Handout 12

Longitudinal Data Analysis.

Longitudinal data analysis is related to repeated measures designs. Consider a few examples below. In each case, you have individuals or subjects being observed over time. For example, in the rat growth data, each rat is measured over 5 weeks. This type of data set is called longitudinal since the observations are taken over time. There is a covariate "mother's weight" (X). The idea is to see how rat weights vary over time since birth. In another example, logarithm of CD4 counts are listed for patients on three different treatments over time. Goal is to investigate how CD4 counts change over time and if age has any effect on this change. Note that in the first example, the times at which measurement are taken are the same for all subjects. In the second case times may be different for different patients. It should be pointed out that there is an extensive literature on longitudinal data analysis and here we will only introduce a few concepts here.

Looking at the rat data, if we make box-plot of weights by week, two things are apparent: the weights increase with week and the variability in weights also seems to increase.

The setting we will here is this: we have s subjects and for each subject a response is measured at each of r_i time points. Thus in Example 1, we can think of a few reasonable models (in all the models below ρ_i's are random subject (called "sub") effect with ρ_i's iid $N(0, \sigma_\rho^2)$ and ε_{ij}'s iid $N(0, \sigma^2)$: These models are related to repeated measures designs. In order to distinguish between week as a factor and week as a quantitative variable, we will write τ_j's as week factor effects and wk_j as a quantitative variable. Thus wk_j takes values $0, \ldots, 4$. One of the distinguishing features of these models and more complicated models given below is to account for the dependence of observations (over time) for the same subject.

Some models for the rat growth data.

(i) $Y_{ij} = \mu + \rho_i + \tau_j + \varepsilon_{ij}$, $\sum \tau_j = 0$. Mother's weight is not included in this model. This is a repeated measures model.

(ii) $Y_{ij} = \mu + \rho_i + \tau_j + \beta_1 X_i + \varepsilon_{ij}$, same as model (i), but mother's weight X_i is included. This is a repeated measures model with a covariate added.

(iii) $Y_{ij} = \mu + \rho_i + \beta_2 wk_j + \beta_1 X_i + \varepsilon_{ij}$, the growth is modeled as a straight line in time (week), thus the model has same slope but different (random) intercepts..

(iv) $Y_{ij} = \mu + \rho_i + \beta_2 wk_j + \gamma_{i2} wk_j + \beta_1 X_i + \varepsilon_{ij}$, the growth is now modeled as a straight line in time, but each subject has its own intercept (random) and its own slope (random). Here γ_{i2} are assumed to mean zero iid normal variables.

(v) $Y_{ij} = \mu + \rho_i + \beta_2 wk_j + \beta_3 wk_j^2 + \beta_1 X_i + \varepsilon_{ij}$, the week effect is now modeled as a quadratic function in wk_j.

(vi) $Y_{ij} = \mu + \rho_i + \beta_2 wk_j + \beta_3 wk_j^2 + \gamma_{i2} wk_j + \gamma_{i3} wk_j^2 + \beta_1 X_i + \varepsilon_{ij}$, where γ_{i2}'s and γ_{i3}'s are mean zero iid normally distributed variables and $Cov(\gamma_{i1}, \gamma_{i2}) = 0$. This is a model with random intercept and random slopes for wk_j and wk_j^2.

Here are the R commands: (you will need to download "lme4" package in R)

(i) lmer(y~wk+(1|sub)),

(ii) lmer(y~wk+X+(1|sub))

(iii) lmer(y~wk+X+(1|sub))

(iv) lmer(y~wk+X+(1|sub)+(0+wk|sub))

(v) lmer(y~wk+wk2+X+(1|sub)), [wk2=wk^2]

(vi) lmer(y~wk+wk2+X+(1|sub)+(0+wk|sub)+(0+wk2|sub)),]

In order to select among these models, we may choose the one with the smallest AIC or BIC value.

A more complicated class of models: If the number of weeks r_i are the same for all subjects ($r_i \equiv r$ for all i), then we may also consider models of the form

$$\begin{aligned} Y_{ij} &= \mu + \tau_j + \beta_1 X_i + \varepsilon_{ij}, \text{ or} \\ Y_{ij} &= \mu + \beta_2 wk_j + \beta_1 X_i + \varepsilon_{ij}, \text{ or} \\ Y_{ij} &= \mu + \beta_2 wk_j + \gamma_{i2} wk_j + \beta_1 X_i + \varepsilon_{ij} \end{aligned}$$

where $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in}), i = 1, \ldots, s$, are iid n-dim normal distribution with mean $(0, \ldots, 0)$ and covariance matrix $\sigma^2 D$. Unlike the models (i)-(vi), ε_{ij}'s are not assumed to be idd. Note that if D is not a diagonal matrix, this amounts to dependence of the observations Y_{ij} over time j for any given subject i. In last model γ_{i2}'s are the same as in Model (iv) above. In these cases, one needs to estimate the covariance matrix $\sigma^2 D$ along with other parameters. We will not investigate these complicated models here.

Analysis of rat grwoth data.

Let us first see what we get using models (ii) and (iv). The plots of the observed vs. fitted along with boxplots are given for both these models.

Here is the R output for Model (ii)

```
Linear mixed model fit by REML
Formula: y ~week + X + (1 | sub)
     AIC            BIC          logLik     deviance   REMLdev
    373.3          388.6         -178.7      376.5      357.3
         Random effects:
   Groups          Name         Variance    Std.Dev
     sub       (intercept)       24.968     4.9968
  Residual                      106.003    10.2958
Number of obs: 50, groups: sub, 10

Fixed Effects:
                 Estimate     Std.Error    t value
 (Intercept)     -31.4063      21.2015      -1.481
   week 1         18.3000      4.6044       3.974
   week 2         53.6000       4.6044     11.641
   week 3         73.1000       4.6044     15.876
   week 4        106.4000       4.6044     23.108
     X             0.5451      0.1290       4.227

Correlation of Fixed Effects:
                   (Intr)      week 1      week 2      week 3      week 4
   week 1         -0.109
   week 2         -0.109       0.500
   week 3         -0.109       0.500       0.500
   week 4         -0.109       0.500       0.500       0.500
     X            -0.985       0.000       0.000       0.000       0.000
```

Here is the R output for Model (iv)

```
Linear mixed model fit by REML
Formula: y ~wk + X + (1 | sub)+(0+wk|sub)
     AIC          BIC        logLik     deviance   REMLdev
    390.1        401.6        -189       381.6      378.1
Random effects:
   Groups         Name        Variance   Std.Dev
    sub       (intercept)      27.819     5.2744
    sub            wk          10.539     3.2464
  Residual                     91.747     9.5784
Number of obs: 50, groups: sub, 10

Fixed Effects:
                 Estimate    Std.Error   t value
 (Intercept)     18.8737      21.0022     0.899
     wk          26.7600       1.4040    19.060
     X            0.5451       0.1290     4.227

Correlation of Fixed Effects:
                  (Intr)        wk
     wk            0.000
     X            -0.995      0.000
```

Example 1: Weights of rats were measured during the first week of birth (week=0) and subsequent four weeks for 10 randomly selected rats. Also recorded were mothers' weights.

	Week					Mother's wt
Rat	0	1	2	3	4	Z
1	61	72	118	130	176	170
2	65	85	129	148	174	194
3	57	68	130	143	201	187
4	46	74	116	124	157	156
5	47	85	103	117	148	155
6	43	58	109	133	152	150
7	53	62	82	112	156	138
8	72	96	117	129	154	154
9	53	54	87	120	138	149
10	72	98	114	144	177	167

Example 2. (CD4 counts) This is a part of a bigger data set posted on the smartsite. This consists of HIV infected males who underwent one of the three treatments. Listed are age, and the times measurements were taken. [CD4 count is an important measure of the immune system. Higher the CD4 higher is the immunity.]

Patient	Treatment	age	week	log(CD4)
1	2	36.4271	0	3.135494216
1	2	36.4271	7.5714	3.044522438
1	2	36.4271	15.5714	2.772588722
1	2	36.4271	23.5714	2.833213344
1	2	36.4271	32.5714	3.218875825
1	2	36.4271	40	3.044522438
4	3	36.5969	0	4.119037175
4	3	36.5969	7.1429	4.110873864
4	3	36.5969	16.1429	4.709530201
4	3	36.5969	32.4286	2.833213344
18	1	51.8275	0	3.931825633
18	1	51.8275	17.8571	3.713572067
18	1	51.8275	32	3.713572067

Appendix.

The general linear mixed model can be described as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Z}_1\gamma_1 + \cdots + \boldsymbol{Z}_q\gamma_q + \boldsymbol{\varepsilon},$$

where Y is the vector of responses, $\boldsymbol{X\beta}$ is the fixed effects part consisting of fixed effects and covariates; $\boldsymbol{Z}_1\gamma_1, \ldots, \boldsymbol{Z}_q\gamma_q$ are the random effects parts and $\boldsymbol{\varepsilon}$ is the vector iid $N(0, \sigma^2)$ variables. Here $\boldsymbol{\gamma}_1$ is a vector of iid $N(0, \sigma_1^2)$ variables, $\boldsymbol{\gamma}_2$ is a vector of iid $N(0, \sigma_2^2)$ variables and so on. It is assumed that $\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_q$ and $\boldsymbol{\varepsilon}$ are all independent. In this setting,

$$E(\boldsymbol{Y}) = \boldsymbol{X\beta},\ \ \boldsymbol{V} = Cov(\boldsymbol{Y}) = \sigma_1^2\boldsymbol{Z}_1\boldsymbol{Z}_1^T + \cdots + \sigma_q^2\boldsymbol{Z}_q\boldsymbol{Z}_q^T + \sigma^2\boldsymbol{I},$$

i.e., $\boldsymbol{Y}$ has a multivariate normal distribution with mean $\boldsymbol{X\beta}$ and covariance matrix $\boldsymbol{V}$. The usual methods of estimation are maximum likelihood (ML) and restricted maximum likelihood (REML). Often it may not be possible to get closed form expressions of the parameter estimates. Iterations are used to obtain the ML or REML estimates.

Model (ii)

Let us describe Model (ii) for the case where $r_i = r$ for all i. Define I_{ijl},$l = 1, \ldots, r-1$, using {1,0,-1} codings for weeks as done for the simpler ANOVA models. Define

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ \vdots \\ Y_{ir} \end{bmatrix}, \boldsymbol{X}_i = \begin{bmatrix} 1 & X_i & I_{i,1,1} & \dots & I_{i,1,r-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_i & I_{i,r,1} & \dots & I_{i,r,r-1} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \tau_1 \\ \vdots \\ \tau_{r-1} \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}, \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \vdots \\ \varepsilon_{ir} \end{bmatrix}$$

Then for the ith subject we may write Model (ii) as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\rho_i + \boldsymbol{\varepsilon}_i, i = 1, \dots, s,$$

or we may combine them as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \vdots \\ \boldsymbol{Y}_s \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \vdots \\ \boldsymbol{X}_s \end{bmatrix}, \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Z}_{s-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{Z}_s \end{bmatrix}, \boldsymbol{\gamma} = \begin{bmatrix} \rho_1 \\ \vdots \\ \vdots \\ \rho_s \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_s \end{bmatrix}$

where the components of the vector$\boldsymbol{\gamma}$ are mean zero iid normally distributed variables. The first portion of the model $\boldsymbol{X\beta}$ combines all the fixed effects terms and $\boldsymbol{\varepsilon}$ is the vector of iid mean zero normal errors. The model in (1) is called a linear mixed effects model. Note that Y has a multivariate Normal distribution with mean $\boldsymbol{X\beta}$ and covariance matrix $\sigma_\gamma^2\boldsymbol{Z}\boldsymbol{Z}^T + \sigma^2\boldsymbol{I}$, where σ_γ^2 is the common variance of $\gamma_1, ..., \gamma_s$.

There are various ways of estimating the parameters of the mixed effects model given in (1): maximum likelihood (ML) and restricted maximum likelihood (REML). Restricted maximum likelihood involves multiplying both sides of (1) by a matrix $\boldsymbol{A}$ which has the property $\boldsymbol{AX} = \mathbf{0}$. It is possible to find such a matrix A of order $(rs - r - 1) \times rs$ since X has rank $r + 1$. In that case we are led to the model

$$\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Z}}\boldsymbol{\gamma} + \tilde{\varepsilon}, \tag{2}$$

with $\tilde{\boldsymbol{Y}} = \boldsymbol{AY}, \tilde{\boldsymbol{Z}} = \boldsymbol{AZ}$ and $\tilde{\varepsilon} = \boldsymbol{A\varepsilon}$. Matrix theory also tells us that it is possible to find the matrix $\boldsymbol{A}$ such that $\boldsymbol{AA}^T = \boldsymbol{I}$ so that the elements of $\tilde{\varepsilon}$ are iid $N(0, \sigma^2)$. Note that the $\tilde{\boldsymbol{Y}}$ is $rs - r - 1$ dimensional vector, unlike $\boldsymbol{Y}$ whose dimension is $rs \times 1$. One can therefore estimate the variance parameters σ_γ^2 and σ^2 using the ML method for the model given in (2). Once σ_γ^2 and σ^2 have been estimated, one can then use these estimates and use a weighted least squares method for the linear mixed model given in (1) in order to estimate $\boldsymbol{\beta}$.

Model (iv)

Let us now describe model (iv) for the case $r_i = r$ for all i. Now week is in the model as a quantitative variable. Define

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ \vdots \\ Y_{ir} \end{bmatrix}, \boldsymbol{X}_i = \begin{bmatrix} 1 & X_i & wk_1 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & X_i & wk_r \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \end{bmatrix}, \ \boldsymbol{Z}_{i1} = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}, \ \boldsymbol{Z}_{i2} = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}, \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \vdots \\ \varepsilon_{ir} \end{bmatrix}$$

Then for the ith subject we may write Model (iv) as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_{i1}\rho_i + \boldsymbol{Z}_{i2}\gamma_{i2} + \boldsymbol{\varepsilon}_i, i = 1, \ldots, s, \tag{3}$$

or we may combine them as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \boldsymbol{Z}_2\boldsymbol{\gamma}_2 + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \vdots \\ \boldsymbol{Y}_s \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \vdots \\ \boldsymbol{X}_s \end{bmatrix}, \boldsymbol{Z}_1 = \begin{bmatrix} \boldsymbol{Z}_{11} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{21} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Z}_{s-1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{Z}_{s,1} \end{bmatrix}, \boldsymbol{Z}_2 = \begin{bmatrix} \boldsymbol{Z}_{12} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}_{22} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Z}_{s-1,2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{Z}_{s,2} \end{bmatrix},$$

$$\boldsymbol{\gamma}_1 = \begin{bmatrix} \rho_1 \\ \vdots \\ \vdots \\ \rho_s \end{bmatrix}, \boldsymbol{\gamma}_2 = \begin{bmatrix} \gamma_{12} \\ \vdots \\ \vdots \\ \gamma_{s2} \end{bmatrix} \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \vdots \\ \boldsymbol{\varepsilon}_s \end{bmatrix}$$

Once again, one may use the ML or the REML method to estimate all the parameters.

Figure 1: Boxplot of rat growth

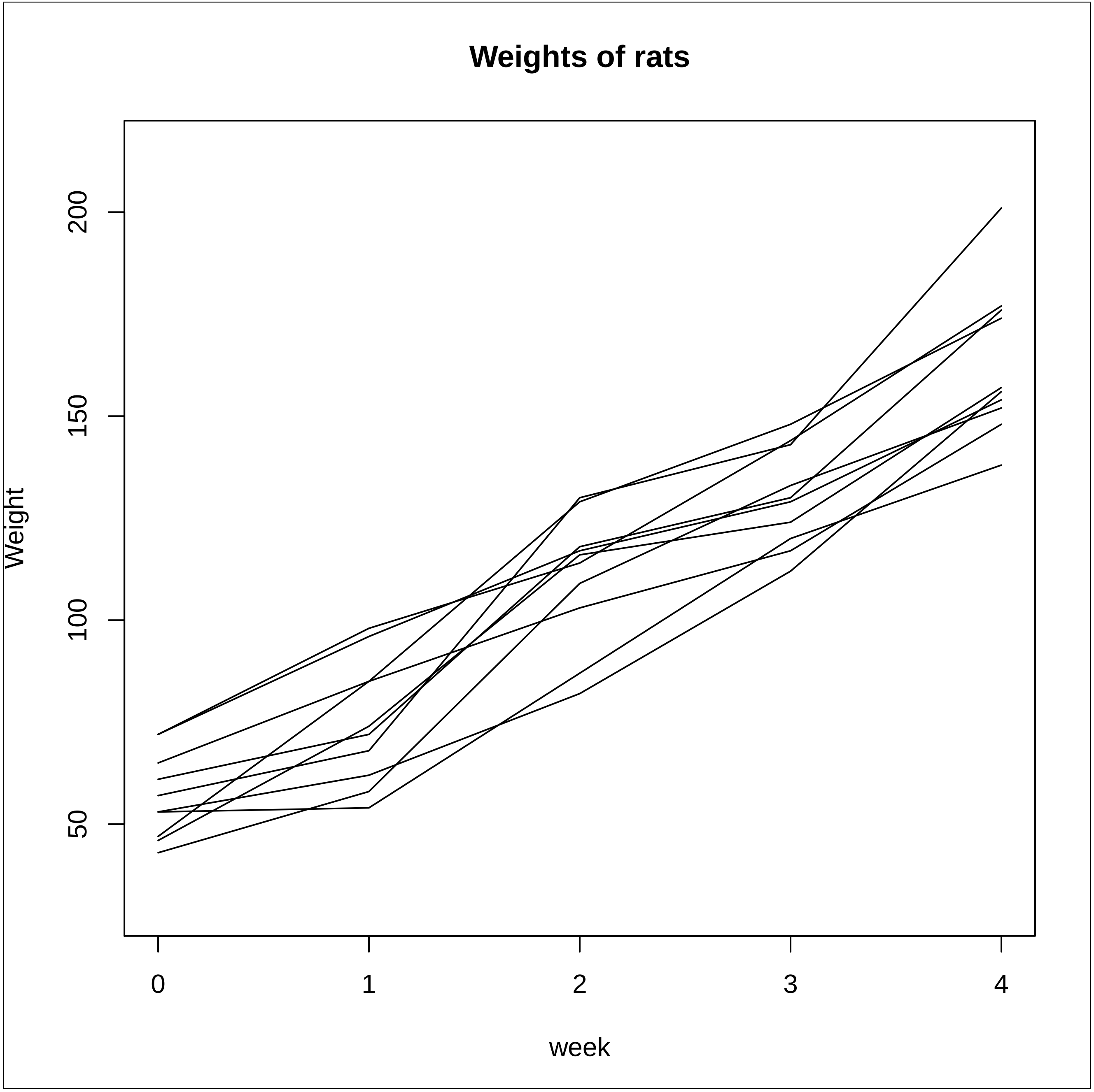


Figure 2: Model (ii): Rat Growth

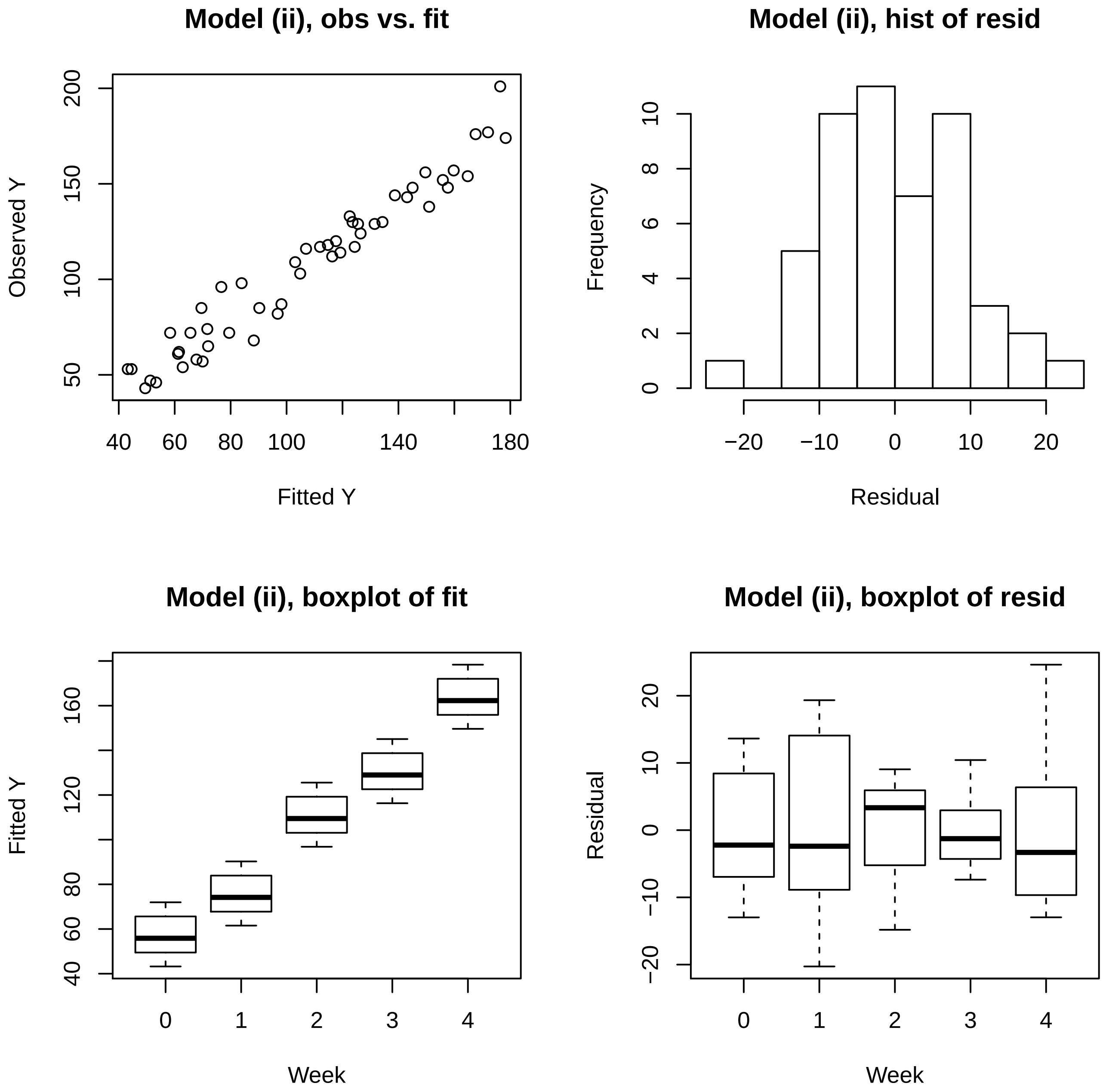
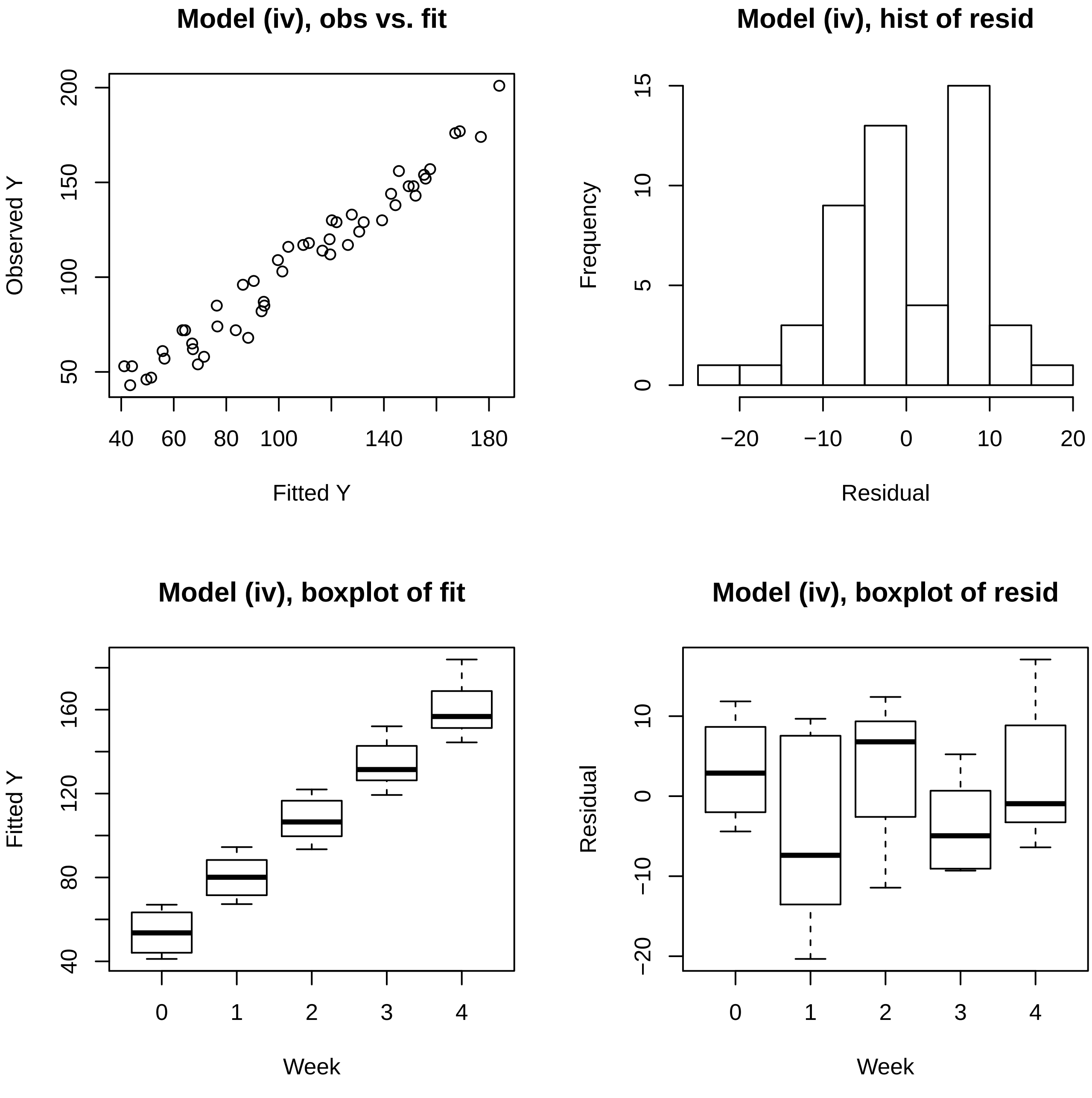


Figure 3: Model (iv): Rat Growth


Model (iv), obs vs. fit
Observed Y
Fitted Y
Model (iv), hist of resid
Frequency
Residual
Model (iv), boxplot of fit
Fitted Y
Week
Model (iv), boxplot of resid
Residual
Week