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
Intro to freqparcoord
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.
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.
Intro to freqparcoord

Overview of freqparcoord:

- Available on CRAN.
- New approach to the parallel coordinates data visualization method. (Examples presented shortly.)
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...
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.
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.
What Is Parallel Coordinates Visualization?

- Very old idea.
What Is Parallel Coordinates Visualization?

- Very old idea.
- If have k variables, draw k vertical axes.
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.
Example: Height/weight/age data.

```r
> dht <- c(71, 66, 68)
> wt <- c(175, 128, 162)
> age <- c(25, 36, 42)
```

Norm Matloff and Yingkang Xie
University of California at Davis

e-mail: matloff@cs.ucdavis.edu
R/stat blog: matloff.wordpress.com
Example: Height/weight/age data.
Example: Height/weight/age data.

> d
  ht    wt  age
  1  71  175  25
  2  66  128  36
  3  68  162  42
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.
Problems with Parallel Coordinates

- Various solutions, e.g. making the lines fainter, or combining them.
  - 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

- Various solutions, e.g. making the lines fainter, or combining them.
- Our solution: Plot only a few “typical” lines,
Problems with Parallel Coordinates

- Various solutions, e.g. making the lines fainter, or combining them
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.
Problems with Parallel Coordinates

- Various solutions, e.g. making the lines fainter, or combining them.
- Our solution: Plot only a few “typical” lines, based on estimated multivariate density.
- E.g., height/weight/age:
Problems with Parallel Coordinates

- Various solutions, e.g. making the lines fainter, or combining them.
- 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?
Problems with Parallel Coordinates

- Various solutions, e.g. making the lines fainter, or combining them.
- 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?
Problems with Parallel Coordinates

- Various solutions, e.g. making the lines fainter, or combining them.
- 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

- Various solutions, e.g. making the lines fainter, or combining them.
- 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

- Various solutions, e.g. making the lines fainter, or combining them.
- 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.**
Problems with Parallel Coordinates

• Highly cluttered, “black screen” problem.
• Various solutions, e.g. making the lines fainter, or combining them.
• 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

- Various solutions, e.g. making the lines fainter, or combining them.
- 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

- Various solutions, e.g. making the lines fainter, or combining them.
- 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.**
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.
UCLA Baseball Player Data

Most typical 25 points for each playing position.

- Catchers heavier, vary widely in height and age.
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.
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.
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?"
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 is 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?"
Application to Regression Diagnostics

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

- Focused vertical axis:
  
  \[
  \text{divergences} = \text{fitted parametric model} - \text{fitted nonparametric model}
  \]
Application to Regression Diagnostics

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

- **Focused vertical axis:**
  
  \[ \text{divergences} = \text{fitted parametric model} - \text{fitted nonparametric model} \]

  (Uses k-NN for nonparametric est.~)

- **The divergences are NOT the residuals (i.e. not actual - fitted parametric).**
Application to Regression Diagnostics

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

- **Focused vertical axis:**
  \[
  \text{divergences} = \text{fitted parametric model} - \text{fitted nonparametric model}
  \]
  (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.
Our `freqparcoord` package includes a function `regdiag()`.

- **Focused vertical axis:**
  
  \[
  \text{divergences} = \text{fitted parametric model} - \text{fitted nonparametric model}
  \]

  (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?”
Example

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

```r
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)

p <- regdiag(l1, tail=0.40)
p
```

```
> p$paramr2
[1] 0.07027561
>
> p$nonparamr2
[1] 0.1286746
```
Example

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

```r
# Example code

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.e.

l1 <- lm(wageinc ~ age + ms + phd + se + sex, data = pg1)

p <- reg.diag(l1, tail = 0.40)

disp(p)

p$paramr2
# parametric adj. R^2

p$nonparamr2
# nonparametric R^2
```
Example

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

```r
> 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
```
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.
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!
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:
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.
Add Quadratic Term

The “typical divergences” plot suggested adding a quadratic term for Age:

\[ \text{pg1} \cdot \text{age2} \cdot \text{pg1} \cdot \text{age}^2 \cdot \text{lm} \cdot \text{wageinc} \cdot \text{age} \cdot \text{age2} \cdot \text{ms} \cdot \text{phd} \cdot \text{se} \cdot \text{sex}, \text{data} = \text{pg1} \]

This brought adj. R\(^2\) up from 0.07 to 0.13.
The “typical divergences” plot suggested adding a quadratic term for Age:
Add Quadratic Term

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

```r
pg1$age2 <- pg1$age^2
l2 <- lm(wageinc ~ age+age2+ms+phd+se+sex, data=pg1)
```
The “typical divergences” plot suggested adding a quadratic term for Age:

```r
go1$age2 <- go1$age^2
go2 <- lm(wageinc ~ age+age2+ms+phd+se+sex, data=go1)
```

This brought adj. $R^2$ up from 0.07 to 0.13.
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 × Gender interaction term.
UCI Adult Data

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

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.
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.
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 × Gender interaction term.
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.
More on Adult Data

• Calls:
  g1 <- glm(gt50 ~ edu + age + gender + mar, data=newadult, family=binomial)
  reg.diag(g1)

• Addition of interaction term:

  • Did NOT improve correct-classification rate (81%).
  • BUT changed \( \hat{\beta} \) 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.
More on Adult Data

• Calls:

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

• Calls:

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

• Addition of interaction term:
More on Adult Data

• Calls:

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

• Addition of interaction term:
  • Did NOT improve correct-classification rate (81%).
More on Adult Data

• Calls:

```r
  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 $\hat{\beta}_{Gender}$ a lot, from 0.351 to 0.610.
    Interaction term -0.006.
More on Adult Data

- Calls:

```r
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 $\hat{\beta}_{Gender}$ 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.
Summary

The `freqparcoord` package plots only "typical" lines, thus avoiding 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. parametric fit - nonparametric fit.

Applies `freqparcoord` to find the most typical divergences, among the most negative and most positive.

Also reports parametric, nonparametric $R^2$ values to see whether the parametric model is "leaving money on the table."

Plots suggest quadratic, interaction terms to add.

Location of these slides: http://heather.cs.ucdavis.edu/freqparcoord/Slides.pdf
Summary

- The `freqparcoord` package plots only “typical” lines, thus avoiding clutter.
Summary

- The `freqparcoord` package plots only “typical” lines, thus avoiding clutter. Can be used for group comparison, outlier hunting, clusters hunting.
Summary

- The `freqparcoord` package plots only “typical” lines, thus avoiding 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 avoiding 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.

Summary

- The `freqparcoord` package plots only “typical” lines, thus avoiding 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 avoiding 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.”
Summary

• The `freqparcoord` package plots only “typical” lines, thus avoiding 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](http://heather.cs.ucdavis.edu/freqparcoord/Slides.pdf)
Summary

- The `freqparcoord` package plots only “typical” lines, thus avoiding 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: