Regression Fit Diagnostics Using freqparcoord

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

useR! 2014
UCLA
July 1, 2014

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Intro to freqparcoord

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Intro to freqparcoord

Overview of **freqparcoord**:

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Intro to freqparcoord

Overview of **freqparcoord**:

- Available on CRAN.

Intro to freqparcoord

Overview of **freqparcoord**:

- Available on CRAN.
- New approach to the parallel coordinates data visualization method.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Intro to freqparcoord

Overview of **freqparcoord**:

- Available on CRAN.
- New approach to the parallel coordinates data visualization method. (Examples presented shortly.)

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Intro to freqparcoord

Overview of **freqparcoord**:

- Available on CRAN.
- New approach to the parallel coordinates data visualization method. (Examples presented shortly.)
- Can also be used for hunting outliers, clusters...

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Intro to freqparcoord

Overview of **freqparcoord**:

- Available on CRAN.
- New approach to the parallel coordinates data visualization method. (Examples presented shortly.)
- Can also be used for hunting outliers, clusters...
- and for regression diagnostics—our topic here.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

What Is Parallel Coordinates Visualization?

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- Very old idea.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- Very old idea.
- If have k variables, draw k vertical axes.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

What Is Parallel Coordinates Visualization?

- Very old idea.
- If have k variables, draw k vertical axes. Each data point is a polygonal line connecting the value of each variable.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Example

Example

Example: Height/weight/age data.

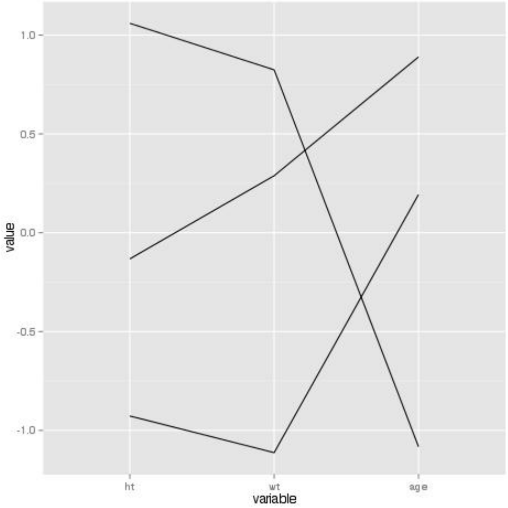
Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Example

Example: Height/weight/age data.

```
> d
  ht  wt age
1 71 175  25
2 66 128  36
3 68 162  42
```



Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few "typical" lines,

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, “black screen” problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.

Problems with Parallel Coordinates

- Highly cluttered, “black screen” problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.
- E.g., height/weight/age:

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, “black screen” problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.
- E.g., height/weight/age:
    - What height/weight/age combinations are typical **overall**?

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, “black screen” problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.
- E.g., height/weight/age:
  - What height/weight/age combinations are typical **overall**?
  - What height/weight/age combinations are typical within groups?

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- E.g., height/weight/age:
  - What height/weight/age combinations are typical **overall**?
  - What height/weight/age combinations are typical within groups? **Group comparison.**

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- E.g., height/weight/age:
    - What height/weight/age combinations are typical **overall**?
    - What height/weight/age combinations are typical within groups? **Group comparison.**
    - What height/weight/age combinations are rare?

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- E.g., height/weight/age:
    - What height/weight/age combinations are typical **overall**?
    - What height/weight/age combinations are typical within groups? **Group comparison.**
    - What height/weight/age combinations are rare? **Outlier hunting.**

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- E.g., height/weight/age:
    - What height/weight/age combinations are typical **overall**?
    - What height/weight/age combinations are typical within groups? **Group comparison.**
    - What height/weight/age combinations are rare? **Outlier hunting.**
    - What height/weight/age combinations are "locally typical"?

Problems with Parallel Coordinates

- Highly cluttered, “black screen” problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.
- E.g., height/weight/age:
  - What height/weight/age combinations are typical **overall**?
  - What height/weight/age combinations are typical within groups? **Group comparison.**
  - What height/weight/age combinations are rare? **Outlier hunting.**
  - What height/weight/age combinations are “locally typical”? **Cluster hunting.**

Problems with Parallel Coordinates

- Highly cluttered, "black screen" problem.
- Various solutions, e.g. making the lines fainte, or combining themr
- Our solution: Plot only a few "typical" lines, based on estimated multivariate density.
- E.g., height/weight/age:
  - What height/weight/age combinations are typical **overall**?
  - What height/weight/age combinations are typical within groups? **Group comparison.**
  - What height/weight/age combinations are rare? **Outlier hunting.**
  - What height/weight/age combinations are "locally typical"? **Cluster hunting.**

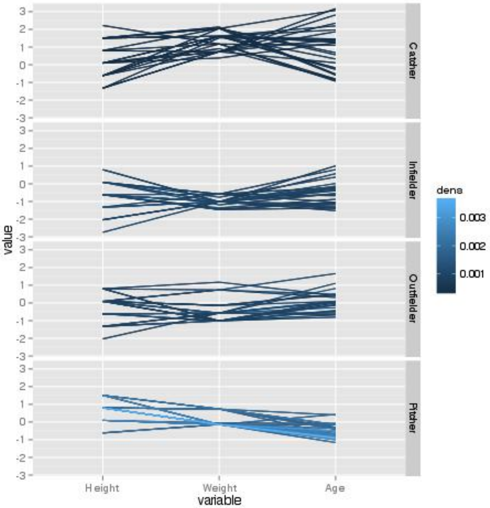
Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCLA Baseball Player Data

UCLA Baseball Player Data

Most typical 25 points for each playing position.



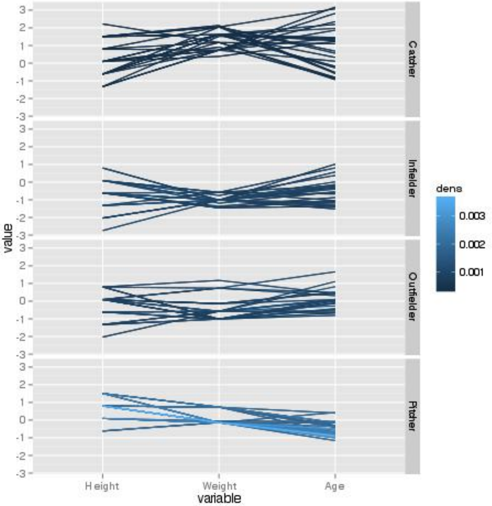
- Catchers heavier, vary widely in height and age.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCLA Baseball Player Data

Most typical 25 points for each playing position.



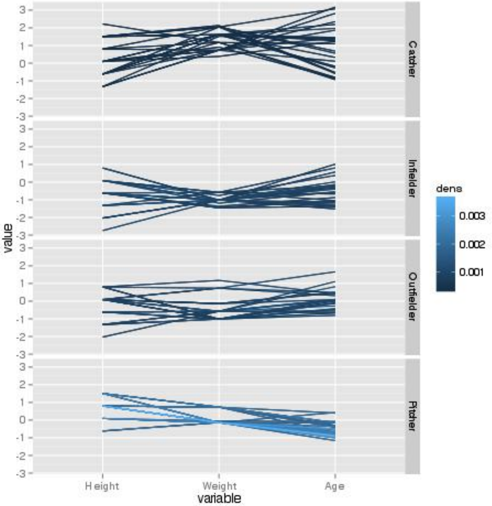
- Catchers heavier, vary widely in height and age.
- Pitchers tall, lighter, less variable in age.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCLA Baseball Player Data

Most typical 25 points for each playing position.



- Catchers heavier, vary widely in height and age.
- Pitchers tall, lighter, less variable in age.
- Infielders vary considerably in height but not weight.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Application to Regression Diagnostics

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Application to Regression Diagnostics

Our **freqparcoord** package includes a function **regdiag()**.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Application to Regression Diagnostics

Our **freqparcoord** package includes a function **regdiag()**.

- Focused vertical axis:

```
divergences = fitted parametric model -
              fitted nonparametric model
```

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Application to Regression Diagnostics

Our **freqparcoord** package includes a function **regdiag()**.

- Focused vertical axis:

```
divergences = fitted parametric model -
              fitted nonparametric model
```

  (Uses k-NN for nonparametric est..)

- The divergences are NOT the residuals (i.e. not actual - fitted parametric).

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Application to Regression Diagnostics

Our **freqparcoord** package includes a function **regdiag()**.

- Focused vertical axis:

```
divergences = fitted parametric model -
              fitted nonparametric model
```

  (Uses k-NN for nonparametric est..)

- The divergences are NOT the residuals (i.e. not actual - fitted parametric).
- What **regdiag()** does it look at the typical values among the most negative and most positive divergences.

Application to Regression Diagnostics

Our **freqparcoord** package includes a function **regdiag()**.

- Focused vertical axis:

```
divergences = fitted parametric model -
              fitted nonparametric model
```

  (Uses k-NN for nonparametric est..)

- The divergences are NOT the residuals (i.e. not actual - fitted parametric).
- What **regdiag()** does it look at the typical values among the most negative and most positive divergences.
- In other words: **regdiag()** asks, "In what region[s] of predictor space is the fit poorer?"

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Example

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Example

Programmers and engineers in Silicon Valley, 2000 Census, 5% PUMS.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Example

Programmers and engineers in Silicon Valley, 2000 Census, 5% PUMS.

```
> data(prgeng)  # fpc. built-in data set
> pg1 <- prgeng
> pg1$ms <- as.integer(pg1$educ == 14)  # MS
> pg1$phd <- as.integer(pg1$educ == 16)  # PhD
> pg1$se <- as.integer(pg1$occ==102)  # s. eng
> l1 <- lm(wageinc ~ age+ms+phd+se+sex,data=pg1
# look at 40% most neg., 40% most pos. divs.
> p <- regdiag(l1,tail=0.40)
> p  # display graph
> p$paramr2  # parametric adj. R2
[1] 0.07027561
> p$nonparamr2  # nonparamr2 R2
[1] 0.1286746
```

Norm Matloff and Yingkang Xie
University of California at Davis

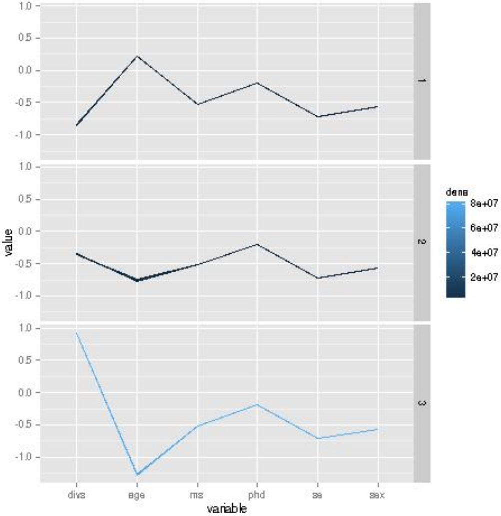
e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Analysis of PUMS Data

Norm Matloff and Yingkang Xie
University of California at Davis

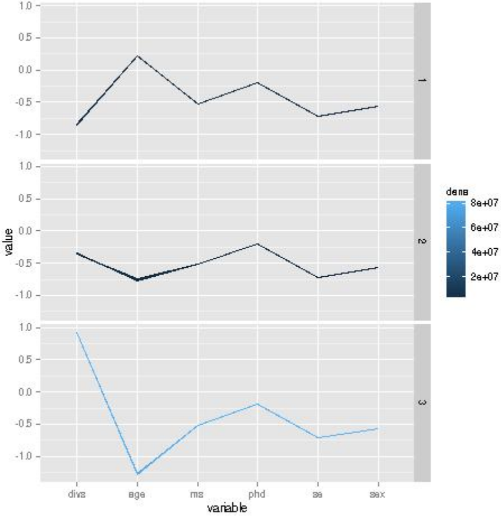
e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Analysis of PUMS Data



- Both R^2 values low, but nonpar. 83% higher.

Analysis of PUMS Data

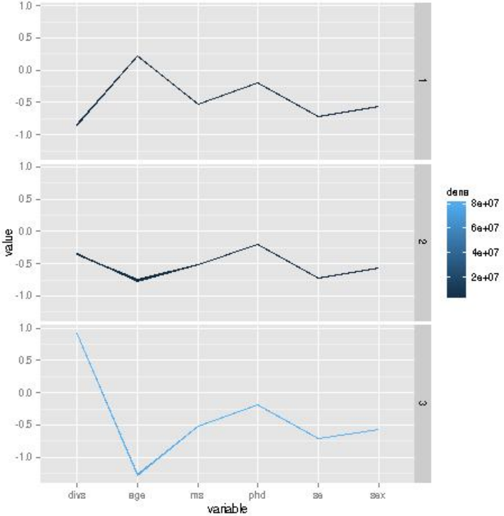


- Both R^2 values low, but nonpar. 83% higher. Room for improvement in param. model!

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Analysis of PUMS Data

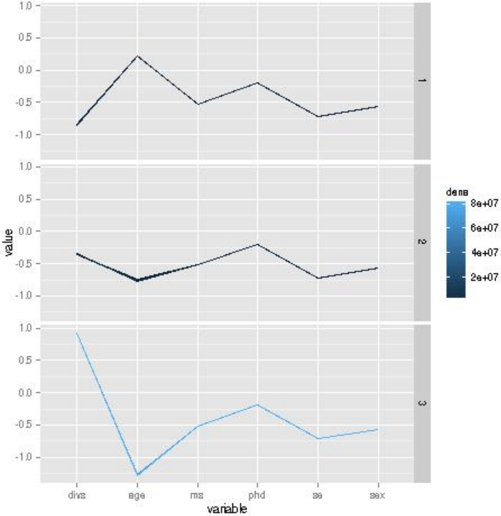


- Both R^2 values low, but nonpar. 83% higher. Room for improvement in param. model!
- The Age variable seems to be the culprit:

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Analysis of PUMS Data



- Both R^2 values low, but nonpar. 83% higher. Room for improvement in param. model!
- The Age variable seems to be the culprit: Overpredict for younger, underpredict for older.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Add Quadratic Term

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Add Quadratic Term

- The "typical divergences" plot suggested adding a quadratic term for Age:

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Add Quadratic Term

- The "typical divergences" plot suggested adding a quadratic term for Age:

```
pg1$age2 <- pg1$age^2
l2 <- lm(wageinc ~
    age+age2+ms+phd+se+sex, data=pg1)
```

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Add Quadratic Term

- The "typical divergences" plot suggested adding a quadratic term for Age:

```
pg1$age2 <- pg1$age^2
l2 <- lm(wageinc ~
    age+age2+ms+phd+se+sex, data=pg1)
```

- This brought adj. R^2 up from 0.07 to 0.13.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCI Adult Data

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCI Adult Data

Can use **regdiag()** for generalized linear models too, e.g. logit.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCI Adult Data

Can use **regdiag()** for generalized linear models too, e.g. logit.

- Predict a binary High Income variable, from Education, Age, Gender, Married.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCI Adult Data

Can use **regdiag()** for generalized linear models too, e.g. logit.

- Predict a binary High Income variable, from Education, Age, Gender, Married.
- The **regdiag()** plot shows younger women overpredicted, men underpredicted.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCI Adult Data

Can use **regdiag()** for generalized linear models too, e.g. logit.

- Predict a binary High Income variable, from Education, Age, Gender, Married.
- The **regdiag()** plot shows younger women overpredicted, men underpredicted.
- Thus, might add Age $\times$ Gender interaction term.

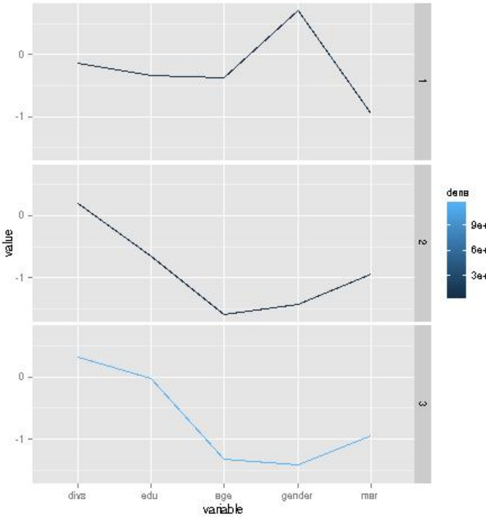
Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

UCI Adult Data

Can use **regdiag()** for generalized linear models too, e.g. logit.

- Predict a binary High Income variable, from Education, Age, Gender, Married.
- The **regdiag()** plot shows younger women overpredicted, men underpredicted.
- Thus, might add Age $\times$ Gender interaction term.



Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

More on Adult Data

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

More on Adult Data

- Calls:

```
g1 <-
   glm(gt50 ~ edu + age + gender + mar,
   data=newadult, family=binomial)
regdiag(g1)
```

More on Adult Data

- Calls:

```
g1 <-
   glm(gt50 ~ edu + age + gender + mar,
   data=newadult, family=binomial)
regdiag(g1)
```

- Addition of interaction term:

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

More on Adult Data

- Calls:

```
g1 <-
    glm(gt50 ~ edu + age + gender + mar,
    data=newadult, family=binomial)
regdiag(g1)
```

- Addition of interaction term:
  - Did NOT improve correct-classification rate (81%).

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

More on Adult Data

- Calls:

```
g1 <-
   glm(gt50 ~ edu + age + gender + mar,
   data=newadult, family=binomial)
regdiag(g1)
```

- Addition of interaction term:
  - Did NOT improve correct-classification rate (81%).
  - BUT changed $\widehat{\beta}_{Gender}$ a lot, from 0.351 to 0.610. Interaction term -0.006.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

More on Adult Data

- Calls:

```
g1 <-
    glm(gt50 ~ edu + age + gender + mar,
    data=newadult, family=binomial)
regdiag(g1)
```

- Addition of interaction term:
  - Did NOT improve correct-classification rate (81%).
  - BUT changed $\widehat{\beta}_{Gender}$ a lot, from 0.351 to 0.610. Interaction term -0.006. Male "advantage" in log-odds ratio now becomes, e.g. 0.46 at age 25, only 0.28 at age 55.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Summary

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
- The package includes a function **regdiag()** that applies these ideas to regression model diagnostics.

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
- The package includes a function **regdiag()** that applies these ideas to regression model diagnostics.
    - Computes "divergences," i.e. par. fit - nonpar. fit.

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
- The package includes a function **regdiag()** that applies these ideas to regression model diagnostics.
  - Computes "divergences," i.e. par. fit - nonpar. fit.
  - Applies **freqparcoord** to find the most typical divergences, among the most neg. and most pos.

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
- The package includes a function **regdiag()** that applies these ideas to regression model diagnostics.
    - Computes "divergences," i.e. par. fit - nonpar. fit.
    - Applies **freqparcoord** to find the most typical divergences, among the most neg. and most pos.
    - Also reports par., nonpar. R^2 values to see whether par. model is "leaving money on the table."

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
- The package includes a function **regdiag()** that applies these ideas to regression model diagnostics.
  - Computes "divergences," i.e. par. fit - nonpar. fit.
  - Applies **freqparcoord** to find the most typical divergences, among the most neg. and most pos.
  - Also reports par., nonpar. R^2 values to see whether par. model is "leaving money on the table."
  - Plots suggest quad., interaction terms to add.

Summary

- The **freqparcoord** package plots only "typical" lines, thus avoding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
- The package includes a function **regdiag()** that applies these ideas to regression model diagnostics.
  - Computes "divergences," i.e. par. fit - nonpar. fit.
  - Applies **freqparcoord** to find the most typical divergences, among the most neg. and most pos.
  - Also reports par., nonpar. R^2 values to see whether par. model is "leaving money on the table."
  - Plots suggest quad., interaction terms to add.
- Location of these slides:
  `http://heather.cs.ucdavis.edu/freqparcoord/Slides.pdf`