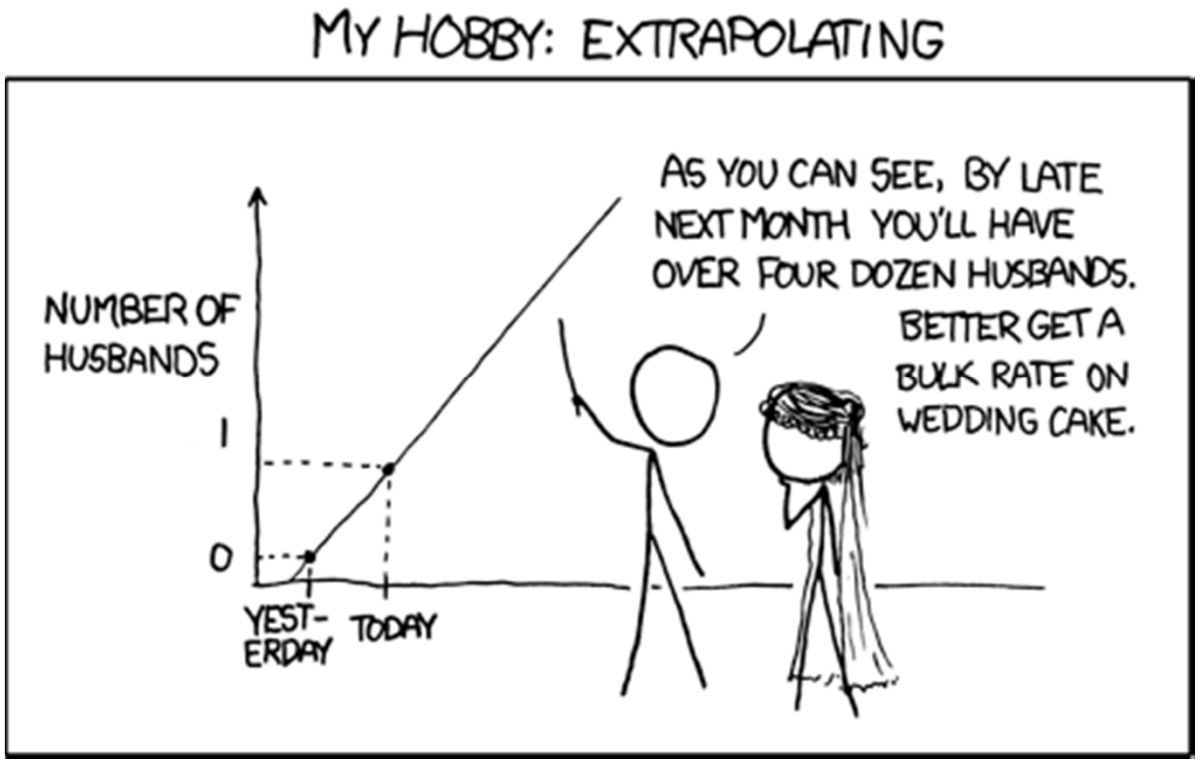
UNIVERSITY of WASHINGTON

Data Science UW
Methods for Data Analysis

Regression Part 2 – Uncertainty and Feature Selection
Lecture 6
Steve Elston


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

Topics

- Review
- Bootstrapping regression models
- Linear Algebra overview
- Decomposition Methods
- SVD regression
- Feature selection, stepwise regression and ANOVA

Review

- Central Limit Theorem
- Linear Regression
- Example of presenting data science results
- Bootstrap hypothesis testing

How good is a regression model

Uncertainty can be significant

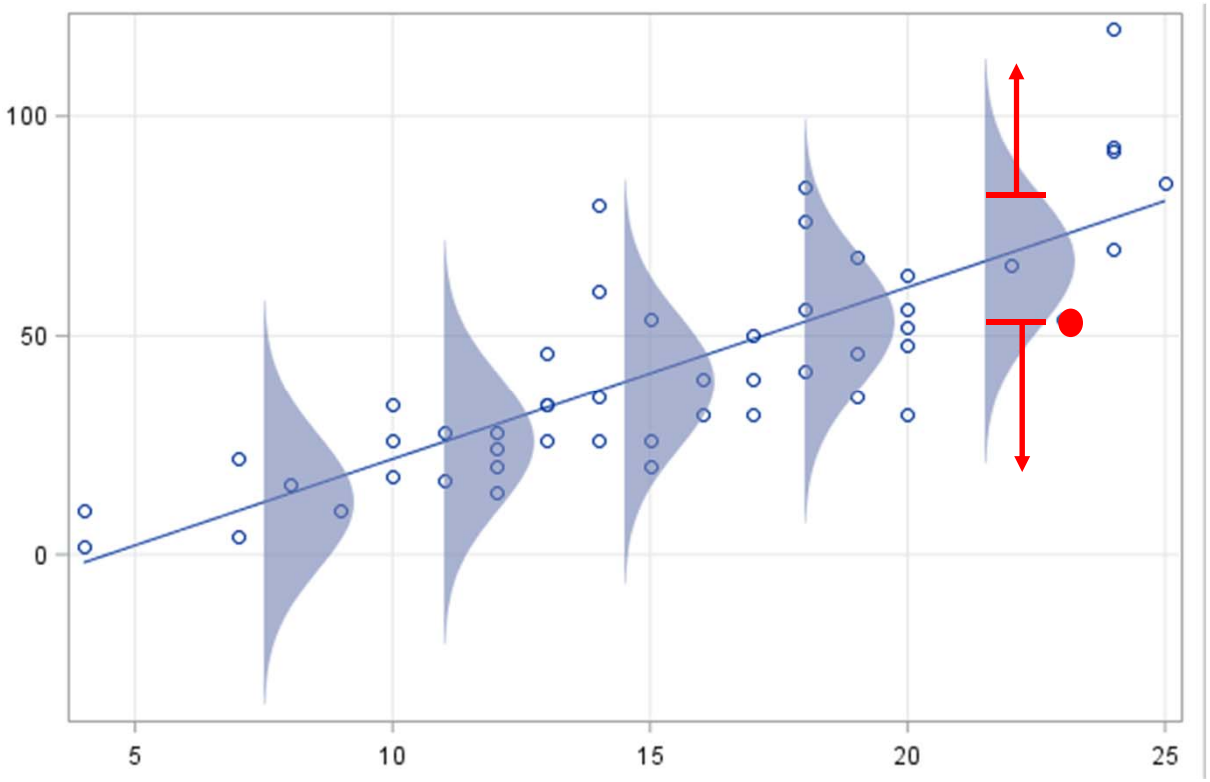
> Important to understand how good your model is
> The predictions are never exact

Uncertainty in Regression

- Given a linear model with known constants:

$$y_i = mx_i + b + N(0, \sigma)$$

- Given a point that comes from that line, we can come up with a probability of observing that point.



Regression Coefficients

Regression coefficients are not exact!

- Uncertainty in regression coefficients leads to uncertainty in predictions
- Coefficients may or may not be significant; e.g. may or may not be indistinguishable from zero
- Examine model uncertainty with bootstrap resampling
- Test Null Hypothesis that coefficient is zero with bootstrap resampling
- R demo

Review of Linear Algebra

Linear Algebra is the Core of Linear Models

- Understanding the linear algebra brings deeper understanding of how models work
- Linear algebra illuminates the ways models can fail

Linear Algebra

- Matrix: a rectangular array of values, with dimensions n by m (n rows, m columns).
- Vector: a one dimensional array of values (n or m = 1).
- Square matrix: a n x n matrix.
- Identity matrix: a square matrix with 1’s on the diagonal and 0’s elsewhere.

Linear Algebra

- Algebraic Properties of Matrices:
  - Add/subtract matrices: Must be of the same dimensions
  - Multiplication of matrices:
    - Inner dimensions must match.

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \times \begin{bmatrix} j & k & l \\ m & n & o \\ p & q & r \end{bmatrix} =$$

$$\begin{bmatrix} aj + bm + cp & ak + bn + cq & al + bo + cr \\ dj + em + fp & dk + en + fq & dl + eo + fr \\ gj + hm + ip & gk + hn + iq & gl + ho + ir \end{bmatrix}$$

  - [n x m] * [m x p] = [n x p]
  - Note that matrix multiplication is not commutative

$$A \times B \neq B \times A$$

Linear Algebra

- Identity matrix: just like 1 is the multiplicative identity.
  - 5*1=5

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \cdot \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

Identity matrix: a square matrix of zeros with 1's on the diagonal. Also written as I_{nxn}

Linear Algebra

- Transpose (given an element in position i,j, the transpose has the same element in position j,i.)
- Inverse:
  - Just like the multiplicative inverse of n is 1/n, matrices also have multiplicative inverses:

$$A_{nxn} \cdot A^{-1}_{nxn} = I_{nxn}$$
$$A^{-1}_{nxn} \cdot A_{nxn} = I_{nxn}$$

$$\begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \cdot \begin{bmatrix} -2 & \frac{3}{2} \\ 1 & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Linear Algebra

- For a 2x2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$= \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ c & d \end{bmatrix}$$

- What if (ad-bc) = 0? Then ad = bc or a/c = b/d.
- If a/c = b/d, then one of the columns is a multiple of the other!
- **A is rank deficient**
- These columns are dependent on each other.
  - If these were columns in our numerical data frame, then one column would be a multiple of the other.
  - Example: Meters and Kilometers as separate predictors.

Linear Algebra

$$A \times X$$

- Given a sequence of data points in a matrix, X, which has dimensions 2xn:

$$X = \begin{bmatrix} 1 & 2 & 0 \\ -2 & 3 & 1 \end{bmatrix}$$

- If we multiply by another matrix A (2x2):

$$A_{2x2} \times \begin{bmatrix} 1 & 2 & 0 \\ -2 & 3 & 1 \end{bmatrix}$$

- Then we have a matrix A that transforms the points in X.

Linear Algebra and Multiple Linear Regression

- Solve linear regression problem with normal equations

$$Ab = x$$

$$b = A^{-1}x$$

where

$$A^{-1}A = I$$

  - In typical case A is long and narrow
  - Solving for A^{-1} is computationally difficult and inefficient
  - May not be unstable if A is rank deficient

- Use an alternative formulation

$$b = (A^T A)^{-1} A^T x$$

  - A^TA is dimension m x m with A of dimension n x m, n >> m
  - **A^TA can still be rank deficient!**

Rank Deficient Problems

What can we do?

- Find full rank approximation of the problem
- Commonly shows up in data science
  - Models with many features
  - Especially, unstructured data

Matrix Decomposition

- Eigenvalues: Given a nxn matrix, A, λ is an eigenvalue if there exists a vector X such that:

$$AX = \lambda X$$

- Finding the eigenvectors of A involves lots of computation.
- If A rotates and shifts a vector X, then we can think of eigenvalues as a geometric hinge on which the 'A' operation acts.
- Eigenvalues have corresponding eigenvectors.
- Rank deficient matrix has eigenvalues that are zero or nearly so
- This may seem insignificant at the moment, but eigenvalues and eigenvectors play an important role in manipulating our data.

Matrix Decomposition

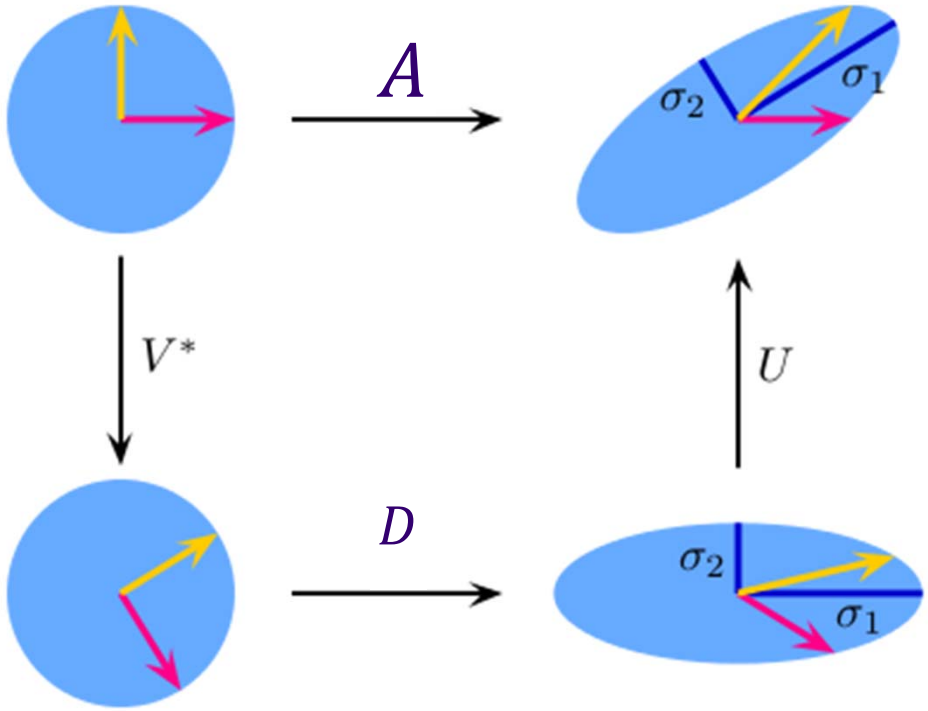
- Matrix Decompositions allow us to write a matrix, A, in many different forms.
- The one that is the most used, is Singular Value Decomposition (SVD).

$$A = UDV^T$$

  - U and V are comprised of orthogonal unit norm **singular vectors**
  - D is a diagonal matrix of **singular value**s
  - A is comprised of the linear combination of singular vectors
  - The singular values define the weights for linear combination
  - Singular values define a **spectrum** of singular vectors

Linear Algebra

> The SVD is a way to express a transformation from one n dimensional space (the space A lies in) to another n dimensional space by rotating and scaling A

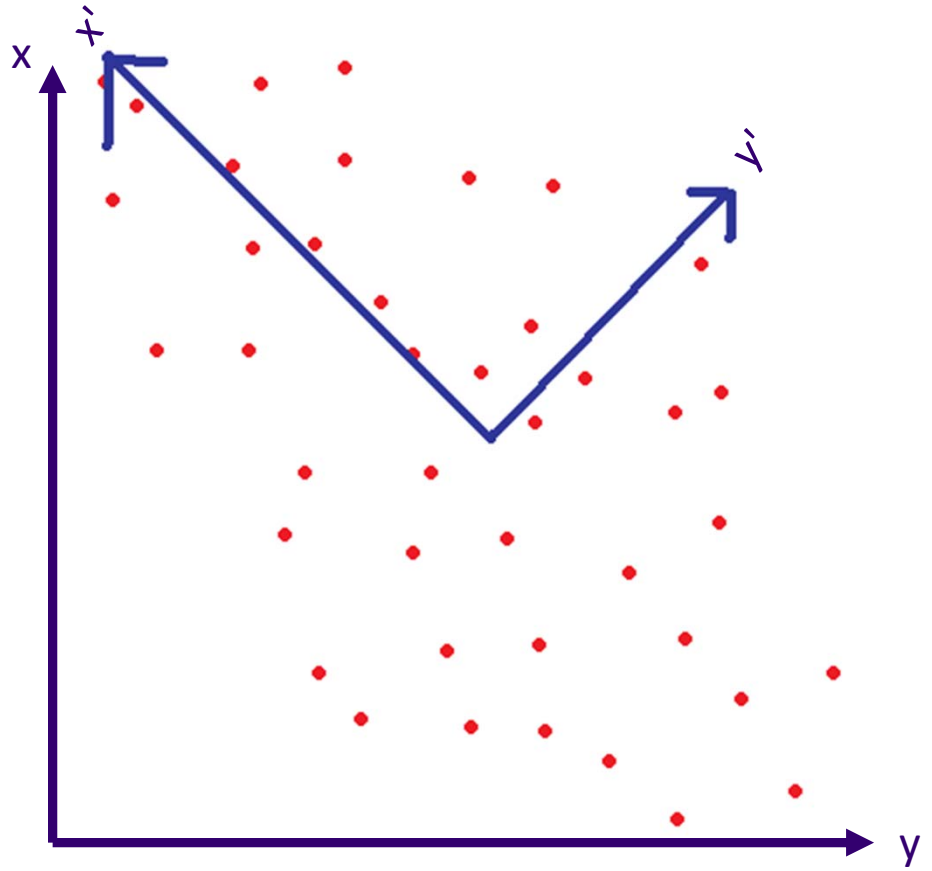


$$A = UDV^T$$

Principle Component Decomposition

PC decomposition same as SVD

- If two variables are correlated, we can transform the data to directions in which they are not correlated.
- These new axes are called the Principal Components.



PCA and SVD

- Know that instead of our original system:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

- We now have the system:

$$y_i = \beta_0 + \beta_1 f_1(x_1, x_2, \ldots) + \beta_2 f_2(x_1, x_2, \ldots) + \cdots$$

- The *f* functions are called our principle components.
- Functions *f* are called the rotations
- The *f* function outputs are guaranteed to be independent of each other.
- We can no longer interpret our linear model coefficients!

Deriving Independent Features from Dependence

Noting that largest singular values are for directions with largest variance

- With larger data sets, we've seen that no matter the quality, we can find a explanatory feature.
- If we consider our data as a matrix, we know that having dependent columns is a problem.
- Transform our data so the columns are the most important and independent:
  - Remove influence of columns that do not contain enough 'information'.
    - Too much missing data.
    - Low Variance.
  - Remove influence of columns that are correlated

Feature Selection with SVD

Smallest singular values correspond to least important directions

- SVD returns the same number of components as number of features.
- Components are orthogonal: the first few components explain much more variance than the last few.
- How do we decide how many to keep?
- We look at the magnitude of the associated singular values for each principal component.
- R-demo.

SVD Feature Selection

Select n orthogonal features with n nonzero singular values

> Example: Select single most important feature

$$U \begin{bmatrix} d_1 & 0 & 0 \\ 0 & 0 & 0 \\ \dots & \dots & \dots \end{bmatrix} V^T$$

> Example Select two most important features

$$U \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ \dots & \dots & \dots \end{bmatrix} V^T$$

SVD

- This seems like an awful lot of work for little improvement and loss of interpretability.
- But note that we lost the dependence in the data set!
- There are other applications as well…

How Do We Find Full Rank Inverse

SVD is ideal for extracting full rank inverse

> Full rank inverse is an approximation
> For SVD we call this the **pseudo inverse**

SVD and the Pseudo Inverse

- Since

$$A = UDV^T$$

- The pseudo inverse of A is:

$$VD^{+}U^T$$

- Where D^+ is the transpose of:

$$\begin{bmatrix} 1/d_1 & 0 & 0 \\ 0 & 1/d_2 & 0 \\ \dots & \dots & \dots \end{bmatrix}$$

- Note that the **smallest singular values have the largest weights** in the pseudo inverse

Pseudo Inverse and Regularization

Regularization is the process of transforming a rank deficient matrix into a full rank approximation

- Regularize the pseudo inverse by selecting the n largest singular values
- Regularization of pseudo inverse is the same as feature selection from the orthogonal projection

Another SVD regularization

- Since

$$A = UDV^T$$

- The pseudo inverse of A is:

$$VD^{+}U^T$$

- Where D^+ inverse transpose of:

$$\begin{bmatrix} d_1 + e & 0 & 0 \\ 0 & d_2 + e & 0 \\ \dots & \dots & \dots \end{bmatrix}$$

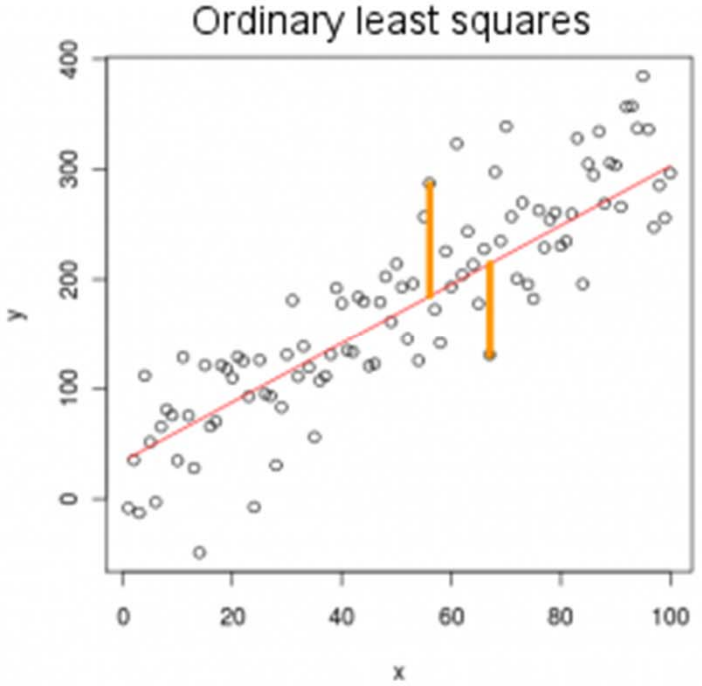
Where e is a small regularization coefficient of bias

- Is the basis of ridge regression

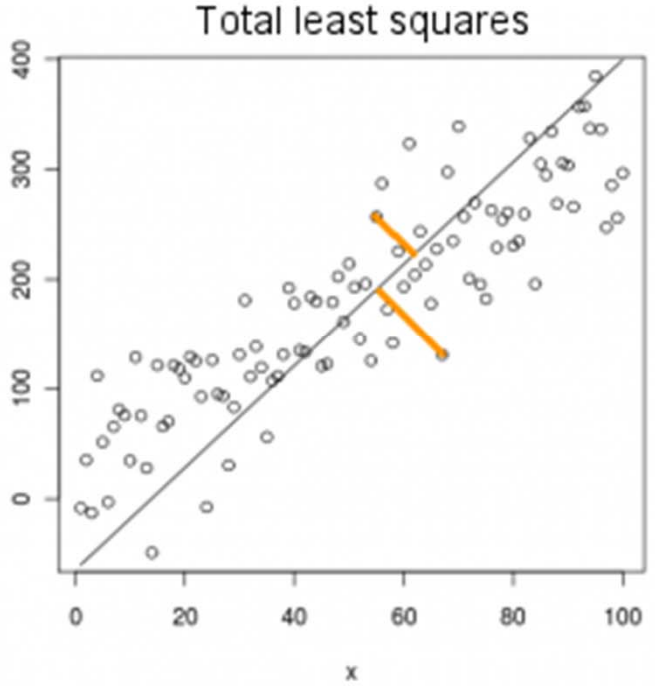
SVD, as a type of regression

> Also, looking at the first principal component, we can consider SVD as a new type of regression, which is called total least squares. (Also called Deming regression or PCA Regression)

Regressing y on x



SVD Primary Principal Component



> R demo

SVD, as a type of regression

- When to use total least squares:
  - If we want to control for error in x as well as y.
  - We are minimizing the distance from the point to the line as opposed to the distance between the y-values.
- R-squared doesn't really apply here, at least in the way we have defined it
- R Demo

SVD, as a way to compress information

- We can group together similar points via SVD
- Store only the first n principle components of the data.
- Widely used for high-dimensional unstructured data
  - Text data
  - Image data

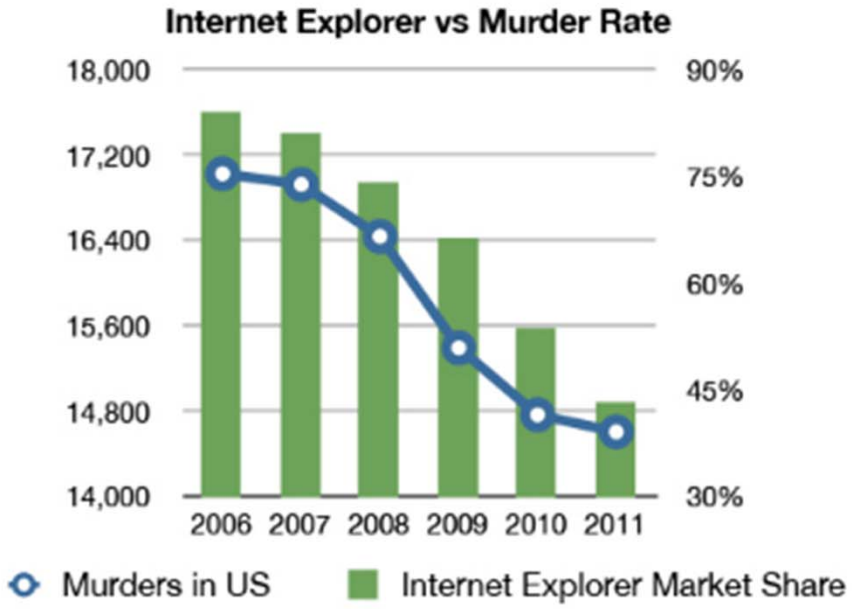
Another Approach to Feature Selection

- SVD has some problems
  - Is approximation to a different problem
  - Loose interpretability
  - Computationally intensive, but can use projection methods
- Are there other ways to measure feature importance?

Multiple Linear Regression

- Throwing in all possible variables to help explain our response is sometimes ***not* a good thing**
  - Variables can be dependent on each other.
  - Variables might not be important to explain the response.
  - Note that the SSE is always larger for reduced models!

Multiple Linear Regression



How do we choose which combinations of independent variables to use?

We might consider looking at the difference in SSE between models and the number of explanatory variables.

Categorical Variables

- Start with a first order model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

- How do we deal with factor/categorical variables?

Gender
F
M
F
F
M
…

Gender
1
0
1
1
0
…

Eye Color
Brown
Brown
Blue
Green
Green
Blue
Brown
…

Brown	Blue
1	0
1	0
0	1
0	0
0	0
0	1
1	0
…	…

"One hot encoding"

DayOfWeek
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
…

DayOfWeek
1
2
3
4
5
6
7
…

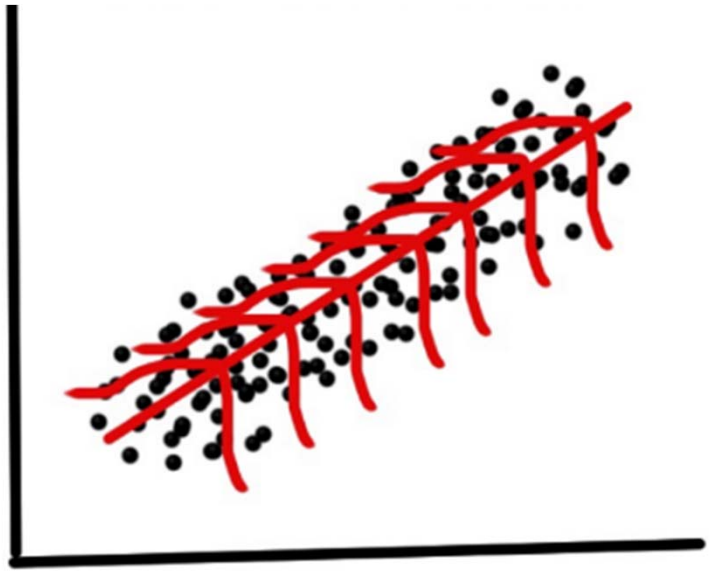
"Factor encoding"

Categorical Variables

- Encoding increases the dimensionality of the model matrix
- Encoding can lead to rank deficiency
  - Categorical variables with many levels
  - Many categorical variables
  - Categorical variables with infrequent cases have near zero variance

Multiple Linear Regression

- Linear Model Likelihood
  - We assume errors are normally distributed. From the resulting set of errors, we can come up with a distribution. We then use each residual point and come up with a total error and calculate the probability of that model given our data. This is the likelihood.



To make the calculations easier, we usually take the logarithm of the model (remember we can do this because it is monotonic). This is called the 'log-likelihood'. We will talk more about this when we get to Bayesian Statistics.

Multiple Linear Regression

- Akaike Information Criterion (AIC)
  - Given a model with k-parameters, and a likelihood of L

$$AIC = 2k - 2\ln(L)$$

  - Note that the more parameters, the higher the AIC.
  - The higher the likelihood, the lower the AIC.
  - Better models have lower AIC values.

Multiple Linear Regression

- How to select the variables in the model?
- Stepwise regression.
  - Forward Selection:
    - Start with no independent variables and add the variables one by one, selecting the variable that improves your criterion the most.
  - Backward Selection
    - Start with all independent variables and remove on at a time. Remove the one that improves the chosen criterion.
- R-demo

Assignment

- Complete Homework 6:
  - Perform SVD regression on the auto price data
    - Use the following features for your initial model: **make, fuel.type, aspiration, body.style, drive.wheels, length, curb.weight, engine.type, num.of.cylinders, engine.size, city.mpg**
    - Apply SVD to a model matrix created with `model.matrix()`, and report the increase in dimensionality
    - Report how many orthogonal features you used for you model
    - Evaluate your model performance with plots and by computing RMS error. Hint see my demo code for plots.
  - Use stepwise regression to select features from the aforementioned set
    - Compare model performance with full model using summary statistics, plots and ANOVA

Assignment Continued

- You should submit:
  - One R-script
    - Follow good coding practice; minimize cut and paste
    - Include comments
  - One document outlining and supporting your conclusions
    - See example I provided last week

> Read Introduction to Data Science, Chapter 16.