.

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

- Predict *yen* from the rest.
- Straight linear model yields $R^2 = 0.89$. Not bad!
- But k-NN yields $R^2 = 0.98$.
- So by using straight linear model we are "leaving money on the table."

Currency data

Currency data, pre-Euro; *franc* and *mark*, plus *pound*, *yen* and Canadian *dollar*.

- Predict *yen* from the rest.
- Straight linear model yields $R^2 = 0.89$. Not bad!
- But k-NN yields $R^2 = 0.98$.
- So by using straight linear model we are "leaving money on the table."
- By exploring what's wrong with the fit, we might gain additional insight.

Currency, cont'd.

Currency, cont'd.

- So, let's do some diagnostic plots.

Currency, cont'd.

- So, let's do some diagnostic plots.
- But not THOSE plots e.g. Y vs. linear fitted $\widehat{Y}$.

Currency, cont'd.

- So, let's do some diagnostic plots.
- But not THOSE plots e.g. Y vs. linear fitted $\widehat{Y}$.
- My **regtools** package includes a number of functions that have one use *nonparametric* estimation to aid in assessing *parametric* fit.

Norm Matloff
University of California at Davis

Currency, cont'd.

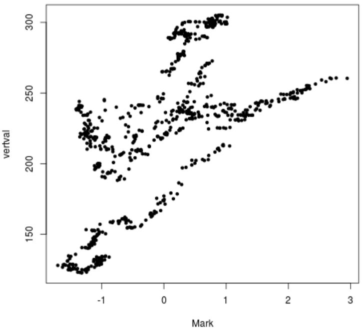
Norm Matloff
University of California at Davis

Currency, cont'd.

E.g., plot nonparametric estimated reg. function (NOT the Y_i) against each predictor $X^{(i)}$, such as

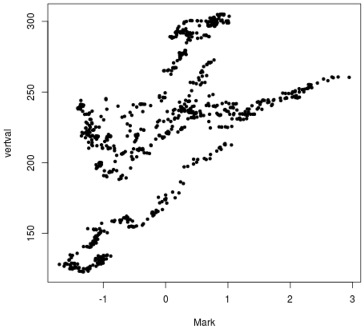
Currency, cont'd.

E.g., plot nonparametric estimated reg. function (NOT the Y_i) against each predictor $X^{(i)}$, such as



Currency, cont'd.

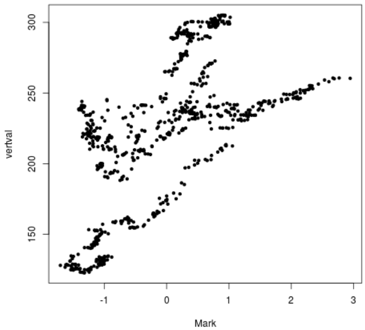
E.g., plot nonparametric estimated reg. function (NOT the Y_i) against each predictor $X^{(i)}$, such as



Whoa! Quite a departure from linear.

Currency, cont'd.

E.g., plot nonparametric estimated reg. function (NOT the Y_i) against each predictor $X^{(i)}$, such as



Whoa! Quite a departure from linear. Need a domain expert to figure out what's happening, but clearly there are some dynamics lurking here that need to be investigated.

Myth #4

Myth #4

Myth #4: "Unbalanced" data in classification applications is a problem (and can be solved).

Myth #4

Myth #4: "Unbalanced" data in classification applications is a problem (and can be solved).

- Say want to predict presence or absence of a disease (Y) from the results of a blood test (X).

Myth #4

Myth #4: "Unbalanced" data in classification applications is a problem (and can be solved).

- Say want to predict presence or absence of a disease (Y) from the results of a blood test (X).
- Say we have a sample of 100 patients, and via followup know the disease status for all.

Myth #4

Myth #4: "Unbalanced" data in classification applications is a problem (and can be solved).

- Say want to predict presence or absence of a disease (Y) from the results of a blood test (X).
- Say we have a sample of 100 patients, and via followup know the disease status for all.
- Say in the sample 8 have the disease, 92 don't.

Myth #4

Myth #4: "Unbalanced" data in classification applications is a problem (and can be solved).

- Say want to predict presence or absence of a disease (Y) from the results of a blood test (X).
- Say we have a sample of 100 patients, and via followup know the disease status for all.
- Say in the sample 8 have the disease, 92 don't.
- Much public angst and handwriting by "experts." Unbalanced data, oh no, what can we do?!

Unbalanced people, cont'd.

Unbalanced people, cont'd.

Think about your goals:

Unbalanced people, cont'd.

Think about your goals:

- If your goal is to maximize your overall rate of correct classification, **there is nothing wrong**.

Unbalanced people, cont'd.

Think about your goals:

- If your goal is to maximize your overall rate of correct classification, **there is nothing wrong**. The data as in its present form is the best you can do.

Unbalanced people, cont'd.

Think about your goals:

- If your goal is to maximize your overall rate of correct classification, **there is nothing wrong**. The data as in its present form is the best you can do.
- Most classification software implicitly assumes the goal as above.

Unbalanced people, cont'd.

Think about your goals:

- If your goal is to maximize your overall rate of correct classification, **there is nothing wrong**. The data as in its present form is the best you can do.
- Most classification software implicitly assumes the goal as above.
- If you wish a better rate for a certain subpopulation (guess disease present when it is), at the expense of other subpopulations, you can TRICK the software, by artificially accentuating the weight of one class or another.

Unbalanced people, cont'd.

Think about your goals:

- If your goal is to maximize your overall rate of correct classification, **there is nothing wrong**. The data as in its present form is the best you can do.
- Most classification software implicitly assumes the goal as above.
- If you wish a better rate for a certain subpopulation (guess disease present when it is), at the expense of other subpopulations, you can TRICK the software, by artificially accentuating the weight of one class or another.
- Fine if you know what you are doing. Note the IF!

Computer Science-ization of Statistics

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

- Someone asked me the other day, "What is a good package for PCA?"

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

- Someone asked me the other day, "What is a good package for PCA?"
- He was not interested in what PCA actually does.

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

- Someone asked me the other day, "What is a good package for PCA?"
- He was not interested in what PCA actually does. He treated it as just another kind of programming.

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

- Someone asked me the other day, "What is a good package for PCA?"
- He was not interested in what PCA actually does. He treated it as just another kind of programming.
- Two others whom I really respect displayed the same attitudes recently.

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

- Someone asked me the other day, "What is a good package for PCA?"
- He was not interested in what PCA actually does. He treated it as just another kind of programming.
- Two others whom I really respect displayed the same attitudes recently.
- Antidote to CS-ization of stat:

Computer Science-ization of Statistics

Contrary to popular opinion, **statistics is not a branch of computer science**.

- Someone asked me the other day, "What is a good package for PCA?"
- He was not interested in what PCA actually does. He treated it as just another kind of programming.
- Two others whom I really respect displayed the same attitudes recently.
- Antidote to CS-ization of stat: My book!