BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 2

LINEAR REGRESSION AND ANALYSIS OF VARIANCE

Linear Regression and Analysis of Variance

As background for our discussion of repeated measures and longitudinal analysis, we review the standard linear regression model for independent observations.

We discuss maximum likelihood (ML) and least squares estimation.

We also gently introduce some vector and matrix notation.

Finally, we consider the close connection between analysis of variance (ANOVA) and linear regression.

Consider a cross-sectional data set of 300 measurements of the logarithm of FEV_1, age, and logarithm of height of children living in Topeka, Kansas.

We will fit a model describing how the value of $\log(FEV_1)$ varies linearly with age and log(height).

The children varied in age from 6 to 18 years.

We will fit a multiple linear regression model, estimate the regression coefficients for age and log(height), and test the hypothesis that these coefficients are not significantly different from 0.

(Note: See the chapter on multiple regression in the folder **Supplementary Reading** on the course website)

Structure of Six Cities Data

Subject	Height	Age	Log(FEV1)
48	1.45	11.2991	0.62058
17	1.17	6.6639	0.28518
166	1.19	8.1396	0.14842
81	1.48	15.3347	0.57661
3	1.60	16.0164	1.08519
218	1.35	9.8015	0.50078
80	1.66	18.5270	0.91629
14	1.27	7.4251	0.37156

Multiple Linear Regression

Multiple linear regression describes how the expected value (mean) of a measured variable depends on a set of measured or categorical covariates, that is, characteristics of the individuals.

Suppose that we have observations on N individuals.

Each individual has a measured outcome,

$$Y_i; \quad i = 1, ..., N$$

Each observation, Y_i, has an associated set of covariates

$$X_{i1}, X_{i2}, ..., X_{ip}$$

The linear regression model for Y_i can be written as

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i$$

Typically, $X_{i1} = 1$ for all individuals, and then β_1 is the *intercept*.

This model says that the expected value of Y (the average for all individuals with the specified covariate values) varies linearly with the values of the covariates.

$$E(Y_i|X_{i1}, \ldots, X_{ip}) = \mu_{y_i|x_{i1},\ldots,x_{ip}} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Specifically, an increase of one unit in X_j (while holding the remaining covariates fixed/constant) produces an increase/decrease of β_j in the mean of Y.

Assumptions of Multiple Linear Regression

1. Individuals represent a random sample from the population of interest.

2. Independence: $Y_1, \ldots, Y_N$ are independent random variables.

3. Linearity: $E(Y|X_1, \ldots, X_p)$ is a linear function of each of the X's.

4. Normality: Given X's, individual observations of the dependent variable, Y_i, have a normal distribution, with means

$$\mu_{y|x_1,\ldots,x_p} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

5. Homoscedasticity: $\text{Var}(e_i)$ is constant, σ^2

   $\Rightarrow$ constant variation about regression line (or "plane")

Estimation

Basic Idea: Among all possible estimates $(\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ of $(\beta_1, \ldots, \beta_p)$ choose the estimates such that the fitted regression model "deviates" the least from the data.

$\Rightarrow$ Least Squares Estimation

"Deviation" of fitted model from the data is defined as:

$$\sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{N} \widehat{e}_i^2$$

where $\widehat{Y}_i = \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_p X_{ip}$.

Least Squares (LS) estimates are those values that minimize these deviations, i.e., minimize the residual sums of squares.

Maximum Likelihood Estimation

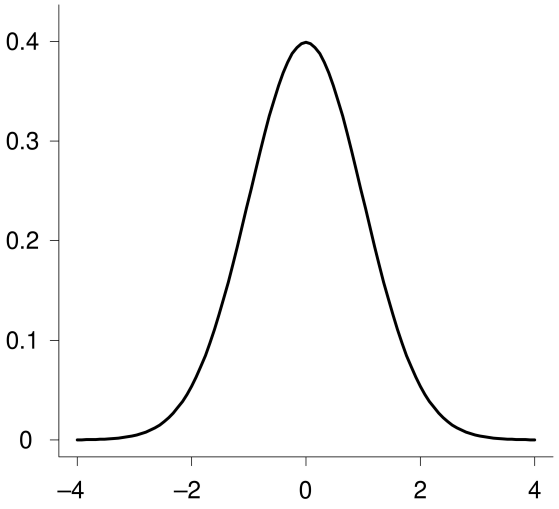
Recall: In the multiple regression model, the values of the covariates are assumed to be fixed.

Only the values of Y_i are random.

The probability distribution corresponding to the linear regression model is given by:

$$f(y_i|X_{i1},...,X_{ip}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

Equivalently, we assume that $e_i \sim N(0, \sigma^2)$



Notable Features of the Normal Distribution

$f(Y_i)$ is completely determined by (μ_i, σ^2)

$f(Y_i)$ depends to a very large extent on

$$(Y_i - \mu_i)^2/\sigma^2$$

The latter can be interpreted as a standardized distance of Y_i from μ_i, relative to σ^2, a measure of the spread of values around μ_i.

Maximum Likelihood Estimation

The main idea behind the method of maximum likelihood (ML) is really quite simple and conveyed by its name:

Use as estimates of $\beta_1, ..., \beta_p$ (and σ^2) the values that are most probable (or "likely") for the data that we have observed.

That is, choose values of $\beta_1, ..., \beta_p$ (and σ^2) that maximize the probability of the response variables evaluated at their observed values.

The resulting values are called the maximum likelihood estimates (MLEs) of $\beta_1, ..., \beta_p$ (and σ^2).

For a single observation, we can be at the "most likely" point on the probability curve

$$f(y_i|X_{i1},...,X_{ip}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

by choosing $\beta_1 X_{i1} + \cdots + \beta_p X_{ip} = y_i$.

However, there is more than one observation.

With N subjects, the likelihood is given by $L(\beta_1,...,\beta_p) =$

$$\prod_{i=1}^{N} f(y_i|X_{i1},...,X_{ip}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y_i - [\beta_1 X_{i1} + \cdots + \beta_p X_{ip}])^2}{2\sigma^2}\right\}$$

In general, it is not possible to choose $\beta_1,\ldots,\beta_p$ that will match every y_i to every $\beta_1 X_{i1} + \cdots + \beta_p X_{ip}$.

Instead, choose $\beta_1,\ldots,\beta_p$ to make the match as close as possible for all subjects.

$\Longrightarrow$ Choose $\beta_1,\ldots,\beta_p$ to maximize $L(\beta_1,\ldots,\beta_p)$.

It happens that ML Estimates = Least Squares Estimates.

These estimates can be written in closed-form using vector and matrix notation.

Linear Regression in Vector Notation

The linear regression model for Y_i

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i$$

can also be written using vector notation

$$\begin{aligned} Y_i &= \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + e_i \\ &= X_i'\beta + e_i \end{aligned}$$

where X_i is a $(p \times 1)$ vector representing the covariates,
$X_i' = (X_{i1}, X_{i2}, ..., X_{ip})$, and $\beta' = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of p regression parameters.

Aside: Matrix Addition and Multiplication

Matrices are like spreadsheets.

Importantly, they allow us to perform several arithmetic operations simultaneously.

They also provide a convenient shorthand notation.

Consider the following two simple examples.

$$\text{Let } A = \begin{pmatrix} 2 & 4 \\ 3 & 5 \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} 1 & 6 \\ 8 & 7 \end{pmatrix}.$$

Then,

$$A + B = \begin{pmatrix} 2+1 & 4+6 \\ 3+8 & 5+7 \end{pmatrix} = \begin{pmatrix} 3 & 10 \\ 11 & 12 \end{pmatrix}$$

$$A * B = \begin{pmatrix} 2*1+4*8 & 2*6+4*7 \\ 3*1+5*8 & 3*6+5*7 \end{pmatrix} = \begin{pmatrix} 34 & 40 \\ 43 & 53 \end{pmatrix}$$

Vectors

Vectors are special cases of matrices with one row or one column.

They follow the rules for matrices.

By convention, when we write a vector as X, we understand it to be a column vector of dimension, say $p \times 1$.

When we want to indicate a row vector, we write X'.

In linear regression we have the product of the following two vectors:

$$X_i'\beta = (X_{i1}, X_{i2}, ..., X_{ip}) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Maximum Likelihood Estimation

With independent observations, the joint density is simply the product of the individual univariate normal densities for Y_i.

Hence, we wish to maximize

$$\prod_{i=1}^{N} f(Y_i|X_{i1},...,X_{ip}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\}$$
$$= \left(2\pi\sigma^2\right)^{-N/2} \exp\left\{-\sum_{i=1}^{N} \frac{(Y_i - X_i'\beta)^2}{2\sigma^2}\right\},$$

evaluated at the observed values of the data, with respect to the regression parameters, β.

This is called maximizing the likelihood function.

Note that maximizing the likelihood is equivalent to maximizing the logarithm of the likelihood.

Hence, we can maximize

$$-\sum_{i=1}^{N} (Y_i - X_i'\beta)^2 / 2\sigma^2$$

by minimizing

$$\sum_{i=1}^{N} (Y_i - X_i'\beta)^2 / 2\sigma^2$$

Note: This is equivalent to finding the least squares estimates of β, i.e., the values that minimize the sum of the squares of the residuals.

The least squares solution can be written as

$$\widehat{\beta} = \left[\sum_{i=1}^{N} (X_i X_i')\right]^{-1} \sum_{i=1}^{N} (X_i Y_i)$$

This least squares estimate is the value that PROC GLM or PROC REG in SAS or any least squares regression program will produce.

Properties of Least Square Estimator

1. For any choice of σ^2, the least squares estimate of β is unbiased, that is

$$E(\widehat{\beta}) = \beta$$

2. The sampling distribution is given by:

$$\text{Cov}(\widehat{\beta}) = \sigma^2 \left[\sum_{i=1}^{N} (X_i X_i')\right]^{-1}$$

$$\widehat{\beta} \sim N\left(\beta,\ \sigma^2 \left[\sum_{i=1}^{N} (X_i X_i')\right]^{-1}\right)$$

Regression using PROC GLM in SAS

```
DATA topeka;
     INFILE 'topeka.txt';
     INPUT id height age logfev;
        loght = log(height);
RUN;

PROC GLM DATA=topeka;
     MODEL logfev = age loght;
RUN;

PROC REG DATA=topeka;
     MODEL logfev = age loght;
RUN;
```


SAS The GLM Procedure

Dependent Variable: logfev

Solution for Fixed Effects

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	29.149	14.5746	876.6	<0.0001
Error	297	4.938	0.0166		
Total	299	34.087			

Parameter	Estimate	Standard Error	t value	Pr > |t|
Intercept	-0.355	0.0319	-11.14	<0.0001
Age	0.020	0.0045	4.42	<0.0001
LogHt	2.295	0.1640	13.99	<0.0001

Interpretation

The fitted model is

$$\log(\text{FEV}_1) = -0.355 + 0.020 * \text{age} + 2.295 * \log(\text{ht})$$

So, a 1 year increase in age is associated with a 0.020 increase in $\log(\text{FEV}_1)$ (while holding height constant).

Similarly, for child number 48 (see earlier slide; age 11.3y height 1.45m), the fitted value of $\log(\text{FEV}_1)$ is

$$\log(\text{FEV}_1) = -0.355 + 0.020 * 11.3 + 2.295 * \log(1.45) = 0.724$$

so that the predicted $\text{FEV}_1 = \exp(0.724) = 2.06$ liters.

Analysis of Variance

ANOVA: Describes how the mean of a continuous dependent variable depends on a nominal (categorical, class) independent variable.

Analyzing samples from each of the p populations, we ask:

- Are there any differences in the p population means?
- If so, which of the means differ?

$\Longrightarrow$ One-Way Analysis of Variance (ANOVA)

Objective: To estimate and test hypotheses about the population group means, $\mu_1, \mu_2, \ldots, \mu_p$.

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_p$
$H_A : \mu_j$'s not all equal

Note: Some of the μ_j's could be equal under H_A

Relationship between ANOVA & Regression

Essentially identical, although often obscured by differences in terminology.

The ANOVA model can be represented as a multiple regression model with dummy (or indicator) variables.

$\Longrightarrow$ A multiple regression analysis with dummy-variable coded factors will yield the same results as an ANOVA.

Dummy or Indicator Variable Coding

Consider a factor with p levels:

Define $X_2 = 1$ if subject belongs to level 2, and 0 otherwise; define $X_3 = 1$ if subject belongs to level 3, and 0 otherwise; and define $X_4, ..., X_p$ similarly.

Note 1: By omission, the first level of the factor is selected as a "reference".
Note 2: Default option in many procedures in SAS is to use last level as a "reference".

Level	X_2	X_3	X_4	...	X_p
1	0	0	0	...	0
2	1	0	0	...	0
3	0	1	0	...	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots\vdots\vdots$	$\vdots$
p	0	0	0	...	1

This leads to a simple way of expressing the ANOVA model:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_p X_{ip} + e_i$$

Note:

$$\begin{aligned} \mu_1 &= \beta_1 \\ \mu_2 &= \beta_1 + \beta_2 \\ \mu_3 &= \beta_1 + \beta_3 \\ &\vdots \\ \mu_p &= \beta_1 + \beta_p \end{aligned}$$

Equivalently:

$$\begin{array}{lcl} \text{Group 2 versus (minus) Group 1} & = & \beta_2 \\ \text{Group 3 versus (minus) Group 1} & = & \beta_3 \\ \vdots & & \vdots \\ \text{Group p versus (minus) Group 1} & = & \beta_p \end{array}$$

Choice of Reference Level

The usual choice of reference group:

(i) A natural baseline or comparison group, and/or

(ii) group with largest sample size

Summary

The regression representation of ANOVA is more attractive because:

- It can handle balanced (i.e., equal cell sizes) and unbalanced data in a seamless fashion.
- In addition to the usual ANOVA table summaries, it provides other useful and interpretable results, e.g., estimates of effects and standard errors.
- Generalizations of ANOVA to include continuous predictors (and interactions among nominal and continuous predictors) are straightforward.