Another Example Use of the Additive Model

Can we estimate average movie ratings and individual customer ratings? Yes under an additive model but no under a cell-means model

customer	movie 1	movie 2	movie 3
1	4	1	?
2	?	3	5
3	?	?	3
4	3	1	?

Can we guess ratings for customer/movie combinations not in the dataset?

Which movie is best?

y_{ij} = customer i's rating of movie j

1 through 5

$$y_{ij} = \mu + c_i + m_j + \epsilon_{ij}$$

end lecture 11 2-14-25

The Linear Model in Vector and Matrix Form

7 observations

$$
\begin{array}{l}
y_{11} \rightarrow \\ y_{12} \rightarrow \\ y_{22} \rightarrow \\ \\ \\ \\ \\
\end{array}
\begin{bmatrix} 4 \\ 1 \\ 3 \\ 5 \\ 3 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix} \overset{\beta}{\begin{bmatrix} \mu \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ m_1 \\ m_2 \\ m_3 \end{bmatrix}} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \end{bmatrix}
$$

$$
\boldsymbol{y} = \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}
$$

Can we estimate means for missing data?

We can estimate all
y_{ij} for which we observed data!

$$E(\boldsymbol{y}) = \boldsymbol{X\beta} = \begin{bmatrix} \mu + c_1 + m_1 \\ \mu + c_1 + m_2 \\ \mu + c_2 + m_2 \\ \mu + c_2 + m_3 \\ \mu + c_3 + m_3 \\ \mu + c_4 + m_1 \\ \mu + c_4 + m_2 \end{bmatrix}$$

Can we estimate
$\mu + c_1 + m_3$?
$\mu + c_2 + m_1$?
$\mu + c_3 + m_1$?
$\mu + c_3 + m_2$?
$\mu + c_4 + m_3$?

$m_1 - m_2$ is estimable because

$$[1, -1, 0, 0, 0, 0, 0]E(\boldsymbol{y}) = [1, -1, 0, 0, 0, 0, 0]\boldsymbol{X\beta} = m_1 - m_2.$$

Can we estimate means for missing data?

$$X\boldsymbol{\beta} = \begin{bmatrix} \mu + c_1 + m_1 \\ \mu + c_1 + m_2 \\ \mu + c_2 + m_2 \\ \mu + c_2 + m_3 \\ \mu + c_3 + m_3 \\ \mu + c_4 + m_1 \\ \mu + c_4 + m_2 \end{bmatrix}$$

Can we estimate

$\mu + c_1 + m_3$?

$\mu + c_2 + m_1$?

$\mu + c_3 + m_1$?

$\mu + c_3 + m_2$?

$\mu + c_4 + m_3$?

Likewise, $m_2 - m_3$ is estimable because

$$[0, 0, 1, -1, 0, 0, 0]E(\boldsymbol{y}) = [0, 0, 1, -1, 0, 0, 0]\boldsymbol{X\beta} = m_2 - m_3.$$

We can estimate all pairwise differences between movie effects.

Because $m_1 - m_2$ and $m_2 - m_3$ are estimable, we can also estimate

$$(m_1 - m_2) + (m_2 - m_3) = m_1 - m_3.$$

This follows because any linear combination of estimable functions is also estimable.

We can estimate the mean underlying the rating for any combination of customer and movie.

It follows that any linear combination of the form

y_{ij}

$$\mu + c_i + m_j$$

can be estimated $\forall\ i = 1, 2, 3, 4$ and $j = 1, 2, 3$ because

$$\mu + c_i + m_j = (\mu + c_i + m_{j^*}) + (m_j - m_{j^*})$$

$$\forall\ i = 1, 2, 3, 4 \text{ and } j, j^* = 1, 2, 3.$$

using indirect information available about m_j based on other customers & customer i

Movie LSMEANS

effects

If our goal is to compare movies to see which is most highly rated, we can accomplish that by estimating the pairwise differences between movie effects.

However, if we want to retain information about the mean rating rather than the difference between mean ratings, it is natural to consider estimating the average (across *all* customers) of the mean rating for each movie.

average / means

Movie LSMEANS

This average for the jth movie is

$$\frac{1}{4}\sum_{i=1}^{4}(\mu + c_i + m_j) = \mu + \frac{1}{4}\sum_{i=1}^{4} c_i + m_j = \mu + \bar{c}. + m_j.$$

a "period"

This average is estimable for each movie in our example because it is a linear combination of estimable functions.

Estimates of $\mu + \bar{c}. + m_j \;\forall\; j = 1, 2, 3$ are movie LSMEANS.

Marginal Means for the Additive Model

	Movie 1	Movie 2	Movie 3	
Cust1	$\mu + c_1 + m_1$	$\mu + c_1 + m_2$	$\mu + c_1 + m_3$	$\mu + c_1 + \bar{m}.$
Cust2	$\mu + c_2 + m_1$	$\mu + c_2 + m_2$	$\mu + c_2 + m_3$	$\mu + c_2 + \bar{m}.$
Cust3	$\mu + c_3 + m_1$	$\mu + c_3 + m_2$	$\mu + c_3 + m_3$	$\mu + c_3 + \bar{m}.$
Cust4	$\mu + c_4 + m_1$	$\mu + c_4 + m_2$	$\mu + c_4 + m_3$	$\mu + c_4 + \bar{m}.$
	$\mu + \bar{c}. + m_1$	$\mu + \bar{c}. + m_2$	$\mu + \bar{c}. + m_3$	$\mu + \bar{c}. + \bar{m}.$

Suppose we consider a different model.

$$
\begin{array}{cc|ccc|}
 & & & \text{movie} & \\
 & & 1 & 2 & 3 \\
\hline
 & 1 & 4 & 1 & ? \\
\text{customer} & 2 & ? & \boxed{3} & \boxed{5} \\
 & 3 & ? & ? & \boxed{3} \\
 & 4 & \underline{3} & \underline{1} & ? \\
\hline
\end{array}
$$

Can we guess ratings for customer/movie combinations not in the dataset?

freely estimable due to the inter-action between c & m

y_{ij} = customer i's rating of movie j

Which movie is best?

$$y_{ij} = \mu_{ij} + \epsilon_{ij}$$

$$= \mu + c_i + m_j + (cm)_{ij}$$

The Linear Model in Matrix and Vector Form

only have data on c_1 for m_1 & m_2

$$
\begin{bmatrix} 4 \\ 1 \\ 3 \\ 5 \\ 3 \\ 3 \\ 1 \end{bmatrix}
=
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{bmatrix}
\begin{array}{c}
\left[\begin{array}{c}
\mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \hline \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \hline \mu_{31} \\ \mu_{32} \\ \mu_{33} \\ \hline \mu_{41} \\ \mu_{42} \\ \mu_{43}
\end{array}\right]
\end{array}
\begin{array}{l}
C_1 \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ C_4 \\ \\
\end{array}
+
\begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{33} \\ \epsilon_{41} \\ \epsilon_{42} \end{bmatrix}
$$

$$
\boldsymbol{y} = \quad \boldsymbol{X} \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\epsilon}
$$

Can we estimate the means that underly the missing table entries? No.

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{22} \\ \mu_{23} \\ \mu_{33} \\ \mu_{41} \\ \mu_{42} \end{bmatrix}$$

Can we estimate

$$\left.\begin{matrix} \mu_{13}? \\ \mu_{21}? \\ \mu_{31}? \\ \mu_{32}? \\ \mu_{43}? \end{matrix}\right\} \text{not under cell-means model}$$

None of the means underlying missing table entries are estimable under the cell means model.

assume additive model:

```
> y=c(4,1,3,5,3,3,1)
>
> X=matrix(c(
+ 1,1,0,0,0,1,0,0,
+ 1,1,0,0,0,0,1,0,
+ 1,0,1,0,0,0,1,0,
+ 1,0,1,0,0,0,0,1,
+ 1,0,0,1,0,0,0,1,
+ 1,0,0,0,1,1,0,0,
+ 1,0,0,0,1,0,1,0
+ ),byrow=T,nrow=7)
```

c_1 c_2 c_3 c_4 m_1 m_2 m_3

$$y_{ij} = \mu + c_i + m_j + \varepsilon_{ij}$$

$$\hat{y} = P_X y$$

Compute P_X next

Computation of P_X

```
> XX=t(X)%*%X
>
> library(MASS)
>
> XXgi=ginv(XX)
>
> Px=X%*%XXgi%*%t(X)
>
> #Px has entries like -1.387779e-16
```

$\hat{y}_{11} = 0.75 y_{11} + 0.25 y_{12} + 0.25 y_{41} - 0.25 y_{42} = 3.75$

y_{22} y_{23} y_{33}

```
> round(Px,2)
      [,1]  [,2] [,3] [,4] [,5]  [,6]  [,7]
[1,]  0.75  0.25    0    0    0  0.25 -0.25
[2,]  0.25  0.75    0    0    0 -0.25  0.25
[3,]  0.00  0.00    1    0    0  0.00  0.00
[4,]  0.00  0.00    0    1    0  0.00  0.00
[5,]  0.00  0.00    0    0    1  0.00  0.00
[6,]  0.25 -0.25    0    0    0  0.75  0.25
[7,] -0.25  0.25    0    0    0  0.25  0.75
```

our best estimate will be the observed y_{ij} itself

```
> fractions(Px)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  3/4  1/4    0    0    0  1/4 -1/4
[2,]  1/4  3/4    0    0    0 -1/4  1/4
[3,]    0    0    1    0    0    0    0
[4,]    0    0    0    1    0    0    0
[5,]    0    0    0    0    1    0    0
[6,]  1/4 -1/4    0    0    0  3/4  1/4
[7,] -1/4  1/4    0    0    0  1/4  3/4
```

Computing $P_X y = \hat{y}$

```
> yhat=Px%*%y
>
> yhat
     [,1]
[1,] 3.75
[2,] 1.25
[3,] 3.00
[4,] 5.00
[5,] 3.00
[6,] 3.25
[7,] 0.75
```

where Observed y_{ij} is our best estimate due to limited amount of information.

One Solution to the Normal Equations

```
> bhat=XXgi%*%t(X)%*%y
>
> bhat
             [,1]
[1,]  1.89473684
[2,]  0.22368421
[3,]  1.97368421
[4,] -0.02631579
[5,] -0.27631579
[6,]  1.63157895
[7,] -0.86842105
[8,]  1.13157895
```

does not have to be unique!

C Matrix for Estimating $\underbrace{\mu + c_i + m_j}_{y_{ij}} + \varepsilon_{ij} \; \forall\, i, j$

$\mu \quad c_1 \quad c_2 \quad \ldots \quad m_1 \quad m_2 \quad m_3$

```
> C=matrix(c(
+ 1,1,0,0,0,1,0,0,
+ 1,1,0,0,0,0,1,0,
+ 1,1,0,0,0,0,0,1,
+ 1,0,1,0,0,1,0,0,
+ 1,0,1,0,0,0,1,0,
+ 1,0,1,0,0,0,0,1,
+ 1,0,0,1,0,1,0,0,
+ 1,0,0,1,0,0,1,0,
+ 1,0,0,1,0,0,0,1,
+ 1,0,0,0,1,1,0,0,
+ 1,0,0,0,1,0,1,0,
+ 1,0,0,0,1,0,0,1
+ ),byrow=T,nrow=12)
```

$y_{ij} + \varepsilon_{ij}$

estimate y_{ij}

$\hat{y}_{ij} = \mu + c_i + m_j$

OLS Estimates of $\mu + c_i + m_j \;\; \forall\, i, j$

```
> Cbhat=C%*%bhat
> Cbhat
      [,1]
 [1,] 3.75
 [2,] 1.25
 [3,] 3.25
 [4,] 5.50
 [5,] 3.00
 [6,] 5.00
 [7,] 3.50
 [8,] 1.00
 [9,] 3.00
[10,] 3.25
[11,] 0.75
[12,] 2.75
```

$C\hat{\beta} = \hat{y}_{ij}$

$\hat{y}_{11}$

$\hat{y}_{12}$

$\hat{y}_{21} = 5.50 > 5$ because we did not put a constraint on the estimation movies

```
> M=matrix(Cbhat,nrow=4,byrow=T)
>
> M
     [,1] [,2] [,3]
[1,] 3.75 1.25 3.25
[2,] 5.50 3.00 5.00
[3,] 3.50 1.00 3.00
[4,] 3.25 0.75 2.75
>
> apply(M,2,mean)
[1] 4.0 1.5 3.5
```

average marginal movie rating

OLS Estimates of $m_j - m_{j^*}$ $\forall\ j \neq j^*$

```
> C=matrix(c(
+ 0,0,0,0,0,1,-1,0,
+ 0,0,0,0,0,1,0,-1,
+ 0,0,0,0,0,0,1,-1
+ ),byrow=T,nrow=3)
>
> Cbhat=C%*%bhat
> Cbhat
      [,1]
[1,]   2.5
[2,]   0.5
[3,]  -2.0
```

$\widehat{m_1 - m_2}$

$\widehat{m_1 - m_3}$

$\widehat{m_2 - m_3}$

y_{11} y_{12} y_{22} y_{23} y_{33} y_{41} y_{42}

$\widehat{\mu_1 - \mu_2}$

```
> round(C%*%XXgi%*%t(X),2)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  0.5 -0.5    0    0    0  0.5 -0.5
[2,]  0.5 -0.5    1   -1    0  0.5 -0.5
[3,]  0.0  0.0    1   -1    0  0.0  0.0
```

$\widehat{\mu_2 - \mu_3}$

		movie		
		1	2	3
customer	1	4	1	?
	2	?	3	5
	3	?	?	3
	4	3	1	?

end lecture 12 02-17-25

(Best to make sense of rows 1 and 3 first and then row 2 follows.)

Alternative Analysis Using the R Full-Rank X Matrix

```
> customer=factor(c(1,1,2,2,3,4,4))
> movie=factor(c(1,2,2,3,3,1,2))
> d=data.frame(customer,movie,y)
>
> d
  customer movie y
1        1     1 4
2        1     2 1
3        2     2 3
4        2     3 5
5        3     3 3
6        4     1 3
7        4     2 1
```

The R Full-Rank $\boldsymbol{X}$ Matrix

```
> o=lm(y~customer+movie,data=d)
> model.matrix(o)

  (Intercept) customer2 customer3 customer4 movie2 movie3
1           1         0         0         0      0      0
2           1         0         0         0      1      0
3           1         1         0         0      1      0
4           1         1         0         0      0      1
5           1         0         1         0      0      1
6           1         0         0         1      0      0
7           1         0         0         1      1      0
```

Marginal Means for the R Full-Rank X Matrix

	Movie 1	Movie 2	Movie 3	
Cust1	μ	$\mu + m_2$	$\mu + m_3$	$\mu + \frac{m_2+m_3}{3}$
Cust2	$\mu + c_2$	$\mu + c_2 + m_2$	$\mu + c_2 + m_3$	$\mu + c_2 + \frac{m_2+m_3}{3}$
Cust3	$\mu + c_3$	$\mu + c_3 + m_2$	$\mu + c_3 + m_3$	$\mu + c_3 + \frac{m_2+m_3}{3}$
Cust4	$\mu + c_4$	$\mu + c_4 + m_2$	$\mu + c_4 + m_3$	$\mu + c_4 + \frac{m_2+m_3}{3}$
	$\mu + \frac{c_2+c_3+c_4}{4}$	$\mu + \frac{c_2+c_3+c_4}{4} + m_2$	$\mu + \frac{c_2+c_3+c_4}{4} + m_3$	$\mu + \frac{c_2+c_3+c_4}{4} + \frac{m_2+m_3}{3}$

$\widehat{\beta}$, $\widehat{y}$, and $y - \widehat{y}$

```
> coef(o)
(Intercept)  customer2  customer3  customer4    movie2    movie3
       3.75       1.75      -0.25      -0.50     -2.50     -0.50

> fitted(o)
   1    2    3    4    5    6    7
3.75 1.25 3.00 5.00 3.00 3.25 0.75

> resid(o)
            1             2     3    4    5
 2.500000e-01 -2.500000e-01     0    0    0
            6             7
-2.500000e-01  2.500000e-01
```

$\widehat{\beta}$, $\widehat{y}$, and $y - \widehat{y}$

```
> o$coe
(Intercept)  customer2  customer3  customer4    movie2    movie3
       3.75       1.75      -0.25      -0.50     -2.50     -0.50

> o$fit
   1    2    3    4    5    6    7
3.75 1.25 3.00 5.00 3.00 3.25 0.75

> o$res
            1             2     3     4     5
 2.500000e-01 -2.500000e-01     0     0     0
            6             7
-2.500000e-01  2.500000e-01
```

OLS Estimates of $m_j - m_{j^*}\ \forall\ j \neq j^*$

```
> -o$coe[5]
movie2
   2.5
> -o$coe[6]
movie3
   0.5
> o$coe[5]-o$coe[6]
movie2
    -2
```

OLS Estimates of $m_j - m_{j^*} \ \forall\ j \neq j^*$

```
> C=matrix(c(
+ 0,0,0,0,-1,0,
+ 0,0,0,0,0,-1,
+ 0,0,0,0,1,-1
+ ),byrow=T,nrow=3)
>
> C%*%o$coe
      [,1]
[1,]   2.5
[2,]   0.5
[3,]  -2.0
```

C Matrix for Estimating $\mu + c_i + m_j \;\; \forall \; i, j$

```
> C=matrix(c(
+ 1,0,0,0,0,0,
+ 1,0,0,0,1,0,
+ 1,0,0,0,0,1,
+ 1,1,0,0,0,0,
+ 1,1,0,0,1,0,
+ 1,1,0,0,0,1,
+ 1,0,1,0,0,0,
+ 1,0,1,0,1,0,
+ 1,0,1,0,0,1,
+ 1,0,0,1,0,0,
+ 1,0,0,1,1,0,
+ 1,0,0,1,0,1
+ ),byrow=T,nrow=12)
```

OLS Estimates of $\mu + c_i + m_j \;\; \forall\, i, j$

```
> matrix(C%*%o$coe,nrow=4,byrow=T)
     [,1] [,2] [,3]
[1,] 3.75 1.25 3.25
[2,] 5.50 3.00 5.00
[3,] 3.50 1.00 3.00
[4,] 3.25 0.75 2.75
```