4. Diagnostics II: Symptoms and Remedies

Assumptions of the simple linear model

In the simple linear model of Y on X

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \text{NID}(0,\sigma^2)$$

the following assumptions are made:

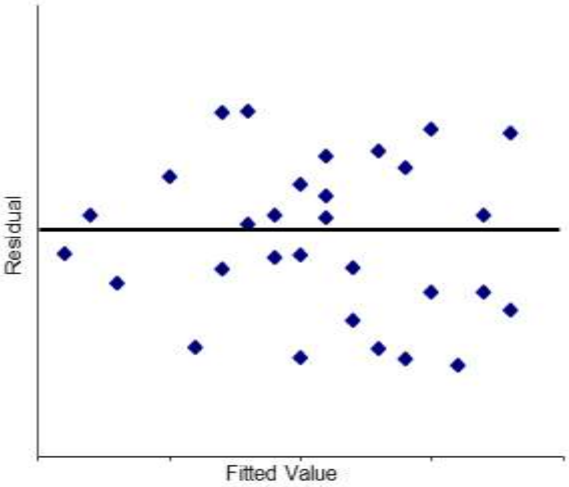
- the relationship between Y and X is **linear**
- the errors e_i (and hence the responses y_i) have **constant variance**
- the errors e_i are **normally distributed**

The residuals, $\hat{e}_i$, provide information about the true errors. In this chapter, we use regression diagnostic plots of these residuals to test these assumptions.

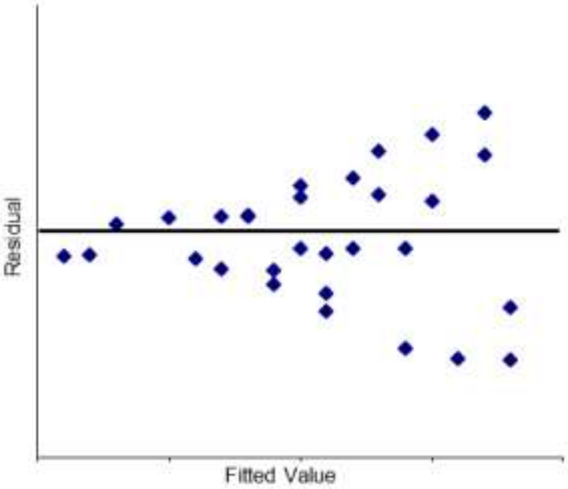
- Plot of **y_i vs. x_i** to assess the assumptions of **linearity** and **constant variance**
- Plot of the residuals vs. the fitted values (**$\hat{e}_i$ vs. $\hat{y}_i$**) to also assess the assumption of **linearity** and **constant variance**
- We study the histogram of the residuals, $\hat{e}_i$, and the normal probability plot of the residuals, $\hat{e}_i$, to test the assumption of **normality**

3 assumptions – 4 plots – 6 conclusions

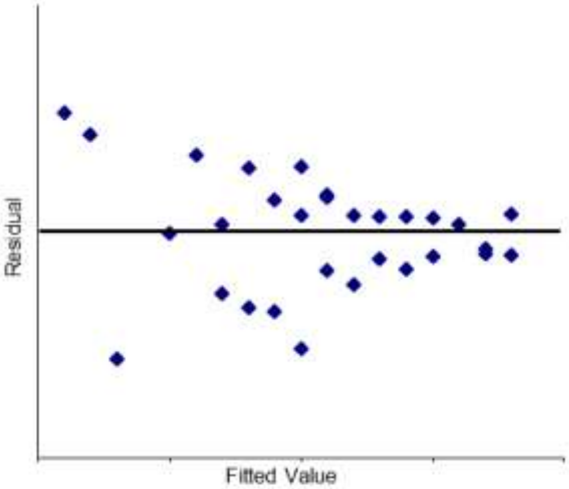
Plots of plot of $\hat{e}_i$ vs. $\hat{y}_i$



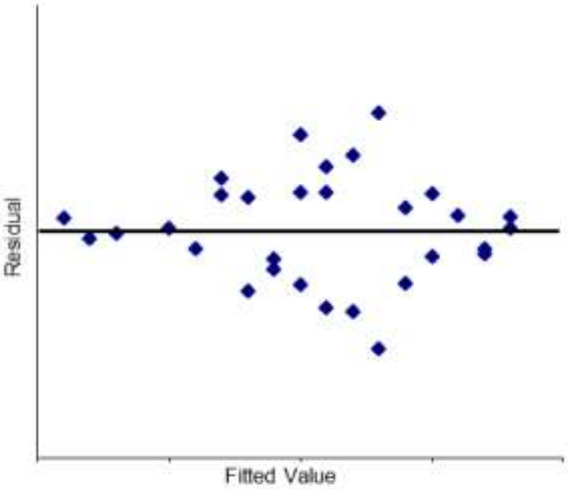
Ideal plot – no pattern



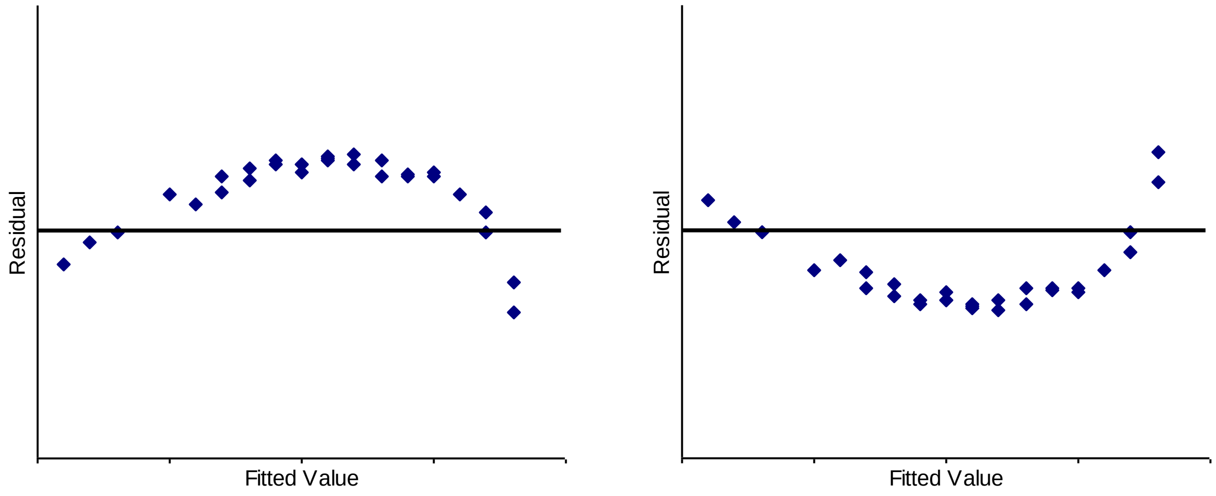
V-shape: Increasing variance
Wide range of Y?
σ proportional to $E(y_i)$?



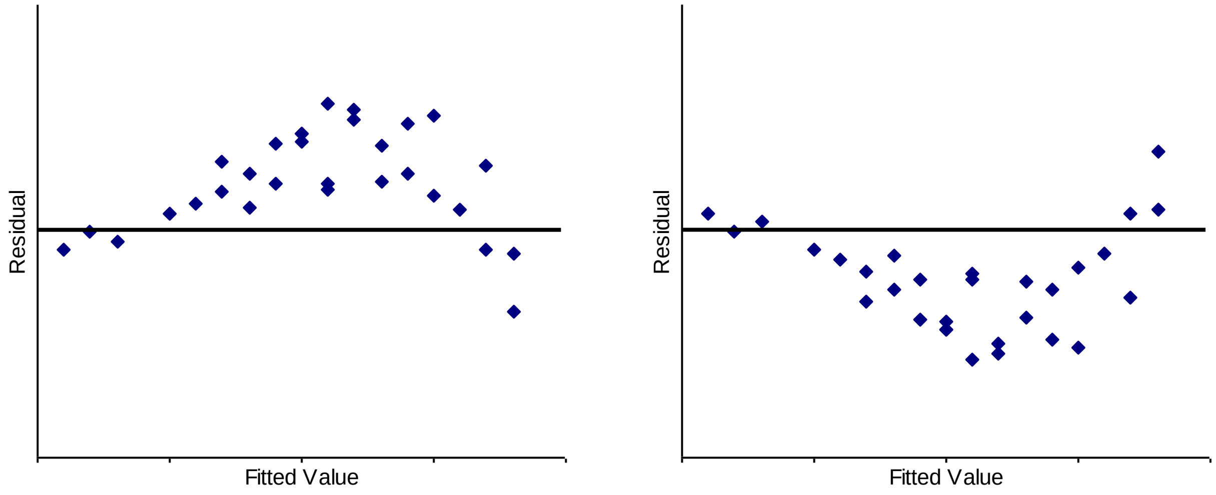
V-shape: Decreasing variances



Diamond shape: Increasing then decreasing variance
Y’s constrained by min and max?
Percentages?



Non-linearity



Non-linearity & non-constant variance

<u>Background Reading</u> on Frequency Distributions and Histograms

Please refer to document “Chapter 4 Supplement” on Canvas.

Normal probability plots

To determine whether a sample of values z_1, z_2, … z_n are a homogeneous sample from some normal distribution, construct a **normal probability plot** as follows:

- Order the z's to get $z_{(1)} \leq z_{(2)} \leq \ldots z_{(n)}$. The $z_{(i)}$'s are called the sample order statistics

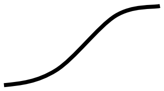
- Calculate $u_{(i)}$, the expected value of the ith sample order statistic from a sample of size n from a **standard** normal distribution ($u_{(i)}$ = ith **rankit**)

- Plot $z_{(i)}$ against $u_{(i)}$.

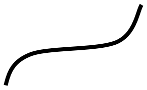
Normal probability plots can be requested in R without the above calculations.

- If the z's are **normal**, then the plot should be a **straight line** – the points should plot on regression line of the $z_{(i)}$ on $u_{(i)}$.

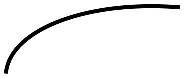
- This shape indicates too few extreme values /tails too short /too little variability:



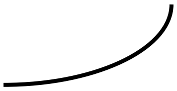
- This shape indicates too many extreme values /tails too long /too much variability:



- This shape indicates a negative skew:



- This shape indicates a positive skew:



There are also formal statistical tests for Normality (e.g. Shapiro-Wilk & Kolmogorov-Smirnov tests). These test tend to be under-power for small samples and over-powered for large samples (*what does power mean here?*).

We will rely on the visual assessment of histogram and the normal probability plot of the residuals.

Note that the Normality assumption is generally regarded as the least important and does not matter for large samples; though confidence intervals may give lower coverage (too narrow) in the case of very non-Normal residuals.

If the regression diagnostic plots indicate that the data fails to satisfy one or more of the model assumptions, then the **aggregate analysis** based on the simple linear model is **invalid**.

Transformations to linearize the model and stabilize the variance

Non-linear:

If the plots indicate that a nonlinear model would be more appropriate, then a quadratic term could be added to the model:

$$y_i = \beta_0 + \beta_1 x_i + \boldsymbol{\beta_2 x_i^2} + e_i$$

Alternatively, we can **transform** *Y*, or *X*, (or both) to linearize the model. For example, suppose the true relationship between the response *Y* and a single predictor *X* is given by

$$Y = \alpha X^{\beta}$$

This relationship is linearized by taking logarithms:

$$\log(Y) = \log(\alpha) + \beta \log(X)$$

The **logarithmic** transformation (log(*Y*), log(*X*)) and the **inverse** transformation
(1/*Y*, 1/*X*) are the most commonly used linearizing transformations

Non-constant variance:

If the plots indicate that the e_i's have non-constant variance, then we can transform the response *Y* via a variance-stabilizing transformation. The most commonly used variance stabilizing transformations are:

the **square root** transformation: $\sqrt{Y}$ (Particularly if Poisson count data)

the **logarithmic** transformation: log(*Y*) (Particularly if σ proportional to $E(y_i)$)

the **inverse** transformation: 1/*Y* (Particularly if Y is time to an event)

Each of these transformations is **more severe** (in stabilizing the variance) than the one before it.

The **square root** transformation is relatively mild and is most appropriate when the y_i's follow a Poisson distribution, usually the first model considered for errors in **counts**.

The **logarithmic** transformation is the **most commonly used** variance stabilizing transformation; the base of the logarithms is irrelevant.
It can only be used if all y_i are strictly **positive**.

If there are any of the y_{i} equal zero, then all y_{i} are replaced by $y_{\mathrm{i}} + 1$ before the logarithmic transformation.

The **inverse** transformation can only be used if all y_i are strictly **positive**.
If there are any of the y_{i} equal zero, then all y_{i} are replaced by $y_{\mathrm{i}} + 1$ before the inverse transformation.

Box and Cox (1964) proposed a systematic approach to the problem of choosing a transformation of the response variable *Y*.
Box and Tidwell (1962) proposed a general method for choosing a transformation of the predictor variable *X*.
We won't study these methods, but instead, we **apply the above transformations** in turn until we arrive at a model satisfying all the model assumptions.

There are potentially 16 transformation/models we could consider:

(X, $\sqrt{X}$, log(*X*) and 1/*X*) combined with (Y, $\sqrt{Y}$, log(*Y*) and 1/*Y*).

However, our judgement and experience will suggest the combinations that are more likely to be successful.

We hope that the **same transformation** will

- linearize the relationship
- stabilize the variance of the residuals
- give normal residuals

These three goals for transformation will not always be met by the same transformation and **compromises** may be required.

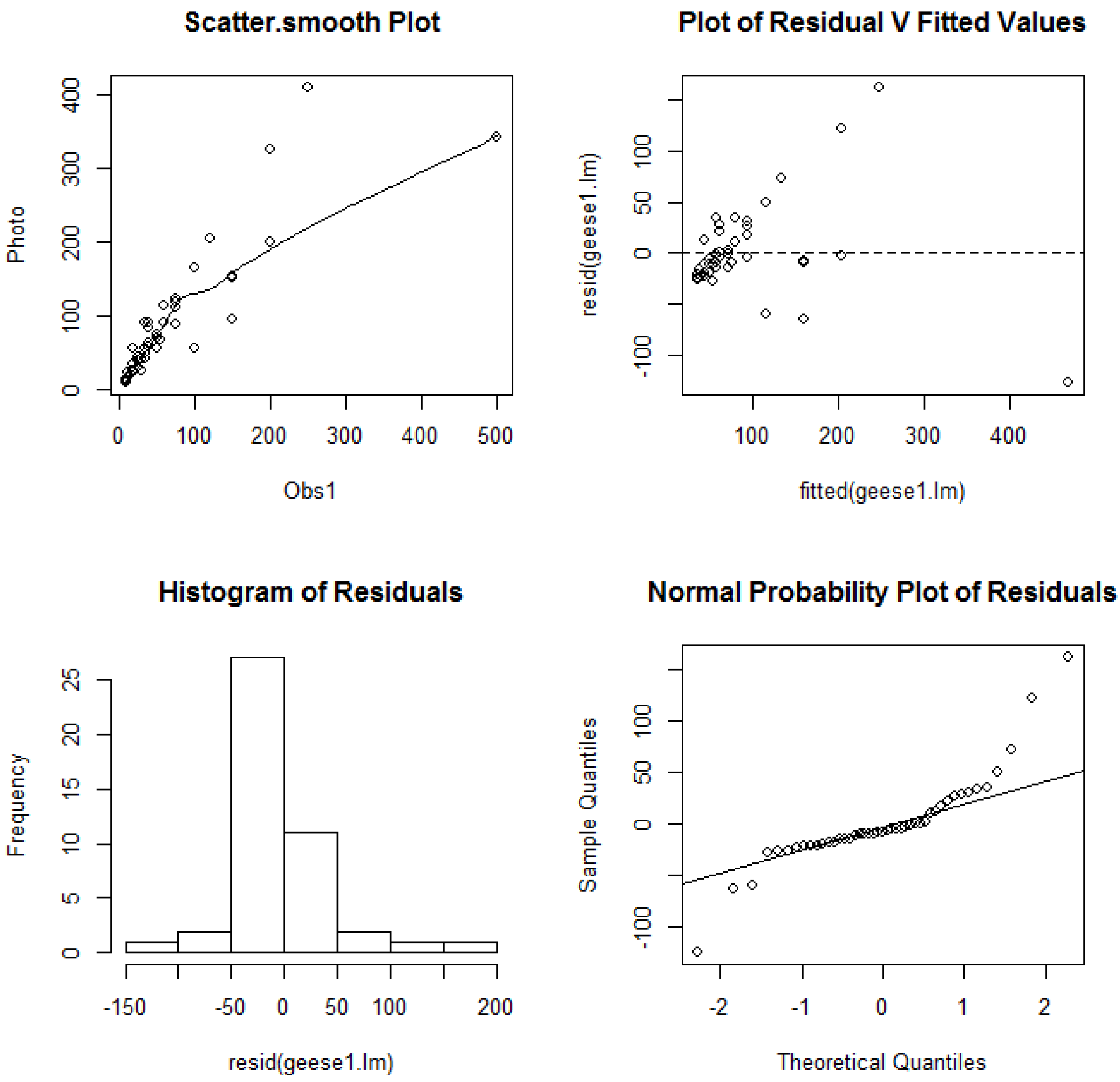
Snow Geese Dataset

45 flocks of geese were photographed and independently counted on site by two observers. The exact count of geese was subsequently determined from the photograph.

We consider the initial model

$$y_i = \beta_0 + \beta_1 x_i + e_i, e_i \sim \mathrm{NID}(0,\sigma^2)$$

where *Y* = Photo and *X* = Obs1. (We omit Obs2 from all models).



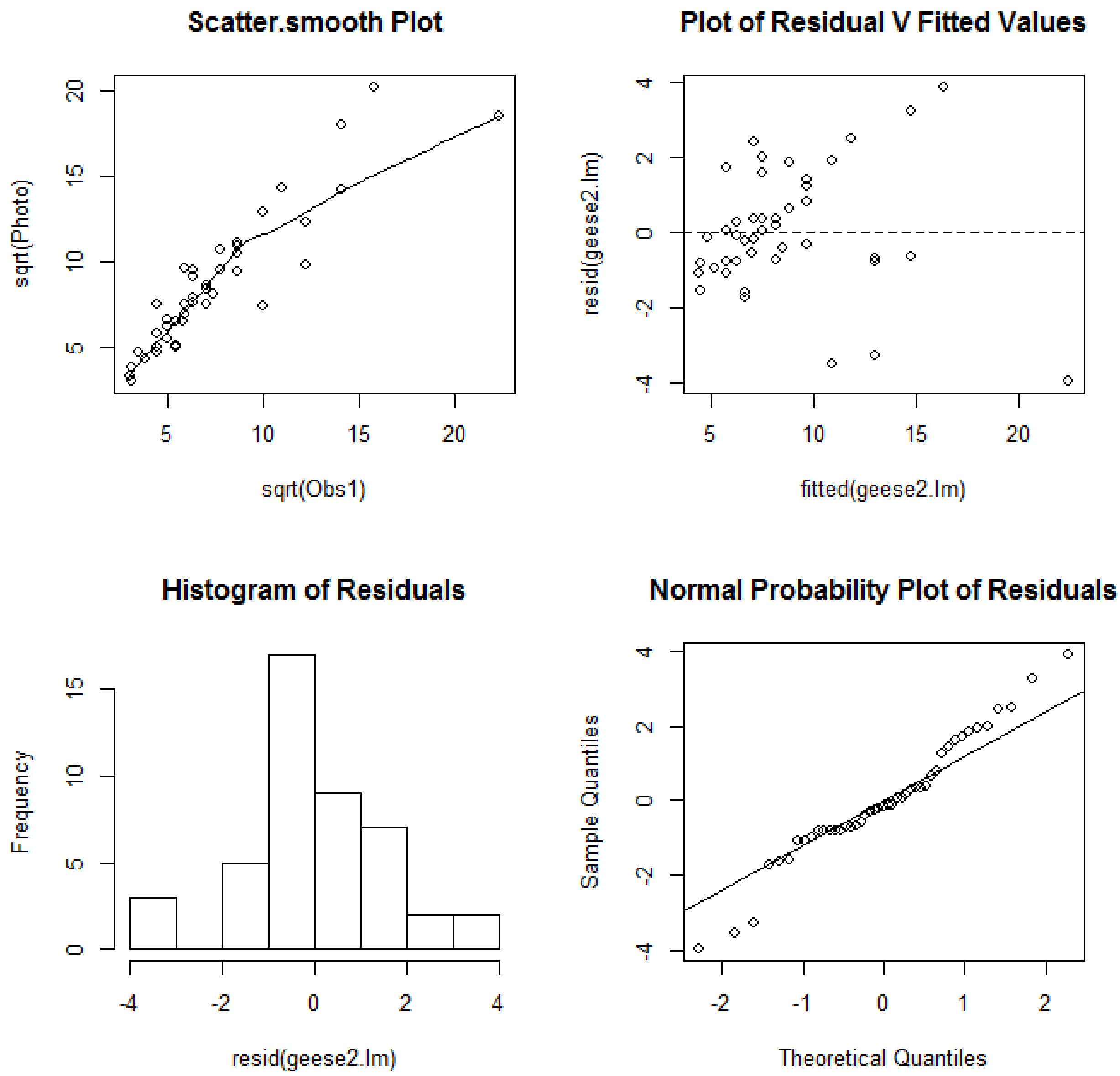
Scatter-plot:	non-linear & variance increasing.
Residuals vs fitted values:	non-linear & variance increasing
Histogram:	non-normal (skewed right)
Normal probability plot:	substantial departures – non-normal

The aggregate analysis for this model is **invalid** since the assumptions of linearity, constant variance and normality are not satisfied.

Because both *X* and *Y* are **counts**, we consider the **square root transformation** for both of these variables. Consequently, the next model considered is

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \text{NID}(0,\sigma^2)$$

where *Y* = sqrt(Photo) and *X* = sqrt(Obs1).



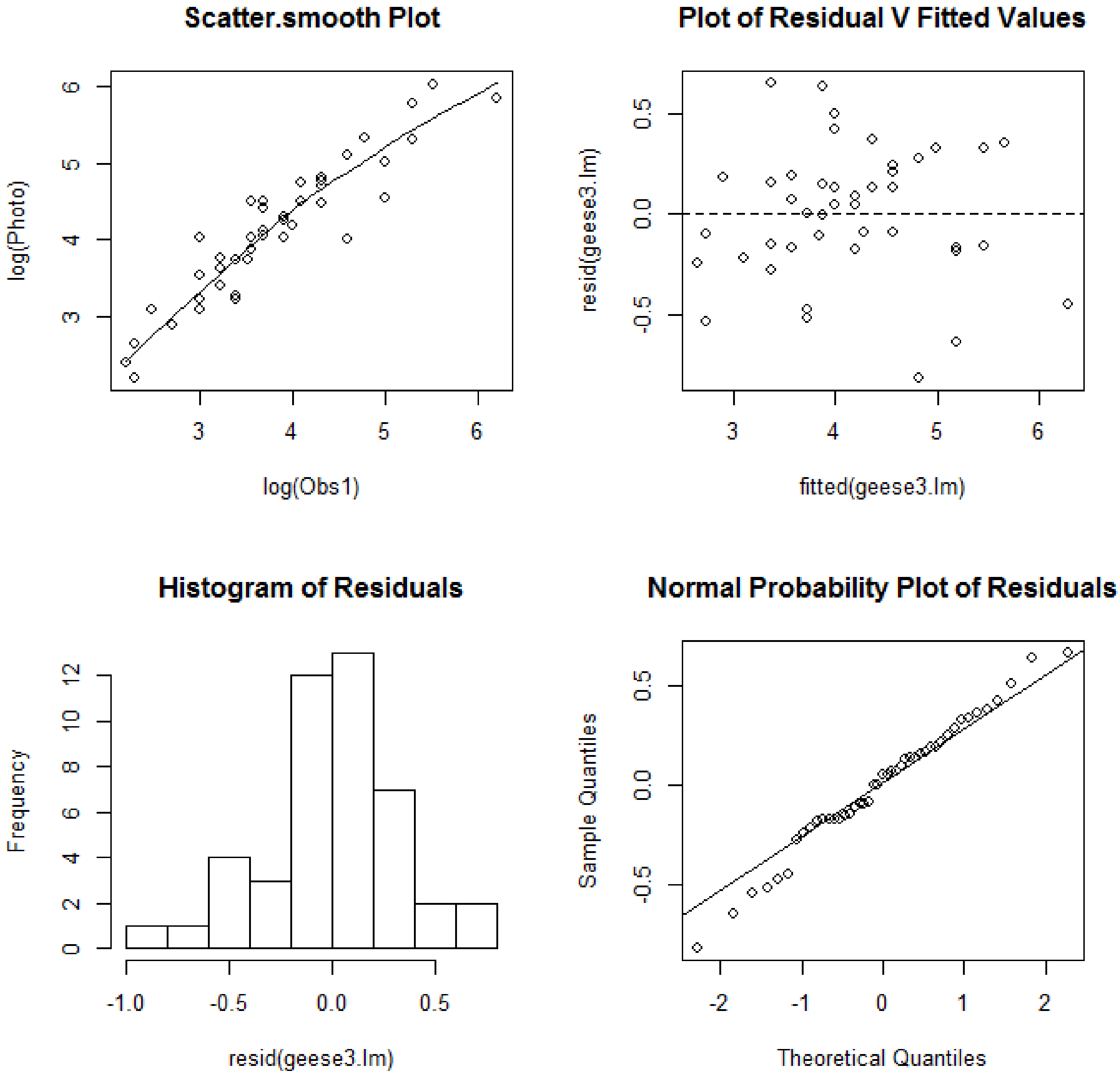
Scatter-plot:	non-linear & variance increasing.
Residuals vs fitted values:	non-linear & variance increasing
Histogram:	approximately normal
Normal probability plot:	several substantial departures – non-normal

This suggests again that, the **stronger** transformations log(Y) and 1/Y should also be assessed.

The next model considered is

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \text{NID}(0,\sigma^2)$$

where Y = log(Photo) and X = log(Obs1).

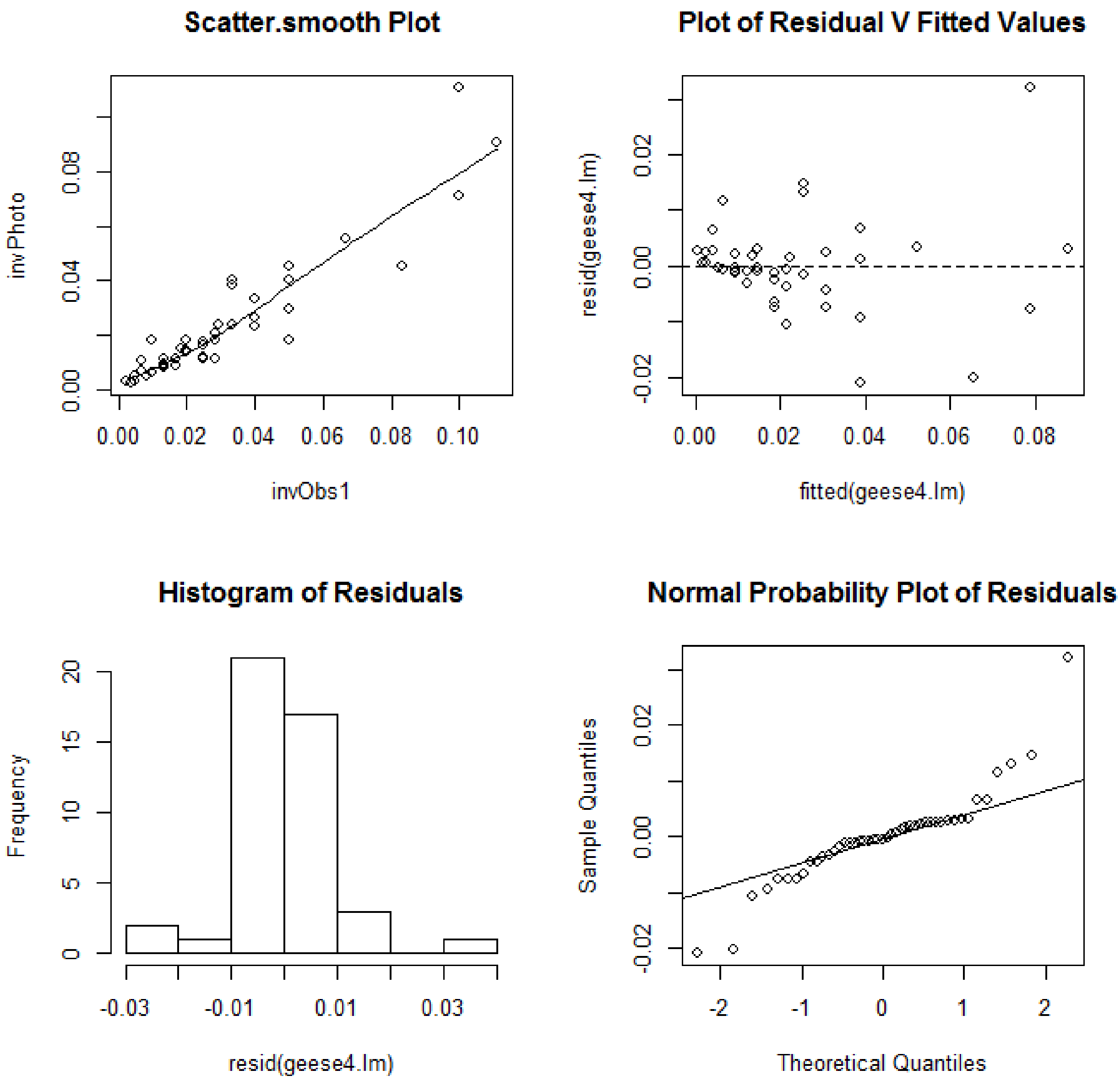


Scatter-plot:	linear & constant variance.
Residuals vs fitted values:	linear & constant variance
Histogram:	non-normal (skewed left)
Normal probability plot:	departures – non-normal

The next model considered is

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \text{NID}(0,\sigma^2)$$

where Y = 1/Photo and X = 1/Obs1.



Scatter-plot:	linear & variance increasing.
Residuals vs fitted values:	non-linear & variance increasing
Histogram:	non-normal (skewed right & too few extreme values)
Normal probability plot:	departures – non-normal

We could continue with other combinations of transformations.
Of the combinations applied so far, which model would you recommend? Why?

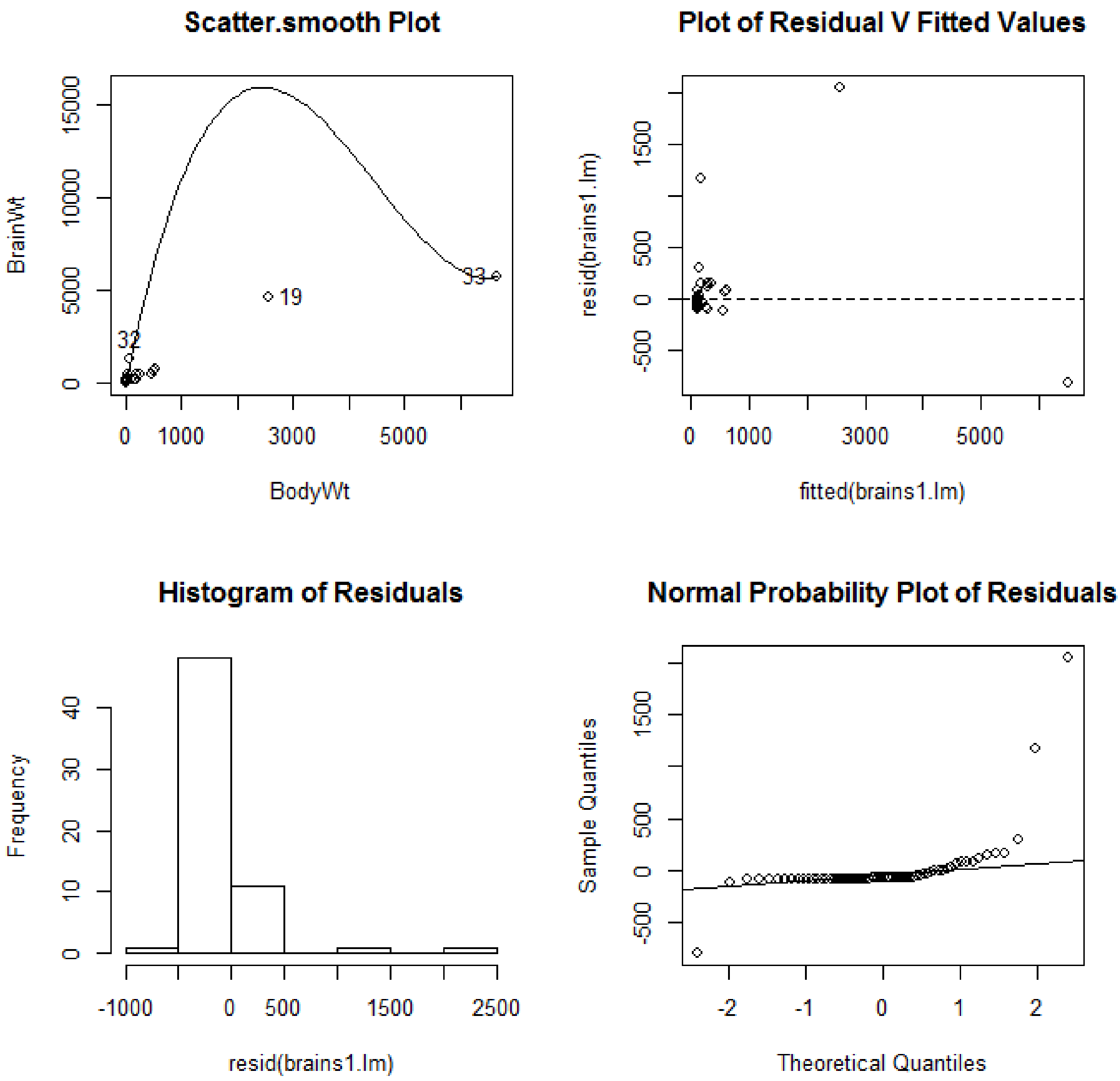
Brains and Body Weights Dataset

The average brain weights (g) and body weights (kg) were calculated for 62 species of mammals. We consider the initial model

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \mathrm{NID}(0,\sigma^2)$$

where Y = BrainWt and X = BodyWt.

It is of interest to see how well the fitted model predicts the brain weight of a human!

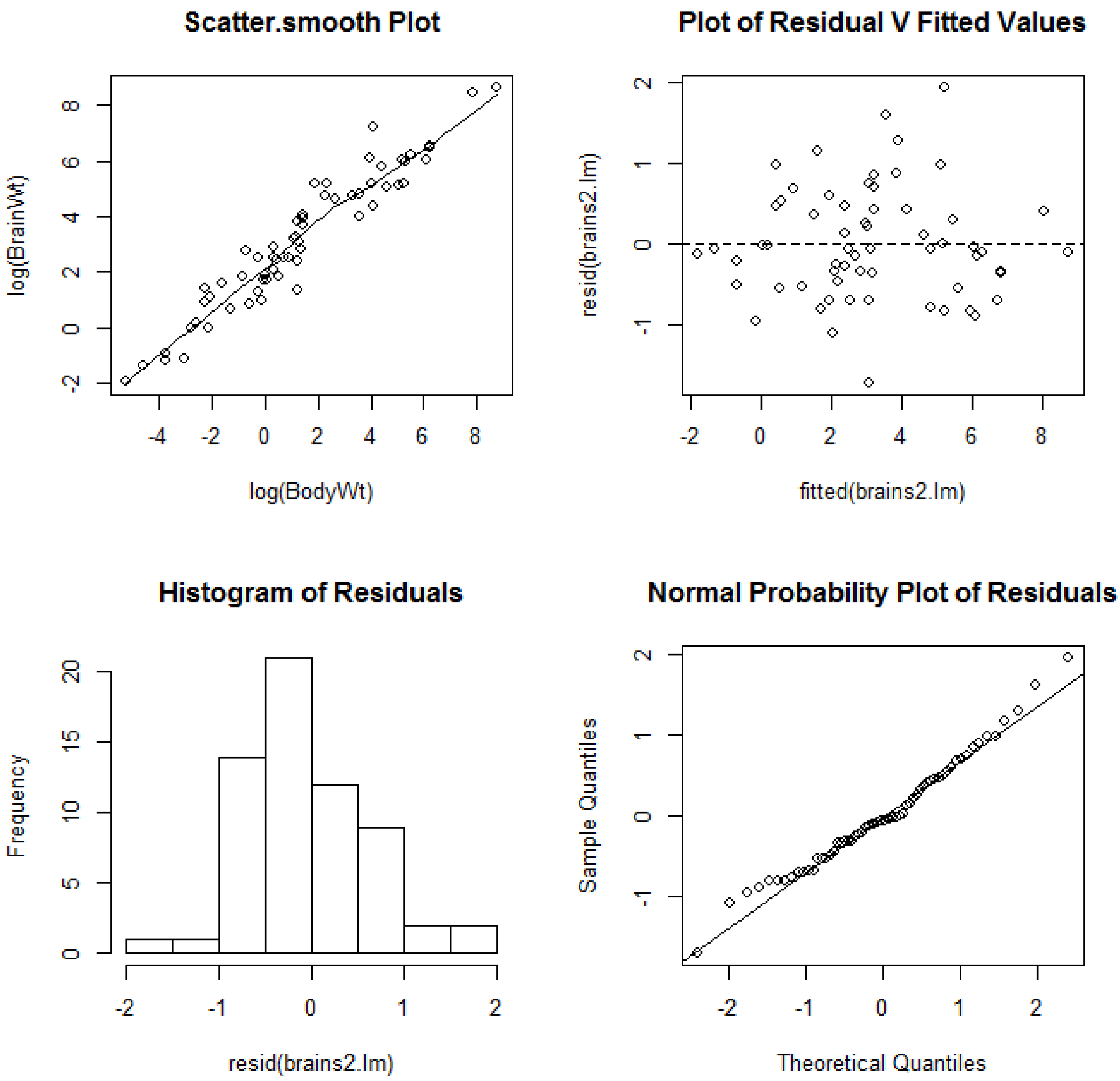


Scatter-plot:	non-linear & inconclusive variance (why?).
Residuals vs fitted values:	non-linear & increasing variance (including the outliers)
Histogram:	non-normal (skewed right)
Normal probability plot:	substantial departures – non-normal

Because of the wide variation in both variables, a logarithmic transformation of both variables is indicated. This leads us to consider the model

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \text{NID}(0,\sigma^2)$$

where $Y =$ log(BrainWt) and $X =$ log(BodyWt).



Scatter-plot:	linear & constant variance.
Residuals vs fitted values:	linear & constant variance
Histogram:	sufficiently normal (despite the slight right skew)
Normal probability plot:	minor departures – approximately normal

The Studentized residual for humans is r_i = 2.848, so that this case is an outlier. Thus humans have a brain weight that is **too large** to be consistent with this model!
Can you identify the data for humans in the graphs above?

The aggregate analysis for this model could now be safely interpreted.

Romanesque Churches

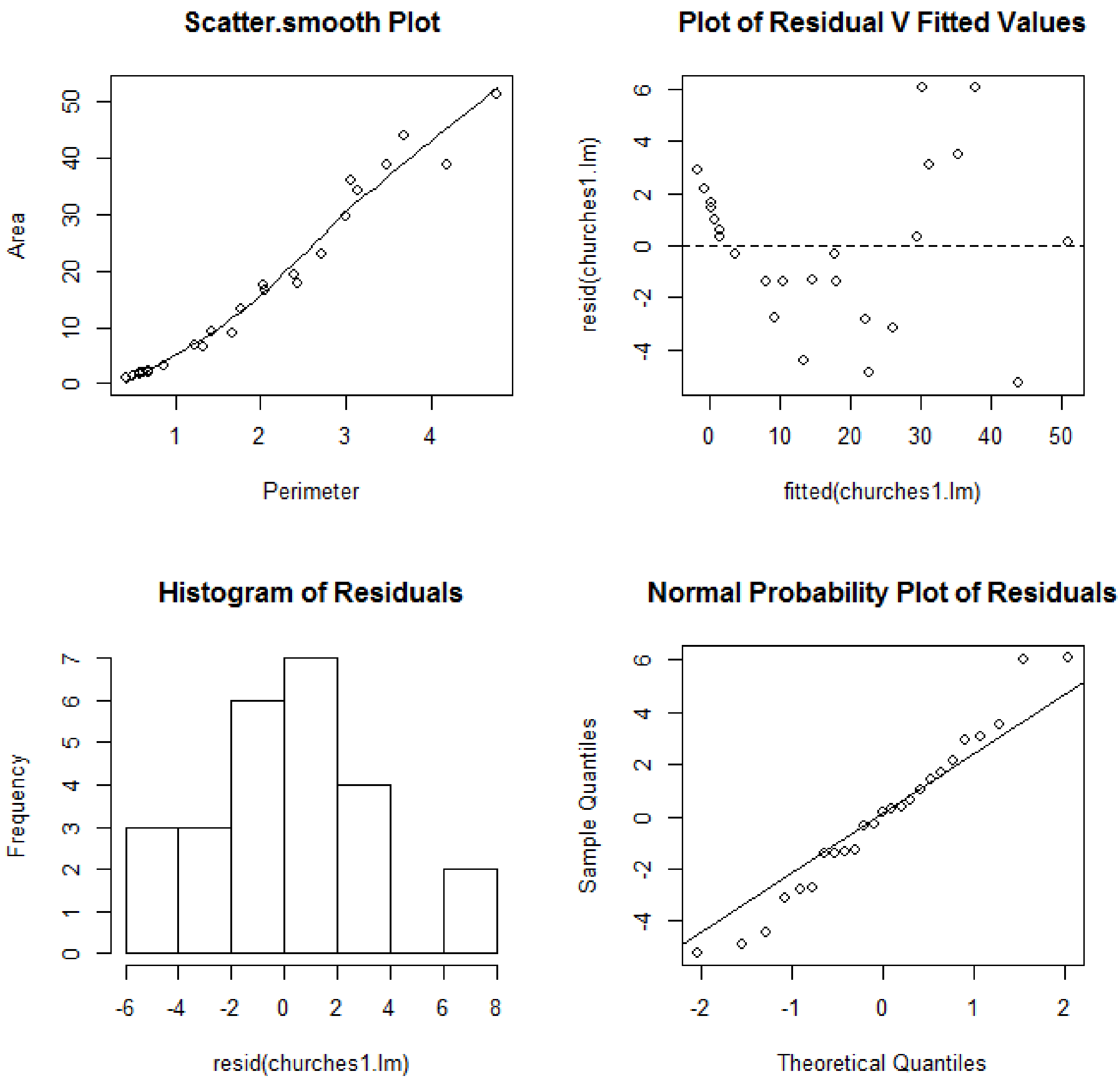
The Perimeter (in hundreds of meters) and the Area (in hundreds of square meters) is recorded for 25 churches. A regression model is to be fitted to predict the floor area from the perimeters of such churches.

If a church was a **square**, then Perimeter = 4(Length) and Area = $(\text{Length})^2$,
so $\sqrt{\text{Area}}$ = Perimeter/4.

This suggests that we may need to use the **square root transformation** of the Area (and/or the Perimeter) to achieve linearity here. We consider the initial model

$$y_i = \beta_0 + \beta_1 x_i + e_i,\ e_i \sim \text{NID}(0,\sigma^2)$$

where Y = Area and X = Perimeter.

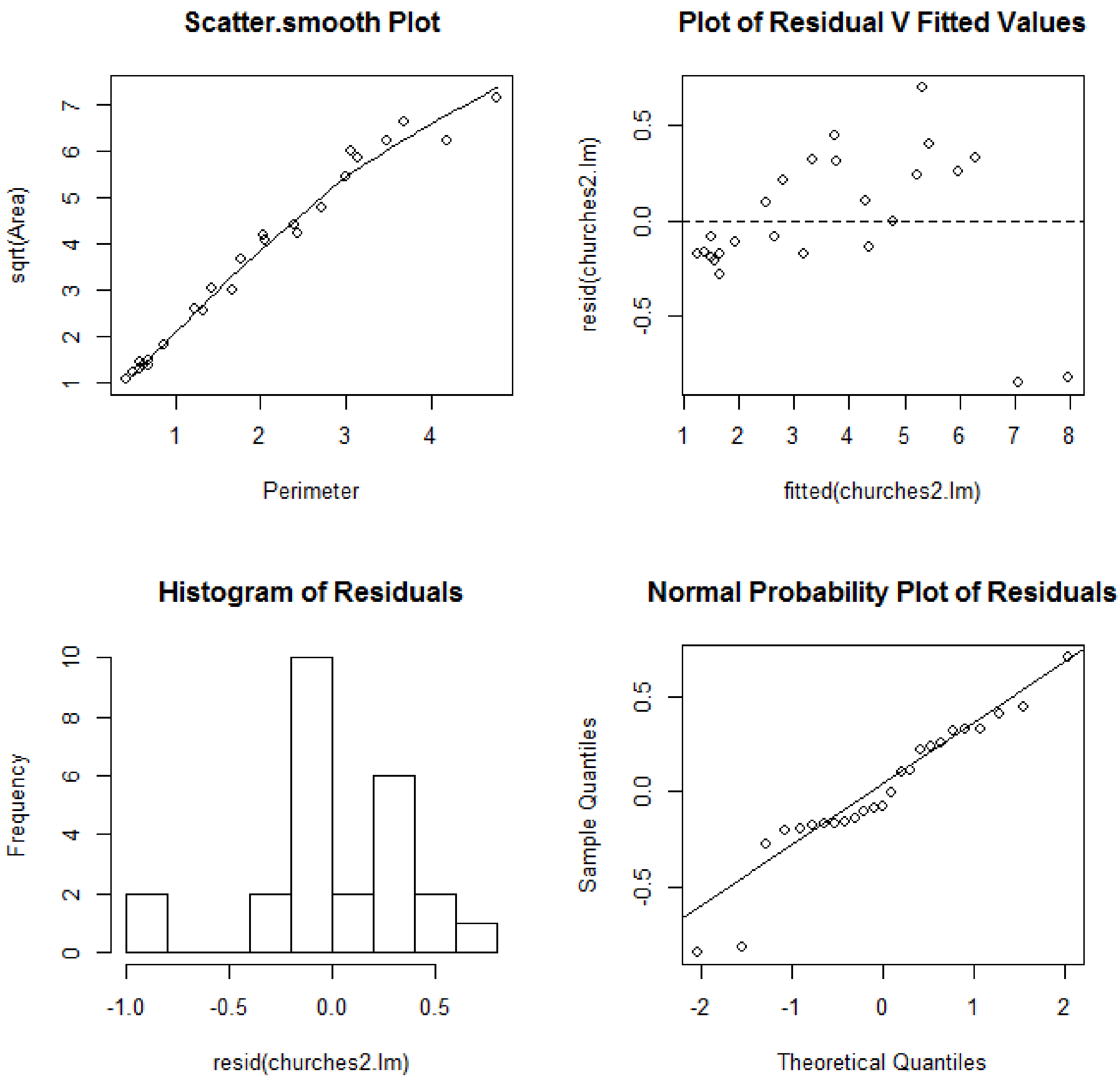


Scatter-plot:	non-linear & increasing variance.
Residuals vs fitted values:	non-linear & increasing variance
Histogram:	non-normal (too few extreme values)
Normal probability plot:	several substantial departures – non-normal

We now consider the model

$$y_i \;=\; \beta_0 \;+\beta_1 x_i + e_i,\; e_i \sim \text{NID}(0,\sigma^2)$$

where Y = sqrt(Area) and X = Perimeter.

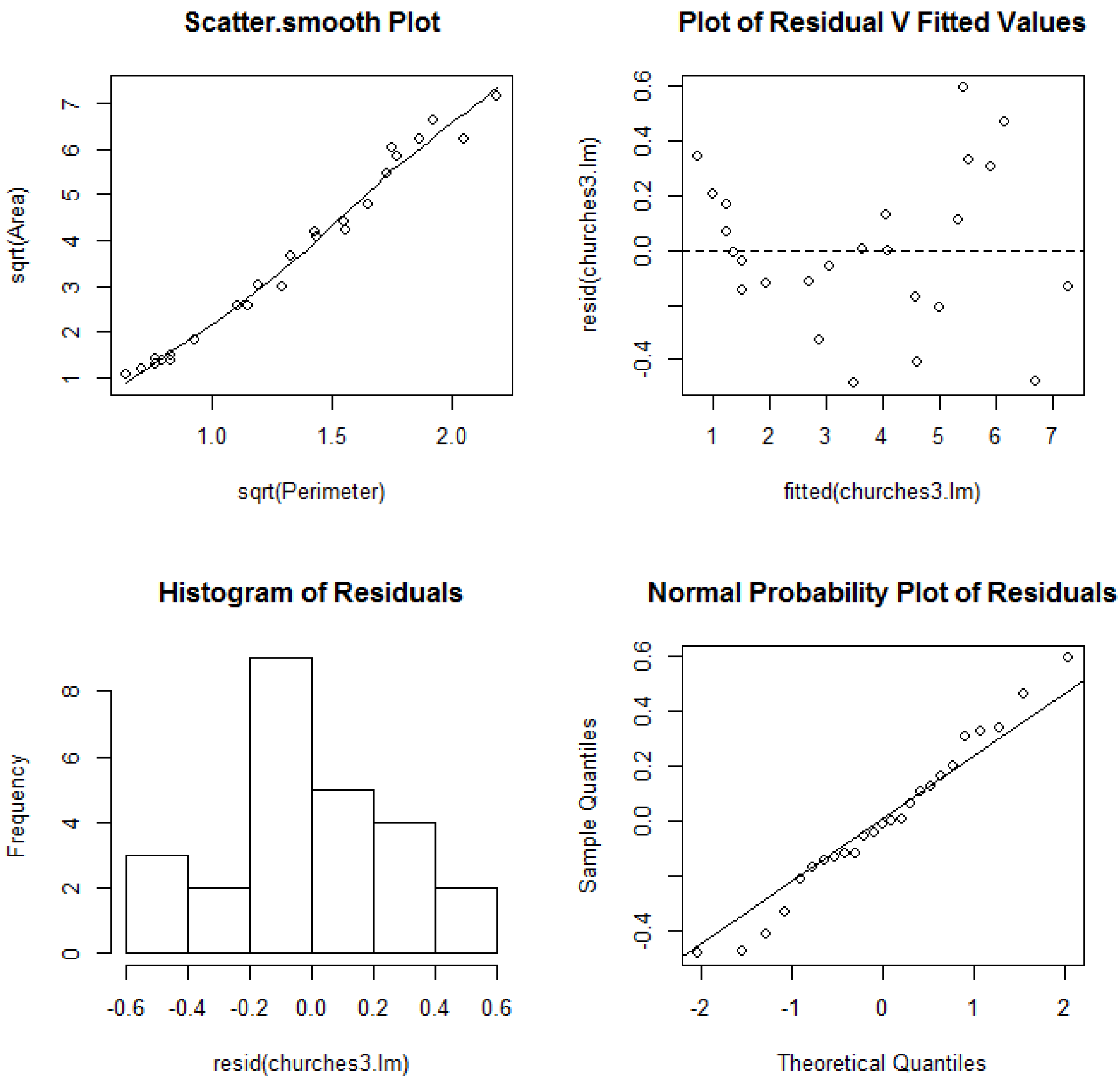


Scatter-plot:	non-linear & constant variance.
Residuals vs fitted values:	non-linear & increasing variance
Histogram:	non-normal (too peaked/ too few extreme values)
Normal probability plot:	substantial departures – non-normal

We now consider the model

$$y_i \;=\; \beta_0 \;+ \beta_1 x_i + e_i,\; e_i \sim \text{NID}(0,\sigma^2)$$

where Y = sqrt(Area) and X = sqrt(Perimeter).

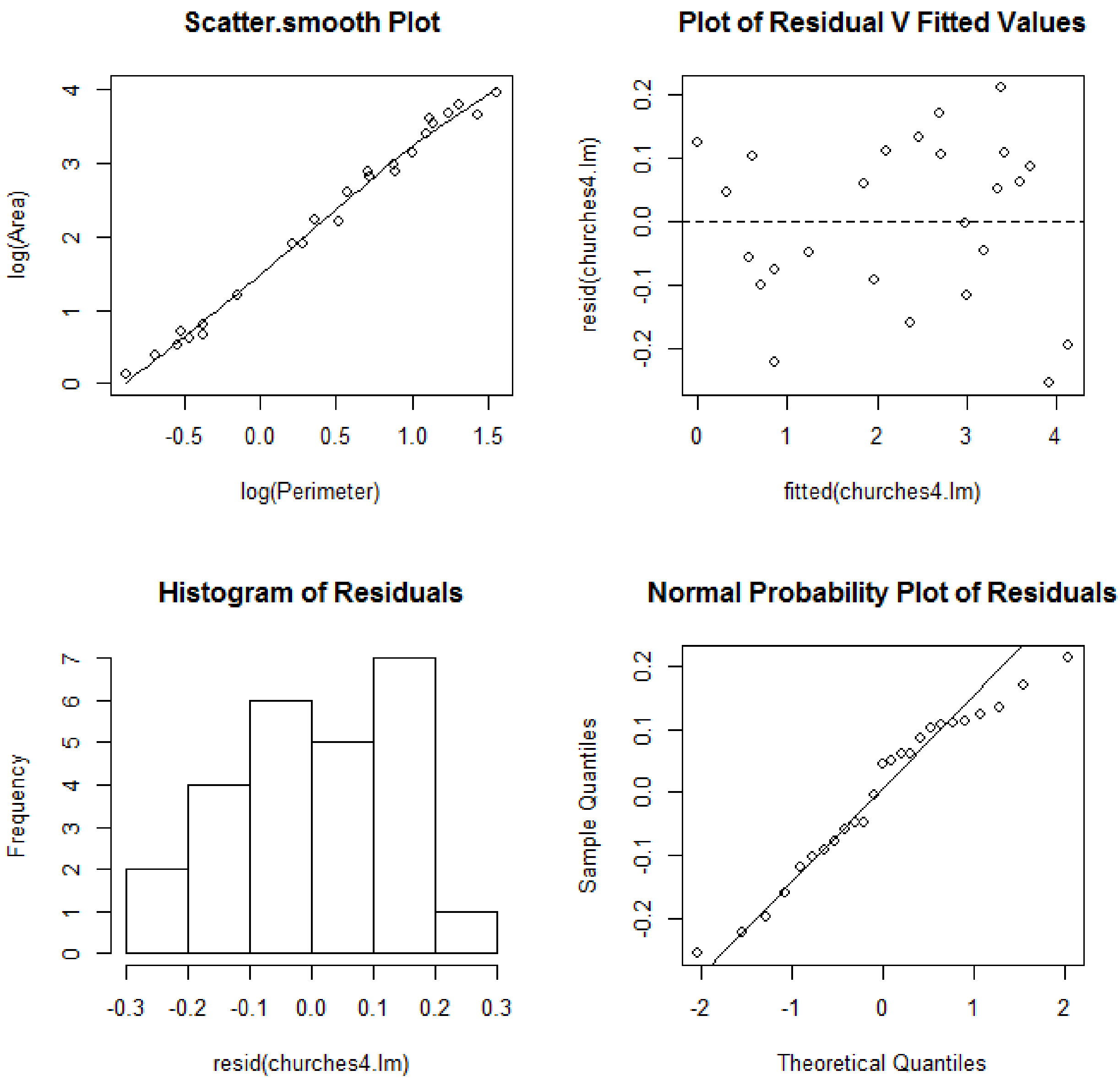


Scatter-plot:	approximately linear & constant variance.
Residuals vs fitted values:	non-linear & increasing variance
Histogram:	approximately normal
Normal probability plot:	moderate departures – approximately-normal

We consider using the next strongest transformation, the logarithmic transformation.

$$y_i = \beta_0 + \beta_1 x_i + e_i, \ e_i \sim \text{NID}(0,\sigma^2)$$

where Y = log(Area) and X = log(Perimeter).



Scatter-plot:	linear & constant variance.
Residuals vs fitted values:	linear & increasing variance
Histogram:	non-normal (slight skewed left)
Normal probability plot:	several moderate departures – non-normal

We consider using the next strongest transformation, the logarithmic transformation.

We could continue with other combinations of transformations.
Of the combinations applied so far, which model would you recommend? Why?

Previous Exam Question 4

For 25 healthy children, data were recorded on the following variables:

Y = Plasma = plasma level of a polyamine

X = Age = age in years

Four different models were fitted to these data:

(a) $Y_i = \beta_0 + \beta_1 X_i + e_i$, $e_i \sim \text{NID}(0, \sigma^2)$, where Y = Plasma and X = Age.

(b) $Y_i = \beta_0 + \beta_1 X_i + e_i$, $e_i \sim \text{NID}(0, \sigma^2)$, where Y = sqrt(Plasma) and X = Age

(c) $Y_i = \beta_0 + \beta_1 X_i + e_i$, $e_i \sim \text{NID}(0,\sigma^2)$, where Y = log(Plasma)and X = Age.

(d) $Y_i = \beta_0 + \beta_1 X_i + e_i$, $e_i \sim \text{NID}(0,\sigma^2)$, where Y = 1/Plasma and X = Age

R plots for these models are shown on the following pages.
Comment on each of the plots for each model.
Which model would you choose for these data and why?

R output for Question 4:

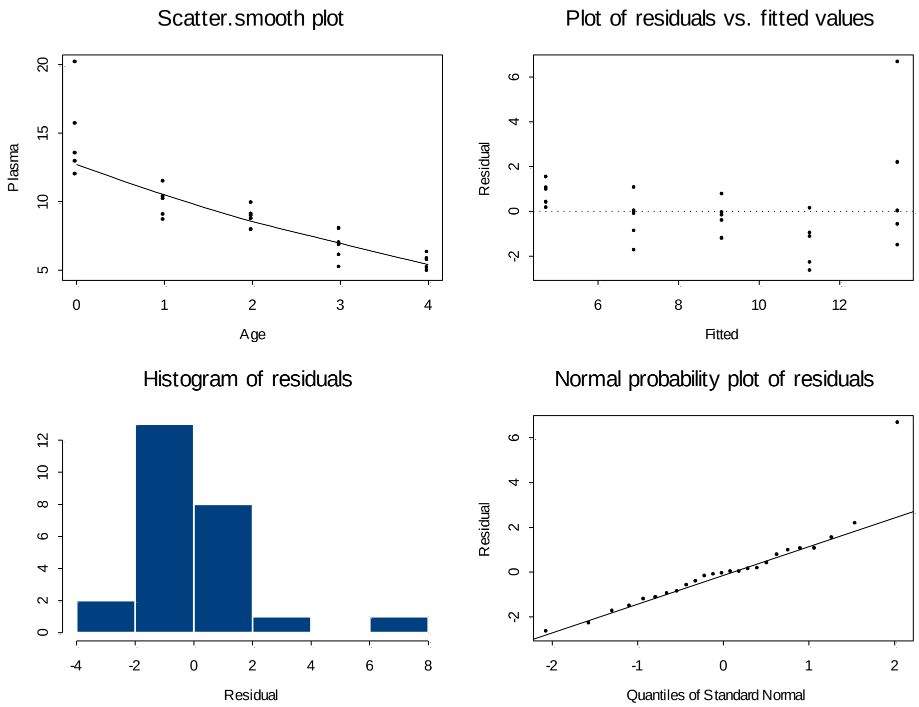


Figure 4.1 Plots for the model of Plasma on Age

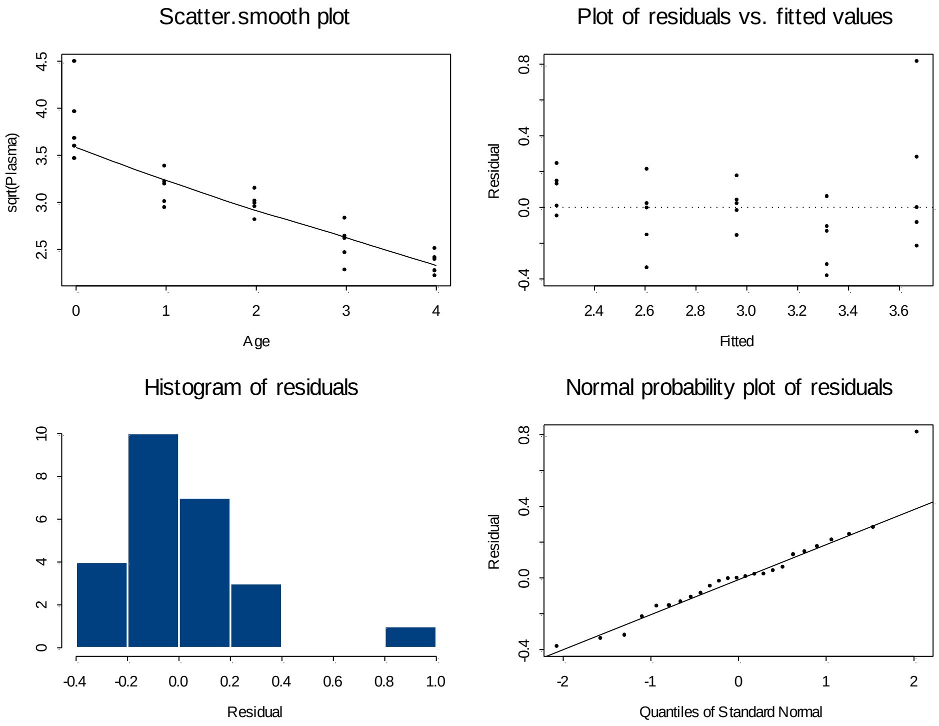


Figure 4.2 Plots for the model of sqrt(Plasma) on Age

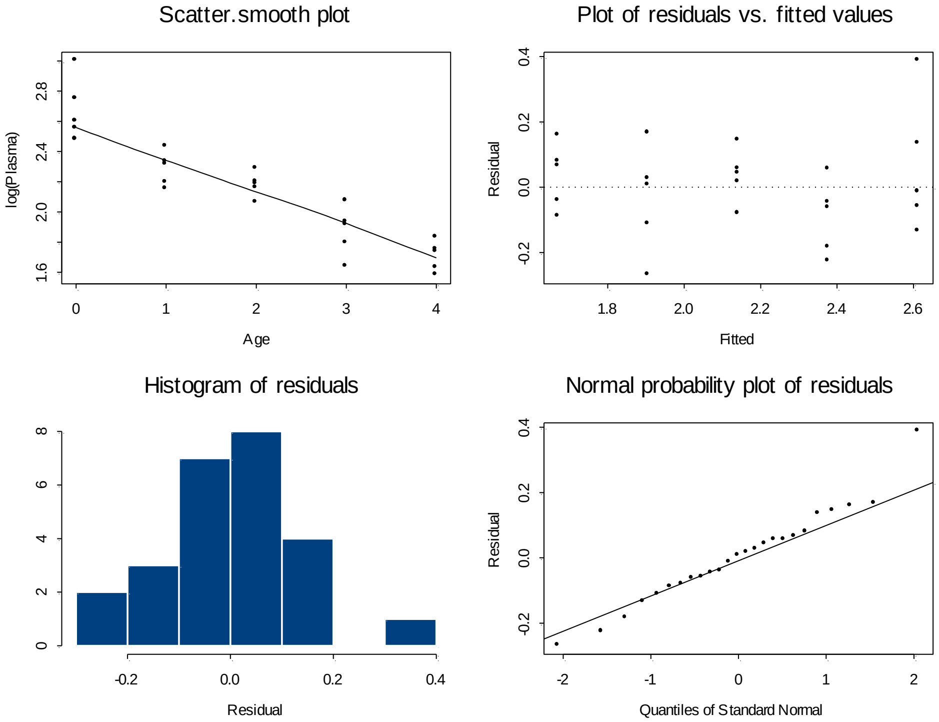


Figure 4.3 Plots for the model of log(Plasma) on Age

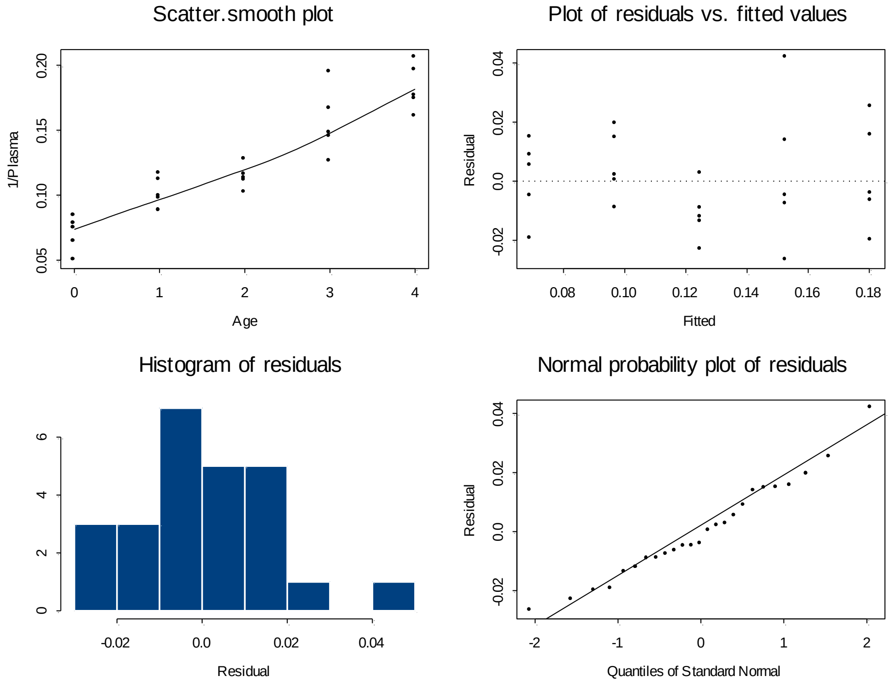


Figure 4.4 Plots for the model of 1/Plasma on Age

Practical (Assignment) 4

Instructions for this practical

- Open the template “Surname Forename Chpt x” from Canvas (in “Practicals”).
- Complete the grid on the first page.
- Save this file (as a Word document) using your own surname, forename and the appropriate chapter number.

- Practice Question:
- Type the commands one by one into R.
- Compare the results in the R text output and graphics with the corresponding results and figures in your notes.
- Use appropriate R output to answer the questions, adapting the R code if necessary.

- Exam Question:
- Adapt the relevant R code you used for the practice question to answer the questions.
- Copy and paste the relevant R text output and graphics into your Word document to support your answers. Change the text font to “Courier New” to align columns.

- Restrict your Word document to a **maximum of 4 pages** (re-sizing graphics and deleting irrelevant R output will help).
- Submit this Word document **via Canvas** by **5.00pm** ____________________ (**STRICT** deadline)
- Note that submitting the practical is a declaration that the practical is your own work. Plagiarism/copying will not be tolerated.

Practice Question (not to be submitted)
Data is studied to determine the relationship between Y = Photo = photo count and X = **Obs2** = count by observer number 2 of flocks of snow geese at a particular location. This data is stored in the data set **geese.txt**.

One of the following models is to be fitted to this data:

(i) $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \mathrm{IN}(0,\sigma^2)$

where Y = Photo and X = Obs2.

(ii) $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \mathrm{IN}(0,\sigma^2)$

where Y = sqrt(Photo) and X = sqrt(Obs2).

(iii) $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \mathrm{IN}(0,\sigma^2)$

where Y = log(Photo) and X = log(Obs2).

(a) Obtain regression diagnostic plots using R to help you decide which model to use. Comment on each of these plots.
(b) Based on your study of these plots, explain which of the above models is preferable for this data.

Exam Question (Winter 2020-21, Question 4) (to be submitted)
Data were collected on a random sample of adults who were undergoing a physical examination. The data are stored in **BMI.txt (on Canvas & P: drive)**.

For a simple linear regression of variable Y = BMI on variable X = Elbow, fit the following three models to the data:

Q4.1.lm: $Y_i = \beta_0 + \beta_1 X_i + e_i$, $e_i \sim \mathrm{NID}(0, \sigma^2)$

Q4.2.lm: $\log(Y_i) = \beta_0 + \beta_1 X_i + e_i$, $e_i \sim \mathrm{NID}(0, \sigma^2)$

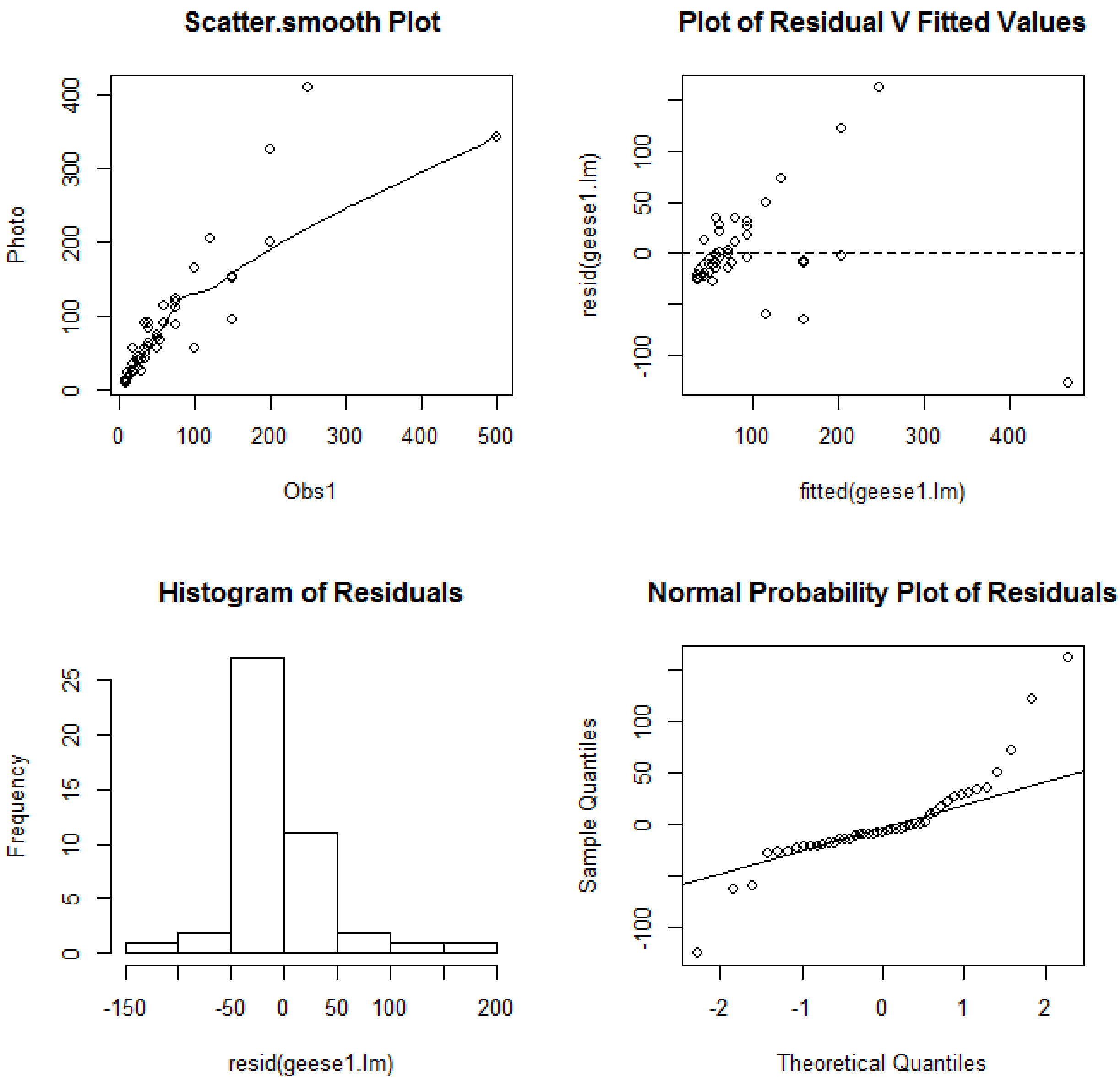
Q4.3.lm: $\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + e_i$, $e_i \sim \mathrm{NID}(0, \sigma^2)$

(a) Provide appropriate diagnostic plots for each model. Comment on each of the diagnostic plots. (3 x 12 marks)

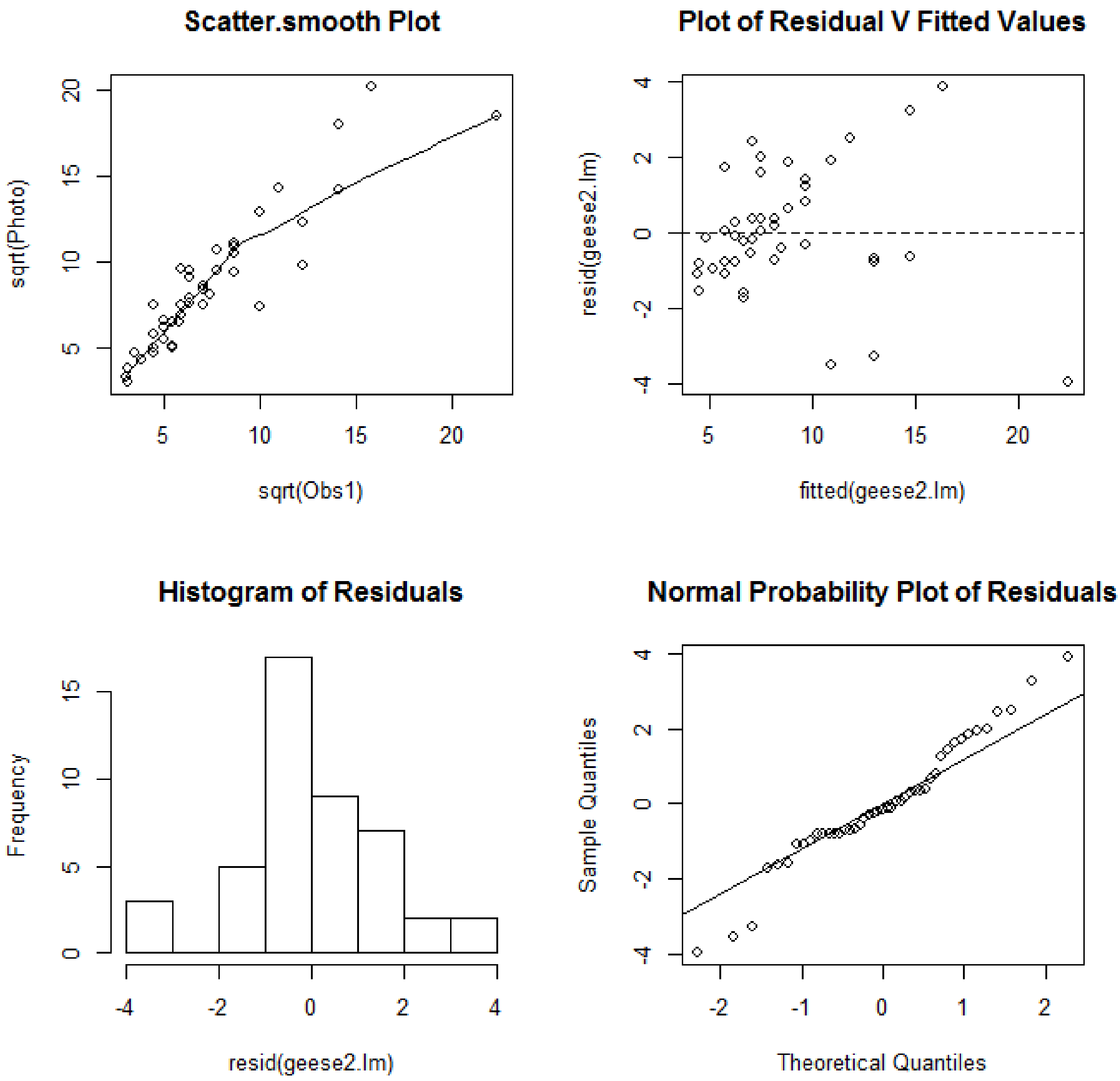
(b) Which of the models would you choose for these data? Explain. (6 marks)

(c) Explain why these particular transformations above (logarithmic) would have been considered. (8 marks)

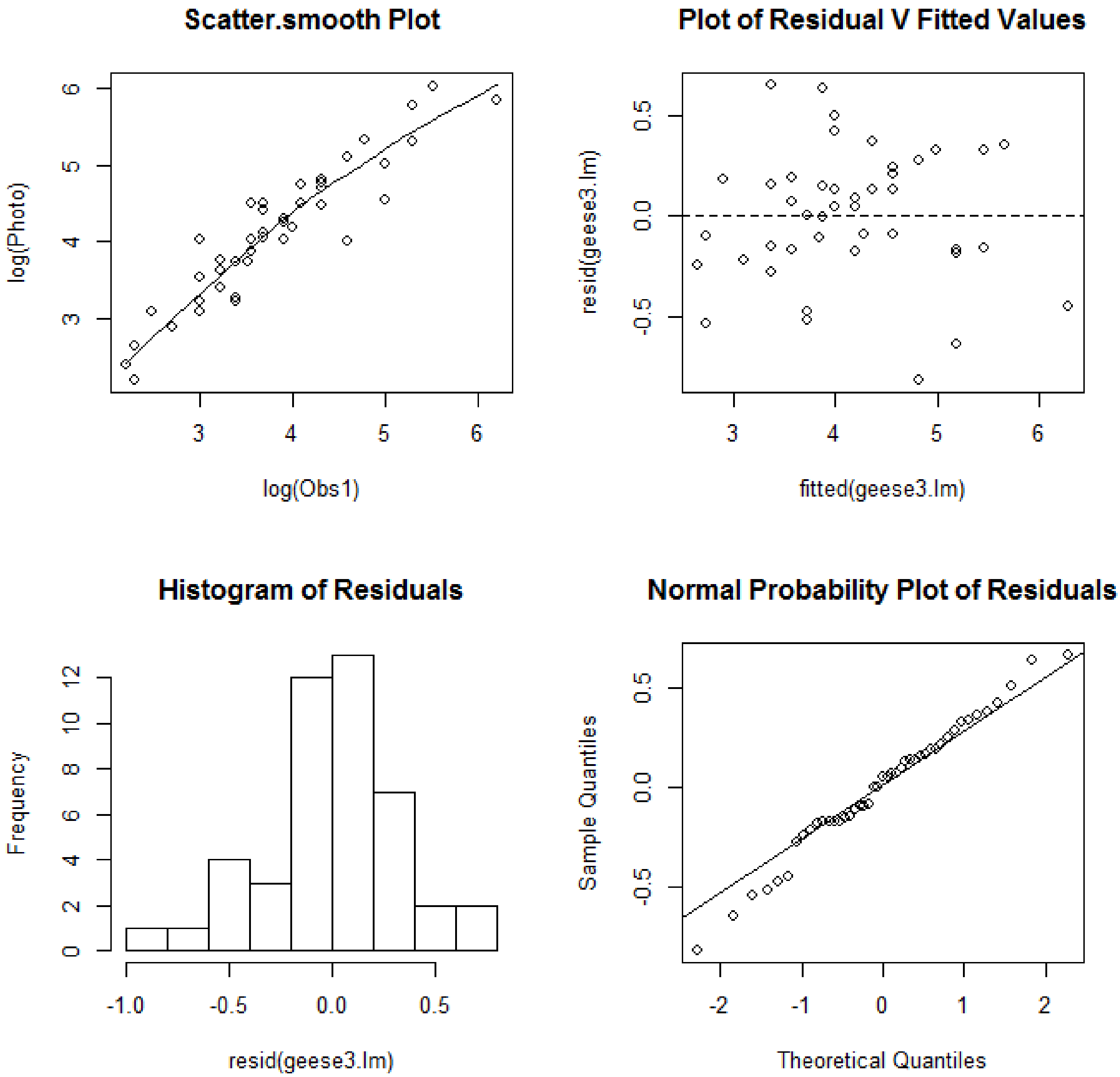
```
> # R code and output for Chapter 4
>
> # Snow Geese data

> geese.df <-read.table("P:\\ST2053\\geese.txt",header=T)
> attach(geese.df)

> # split the graphics window into a 2x2 grid
> par(mfrow=c(2,2))

> # smoothed scatter-plot
> scatter.smooth(Obs1, Photo,main="Scatter.smooth Plot")

> # fit linear model
> geese1.lm<-lm(Photo ~ Obs1)

> # plot of residuals versus fitted values
> plot(fitted(geese1.lm),resid(geese1.lm),main="Plot of
Residual V Fitted Values")
> # horizontal reference line
> abline(h=0,lty=2)

> # histogram of residuals
> hist(resid(geese1.lm),main="Histogram of Residuals")

> # normal probability plot of residuals
> qqnorm(resid(geese1.lm),main="Normal Probability Plot of
Residuals")
> # normal probability plot reference line
> qqline(resid(geese1.lm))

> # return graphics window to default 1x1 grid
> par(mfrow=c(1,1))
```


Scatter.smooth Plot
Photo
0
100
200
300
400
Obs1
0
100
200
300
400
500
Plot of Residual V Fitted Values
resid(geese1.lm)
-100
0
50
100
fitted(geese1.lm)
100
200
300
400
Histogram of Residuals
Frequency
0
5
10
15
20
25
resid(geese1.lm)
-150
-50
0
50
100
200
Normal Probability Plot of Residuals
Sample Quantiles
-100
0
50
100
Theoretical Quantiles
-2
-1
0
1
2

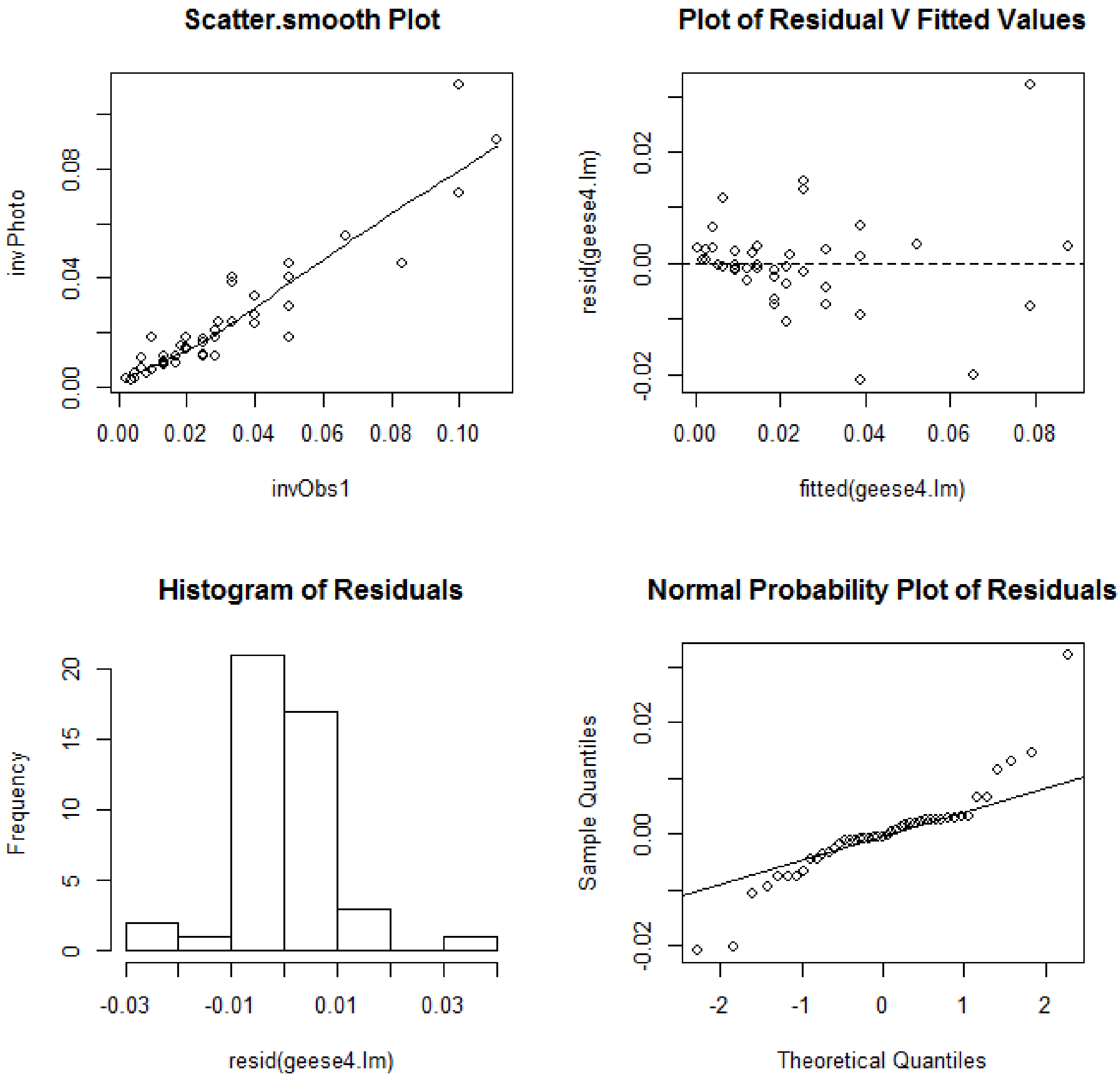
```
> # sqrt(Photo) vs. sqrt(Obs1)
> par(mfrow=c(2,2))
> scatter.smooth(sqrt(Obs1), sqrt(Photo),
main="Scatter.smooth Plot")
> geese2.lm<-lm(sqrt(Photo) ~ sqrt(Obs1))
> plot(fitted(geese2.lm),resid(geese2.lm),main="Plot of
Residual V Fitted Values")
> abline(h=0,lty=2)
> hist(resid(geese2.lm),main="Histogram of Residuals")
> qqnorm(resid(geese2.lm),main="Normal Probability Plot of
Residuals")
> qqline(resid(geese2.lm))
```



```
> # log(Photo) vs. log(Obs1)
> scatter.smooth(log(Obs1), log(Photo),main="Scatter.smooth
Plot")
> geese3.lm<-lm(log(Photo) ~ log(Obs1))
> plot(fitted(geese3.lm),resid(geese3.lm),main="Plot of
Residual V Fitted Values")
> abline(h=0,lty=2)
> hist(resid(geese3.lm),main="Histogram of Residuals")
> qqnorm(resid(geese3.lm),main="Normal Probability Plot of
Residuals")
> qqline(resid(geese3.lm))
```



```
> # 1/Photo vs. 1/Obs)
> invPhoto <- 1/Photo
> invObs1 <- 1/Obs1
> scatter.smooth(invObs1, invPhoto, main="Scatter.smooth
Plot")
> geese4.lm<-lm(invPhoto ~ invObs1)
> plot(fitted(geese4.lm),resid(geese4.lm),main="Plot of
Residual V Fitted Values")
> abline(h=0,lty=2)
> hist(resid(geese4.lm),main="Histogram of Residuals")
> qqnorm(resid(geese4.lm),main="Normal Probability Plot of
Residuals")
> qqline(resid(geese4.lm))
```



Modify the code used above for the Snow Geese data to reproduce the graphs presented in the lectures for the **Brain and Body Weight** data and the **Romanesque Churches** data.

```
> brains.df <-read.table("P:\\ST2053\\brains.txt",header=T)
> attach(brains.df)

> churches.df <
read.table("P:\\ST2053\\churches.txt",header=T)
> attach(churches.df)
```