CS 147:
Computer Systems Performance Analysis

Advanced Regression Techniques

CS 147:
Computer Systems Performance Analysis
Advanced Regression Techniques

Overview

Overview
Curvilinear Regression
Common Transformations
General Transformations
Handling Outliers
Common Mistakes

Curvilinear Regression

- Linear regression assumes a linear relationship between predictor and response
- What if it isn't linear?
- You need to fit some other type of function to the relationship

Curvilinear Regression
- Linear regression assumes a linear relationship between predictor and response
- What if it isn't linear?
- You need to fit some other type of function to the relationship

When To Use Curvilinear Regression

- Easiest to tell by sight
- Make a scatter plot
  - If plot looks non-linear, try curvilinear regression
- Or if non-linear relationship is suspected for other reasons
- Relationship should be convertible to a linear form

When To Use Curvilinear Regression

- Easiest to tell by sight
- Make a scatter plot
  - If plot looks non-linear, try curvilinear regression
- Or if non-linear relationship is suspected for other reasons
- Relationship should be convertible to a linear form

Types of Curvilinear Regression

- Many possible types, based on a variety of relationships:
  - $y = ax^b$
  - $y = a + b/x$
  - $y = ab^x$
  - Etc., ad infinitum

Types of Curvilinear Regression

- Many possible types, based on a variety of relationships:
  - $y = ax^b$
  - $y = a + b/x$
  - $y = ab^x$
  - Etc., ad infinitum

Transform Them to Linear Forms

- Apply logarithms, multiplication, division, whatever to produce something in linear form
- I.e., $y = a + b \times$ something
  - Or a similar form
- If predictor appears in more than one transformed predictor variable, correlation is likely!

Transform Them to Linear Forms

- Apply logarithms, multiplication, division, whatever to produce something in linear form
- I.e., $y = a + b \times$ something
  - Or a similar form
- If predictor appears in more than one transformed predictor variable, correlation is likely!

Sample Transformations

- For $y = ae^{bx}$ take logarithm of y, do regression on $\log y = b_0 + b_1 x$, let $b = b_1$, $a = e^{b_0}$
- For $y = a + b \log x$, take log of x before fitting parameters, let $b = b_1$, $a = b_0$
- For $y = ax^b$, take log of both x and y, let $b = b_1$, $a = e^{b_0}$

Sample Transformations

- For $y = ae^{bx}$ take logarithm of y, do regression on $\log y = b_0 + b_1 x$, let $b = b_1$, $a = e^{b_0}$
- For $y = a + b \log x$, take log of x before fitting parameters, let $b = b_1$, $a = b_0$
- For $y = ax^b$, take log of both x and y, let $b = b_1$, $a = e^{b_0}$

Corrections to Jain p. 257

(Early Editions)

Nonlinear	Linear
$y = a + b/x$	y = a+b(1/x)
$y = 1/(a + bx)$	$(1/y) = a + bx$
$y = x(a + bx)$	$(x/y) = a + bx$
$y = ab^x$	$\ln y = \ln a + x \ln b$
$y = a + bx^n$	$y = a + b(x^n)$

Corrections to Jain p. 257
(Early Editions)

Nonlinear	Linear
$y = a + b/x$	y = a+b(1/x)
$y = 1/(a + bx)$	$(1/y) = a + bx$
$y = x(a + bx)$	$(x/y) = a + bx$
$y = ab^x$	$\ln y = \ln a + x \ln b$
$y = a + bx^n$	$y = a + b(x^n)$

General Transformations

- Use some function of response variable y in place of y itself
- Curvilinear regression is one example
- But techniques are more generally applicable

General Transformations

- Use some function of response variable y in place of y itself
- Curvilinear regression is one example
- But techniques are more generally applicable

When To Transform?

- If known properties of measured system suggest it
- If data's range covers several orders of magnitude
- If homogeneous variance assumption of residuals (homoscedasticity) is violated

When To Transform?

- If known properties of measured system suggest it
- If data's range covers several orders of magnitude
- If homogeneous variance assumption of residuals (homoscedasticity) is violated

Transforming Due To (Lack of) Homoscedasticity

- If spread of scatter plot of residual vs. predicted response isn't homogeneous,
- Then residuals are still functions of the predictor variables
- Transformation of response may solve the problem

Transforming Due To (Lack of) Homoscedasticity

- If spread of scatter plot of residual vs. predicted response isn't homogeneous,
- Then residuals are still functions of the predictor variables
- Transformation of response may solve the problem

What Transformation To Use?

- Compute standard deviation of residuals
  - Plot as function of mean of observations
    - Assuming multiple experiments for single set of predictor values
  - Check for linearity: if linear, use a log transform
- If *variance* against mean of observations is linear, use square-root transform
- If standard deviation against mean *squared* is linear, use inverse ($1/y$) transform
- If standard deviation against mean *to a power* is linear, use power transform
- More covered in the book

What Transformation To Use?

- Compute standard deviation of residuals
  - Plot as function of mean of observations
    - Assuming multiple experiments for single set of predictor values
  - Check for linearity: if linear, use a log transform
- If *variance* against mean of observations is linear, use square-root transform
- If standard deviation against mean *squared* is linear, use inverse ($1/y$) transform
- If standard deviation against mean *to a power* is linear, use power transform
- More covered in the book

General Transformation Principle

For some observed relation between standard deviation and mean, $s = g(\overline{y})$:

let $h(y) = \int \frac{1}{g(y)}\, dy$

transform to $w = h(y)$ and regress on w

General Transformation Principle

For some observed relation between standard deviation and mean, $s = g(\overline{y})$:

let $h(y) = \int \frac{1}{g(y)}\, dy$

transform to $w = h(y)$ and regress on w

Example: Log Transformation

If standard deviation against mean is linear, then $g(y) = a\overline{y}$

So $h(y) = \int \frac{1}{ay}\, dy = \frac{1}{a} \ln y$

Example: Log Transformation

If standard deviation against mean is linear, then $g(y) = a\overline{y}$

So $h(y) = \int \frac{1}{ay}\, dy = \frac{1}{a} \ln y$

Confidence Intervals for Nonlinear Regressions

- For nonlinear fits using general (e.g., exponential) transformations:
  - Confidence intervals apply to transformed parameters
  - **Not** valid to perform inverse transformation before calculating intervals
  - Must express confidence intervals in transformed domain

Confidence Intervals for Nonlinear Regressions

- For nonlinear fits using general (e.g., exponential) transformations:
  - Confidence intervals apply to transformed parameters
  - **Not** valid to perform inverse transformation before calculating intervals
  - Must express confidence intervals in transformed domain

Outliers

- Atypical observations might be outliers
  - Measurements that are not truly characteristic
  - By chance, several standard deviations out
  - Or mistakes might have been made in measurement
- Which leads to a problem:
  Do you include outliers in analysis or not?

Outliers

- Atypical observations might be outliers
  - Measurements that are not truly characteristic
  - By chance, several standard deviations out
  - Or mistakes might have been made in measurement
- Which leads to a problem:
  Do you include outliers in analysis or not?

Deciding How To Handle Outliers

1. Find them (by looking at scatter plot)
2. Check carefully for experimental error
3. Repeat experiments at predictor values for each outlier
4. Decide whether to include or omit outliers
   - Or do analysis both ways

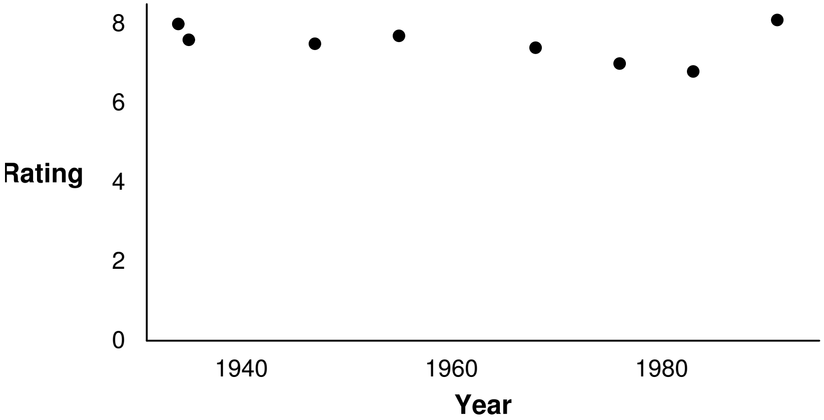
Question: Is last point in last lecture's example an outlier on rating vs. year plot?

Deciding How To Handle Outliers

1. Find them (by looking at scatter plot)
2. Check carefully for experimental error
3. Repeat experiments at predictor values for each outlier
4. Decide whether to include or omit outliers
   - Or do analysis both ways

Question: Is last point in last lecture's example an outlier on rating vs. year plot?

Rating vs. Year

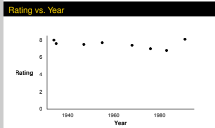


2015-06-15

CS147

Handling Outliers

Rating vs. Year



Common Mistakes in Regression

- Generally based on taking shortcuts
- Or not being careful
- Or not understanding some fundamental principle of statistics

Common Mistakes in Regression

- Generally based on taking shortcuts
- Or not being careful
- Or not understanding some fundamental principle of statistics

Not Verifying Linearity

- Draw the scatter plot
- If it’s not linear, check for curvilinear possibilities
- Misleading to use linear regression when relationship isn’t linear

Not Verifying Linearity
- Draw the scatter plot
- If it’s not linear, check for curvilinear possibilities
- Misleading to use linear regression when relationship isn’t linear

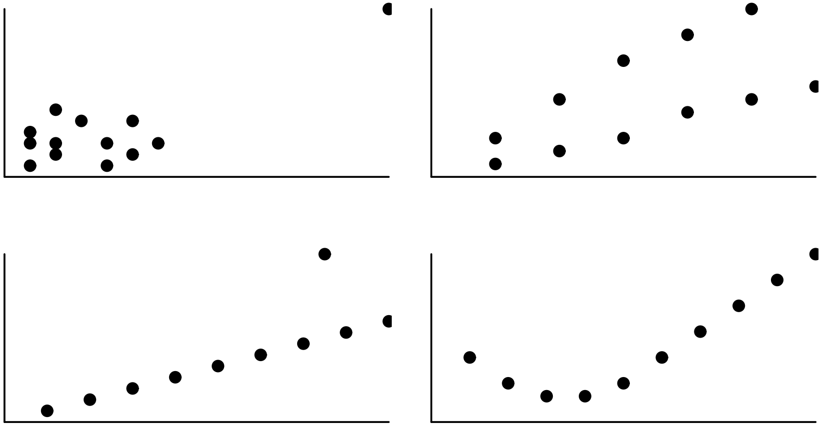
Relying on Results Without Visual Verification

- Always check scatter plot as part of regression
  - Examine predicted line vs. actual points
- Particularly important if regression is done automatically

Relying on Results Without Visual Verification

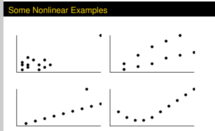
- Always check scatter plot as part of regression
  - Examine predicted line vs. actual points
- Particularly important if regression is done automatically

Some Nonlinear Examples



2015-06-15

CS147
└Common Mistakes
└Some Nonlinear Examples



Attaching Importance to Parameter Values

- Numerical values of regression parameters depend on scale of predictor variables
- So just because a particular parameter's value seems "small" or "large," not necessarily an indication of importance
- E.g., converting seconds to microseconds doesn't change anything fundamental
  - But magnitude of associated parameter changes

Attaching Importance to Parameter Values
- Numerical values of regression parameters depend on scale of predictor variables
- So just because a particular parameter's value seems "small" or "large," not necessarily an indication of importance
- E.g., converting seconds to microseconds doesn't change anything fundamental
  - But magnitude of associated parameter changes

Not Specifying Confidence Intervals

- Samples of observations are random
- Thus, regression yields parameters with random properties
- Without confidence interval, impossible to understand what a parameter really means

Not Specifying Confidence Intervals
- Samples of observations are random
- Thus, regression yields parameters with random properties
- Without confidence interval, impossible to understand what a parameter really means

Not Calculating Coefficient of Determination

- Without R^2, difficult to determine how much of variance is explained by the regression
- Even if R^2 looks good, safest to also perform an F-test
- Not that much extra effort

Not Calculating Coefficient of Determination

- Without R^2, difficult to determine how much of variance is explained by the regression
- Even if R^2 looks good, safest to also perform an F-test
- Not that much extra effort

Using Coefficient of Correlation Improperly

- Coefficient of determination is R^2
- Coefficient of correlation is R
- R^2 gives percentage of variance explained by regression, not R
- E.g., if R is .5, R^2 is .25
- And regression explains 25% of variance
- Not 50%!

Using Coefficient of Correlation Improperly

- Coefficient of determination is R^2
- Coefficient of correlation is R
- R^2 gives percentage of variance explained by regression, not R
- E.g., if R is .5, R^2 is .25
- And regression explains 25% of variance
- Not 50%!

Using Highly Correlated Predictor Variables

- If two predictor variables are highly correlated, using both degrades regression
- E.g., likely to be correlation between an executable's on-disk and in-core sizes
  - So don't use both as predictors of run time
- Means you need to understand your predictor variables as well as possible

Using Highly Correlated Predictor Variables

- If two predictor variables are highly correlated, using both degrades regression
- E.g., likely to be correlation between an executable's on-disk and in-core sizes
  - So don't use both as predictors of run time
- Means you need to understand your predictor variables as well as possible

Using Regression Beyond Range of Observations

- Regression is based on observed behavior in a particular sample
- Most likely to predict accurately within range of that sample
  - Far outside the range, who knows?
- E.g., regression on run time of executables smaller than size of main memory may not predict performance of executables that need VM activity

Using Regression Beyond Range of Observations

- Regression is based on observed behavior in a particular sample
- Most likely to predict accurately within range of that sample
  - Far outside the range, who knows?
- E.g., regression on run time of executables smaller than size of main memory may not predict performance of executables that need VM activity

Measuring Too Little of the Range

- Converse of prevoius mistake
- Regression only predicts well near range of observations
- If you don't measure commonly used range, regression won't predict much
- E.g., if many programs are bigger than main memory, only measuring those that are smaller is a mistake

Measuring Too Little of the Range

- Converse of prevoius mistake
- Regression only predicts well near range of observations
- If you don't measure commonly used range, regression won't predict much
- E.g., if many programs are bigger than main memory, only measuring those that are smaller is a mistake

Using Too Many Predictor Variables

- Adding more predictors does not necessarily improve model!
- More likely to run into multicollinearity problems
- So what variables to choose?
  - It's an art
  - Subject of much of this course

Using Too Many Predictor Variables

- Adding more predictors does not necessarily improve model!
- More likely to run into multicollinearity problems
- So what variables to choose?
  - It's an art
  - Subject of much of this course

Assuming a Good Predictor Is a Good Controller

- Often, a goal of regression is finding control variables
- But correlation isn't necessarily control
- Just because variable *A* is related to variable *B*, you may not be able to control values of *B* by varying *A*
- E.g., if number of hits on a Web page is correlated to server bandwidth, you might not boost hits by increasing bandwidth

Assuming a Good Predictor Is a Good Controller

- Often, a goal of regression is finding control variables
- But correlation isn't necessarily control
- Just because variable *A* is related to variable *B*, you may not be able to control values of *B* by varying *A*
- E.g., if number of hits on a Web page is correlated to server bandwidth, you might not boost hits by increasing bandwidth