History/Course Content

- Theory of Linear Models → 2nd MS-level Methods Course
  - Linear Models
  - "Nonlinear" Models
  - Mixed (and Random Effects) Linear Models
  - Bootstrap Methods
  - Smoothing Methods (Generalized Additive Models)

E-mail — read carefully + store

Exams
Computing — R

"Blackboard" notation defaults
- caps are matrices/vectors A, X etc.
- l.c. are scalars c, d
- to make something l.c. into a vector or matrix I'll use an underline ~
- vectors are column vectors

Books — Roucher
  (Christensen)
  (Neter et al.)
  etc.

Koehler's Notes
Taped Outlines 2003
Handwritten Notes 2004
2003 Exams, 511 questions from MS exams
HW Problems + Keys

Basic "Linear Model" structure

\[ Y = X\beta + \varepsilon \]

\[ Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \]

I-dimensional random variables

(observable)
\[ X \text{ a matrix of (known) constants} \]

\[ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \text{ an unknown vector of parameters (constants)} \]

\[ \varepsilon \sim \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ an unobservable vector of random variables} \]

Almost Always people assume \( \mathbb{E} \varepsilon = 0 \)

Often \( \text{Var} \varepsilon = \sigma^2 I \)

Often \( \varepsilon \sim \text{MVN}_n \)

(in the presence of \( \text{Var} \varepsilon = \sigma^2 I \) this is \( \varepsilon_i \) are iid \( \mathcal{N}(0, \sigma^2) \))

Examples:

a) Multiple regression (2 predictors)

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \]

\[ \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \]

b) One-Way ANOVA model

Version 1: \( y_{ij} = \mu + \tau_i + \varepsilon_{ij} \)

Version 2: \( y_{ij} = \mu + T_i + \varepsilon_{ij} \)

Version 1 (3 treatments, 2 observations/treat)

\[ \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix} \]

Version 2

\[ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ T_1 \\ T_2 \\ T_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix} \]
More examples of $Y = X\beta + \varepsilon$ structure

c) One-way ANOVA model

Version 1: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$

Version 2: $y_{ij} = \mu + \tau_i + \epsilon_{2ij}$

Version 1 (3 treatments, 2 observations/treatment)

$\begin{pmatrix}
y_{11} \\
y_{12} \\
y_{21} \\
y_{22} \\
y_{31} \\
y_{32}
\end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & x_{11} \\
1 & 0 & 0 & x_{12} \\
0 & 1 & 0 & x_{21} \\
0 & 1 & 0 & x_{22} \\
0 & 0 & 1 & x_{31} \\
0 & 0 & 1 & x_{32}
\end{pmatrix} \begin{pmatrix} \mu \\
\tau_1 \\
\tau_2 \\
\epsilon_{11} \\
\epsilon_{12} \\
\epsilon_{21} \\
\epsilon_{22} \\
\epsilon_{31} \\
\epsilon_{32}
\end{pmatrix}$

Version 2 (3 treatments, 2 observations/treatment)

$\begin{pmatrix}
y_{11} \\
y_{12} \\
y_{21} \\
y_{22} \\
y_{31} \\
y_{32}
\end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & x_{11} \\
1 & 0 & 0 & x_{12} \\
0 & 1 & 0 & x_{21} \\
0 & 1 & 0 & x_{22} \\
0 & 0 & 1 & x_{31} \\
0 & 0 & 1 & x_{32}
\end{pmatrix} \begin{pmatrix} \mu \\
\tau_1 \\
\tau_2 \\
\epsilon_{11} \\
\epsilon_{12} \\
\epsilon_{21} \\
\epsilon_{22} \\
\epsilon_{31} \\
\epsilon_{32}
\end{pmatrix}$

d) 2-way ANOVA model (w/ interaction)

$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$

e.g. $I = J = 2$ and 2 observations/treatment

All (fixed-effects) standard models of statistical methods can be written in this form: $Y = X\beta + \varepsilon$

- estimate $X\beta$ $(\hat{\varepsilon}'Y)$
- make sensible point estimates of $\sigma^2$, elements of $\hat{\beta}$, $\hat{\epsilon}'\hat{\beta}$ for interesting vectors of constants $\hat{\omega}$
- make C.I.'s for $\sigma^2$, $e^\beta$
- make prediction intervals for new responses for a given set of conditions/predictors
- test hypotheses $H: \beta = \beta_0$, $H: \beta_k = 0$ (and generalizations)

The major complication that arises getting to these goals concerns the possibility that $X$ is not "full rank!" when it is not, some annoying (but not essential) abiguities arise.

there is no way to tell them apart based on data... This is related to the fact that

\[
X = \begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1
\end{pmatrix}
\]

doesn't have linearly independent columns—it doesn't have rank 4, it has rank 3—note: 1st column is the sum of the last 3

Example b Version 2

\[
\begin{pmatrix}
\mu \\
T_1 \\
T_2 \\
T_3
\end{pmatrix} = \begin{pmatrix}
0 \\
1 \\
2 \\
3
\end{pmatrix}
\] and
\[
\begin{pmatrix}
\mu \\
T_1 \\
T_2 \\
T_3
\end{pmatrix} = \begin{pmatrix}
-5 \\
6 \\
7 \\
8
\end{pmatrix}
\]

both produce

\[
EY = \begin{pmatrix}
1 \\
2 \\
3
\end{pmatrix}
\]
i.e., they produce the same set of mean values for the observables

Notice that the 2 versions of the 1-way ANOVA model (Example b)
have $X$ matrices with the same "column space" $C(X)$

\[
C(X) = \{ \text{all vectors that can be constructed as linear combinations of the columns of } X \}
\]

In example b)

\[
C(X) = \{ \begin{pmatrix}
a \\
b \\
c
\end{pmatrix} \mid a,b,c \text{ real } \}
\]
as a matter of fact, what is essential/ 

fundamental about the statement 

\[ Y = X \beta + \varepsilon \]

(at least as regards \( EY \)) is this 

\( C(X) \) - and in fact, a way to 

interpret or rephrase this 

is to say \((E \in \mathbb{R}) \) \( EY \in C(X) \)

---

\( \text{\textit{as it turns out}} \)

\( \text{rank}(X) = \text{"dimension of } C(X)\text{"} \)

= maximum # of vectors in 

\( C(X) \) that are linearly 

independent

(Ordinary) Least Squares Estimation of \( EY = X \beta \) some point in \( C(X) \)... to 

estimate it, it is plausible to try to 

minimize 

\( (Y - \hat{Y})' (Y - \hat{Y}) = \sum (y_i - \hat{y}_i)^2 \)

over choices of \( \hat{Y} \in C(X) \)

\[ C(X) \] is a "subspace" of \( \mathbb{R}^n \)

Geometry?

For \( n = 3 \)

\( \mathbb{R}^3 = \{ (u,v,w)' \} \)

subspaces are:

- all of \( \mathbb{R}^3 \)
- a plane 
  containing \( Q \)
- a line Through \( Q \)
  a one-dimensional 
  subspace

\[ \text{Example} \]

One Way model (3 groups, 
2 observations/treatment)

\[ C(X) = \begin{cases} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \mid a, b, c \text{ real} \end{cases} \]

\[ Y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} \quad \hat{Y} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \]
\[(y-\hat{y})(y-\hat{y}) = (y_1-x)^2 + (y_2-a)^2
\]
\[+ (y_3-b)^2 + (y_4-c)^2
\]
and I wish to minimize this by choice of \(a, b, c\). This is accomplished by the choice \(a = \bar{y}_1, b = \bar{y}_2, c = \bar{y}_3\).

Note that \((y-\hat{y})(y-\hat{y})\) is the squared distance (in \(\mathbb{R}^n\)) between \(y\) and \(\hat{y}\).

\(\hat{y}\) is the "perpendicular projection of \(y\) onto \(c(x)\)"

**Exeuction?** Koehler's notes around panel 175

Christensen's book Appendix B.3

**Facts**

1) \(\exists\) a unique \(n \times n\) matrix \(P_X\) such that \(\forall y \in \mathbb{R}^n\), \(P_Xy\) is the perpendicular projection of \(y\) onto \(c(x)\)

2) \(P_X = X(X'X)^{-1}X'\) for any \(X'X\)

---

So the prescription "choose \(\hat{y} \in c(x)\) to minimize \((y-\hat{y})(y-\hat{y})\)" is equivalently "choose \(\hat{y} \in c(x)\) closest to \(y\)".

Geometry for \(R^3\) in \(c(x)\)

\(\hat{y}\) is the closest point in \(c(x)\) to \(y\)

- OLS estimate of \(E\)

(All choices of \((XX)^{-}\) produce the same \(P_X\))

**Generalized Inverses**

1) A generalized inverse of \(A\) is any matrix \(G\) such that \(AGA = A\)

2) In general there can be more than 1 generalized inverse and there are a variety of algorithms for computing such (see panels 152-161)

3) A square and nonsingular \(\Rightarrow \exists\) a unique generalized inverse \(A^{-}\), and \(A^{-} = A^{-1}\)
4) A square + symmetric $\Rightarrow$ 

$\exists$ at least one symmetric $A$.

$P_X$ is sometimes called "the hat matrix" and written $\hat{H}$.

Properties of projection

$P_X' = P_X$ ( $P_X$ is symmetric) 
and $P_XP_X = P_X$ ( $P_X$ is idempotent) 

This makes sense because

$P_Xy \in C(X)$ and thus

$P_X(P_Xy) = P_Xy \quad \forall y$

Notice

$e = y - \hat{y} = (I - P_X)y$

... and we need to think about residuals. This matrix $I - P_X$, what it's doing and more about the geometry of least squares...

A bit more matrix/linear algebra background

Two vectors $u$ and $v$ are perpendicular provided $u'v = 0$

often people write $u \perp v$. 
