Regression Models for Quantitative and Qualitative Predictors

Paweł Polak

November 1, 2017

Linear Regression Models - Lecture 10

Content:

- Polynomial Regression Models
- Interaction Regression Models
- Categorical Explanatory Variables
- ANOVA and ANCOVA
- Two way ANOVA

General Linear Regression Model

- *Independent responses* of the form $Y_i \sim N(\mu_i, \sigma^2)$, where

$$\mu_i = \mathbf{X}_i^\top \boldsymbol{\beta}$$

for some known vector of *explanatory* variables $\mathbf{X}_i^\top = (X_{i1}, \ldots, X_{ip})$.

- Unknown *parameter* vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{P-1})^\top$, where $P < N$.

- This is the *linear model* and is usually written as

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

(in vector notation) where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{P-1} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix},$$

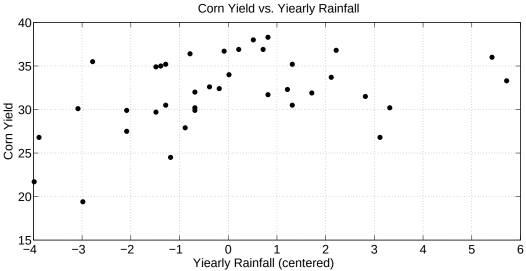
where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, for $i = 1, 2, \ldots, N$.

Building Regression Models

- One of the first steps in the construction of a regression model is to *hypothesize* the form of the regression function.
- We can dramatically expand the scope of our regression models by including specially constructed explanatory variables.
- These include *indicator* variables, *interaction terms*, *transformed* variables, and *higher order* terms.

Yield and Rainfall

- Data was collected on the yearly rainfall and corn yield at a farm during a 38 year period.



- There exits a *curvilinear* relationship between the variables.

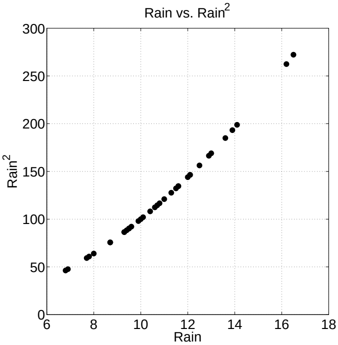
- The relationship appears to be quadratic.

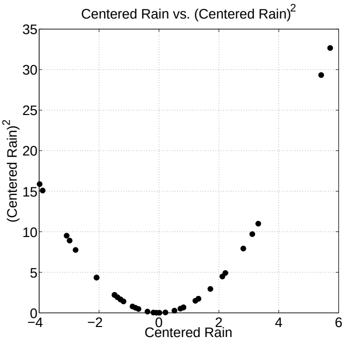
Polynomial Regression Models

- Polynomial regression models are useful when there is reason to believe the relationship between two variables is *curvilinear*.
- The general form for the polynomial regression model with one explanatory variable is:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_P X_i^p + \varepsilon_i$$

- The *order* of the model, P, is the highest power used for the explanatory variable.
- Prior to performing polynomial regression it is recommended to *center* the observations by removing their mean, i.e., exchange X_i with $X_i - \bar{X}$ to minimize problems with multicollinearity.





Alternative format

- Regression coefficients in polynomial regression are often written in an alternative format.

- We write
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$
as
$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

Yield and Rainfall

- Let Y be the *yield* and X the *yearly rainfall*.

- Since the relationship is *quadratic*, we use poly. regression of order 2.

- Predicted response:
$$\begin{aligned}\hat{Y}_i &= \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i \\ \hat{Y} &= 33.06 + 1.06X - 0.23X^2\end{aligned}$$

Re-expression

- After fitting a polynomial regression model, we often re-express it using the original variables.
- The fitted model:

$$\hat{Y}_i = b_0 + b_1 X_i + b_{11} X_i^2$$

becomes

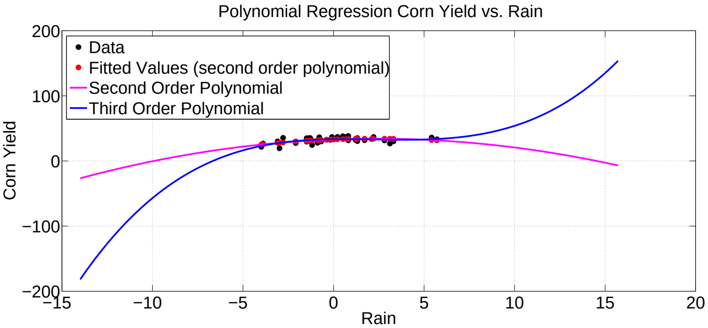
$$\hat{Y}_i = b_0' + b_1' X_i + b_{11}' X_i^2$$

where

$$b_0' = b_0 - b_1 \bar{X} + b_{11} \bar{X}^2, \quad , b_1' = b_1 - 2 b_{11} \bar{X}, \quad b_{11}' = b_{11}$$

Comments

- Be careful when choosing the order of the polynomial regression model, as it is easy to *over-fit* the model.
- For a problem with N data points, a polynomial of order $N-1$ will pass through all N points.
- However, such a model will not be useful for predicting future values.
- Extrapolation is particularly hazardous when using polynomial regression.
- Polynomial regression may provide good fits for the data at hand but may turn in unexpected directions when extrapolated beyond the range of the data.



Interaction Regression Models

- A regression model with $P-1$ explanatory variables contains *additive effects* if the regression function can be written in the form:

$$\mathbb{E}(Y) = f(X_1) + f(X_2) + \ldots + f(X_{P-1}).$$

- Example:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2.$$

- Two explanatory variables are said to *interact* if the effect that one of them has on the mean response *depends* on the value of the other.

- A simple way of modelling interaction is by including a *bilinear interaction term* (e.g., $X_1 X_2$).

Interaction Regression Models

- For example:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- The interpretation of the coefficients β_1 and β_2 differ due to the inclusion of the interaction term.
- The change in mean response with a unit increase in X_1, when X_2 is fixed, is $\beta_1 + \beta_3 X_2$.
- Hence, the effect of X_1, for a given level of X_2, will depend on the value of X_2.
- The same relationship holds for X_1.

Polynomial Regression Models

- The regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_{11} X_{1i}^2 + \beta_2 X_{2i} + \beta_{22} X_{2i}^2 + \beta_{22} X_{1i} X_{2i} + \varepsilon_i$$

where

$$X_{1i} \to X_{1i} - \bar{X}_1, \qquad X_{2i} \to X_{2i} - \bar{X}_2$$

is a *second order* model with *two* explanatory variables.

Categorical Explanatory Variables

- So far we have only used quantitative explanatory variables in our regression models.
- However, often the explanatory variables we are interested in are categorical (e.g., gender, weekday, hair color).
- We can use *indicator* variables, or *dummy* variables to denote the values of the categorical variable.
- There are a number of ways of quantitatively identifying the classes of a categorical variable.
- Often the most appropriate is to use indicator variables that take on the values 0 and 1, i.e., $X_i = 1$ if the observation belongs to group A, and 0 otherwise.

Illustration

- Suppose we have data on two variables X_1 and Y collected for two separate groups (A and B).
- Define X_2 to be an *indicator* variable that is equal to 1 if the observation belongs to group A and 0 if it belongs to group B.
- Consider the regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

- The mean response for

$$\begin{aligned} \text{Group } A: &\quad \mathbb{E}(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \\ \text{Group } B: &\quad \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 \end{aligned}$$

- The groups are allowed to have different intercepts, but must have the same slope.

Inference

- To determine whether the mean of Y differs between the two groups, after *controlling* for the other explanatory variable, test:

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_a : \beta_2 \neq 0$$

- If we reject H_0, there is evidence of a significant difference in means between the groups.

Insulating foam

- Data was collected to see whether a certain type of *insulating foam* had an effect on the ambient *formaldehyde* (CH_2O) concentration inside a house.
- As the amount of CH_2O was also influenced by the amount of air that can move through the house via windows and cracks, an *air tightness* rating (between 0-10) was determined for each house.
- Let Y be the CH_2O concentration, X_1 the air tightness of the house and X_2 equal to 1 if foam is present in the house and 0 otherwise.
- Model: $Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.
- Is there a *difference* in the average concentration of CH_2O between homes of equal air tightness but different insulation?

- Predicted Response: $\hat{Y} = 31.37 + 2.85X_1 + 9.31X_2$.
- Test: $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$
- From the output: $t = 4.37$, p-value $= 0.0003$
- There is *strong* evidence that homes with foam insulation have *higher* CH_2O concentration.

Varying Slopes and Intercepts

- In the previous example we used an indicator variable to model differences in the *intercept* between groups.
- Sometimes we also want the *slopes* of the regression model to differ between groups.
- This can be done by including an *interaction* term together with an indicator variable in the model.

Illustration

- Suppose we have data on two variables X_1 and Y collected for *two groups* (A and B).
- Let X_2 be equal to 1 if the observation belongs to group A and 0 if it belongs to group B.
- Consider a regression model with *interactions*:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i.$$

- The response surface:

$$\begin{aligned} \text{Group } A: &\qquad \mathbb{E}(Y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \\ \text{Group } B: &\qquad \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 \end{aligned}$$

- Picture! Both the *intercept* and *slope* are allowed to vary.

Inference

- Testing whether the two regression equations are *identical* involves the following hypothesis:

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{versus} \quad H_1 : \text{ Both not equal to 0.}$$

- Perform this test using a t-test.

- Perform this test using a general linear F-test.

Varying slopes

- We have looked at regression models where:
  - the *intercept* is allowed to *vary* between groups.
  - the *intercept* and *slope* are allowed to vary across groups.
- How about the case where the slope varies but not the intercept?

Illustration

- Suppose we have data on two variables X_1 and Y collected for two groups (A and B).
- Let X_2 be equal to 1 if the observation belongs to group A and 0 if it belongs to group B.
- Consider a regression model with interactions:

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i} X_{2i} + \varepsilon_i.$$

- The response surface:

$$\begin{aligned} \text{Group } A: &\qquad \mathbb{E}(Y) = \beta_0 + (\beta_1 + \beta_2) X_1 \\ \text{Group } B: &\qquad \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 \end{aligned}$$

- Picture! The *slopes* are allowed to *vary*, but *not* the *intercepts*.

Inference

- Testing whether the two regression equations are identical involves the following hypothesis:

$$H_0 : \beta_2 = 0 \quad \text{versus} \quad H_1 : \beta_2 \neq 0.$$

- Perform this test using a t-test.

- To determine whether the effect of the foam depends on air tightness include an interaction term.

- Predicted Response: $\hat{Y} = 30.00 + 3.12X_1 + 12.48X_2 - 0.62X_1X_2$

Results

- Test: $H_0 : \beta_2 = \beta_3 = 0$ versus $H_1 :$ Both not equal to 0.
- Reject H_0.
- There is *strong evidence* that the *foam insulation* has an effect on the CH_2O concentration.
- Test individual regression coefficients:

$$
\begin{array}{rl}
H_0 : \beta_2 = 0 & \quad H_0 : \beta_3 = 0 \\
H_1 : \beta_2 \neq 0 & \quad H_1 : \beta_3 \neq 0. \\
\text{From the output:} & \quad \text{From the output:} \\
t = 2.79 & \quad t = -0.81 \\
p - value = 0.0113 & \quad p - value = 0.4292
\end{array}
$$

- The intercept appears to differ, but not slope.

Multiple classes

- Sometimes a categorical variable can take *more* than 2 possible values.
- A categorical variable with c classes is best represented using $c-1$ separate indicator variables.
- This provides a more flexible model than coding the different classes using a single variable.
- Illustration: Create a model relating profit, Y, to bank size, X_1, and bank type (Commercial, Mutual savings or Savings and loans).

Model I

- If *bank type* is coded as a variable X_2 with values

$$\begin{aligned} \text{Comercial:} &\quad 1 \\ \text{Mutual Savings:} &\quad 2 \\ \text{Savings \& loan:} &\quad 3 \end{aligned}$$

- The regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$.

- The mean response for

$$\begin{aligned} \text{Comercial:} &\quad \mathbb{E}(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \\ \text{Mutual Savings:} &\quad \mathbb{E}(Y) = (\beta_0 + 2\beta_2) + \beta_1 X_1 \\ \text{Savings \& loan:} &\quad \mathbb{E}(Y) = (\beta_0 + 3\beta_2) + \beta_1 X_1 \end{aligned}$$

- This approach is not very effective and/or very realistic.

- The difference in profit between Commercial and Mutual savings banks is β_2. Similarly, the difference between Savings & loan and Mutual saving banks must also be equal to β_2.

Model II

- If bank type is coded as *two* variables X_2 and X_3 with values

	X_2	X_3
Comercial:	1	0
Mutual Savings:	0	1
Savings & loan:	0	0

- The regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$.

- The mean response for

$$\begin{aligned}
\text{Comercial:} \quad & \mathbb{E}(Y) = (\beta_0 + \beta_2) + \beta_1 X_1 \\
\text{Mutual Savings:} \quad & \mathbb{E}(Y) = (\beta_0 + \beta_3) + \beta_1 X_1 \\
\text{Savings \& loan:} \quad & \mathbb{E}(Y) = \beta_0 + \beta_1 X_1
\end{aligned}$$

- This approach is more flexible.

ANOVA and ANCOVA

- In analysis of variance (ANOVA) models *all* explanatory variables are *categorical*.
- In analysis of covariance (ANCOVA) models there are *both* quantitative and categorical variables. The explanatory variable of interest is categorical and the quantitative variables are included primarily to reduce variation.

Analysis of Variance

- The *one-way* analysis of variance (*ANOVA*) model is given by

$$Y_{jk} = \mu_j + \varepsilon_{jk},$$

  for $j = 1, \ldots, J$ and $k = 1, \ldots, N_j$, where the ε_{jk}'s are i.i.d. and follow a $N(0, \sigma^2)$ distribution.

- It can be used to test:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_J \quad \text{versus} \quad H_a : \text{ Not all means equal.}$$

- Example: Consider measuring yields of plants under a *control* condition and $J - 1$ different *treatment* conditions.

- The *explanatory* variable (factor) has J levels, and the response variables at level j are $Y_{j1}, \ldots, Y_{jn_j}$.

- The model that the responses are independent with

$$Y_{jk} \sim N(\mu_j, \sigma^2), \qquad j = 1, \ldots, J; \quad k = 1, \ldots, N_j$$

is of linear model form, with

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1N_1} \\ Y_{21} \\ \vdots \\ Y_{2N_2} \\ \vdots \\ Y_{J1} \\ \vdots \\ Y_{Jn_J} \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} N_1 \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} N_2 \\ \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} N_J \end{array} \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}.$$

Cereal grain example

- Six samples of each of *four* types of cereal grain were analyzed to determine the *thiamin* content, resulting in the following data:

Wheat	5.2	4.5	6.0	6.1	6.7	5.8
Barley	6.5	8.0	6.1	7.5	5.9	5.6
Maize	5.8	4.7	6.4	4.9	6.0	5.2
Oats	8.3	6.1	7.8	7.0	5.5	7.2

- Is there evidence of a *difference* in mean thiamin content between the grain types?

ANOVA and Regression

- ANOVA can be formulated and performed within the multiple regression framework.
- The variable 'grain type' can be included in the regression model using a series of indicator variables.
- If a variable has K levels, we need $K-1$ indicator variables in order to represent it properly.

Cereal grain

- Since there are 4 levels we need to define 3 indicator variables:

Groups	X_1	X_2	X_3
Wheat:	1	0	0
Barley	0	1	0
Maize	0	0	1
Oats	0	0	0

- The regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$.

- The mean response for

$$\begin{array}{rl} \text{Wheat} & \mathbb{E}(Y) = \beta_0 + \beta_1 \\ \text{Barley} & \mathbb{E}(Y) = \beta_0 + \beta_2 \\ \text{Maize} & \mathbb{E}(Y) = \beta_0 + \beta_3 \\ \text{Oats} & \mathbb{E}(Y) = \beta_0 \end{array}$$

- Each group has its *own mean response*.

- The standard ANOVA null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ is equivalent to testing the hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

  in the regression model.

- For the above test, $F = 3.957$, p-value $= 0.02293$.

- Moderately strong evidence of a *difference* in mean thiamin content between the four grain types.

- An alternative parameterization, emphasizing the differences between treatments, is

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}, \qquad j = 1, \ldots, J; \quad k = 1, \ldots, N_j$$

  where

  - μ is the *baseline* or *mean effect*
  - α_j is the effect of the j^{th} *treatment* (or the control $j = 1$).

- Notice that the parameter vector $(\mu, \alpha_1, \alpha_2, \ldots, \alpha_J)^\top$ is not *identifiable*, since replacing μ with $\mu + 10$ and α_j by $\alpha_j - 10$ gives the same model. Either a

  - *corner point* constraint $\alpha_1 = 0$ is used to emphasise the differences from the control, or the

  - *sum–to–zero* constraint $\sum_{j=1}^{J} N_j \alpha_j = 0$

  can be used to make the model identifiable.

- R uses corner point constraints.

- If $N_j = K$, say, for all j, the data are said to be *balanced*.

- We are usually interested in comparing the null model

$$H_0 : Y_{jk} = \mu + \varepsilon_{jk}$$

with that given above, which we call H_1; i.e., we wish to test whether the treatment conditions have an effect on the plant yield:

$$H_0 : \boldsymbol{\alpha} = 0, \text{ where } \alpha = (\alpha_1, \ldots, \alpha_J), \text{ against } \quad H_1 : \boldsymbol{\alpha} \neq 0.$$

- Check that the MLE fitted values, under H_1, are

$$\hat{Y}_{jk} = \bar{Y}_j \equiv \frac{1}{N_j} \sum_{k=1}^{N_j} Y_{jk},$$

whatever parameterization is chosen, and, under H_0, are

$$\hat{\hat{Y}}_{jk} = \bar{Y} \equiv \frac{1}{N} \sum_{j=1}^{J} N_j \bar{Y}_j, \qquad \text{where } N = \sum_{j=1}^{J} N_j.$$

- Our linear model theory says that we should test H_0 by referring

$$F = \frac{\frac{1}{J-1}\sum_{j=1}^{J} N_j(\bar{Y}_j - \bar{Y})^2}{\frac{1}{N-J}\sum_{j=1}^{J}\sum_{k=1}^{N_j}(Y_{jk} - \bar{Y}_j)^2} \equiv \frac{\frac{1}{J-1}S_2}{\frac{1}{N-J}S_1}$$

to $F_{J-1,N-J}$, where S_1 is the "within groups" *sum of squares* and S_2 is the "between groups" sum of squares.

- In (familiar) tabular form

Source of variation	Degrees of freedom	Sum of squares	F–statistic
Between groups	$J-1$	S_2	$F = \frac{\frac{1}{J-1}S_2}{\frac{1}{N-J}S_1}$
Within groups	$N-J$	S_1	
Total	$N-1$	$\sum_{j=1}^{J}\sum_{k=1}^{N_j}(y_{jk} - \bar{y})^2$	

Two–way ANOVA

- Suppose now that we have *two factors* having I, J levels respectively, and that our model for independent responses $\{Y_{ijk}\}$ is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

for $i = 1, \ldots, I; \;\; j = 1, \ldots, J; \;\; k = 1, \ldots, N_{ij}$, where $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

- For example, Y_{ijk} might represent the exam score of the k^{th} individual of sex $i \in \{\text{M,F}\}$ taking course j.

- This model is called an *additive two–way ANOVA model* because it is assumed that the effects of the different factors are *additive*.

- Possible identifiability constraints are

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0 \qquad \text{or} \qquad \alpha_1 = \beta_1 = 0.$$

Again, R uses the latter corner point constraint.

- Let this model correspond to the hypothesis H_3. We might be interested in testing

$$\begin{aligned} &H_0 : \alpha_i = \beta_j = 0 && \text{for all} && i = 1, \dots, I; \;\; j = 1, \dots, J \\ &H_1 : \alpha_i = 0 && \text{for all} && i = 1, \dots, I \\ &H_2 : \beta_j = 0 && \text{for all} && j = 1, \dots, J. \end{aligned}$$

For simplicity, assume that $N_{ij} = K$, say.

- The expressions for the MLE under each model depends on the identifiability constraint imposed, but the *fitted* values are the *same* and the residual sum of squares in each case is:

$$\begin{aligned} SSE(H_0) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y})^2 && \textbf{where} && \bar{Y} \equiv \bar{Y}_{+++} = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk} \\ SSE(H_1) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{+j+})^2 && \textbf{where} && \bar{Y}_{+j+} = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk} \\ SSE(H_2) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{i++})^2 && \textbf{where} && \bar{Y}_{i++} = \frac{1}{JK}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk} \\ SSE(H_3) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \tilde{Y}_{ijk})^2 && \textbf{where} && \tilde{Y}_{ijk} = \bar{Y}_{i++} + \bar{Y}_{+j+} - \bar{Y}. \end{aligned}$$

These can be used to calculate F–*statistics* in a way similar to two–way ANOVA.

Interactions

- In the two–way ANOVA model we assumed that the effects of the factors were *additive*.

- We may also want to check for the presence of *interaction* between the two effects, using the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

  Sometimes γ_{ij} is notated as $(\alpha\beta)_{ij}$ to more explicitly denote the interaction of α and β.

- Possible identifiability constraints include
  1. $\alpha_1 = \beta_1 = 0$, $\gamma_{1j} = 0$ for all j, and $\gamma_{i1} = 0$ for all i, or alternatively
  2. $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$, $\sum_{i=1}^{I} \gamma_{ij} = 0$ for each j, and $\sum_{j=1}^{J} \gamma_{ij} = 0$ for each i.

One can show that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk}-\bar{Y}_{+++})^2 = JK\sum_{i=1}^{I}(\bar{Y}_{i++}-\bar{Y}_{+++})^2 + IK\sum_{j=1}^{J}(\bar{Y}_{+j+}-\bar{Y}_{+++})^2$$
$$+ K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{ij+}-\bar{Y}_{i++}-\bar{Y}_{+j+}+\bar{Y}_{+++})^2$$
$$+ \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk}-\bar{Y}_{ij+})^2$$

That is the total sum squares is decomposed into that due to row differences, that due to column differences, that due to interaction, and that within cells. The test

$$H_0 : \gamma_{ij} = 0 \text{ for all } i,j \text{ vs. } H_1 : \gamma_{ij} \neq 0 \text{ for some } i,j$$

is based upon an $F_{(I-1)(J-1),IJ(K-1)}$ distributed test statistic given by

$$F = \frac{\left[K\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\bar{Y}_{ij+}-(\bar{Y}_{i++}+\bar{Y}_{+j+}-\bar{Y}_{+++})\right)^2\right]/[(I-1)(J-1)]}{\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(Y_{ijk}-\bar{Y}_{ij+}\right)^2\right]/[IJ(K-1)]}$$

- If an interaction is present, the interpretation is that the effect of the first factor on the response depends on the level of the second factor.
- For example, the response might be a "tastiness score" for a cake which depends on the factors of (1) baking time and (2) baking temperature.
- Interaction effects are most easily seen via plots, e.g., plot the responses Y_{ijk} against j, for each level of i. The statistical way is via an F–test (see `anova2_int.R`).