Other Regression Models

Dr. John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@cs.rice.edu

Goals for Today

Understand

- **Limitations of simple linear regression**
- **How to apply regression analysis more broadly**
  - —multiple linear regression
  - —curvilinear regression
  - —transformations
  - —categorical predictors
- **Common mistakes in regression analysis & how to avoid them**

Limitations of Simple Linear Regression

Three key restrictions

- **Only one predictor variable is allowed**
- **The predictor variable must be quantitative**
- **The response must be a linear function of the predictor**

Towards Broader Applicability

Relaxing some of the restrictions

- **Allow more than one predictor variable**
  - —multiple linear regression
- **Allow categorical variables, e.g. CPU type**
  - —categorical predictors
- **Allow non-linear relationship between response and predictors**
  - —curvilinear regression
- **Use transformations to cope with**
  - —errors that are not normally distributed
  - —variance that is not homogeneous

Multiple Linear Regression (MLR)

Predict a response variable from k predictor variables

$$y = b_0 + \left(\sum_{i=1}^{k} b_i x_i \right) + e$$

—b_0, b_1, … b_k **are fixed parameters**

—**e is the error term**

- **In vector notation**

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \quad \text{or} \quad \mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Estimating Model Parameters for MLR

- Given n sets of observations of a y response variable given a vector of x predictor variables
- Solve for the b model parameters in $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$
- Solve using matrix notation

$$\mathbf{y} = \mathbf{X}\mathbf{b}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{b}$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{I}\mathbf{b}$$

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{b}$$

matrix solution technique works great for simple linear regression as well!

Allocating Variation for MLR

- **Quantifying variation**

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad \text{(total sum of squares)}$$

- **Key questions**
  - **—how much variation is unexplained?**

$$SSE = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \qquad \text{(sum of squares error)}$$

  - **—how much variation is accounted for by the regression?**

$$SSR = SST - SSE \qquad \text{(sum of squares regression)}$$

same as simple linear regression

Coefficient of Determination for MLR

Measuring the quality of a regression model

coefficient of determination
$$= R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

same as linear regression through here

- **Coefficient of multiple correlation = R**

Standard Deviation of Errors for MLR

- **Variance of errors = SSE/(degrees of freedom) (aka MSE)**

$$s_e^2 = \frac{SSE}{n-k-1}$$

- **Why n-k-1 degrees of freedom for SSE?**
  - **—SSE computed after calculating k+1 regression parameters**
- **Degrees of freedom and multiple linear regression**

SST	**=**	**SSR**	**+**	**SSE**
(n-1)	**=**	**k**	**+**	**(n-k-1)**

- **Standard deviation of errors**

$$s_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-k-1}}$$

Analysis of Variance (ANOVA)

- MSR = SSR/k
- MSE = SSE/(n-k-1)
- Assuming that
  - —errors are independent, normally distributed with same μ, σ^2
  - —then y's are normally distributed since x's are non-stochastic
- The sum of squares of a normal variate has a χ^2distribution
  - —see section 29.4 in Jain
- $F_{[1-\alpha;k,n-k-1]}$ distribution: models ratio of sample variances (each χ^2)
  - —k = DOF numerator (MSR); n-k-1 = DOF denominator (MSE)
  - —α is the significance level

described in Jain pp. 489-490,634

- Regression is significant if MSR/MSE > $F_{[1-\alpha;k,n-k-1]}$
  - —if MSR/MSE is greater than the F-ratio, the predictor variables are assumed to explain a significant fraction of the response variation
  - —F-test is equivalent to testing that y does not depend on any x_j
    - if computed ratio is < F value, hypothesis that $b_1=b_2=\ldots=b_k=0$ cannot be rejected

Comparing to MLR ANOVA to Simple Regression

- **For simple linear regression, F-test reduces to testing if $b_1 = 0$**
- **F-test not necessary for simple linear regression**
  - —If the confidence interval of b_1 does not include 0
  - —then the regression explains a significant portion of the response variation

Curvilinear Regression

- **Linear regression model can only be used if response variable is linear function predictor variables**

  —that's why we first check a scatter diagram of y vs x!

- **If the relationship between y and x appears (or is known) to be non-linear, must use a non-linear regression model**
- **If non-linear form can be converted into a linear one, we can use simple or multiple linear regression techniques**

Non-linear forms	Linear forms
$y = a + b/x$	$y = a + b(1/x)$
$y = 1/(a + bx)$	$1/y = a + bx$
$y = x/(a + bx)$	$x/y = a + bx$
$y = ab^x$	$\ln y = \ln a + (\ln b)x$
$y = bx^a$	$y = \ln b + a \ln x$
$y = a + bx^n$	$y = a + b(x^n)$

Curvalinear Regression Issues

- **If predictor variable appears in more than one transformed predictor variable**
  - —transformed variables are likley to be correlated
  - —causes problem of multi-colinearity
- **Strategy for minimizing multi-colinearity**
  - —avoid using unnecessary predictor variables
  - —use smallest subset that gives significant parameters and explains a high percent of observed variation

Transformations

- **The term transformation is used when some function of the measured response variable y is used in place of y in a model**
- **For example**

$$\sqrt{y} = b_0 + \left(\sum_{i=1}^{k} b_i x_i \right) + e$$

- **Three cases where transformations should be investigated**
  - **—if known from physical considerations that f(y) would yield better model than y**
    - **e.g. if requests per unit time (1/y) has linear relationship with predictor, then use 1/y rather than y**
  - **—if data covers several orders of magnitude and sample size is small**
  - **—if homogeneous variance (homoscedasticity) assumption of residuals is violated**
    - **if this is true, residuals are still functions of predictor variables**

Regression with Categorical Predictors

- **Categorical variable = non-numerical variable**
  - **e.g. CPU type**
- **Regressions can still be used if one or more of the predictor variables is categorical**
- **Coding binary categorical variables**
  - **e.g. CPU A vs. CPU B**

$$x_j = \begin{cases} -1 & \Rightarrow \text{first value} \\ +1 & \Rightarrow \text{second value} \end{cases}$$

  - with this coding, value of parameter b_j represents the average difference in response for each level
  - difference of effect of the 2 levels is $2b_j$
- **What about multi-valued categorical variables?**

Multi-valued Categorical Predictors - I

- **Coding multi-valued categorical variables**
  - —**e.g. CPU A vs. CPU B vs. CPU c**
    - **one approach**

$$\cancel{x_j = \begin{cases} 1 \texttt{ => type A} \\ 2 \texttt{ => type B} \\ 3 \texttt{ => type C} \end{cases}}$$

problem: this coding implies ordering among variables

- **a better approach**

$$x_1 = \begin{cases} 1 \texttt{ => type A} \\ 0 \texttt{ => otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \texttt{ => type B} \\ 0 \texttt{ => otherwise} \end{cases}$$

$(x_1,x_2) = (1,0)$ => type A
$(x_1,x_2) = (0,1)$ => type B
$(x_1,x_2) = (0,0)$ => type C

this coding implies no ordering

Multi-valued Categorical Predictors - II

$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

$(x_1,x_2) = (1,0) \Rightarrow$ type A
$(x_1,x_2) = (0,1) \Rightarrow$ type B
$(x_1,x_2) = (0,0) \Rightarrow$ type C

$$\bar{y}_A = b_0 + b_1$$

$$\bar{y}_B = b_0 + b_2$$

$$\bar{y}_C = b_0$$

- **parameter b_0 represents diff betw/ avg responses type A & C**
- **parameter b_1 represents diff betw/ avg responses type B & C**

Multi-valued Categorical Predictors - III

- **In general, with a categorical variable of k levels**
  - **need k - 1 binary variables defined as follows**

$$x_i = \begin{cases} 1 \Rightarrow \text{if } i^{th} \text{ value} \\ 0 \Rightarrow \text{otherwise} \end{cases}$$

  - **k^{th} value defined by $x_1 = x_2 = \ldots = x_{k-1} = 0$**
  - **regression parameter b_0 represents average response with k^{th} alternative**
  - **regression parameter b_i represents difference between the average response for the alternatives i and k**

Outliers

- Definition: atypical observations
- Dilemma
  - —including outliers in analysis may significantly change conclusions
  - —excluding outliers may lead to a misleading conclusion if outlier represents correct operation of the system
- Easiest way to identify outliers is to look at scatter plot
- Any value significantly different from others should be investigated for experimental error
- Once possibility of experimental error has been eliminated
  - —can decide whether to include outlier or not using intuition

Common Mistakes - I

- **Not verifying relationship is linear**
  - —check scatter plot
  - —if non-linear, consider curvalinear regression
- **Relying on automated results without visual verification**
  - —check your scatter plots, even if using an automated analysis package!
- **Attaching numerical importance to regression parameters**
  - —small regression parameters may be meaningful
  - —changing units (e.g. seconds to milliseconds) can change their magnitude dramatically
- **Not specifying confidence intervals for regression parameters**
  - —remember they are derived from a *sample*, not the *population*!
- **Not specifying the coefficient of determination**
  - —without R^2 it is hard to understand the quality of a regression

Common Mistakes - II

- Confusing the coefficient of determination with the coefficient of correlation
  - —COD = R^2 = % explained variance
  - —COC = R does not
- Using highly correlated predictor variables
  - —both should be included only if there is a considerable increase in significance of regression
- Using regression to predict far beyond the measured range
  - —measurements in one operating range may not apply elsewhere
- Too many predictor variables
  - —try subsets of predictor variables to look for most parsimonious and accurate prediction
- Measuring only a small subset of range of operation
  - —may miss non-linearity

Common Mistakes - III

- **Assuming that a good predictor is also a good control variable**
  - —regression model can predict performance
  - —if goal is to improve performance, regression only helpful if predictors are also control variables
  - —e.g. CPU time is a predictor but not a controller of number of disk I/Os