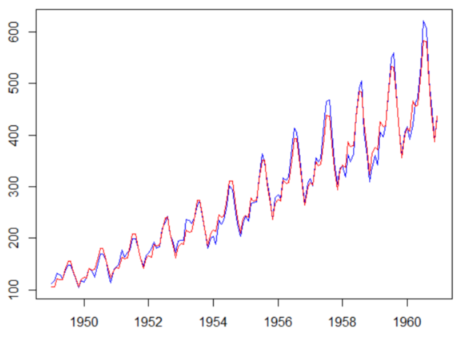
# S&DS 361/661: Data Analysis

## Iterative fitting


Dr. Brinda

Yale University


02/20/2019



What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

# Discussion

What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

- least-squares fitting

# Discussion

What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

- least-squares fitting
- model selection criterion

What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

- least-squares fitting
- model selection criterion
- regularization

# Discussion

What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

- least-squares fitting
- model selection criterion
- regularization
- cross-validation

What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

- least-squares fitting
- model selection criterion
- regularization
- cross-validation
- linear model fitting with hypothesis tests

What (if anything) keeps each of the following techniques from selecting an overly complicated regression function?

- least-squares fitting
- model selection criterion
- regularization
- cross-validation
- linear model fitting with hypothesis tests
- iterative fitting

# Iterative fitting steps

**1. Transformation**

**1. Transformation**

Draw a boxplot or histogram for each variable. Any variable that is highly skewed should be transformed toward a more symmetric distribution. Two common transformations for positive right-skewed data are:

**1. Transformation**

Draw a boxplot or histogram for each variable. Any variable that is highly skewed should be transformed toward a more symmetric distribution. Two common transformations for positive right-skewed data are:

- $log(\cdot)$ (or a shifted version such as $log(1 + \cdot)$ if there are values at or near zero)

**1. Transformation**

Draw a boxplot or histogram for each variable. Any variable that is highly skewed should be transformed toward a more symmetric distribution. Two common transformations for positive right-skewed data are:

- $log(\cdot)$ (or a shifted version such as $log(1 + \cdot)$ if there are values at or near zero)
- square-root

**2. Selection of the best variable**

**2. Selection of the best variable**

For each explanatory variable, draw a scatterplot with the response variable. Identify the most predictive explanatory variable (if any), and devise a function of that variable (e.g. least-squares line, least-squares quadratic, ...) to predict the response.

**2. Selection of the best variable**

For each explanatory variable, draw a scatterplot with the response variable. Identify the most predictive explanatory variable (if any), and devise a function of that variable (e.g. least-squares line, least-squares quadratic, ...) to predict the response.

- Select the "simplest" function whose residuals show no (predictive) pattern. If you'd like you can place your residuals in a null-lineup; if they stand out, that indicates that there's likely a real pattern remaining. (But with experience, you'll get a better feel for whether or not there's a meaningful pattern. Then you might only draw a null line-up in tough cases or in order to convince a reader of your analysis.)

**2. Selection of the best variable**

For each explanatory variable, draw a scatterplot with the response variable. Identify the most predictive explanatory variable (if any), and devise a function of that variable (e.g. least-squares line, least-squares quadratic, ...) to predict the response.

- Select the "simplest" function whose residuals show no (predictive) pattern. If you'd like you can place your residuals in a null-lineup; if they stand out, that indicates that there's likely a real pattern remaining. (But with experience, you'll get a better feel for whether or not there's a meaningful pattern. Then you might only draw a null line-up in tough cases or in order to convince a reader of your analysis.)
- You might even use "iterative fitting" to find a good function for a single variable by finding one relationship at a time (recall the AirPassengers data).
- If it's not clear which variable is most predictive, you might compare the R-squared values of their (not necessarily linear) fits.

# Iterative fitting steps

**3. Repeat**

# Iterative fitting steps

**3. Repeat**

Treat the residuals of your fit as your new response variable, and repeat step 2 with the remaining explanatory variables.

**3. Repeat**

Treat the residuals of your fit as your new response variable, and repeat step 2 with the remaining explanatory variables.

- If convenient, recalculate the parameters of your regression function using your original response variable with all the terms you've introduced so far. You might find that some terms no longer seem valuable when others have been included (e.g. check the "summary" object); these terms can be dropped out of the regression function.

**3. Repeat**

Treat the residuals of your fit as your new response variable, and repeat step 2 with the remaining explanatory variables.

- If convenient, recalculate the parameters of your regression function using your original response variable with all the terms you've introduced so far. You might find that some terms no longer seem valuable when others have been included (e.g. check the "summary" object); these terms can be dropped out of the regression function.
- If recalculating the parameters is inconvenient, you can instead use the sum of the individual regression functions as your overall regression function.

**3. Repeat**

Treat the residuals of your fit as your new response variable, and repeat step 2 with the remaining explanatory variables.

- If convenient, recalculate the parameters of your regression function using your original response variable with all the terms you've introduced so far. You might find that some terms no longer seem valuable when others have been included (e.g. check the "summary" object); these terms can be dropped out of the regression function.
- If recalculating the parameters is inconvenient, you can instead use the sum of the individual regression functions as your overall regression function.
- Repeat this process again with the new residuals and the remaining explanatory variables, and stop when none of the remaining explanatory variables have a (predictive) relationship with the residuals.

**4. Check for interactions**

**4. Check for interactions**

Treat the product of each pair of terms in your regression function as a new explanatory variable, and repeat the process (steps 2 and 3) with these new variables.

**4. Check for interactions**

Treat the product of each pair of terms in your regression function as a new explanatory variable, and repeat the process (steps 2 and 3) with these new variables.

- If you already have a lot of terms, you might have a lot of potential interactions to check. (With $k$ terms, there are $k$ choose 2 interaction terms.) In that case, you might not want to bother with this step.