Chapter 3 Diagnostics and Remedial Measures

Review:

Normal error regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- y_i: observed response in the ith trial
- x_i: a known constant, the level of the predictor variable in the ith trial
- β_0 and β_1: parameters
- $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$
- $E(y_i) = \beta_0 + \beta_1 x_i$, $\text{Var}(y_i) = \sigma^2$

Normal Error in Simple Linear Regression Model

- To do statistical inference, testing hypothesis and to construct confidence interval, we need to make an assumption about the distribution of ϵ in the regression model

- Common assumption: ϵ is a normal distribution

- Why assume normal distribution of errors?

  —Sometimes the errors have approximately normal distributions

  —We get nice methods for statistical inferences

  —If the errors are only approximately normal, the methods developed assuming normality still perform approximately as we would expect

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Possible problems need to check

- Linearity Assumption: Does the linearity assumption between $E(Y)$ and X make sense?
- Error Assumption: Are errors independent, normal random variables with common variance?
- Outlier Detection: Are there outliers i.e a response that is vastly different from other responses?
- Predictor Range: Are there one or more important predictor variables that have been omitted from the model?

Section 3.2-3.3: Diagnostics for Residuals

Properties of Residuals

- $\sum_{i=1}^{n} e_i = 0$, so $\bar{e} = 0$
- $\sum_{i=1}^{n} e_i^2/(n-2) = MSE$, if the model is appropriate, MSE is an unbiased estimator of the variance σ^2
- $\sum_{i=1}^{n} x_i e_i = 0$, x_is and e_is are uncorrelated (independent under normality)
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$, $\hat{y}_i$s and e_is are uncorrelated (independent under normality)
- Residuals are not independent
- Generally, the correlation of residuals is small and ignored

Type of Residuals

- Raw residual: $e_i = y_i - \hat{y}_i$
- Semi-studentized residual: $\frac{e_i}{\sqrt{MSE}}, (i = 1, \cdots n)$, MSE is an approximation of the standard deviation of e_i
- Standardized Residual: $\frac{e_i}{s(e_i)}, i = 1, 2, \cdots n$

  where $s(e_i) = \sqrt{MSE(1 - h_{ii})}$, h_{ii} is the diagonal element of hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\hat{\mathbf{Y}} = \mathbf{HY}$
- Studentized residual: $\frac{e_i}{s_{(i)}(e_i)}, i = 1, 2, \cdots n$

  ——If the residual is standardized with an independent estimate of σ^2, the result has a Student's t distribution if the data satisfy the normality assumption

  ——We estimate σ^2 by $s^2_{(i)}$, the estimate of σ^2 obtained after deleting the ith observation, the result is a studentized residual

Graphical Analysis of Residuals

1. Assessing nonlinearity of Regression Functions

Plot Y v.s X

Plot e_i v.s x_i or e_i v.s $\hat{y}_i$

Figure 1: Plot of y v.s x, data were generated from $y = x * x + 10 * x + 30 + N(0, 25)$

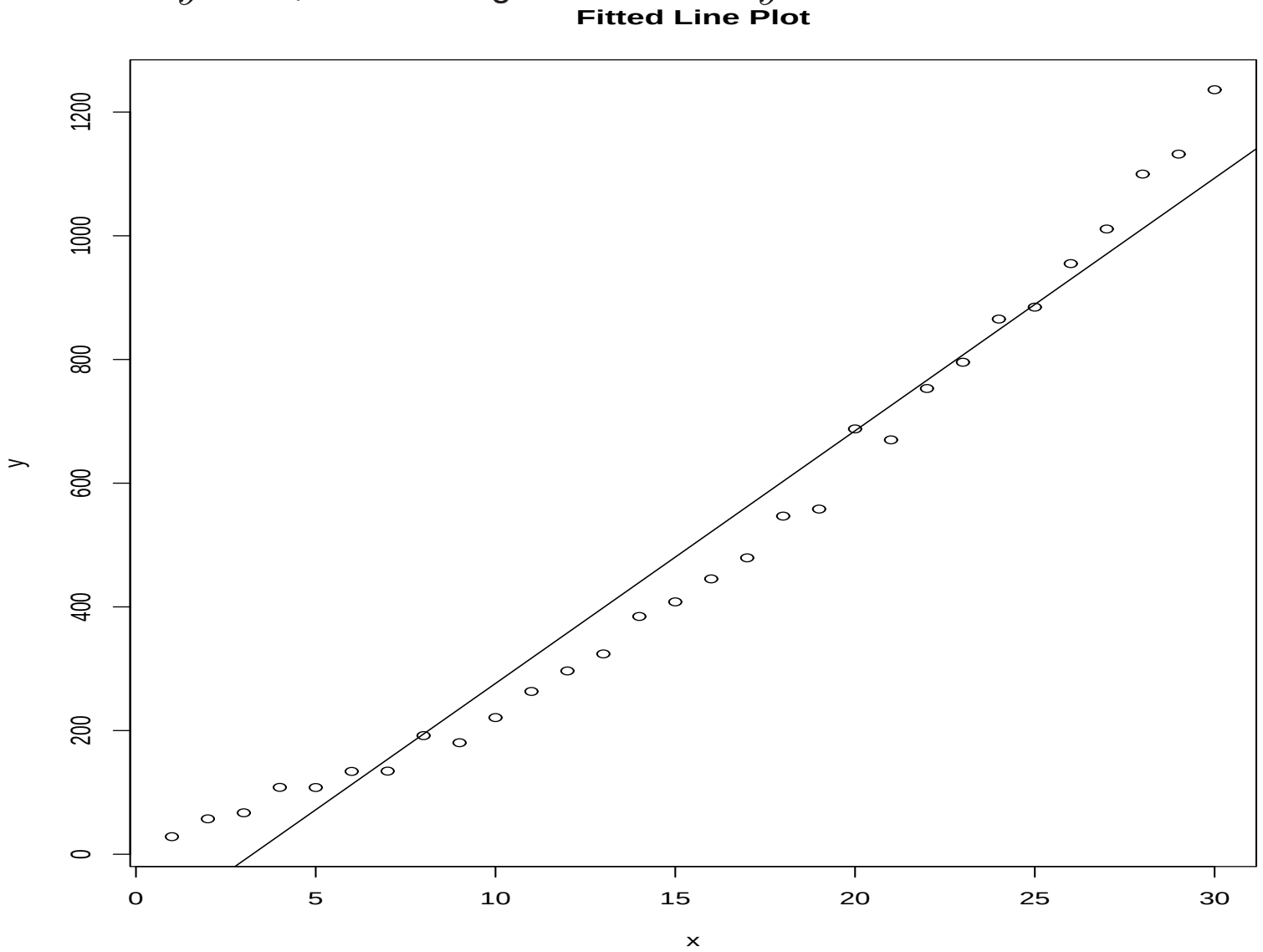
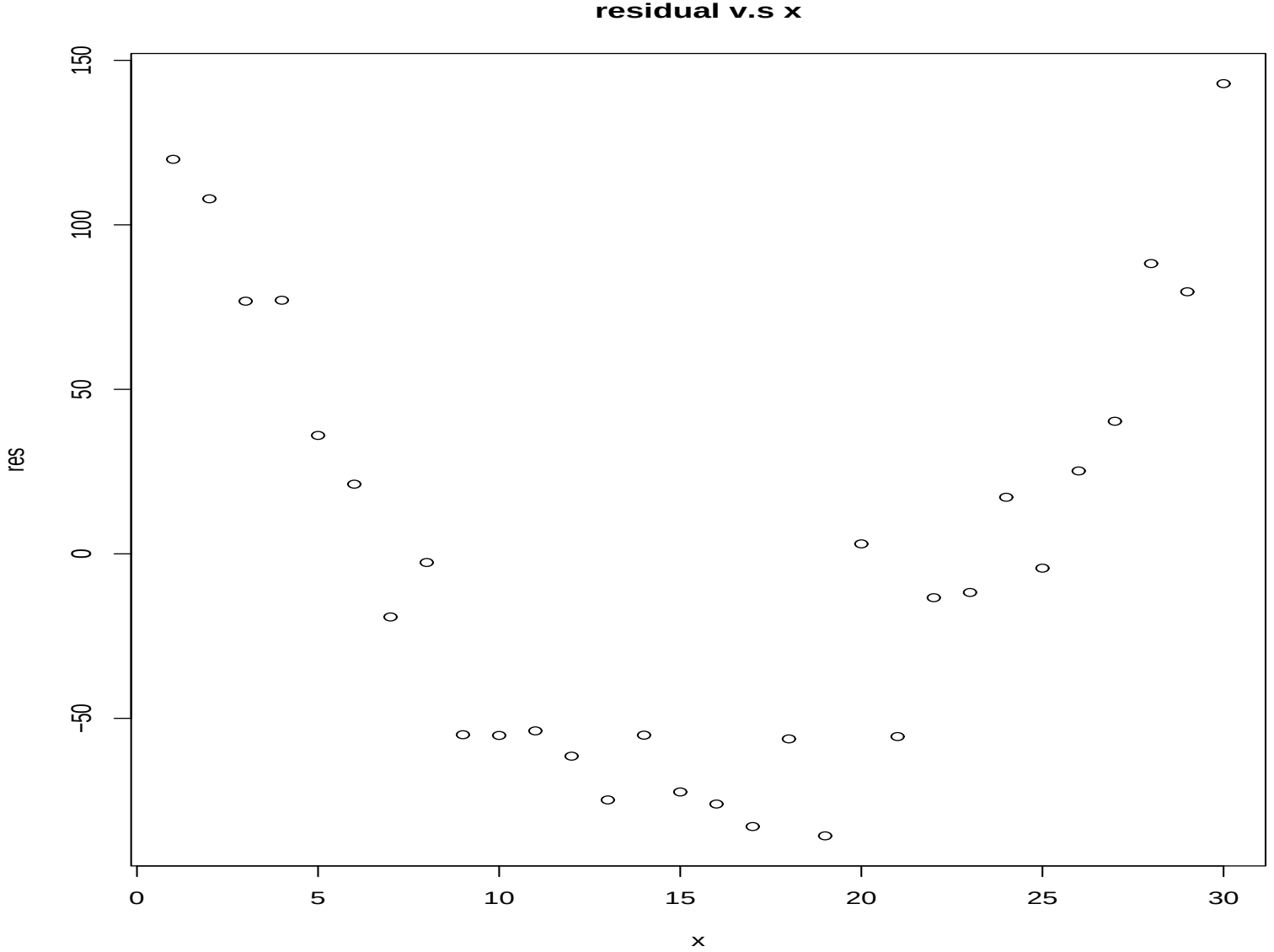


Figure 2: Plot of residual v.s x,data were generated from $y = x * x + 10 * x + 30 + N(0, 25)$



- If the linear regression function is not appropriate, then the residual plots might show a trend
- If the linear regression function is appropriate, there should be no obvious trend in the residual plots

Figure 3: Plot of y v.s x, data were generated from $y = 10 * x + 30 + N(0, 25)$

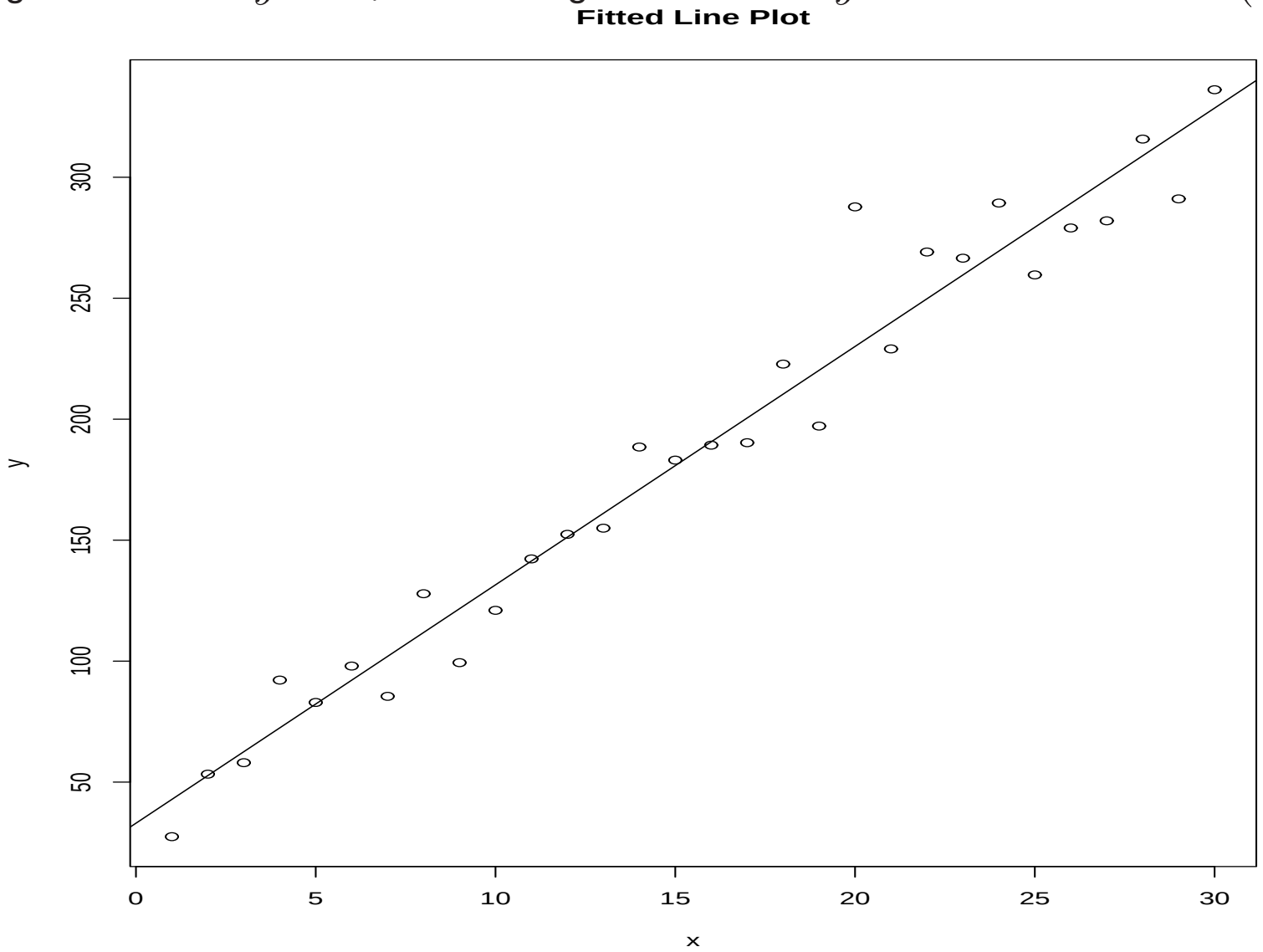
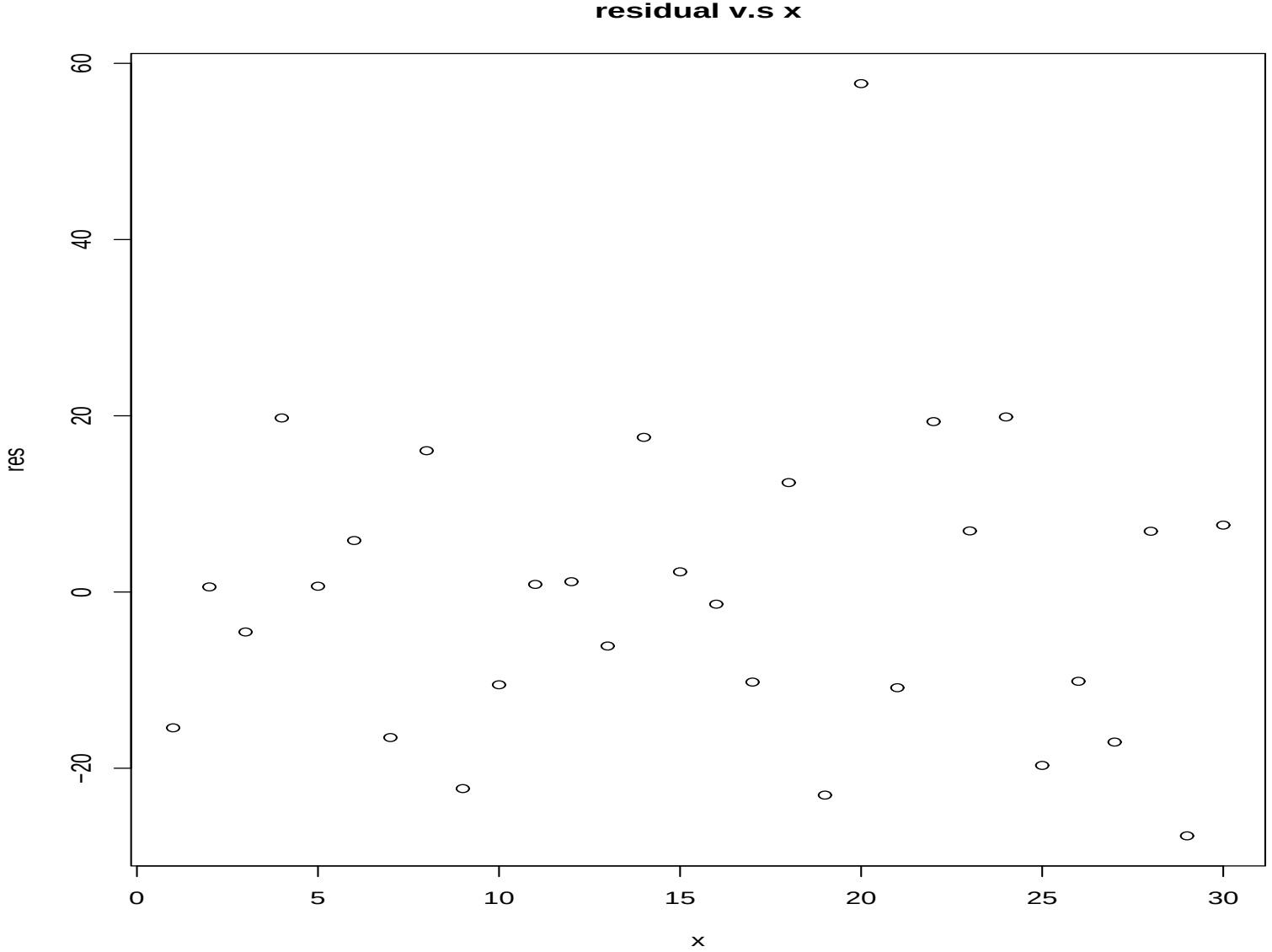


Figure 4: Plot of residual v.s x,data were generated from $y = 10 * x + 30 + N(0, 25)$



- Comments:

  (1) For simple linear regression, plotting e_i v.s x_i gives essentially the same graph as e_i v.s $\hat{y}_i$ since $\hat{y}_i = b_0 + b_1 x_i$

  (2) $\sum_{i=1}^{n} x_i e_i = 0$

  (3) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

  (4) Don't plot e_i v.s y_i since they are correlated. The plot of e_i v.s y_i will show a trend.

2. Checking for non-constant error variance

Plot e_i v.s $\hat{y}_i$ or e_i v.s x_i

- If error variance is constant, the plot should show no trends, but a random scatter of points around the horizontal $e = 0$ axis.
- If there is a problem with the constancy of variance assumption, e_i's will show some trend.

- Example of non-constant variance:

  $Y = 30 + 10 * X + N(0, 10X^2)$, variance changes with X

Figure 5: Example of nonconstant variance

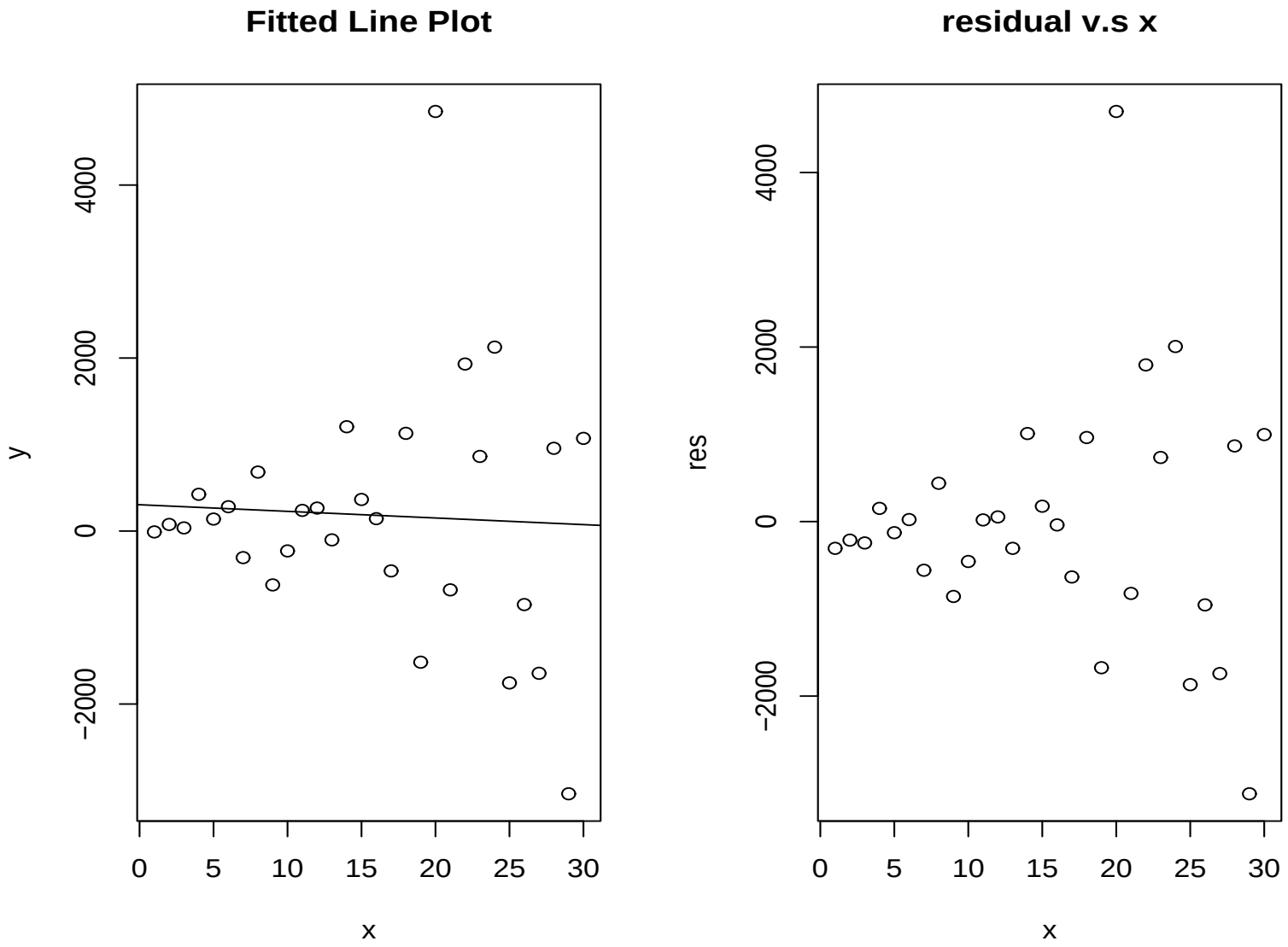
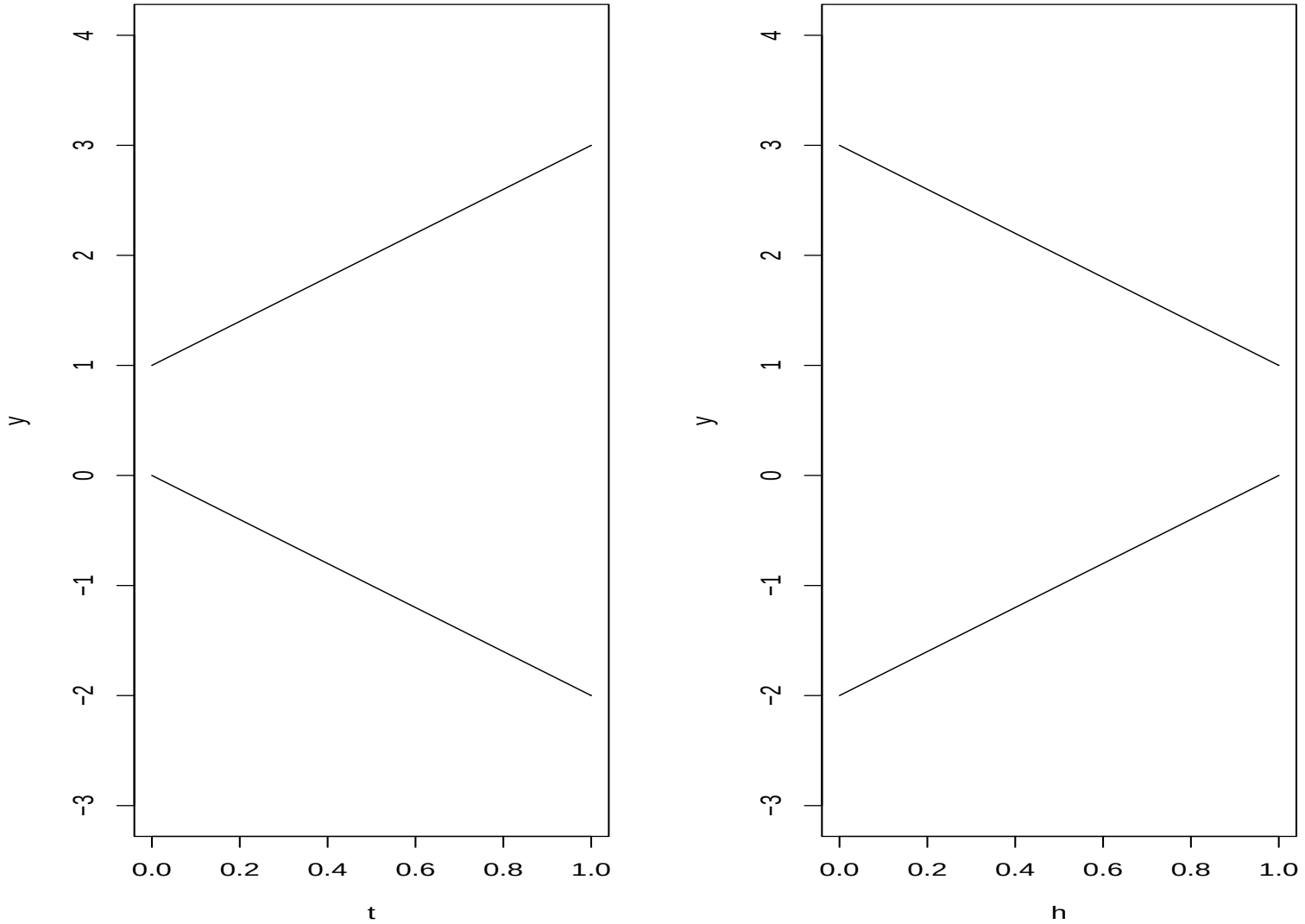


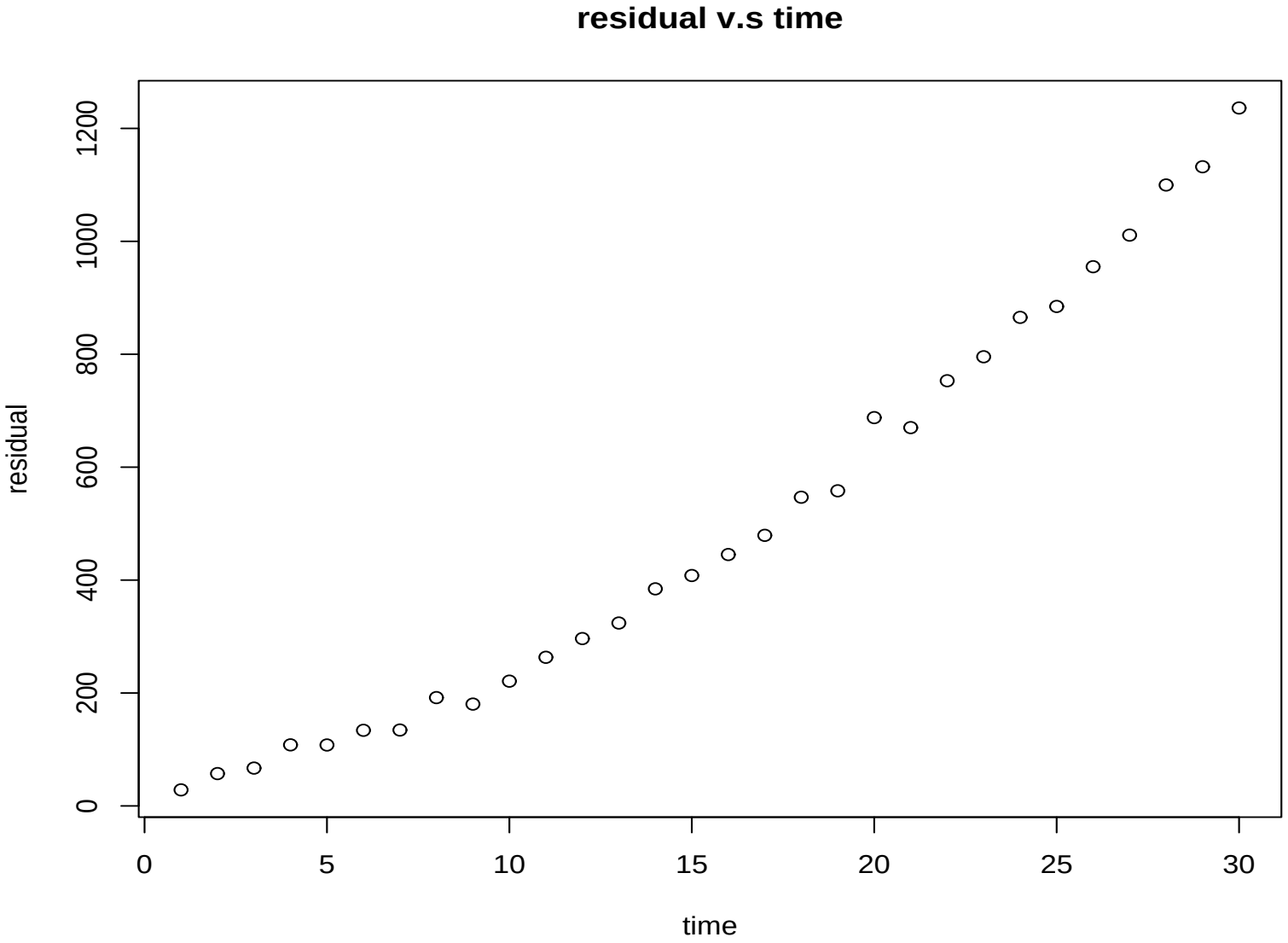
Figure 6: A residual plot that displays an increasing variance looks roughly like a horn opening to the right. A residual plot indicating a decreasing variance is a horn opening to the left.



3. Checking for independence of Error terms

- This is hard to do. It is common for nonindependence to be related to the sequence in which the observations are obtained.
- Plot e_i v.s time sequence (order in which data is collected)
- Independence would result in a random scatter of points.
- Non independence would result in a trend in the plot

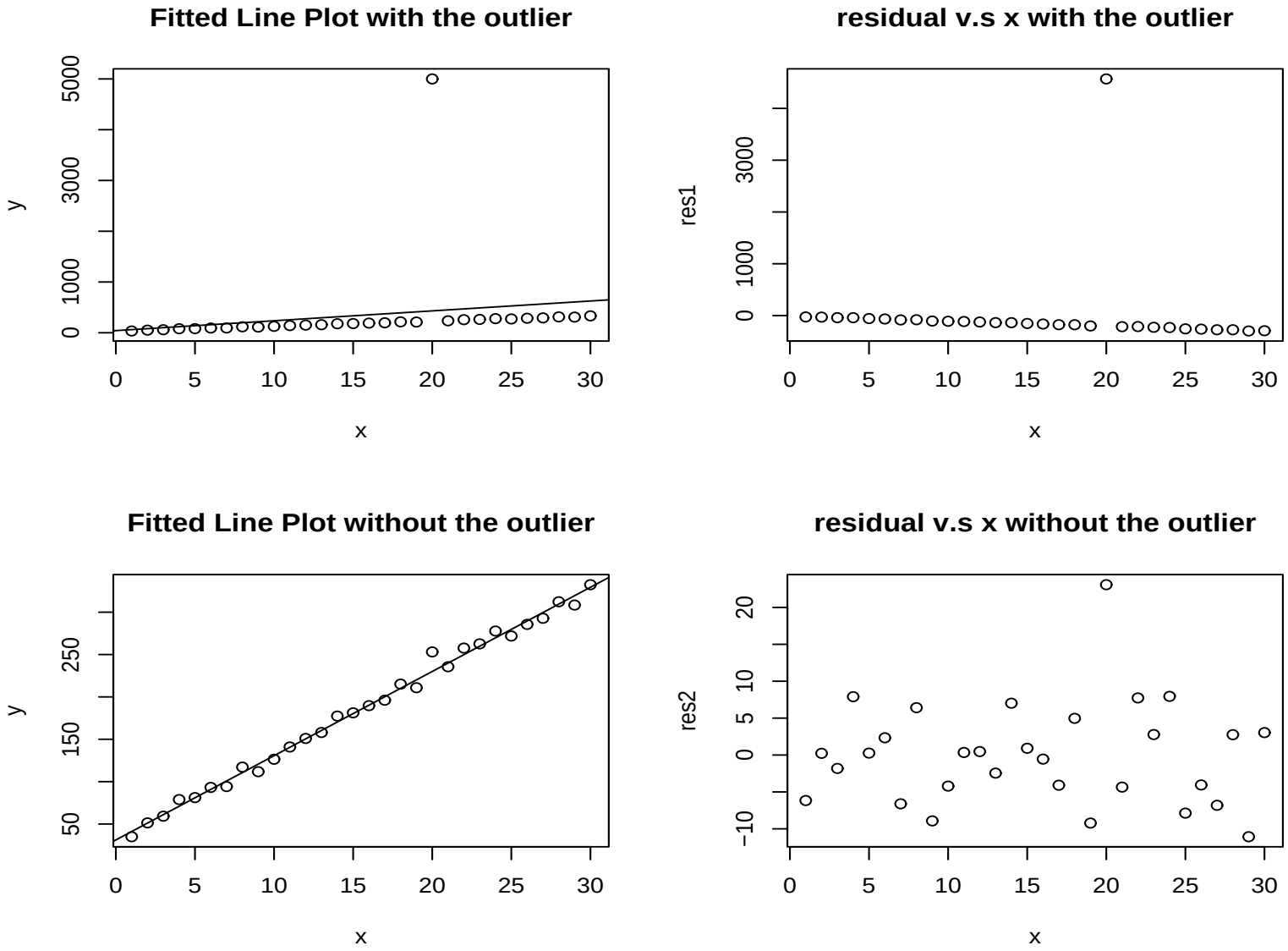
Figure 7: Plot of residual v.s time



4. Checking for Outliers

- Outliers are observations whose values are far from “typical” values in the sample
- Plot Y vs X, plot e_i vs x_i
- Outliers (y observations) can easily be spotted on a residual plot, especially if studentized residuals are used. Look for residuals that are far from the main set of residual values.

Figure 8: outliers, data were generated from $y = 30 + 10 * x + N(0, 10)$ with the No.20 observation replaced by 5000



X outliers and Y outliers

- Leverages (h_{ii}) are values between 0 and 1, $0 < h_{ii} < 1$, that measure how bizarre x value is relative to the other x values in the data

$$\sum_{i=1}^{n} h_{ii} = \text{number of predictor variables} + 1$$

In a simple linear regression

$$\sum_{i=1}^{n} h_{ii} = 2, \sum_{i=1}^{n} h_{ii}/n = 2/n$$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2, \text{ by } \mathbf{H}^2 = \mathbf{H}$$

Average of leverages in a simple linear regression is 2/n. The farther x_i is from the center of the data (as measured by the sample mean $\bar{x}$), the higher the leverage.

- Points with leverages larger than $2 * 2/n$ cause concern and leverage above $3 * 2/n$ cause considerable concern.
- Y outlier, examine the largest absolute standardized deleted residual, the appropriate α level test rejects if

$$\text{max}|(t_h)| \geq t(1 - \frac{\alpha}{2n}, dfE - 1).$$

- Comments:

  There are many possible reason for the presence of outliers, A random occurrence of an outlier, measurement error, non normal distributions for y's.

  —Outliers may be discarded from an analysis. This is reasonable if the data point resulted from a recording error, miscalculation, equipment failure or others

  —Otherwise, discarding an outlier, might not be wise, the results of the analysis might be biased

  —Try to determine why a data point is an outlier

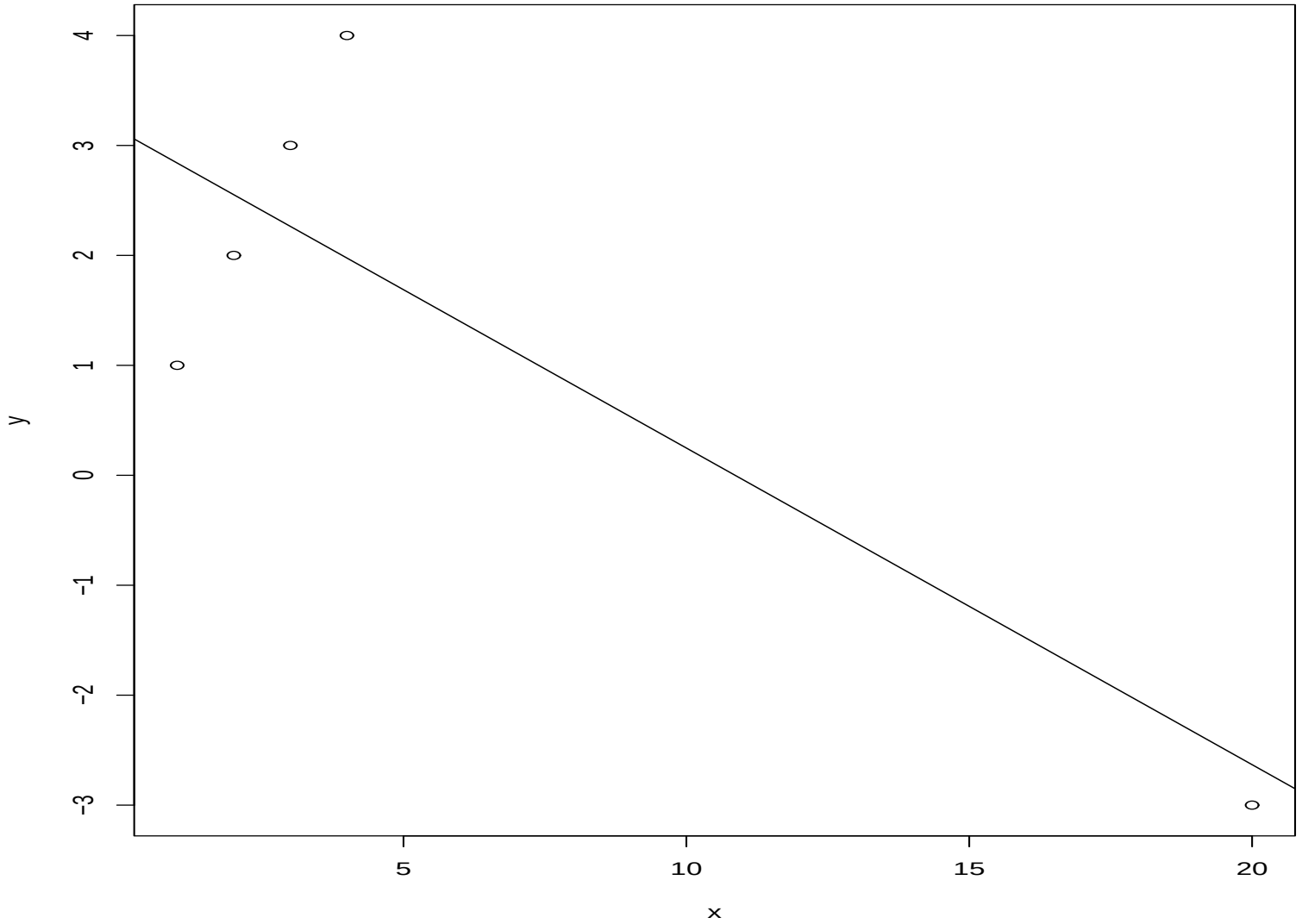
  —An outlier can greatly affect the results of the method of least squares. The fitted line is pulled disproportionately toward the outlier. The result can be a misleading the fit of the data.

Effects of high leverage

Example. The actual data are given below along with their leverages.

Case	1	2	3	4	5
x	1	2	3	4	20
y	1	2	3	4	-3
Leverage	0.30	0.26	0.24	0.22	0.98

Figure 9: Plot of data points along with fitted regression line

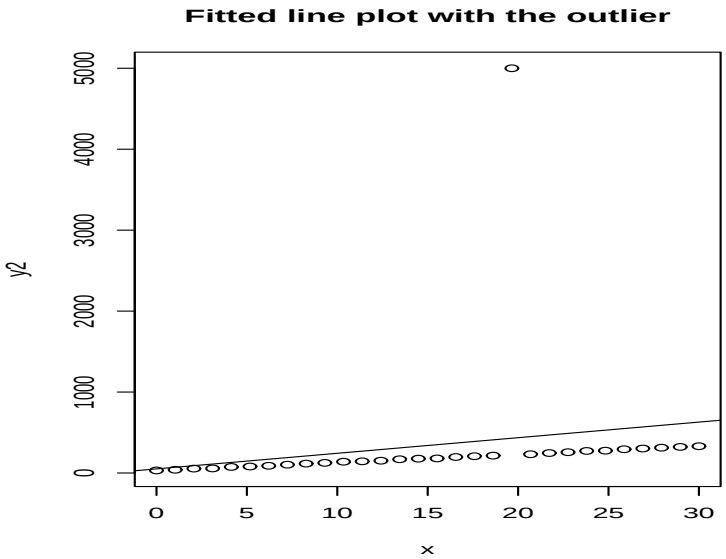


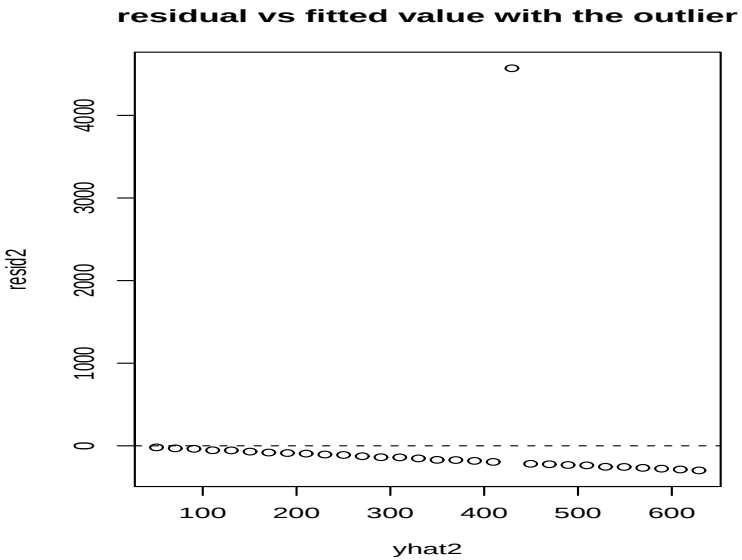
Comments:

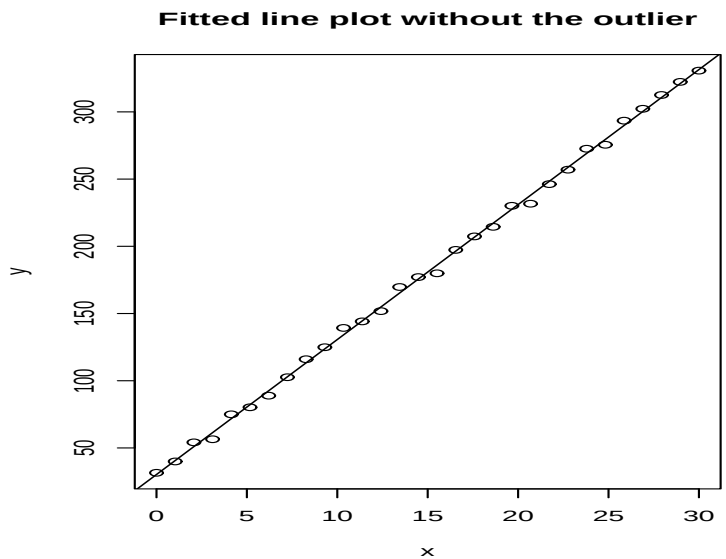
- The four points on the left form a perfect line with slope 1 and intercept 0. There is one high leverage point far away to the right.
- The estimated regression line is forced to go very nearly through this high leverage point.
- The plot has two clusters of points that are very far apart, so the estimated line is the line that goes through the mean x and y values for each of the two clusters. The single point on the right dominates the estimated straight line. This happens regardless of the fact that the four cases on the left follow a perfect straight line.

Effects of Y outliers

Figure 10: data were generated from $y = 30 + 10 * x + N(0, 10)$ seed=16, with the No.20 observation replaced by 5000







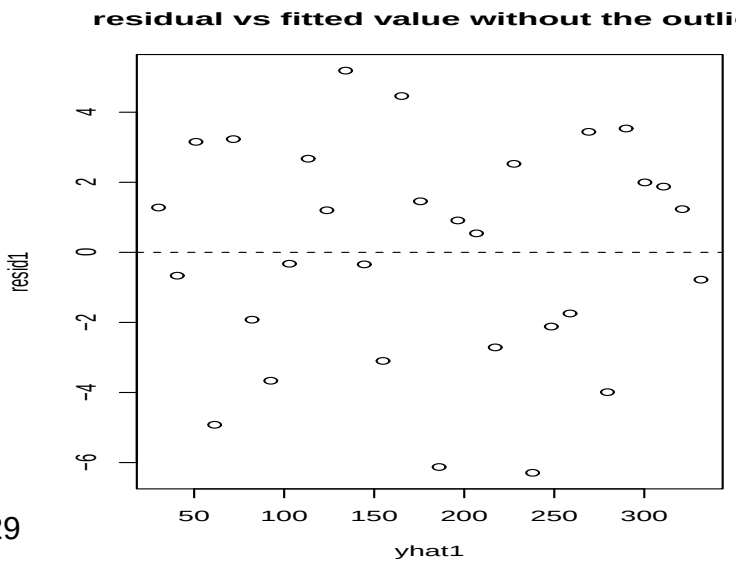
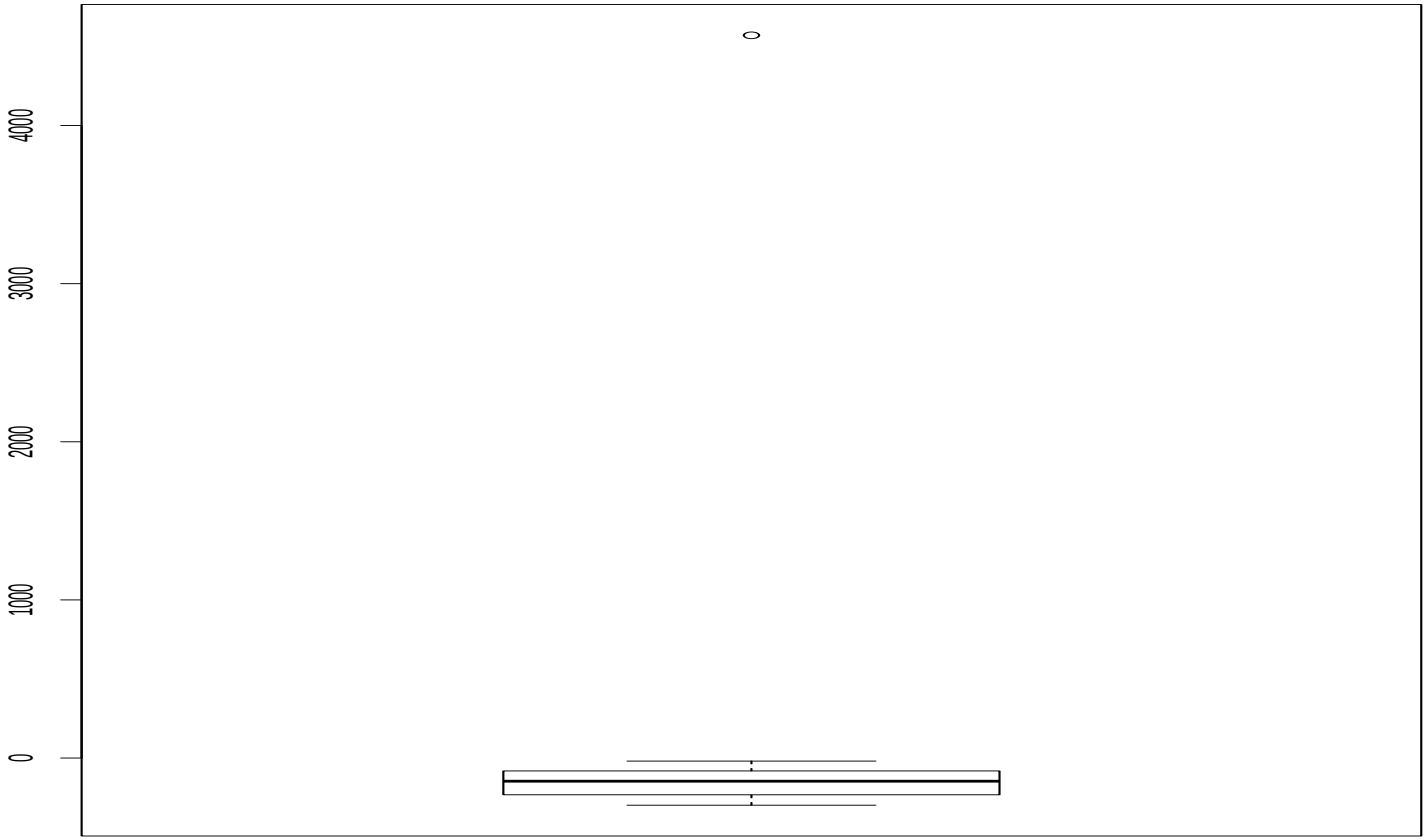


Figure 11: Boxplot



- rstudent value for No.20 observation is 1466.30

$$\begin{aligned} 1466.30 &>> qt(1 - 0.05/(2 * n), df(E) - 1) \\ &= qt(1 - 0.05/(2 * 30), 28 - 1) \\ &= 3.49 \end{aligned}$$

Coefficients	Estimate	SE	t-value	p-value
Without Outlier				
Intercept	30.22682	1.1276	26.81	$< 2e-16$
x	10.04131	0.06455	155.56	$< 2e-16$
With Outlier				
Intercept	50.74	314.40	0.161	0.873
x	19.27	18.00	1.071	0.293

- regression lines are

  $y = 30.23 + 10.04x$ without the outlier

  $y = 50.74 + 19.27x$ with the outlier

  which is pulled up disproportionately toward the outlier.

- the outlier inflated the standard errors.

5. Checking for normality

- Whether the distribution of the errors is far enough away from normal to invalidate our confidence intervals and significance tests. Look at the residuals' distribution. Use a normal quantile plot (qqplot) and a histogram plot.

- Normal quantile plot: A graph of the residuals versus the expected order statistics of the standard normal distribution. It plots quantiles of the data versus quantiles of a distribution. The Q-Q plot may be constructed using raw, standardized or jack-knifed residuals

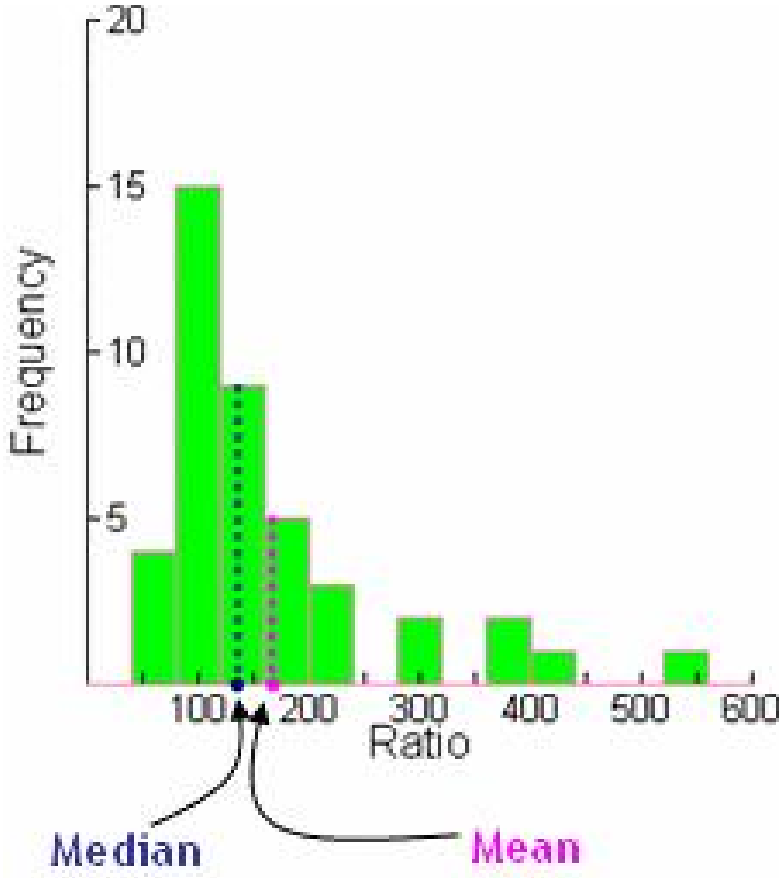
  —-If the observations come from a normal distribution we would expect the observed order statistics to be reasonably close to the expected order statistics. We should get approximately a straight line

  —-In general, Q-Q plots showing curvature indicate skew distributions, with downward concavity corresponding to negative skewness (long tail to the left) and upward concavity indicating positive skewness. On the other hand, S-shaped Q-Q plots indicate heavy tails, or an excess of extreme values, relative to the normal distribution.

Skewness: is a measure of the symmetry of the distribution of data values

——In a histogram, skew occurs if the values on one side of the histogram tend to extend further from the ”middle” than the values on the other side

——Right-skewed (positive skew): if the right (higher value) tail is longer or fatter. The mean is to the right, thus bigger than the median


20
15
10
5
Frequency
100
200
300
400
500
600
Ratio
Median
Mean

——Left-skewed (negative skew): if the left (lower value) tail is longer or fatter. The mean is to the left, thus smaller than the median

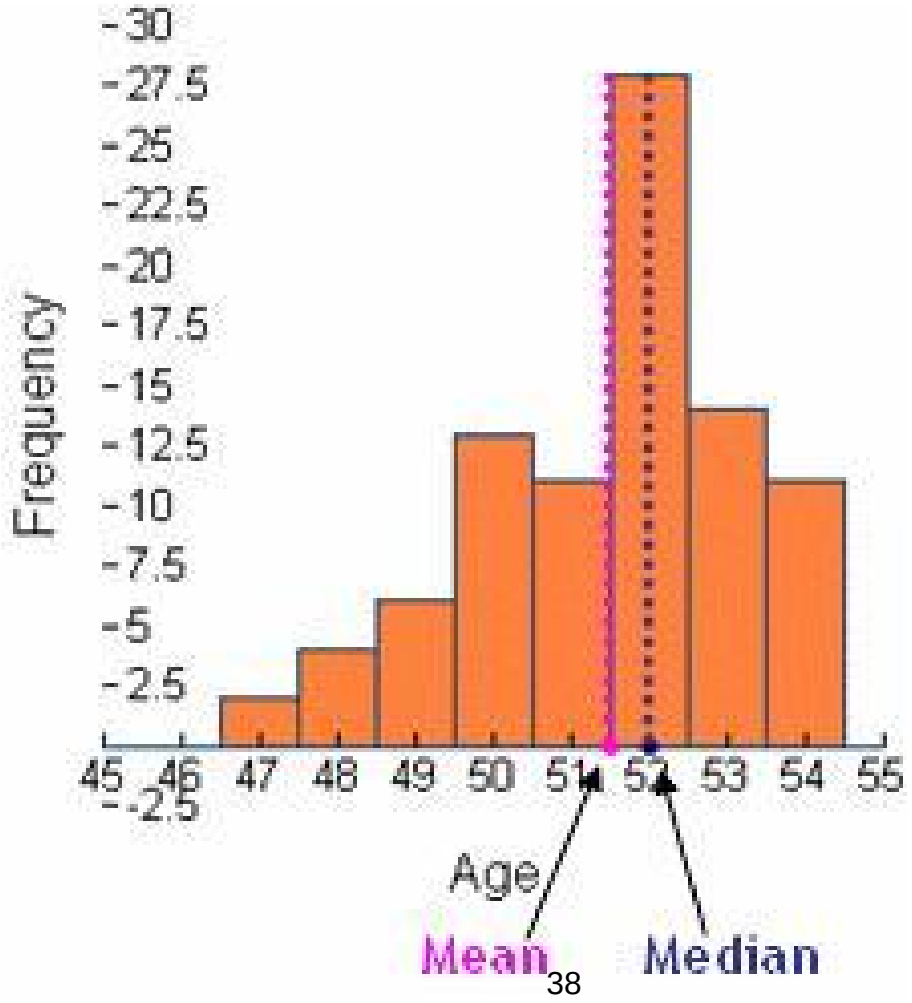

30
27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
2.5
Frequency
45
46
47
48
49
50
51
52
53
54
55
-2.5
Age
Mean
Median

Figure 12: check normality, data were generated from $y = 30 + 10 * x + N(0, 10)$

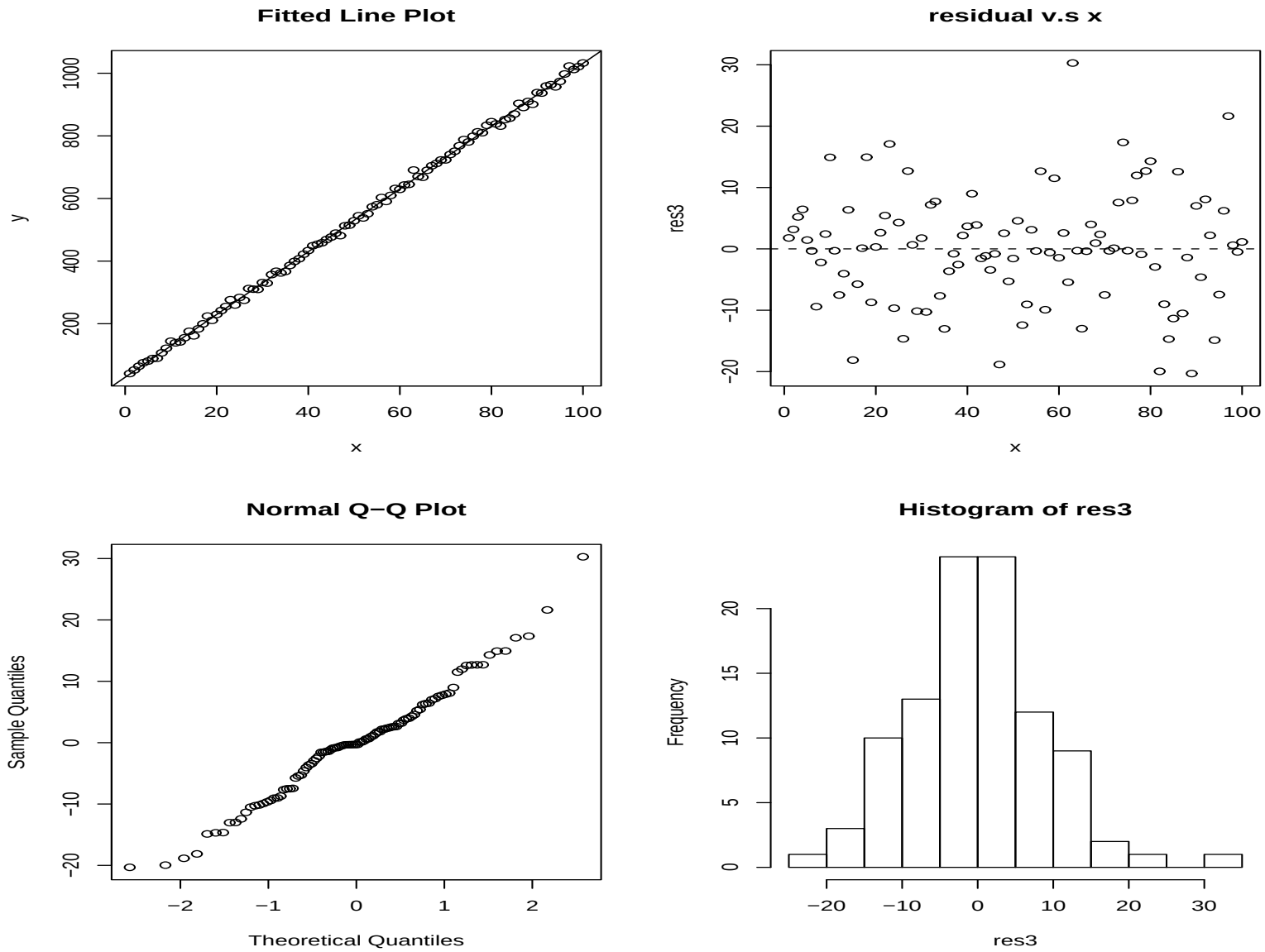


Figure 13: nonnormality, data were generated from $y = 30 + 10 * x + \chi^2(1)$

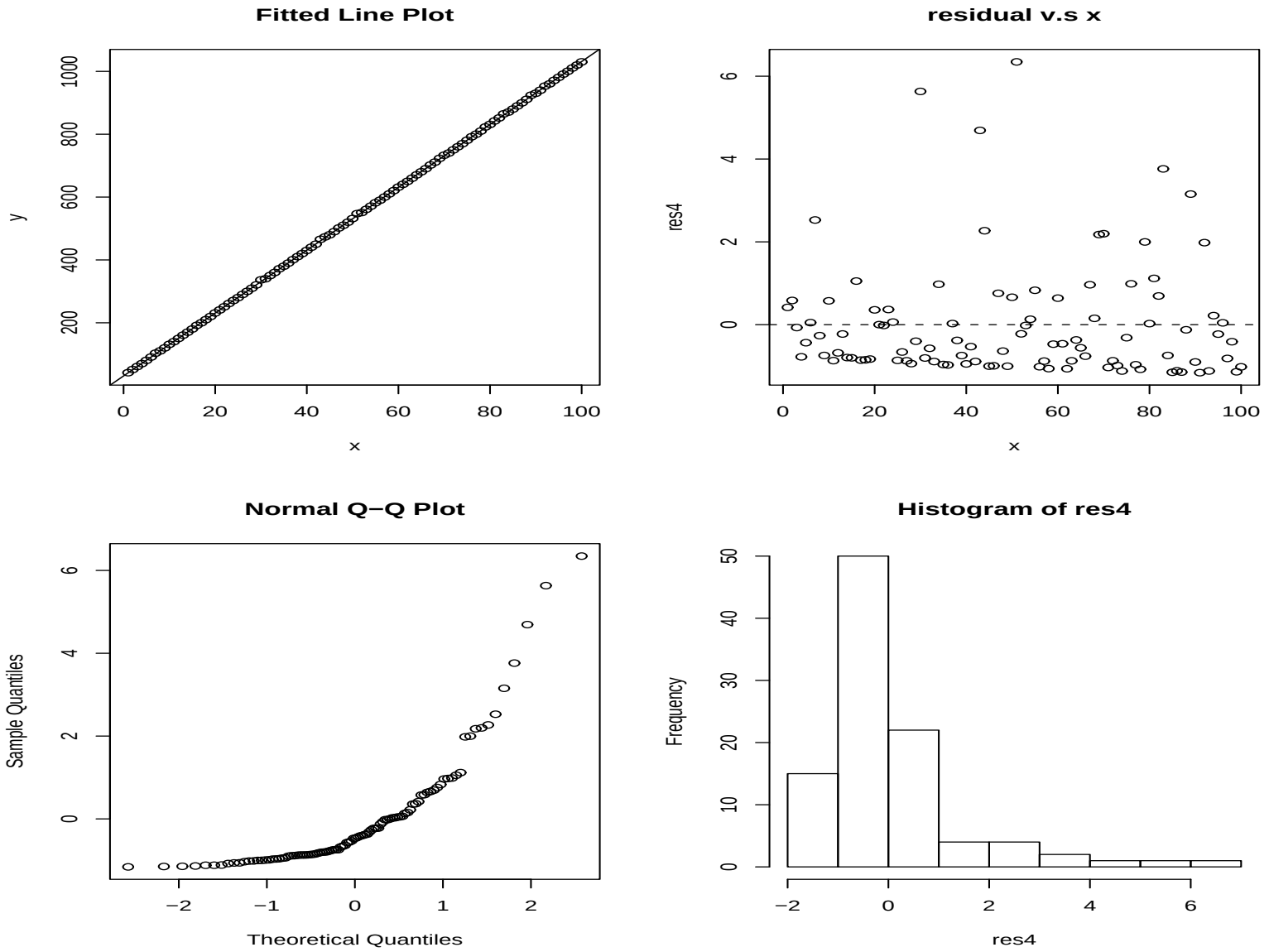
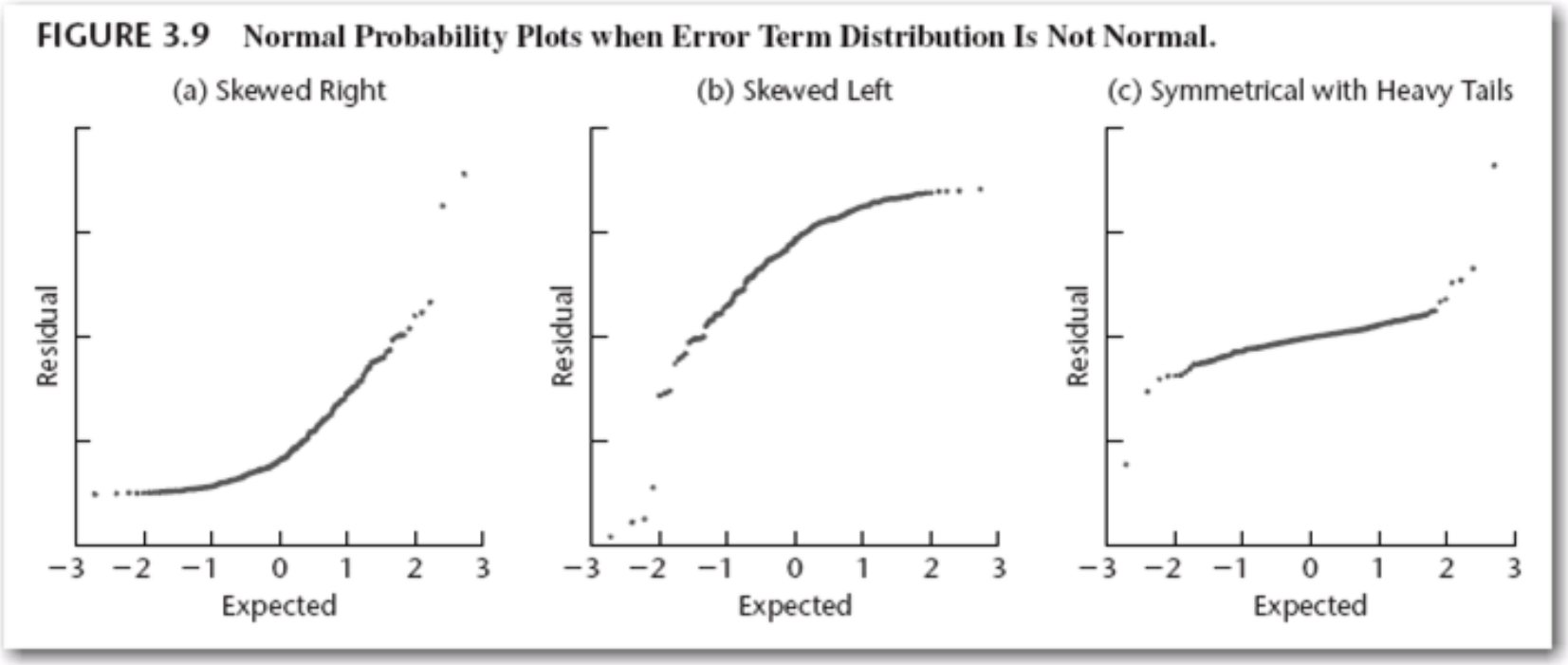


Figure 14: Normal probability plots when error term distribution is not normal



6. Predictor Variable

- The purpose is to determine whether there are any other key variables that could provide important additional descriptive and predictive power to the model
- Plot residuals vs. any variable omitted from the regression model. Any trend in this plot would indicate that you should consider including the variable in the regression model.

7. Summary

We discussed how to use plots to examine departures from the important assumptions such as linear relationship, constant variance, normal errors and independence.

- Plot, plot, plot, always check the plots first. Although graphic analysis of residuals is only an informal method of analysis, in many cases it suffices for examining the aptness of a model. Use the significance tests if you are uncertain what to conclude after examining the plots. Tests are not a replacement for the plots, but a supplement to them.
- plot Y vs. X (check for linearity, outliers)

- plot residuals vs. X or residuals vs. $\hat{y}$(check for constant variance, outliers, linearity, normality)
- qqplot and histogram of residuals (check normality)
- In practice, several types of departures may occur together.
- The basic approach to residual analysis explained here applies not only to simple linear regression but also to more complex regression and other types of statistical models

Some Tests:

Correlation Test for Normality

- Correlation between residuals e_i and their expected values under normality. Reject H_0 : Data is normally distributed if corr(obs, exp) is too small.

- See Table B.6 for critical values for this correlation test, for Toluca company example, the critical value for $n = 25, \alpha = .05$ is .959. The coefficient of correlation between the ordered residuals and their expected values under normality is .991. So we support the conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

Two formal test for constance of error variance

(1) Modified Levene test (Brown-Forsythe Test), more appropriate for analysis of variance (leave as homework)

(2) Breusch-Pagan test

- Assume that the error terms are independent and normally distributed and variance of the error term ε_i, denoted by σ_i^2, is related to the level of X by

$$\ln\sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

- σ_i^2 either increases or decreases with the level of X.

- constant variance indicates that $\gamma_1 = 0$, that is

$$\ln\sigma_i^2 = \gamma_0 = \text{constant.}$$

- If $\gamma_1 \neq 0$, then the variance is not constant.
- If we could estimate σ_i^2 at each x_i, we could use regression analysis to fit the model and test for the significance of γ_1
- The proposed test estimate the σ_i^2 by e_i^2, the test then obtain SSR^* from the regression of e_i^2 on x_i's. The test statistic is

$$X_{BP}^2 = \frac{SSR^*}{2} \Big/ \left(\frac{SSE}{n}\right)^2$$

where SSE is the sum of squares error from regressing Y on X.

- To test $H_0 : \gamma_1 = 0$ vs $H_\alpha : \gamma_1 \neq 0$, we reject H_0 if $X^2_{BP} > \chi^2(1-\alpha;1)$ for large n.

- Example: BP test for the Toluca Company example. To discover the relationship between lot size and labor hours required to produce the lot. Data on lot size and work hours for 25 recent production runs were utilized.

- Regress Y on X and obtain $SSE = 54825$.
  Regress the squared residuals e_i^2 against X and obtain $SSR^* = 7896128$.

$$X^2_{BP} = \frac{7896128}{2} / \left(\frac{54825}{25}\right)^2 = .821.$$

- $\chi^2(.95;1) = 3.84$, $X^2_{BP} = .821 \leq 3.84$, we conclude H_0, that the error variance is constant.

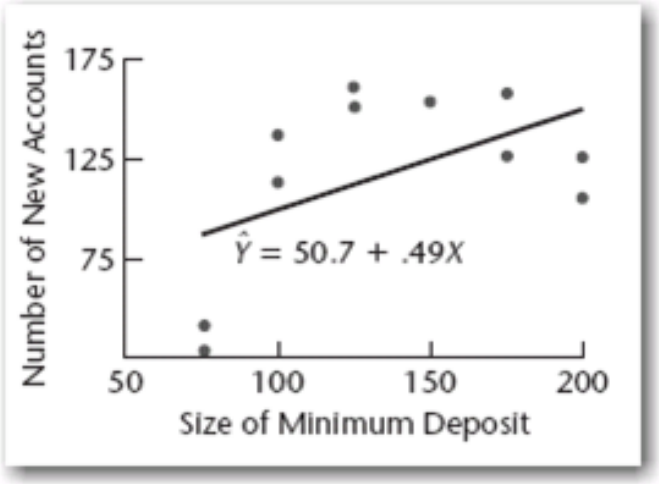
Testing Lack of Fit

- Lack of fit involves an initial model that does not fit the data adequately
- Look for models that fit significantly better than the initial model

Fisher's test for lack of fit

Example: (page 120) 12 suburban branch offices of a commercial bank. Holders of checking accounts were offered gifts for setting up money market accounts. Minimum deposits were required and the value of the gift was directly proportional to the minimum deposit. 6 levels were used. One bank dropped out. Linear fit is bad. How can we test it?

Figure 15: Scatter plot and fitted regression line-Bank example



The lack of fit test requires that the observations Y for given X are

- Independent
- Normally distributed
- The distributions of Y have the same variance σ^2
- You have replicate (repeat) values (two or more observations with exactly the same X value(s) at one or more X levels

Notation

- $X_1, X_2, \cdots, X_c$ are c unique X values
- The number of replicates for the jth level of X is n_j
- $n_1 + n_2 + \cdots n_c = n$
- At X_j we have $n_j \geq 1$ observations denoted, say Y_{1j}, Y_{2j}.
- At X_j, the average response is $\bar{Y}_j = (Y_{1j} + Y_{2j} + \cdots Y_{n_j j})/n_j$

General linear test approach

Full model("pure" model):

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

- μ_j are parameters $j = 1, \cdots, c$
- ϵ_{ij} are independent $N(0, \sigma^2)$

Estimated full model:

$$\hat{y}_{ij} = \hat{\mu}_j = \bar{y}_j$$

SSE(Full) = SS for "pure" error

$$SSE(F) = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2 = SSPE$$

DF(Full):

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c$$

Reduced model = regression model:

$$y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}$$

Estimated reduced model:

$$\hat{y}_{ij} = b_0 + b_1 x_j$$

SSE(Reduced) =SSE for regression model:

$$\begin{aligned} SSE(R) &= \sum\sum [y_{ij} - (b_0 + b_1 x_j)]^2 \\ &= \sum\sum (y_{ij} - \hat{y}_{ij})^2 \\ &= SSE \end{aligned}$$

$$df(R) = n - 2$$

General linear test

(1) Hypotheses:

$$H_0 : E(y) = \beta_0 + \beta_1 x$$
$$H_\alpha : E(y) \neq \beta_0 + \beta_1 x$$

(2) Test statistic:

$$\begin{aligned} F^* &= \frac{SSE - SSPE}{(n-2)-(n-c)} \div \frac{SSPE}{n-c} \\ &= \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE} \end{aligned}$$

(3) Decision Rule:

$$F^* \leq F(1-\alpha; c-2, n-c), \text{conclude} H_0$$
$$F^* \geq F(1-\alpha; c-2, n-c), \text{conclude} H_\alpha$$

Bank Example: ANOVA table

Source of Variation	SS	df	MS
Regression	5141.3	1	5141.3
Error	14741.6	9	1638.0
Lack of fit	13593.6	4	3398.4
Pure error	1148.0	5	229.6
Total	19882.9	10	

$F^* = MSLF/MSPE = 3398.4/229.6 = 14.80$, let $\alpha = .01$, then $F(.99; 4, 5) = 11.4$, $F^* = 14.8 > 11.4$, P value for the test is 0.006.

we conclude H_α, that the regression function is not linear.

Comments on general linear approach for lack of fit test

- Not all x's need to be replicated
- Test can be applied with more than one predictor; if there are p parameters the numerator degrees of freedom will be $c - p$, denominator degrees of freedom will be $n - p$
- Doesn't tell you what the right model is if it rejects; but you can believe your model if it accepts.

Remedial Measures for Inappropriate Regression Models:

- suppose one or more model assumptions appear to be violated, what to do?

  (1) Abandon the model and search for a more appropriate one

  (2) Use transformations on the data (y and/or x), so that the transformed data is fit well by your simple model.

Assumption Violation:

1. Nonlinearity of regression function

- Modify the regression function

  Example:

$$y = \beta_0 + \beta_1 x^2 + \varepsilon$$

$$y = \beta_0 + \beta_1 \log x + \varepsilon,$$

  consider x^2 and $\log x$ as individual quantities. Consider model

$$y = \beta_0 + \beta_1 x' + \varepsilon,$$

  where $x' = x^2$ or $x' = \log x$.

- Transform the data, so that the transformed data follows a linear relationship. Power transformation, Box-Cox transformation

2. Nonconstancy of Error Variance

- Modify the model so that nonconstant variances are allowed. Use the method of weighted least squares (we need to know how the variances are changing with x to use this method)
- Sometimes a simple transformation of the response will make the variances constant

3. Nonindependence of Error terms

- If the error terms are correlated and we know the structure of the correlation, then a model with correlated errors can be used and analyzed with generalized least squares.
- Sometimes a simple transformation of the response, such as differences between successive observations will eliminate the correlation between observations.

4. outliers

- A more appropriate regression function might remedy problems with outliers
- Transformations at times will tend to “pull in” the outliers

5. Non normality of error terms

- Could consider a model with nonnormal error terms. But it is difficult to do so.
- Transform the response, so that the distribution of the error terms is normal. Sometimes transformations of the response can get rid of problems with nonconstant variance and nonnormality.
- Nonnormality (at least not severe) is not as serious as nonconstant variance or the wrong regression function.

Intrinsically Linear Model:

There are nonlinear model that is intrinsically linear.

Nonlinear	Transformation	Linear
$y = \gamma_0 x^{\gamma_1}$	$y' = \ln y, x' = \ln x$	$y' = \ln\gamma_0 + \gamma_1 x'$
$y = \gamma_0 e^{\gamma_1 x}$	$y' = \ln y$	$y' = \ln\gamma_0 + \gamma_1 x$
$y = \frac{1}{\gamma_0 + \gamma_1 x}$	$y' = \frac{1}{y}$	$y' = \gamma_0 + \gamma_1 x$
$y = \frac{x}{\gamma_0 x - \gamma_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \gamma_0 - \gamma_1 x'$
$y = \frac{1}{1 + exp(\gamma_0 + \gamma_1 x)}$	$y' = \ln(\frac{1}{y} - 1)$	$y' = \gamma_0 + \gamma_1 x$

Transformations:

If the residuals show a problem with

- lack of fit (having the wrong model for the mean)
- heteroscedasticity
- nonnormality

  Try y transformation or x transformation or both

- y transformation is more common
- only works when $y_{\text{max}}/y_{\text{min}}$ is reasonably large
- choose a transformation to stabilize variance
- log or square transformations can solve many problems

Table 1: Variance stabilizing transformations

Data	Distribution	Mean, Variance Relationship	Transformation
Count	Poisson	$\mu_h \propto \sigma_h^2$	$\sqrt{y_h}$
Amount	Gamma	$\mu_h \propto \sigma_h$	$\log(y_h)$
Proportion	Binomial/N	$\mu_h(1-\mu_h)/N \propto \sigma_h^2$	$\sin^{-1}(\sqrt{y_h})$

Figure 16: Circle of Transformations

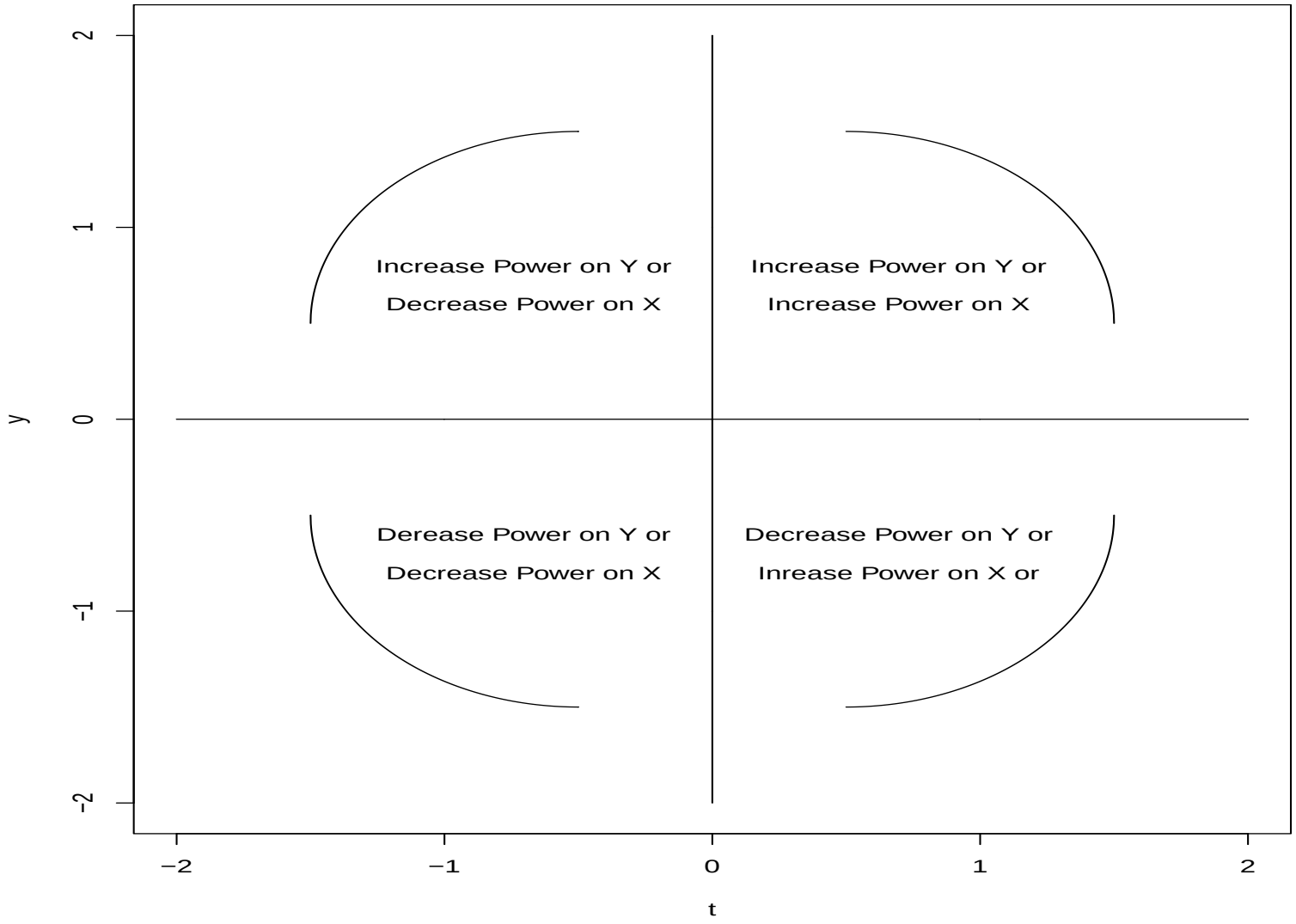


Figure 17: Curved x, y plot ($y = \cos x$ in the first quadrant. According to figure 16, need to increase power of both x and y

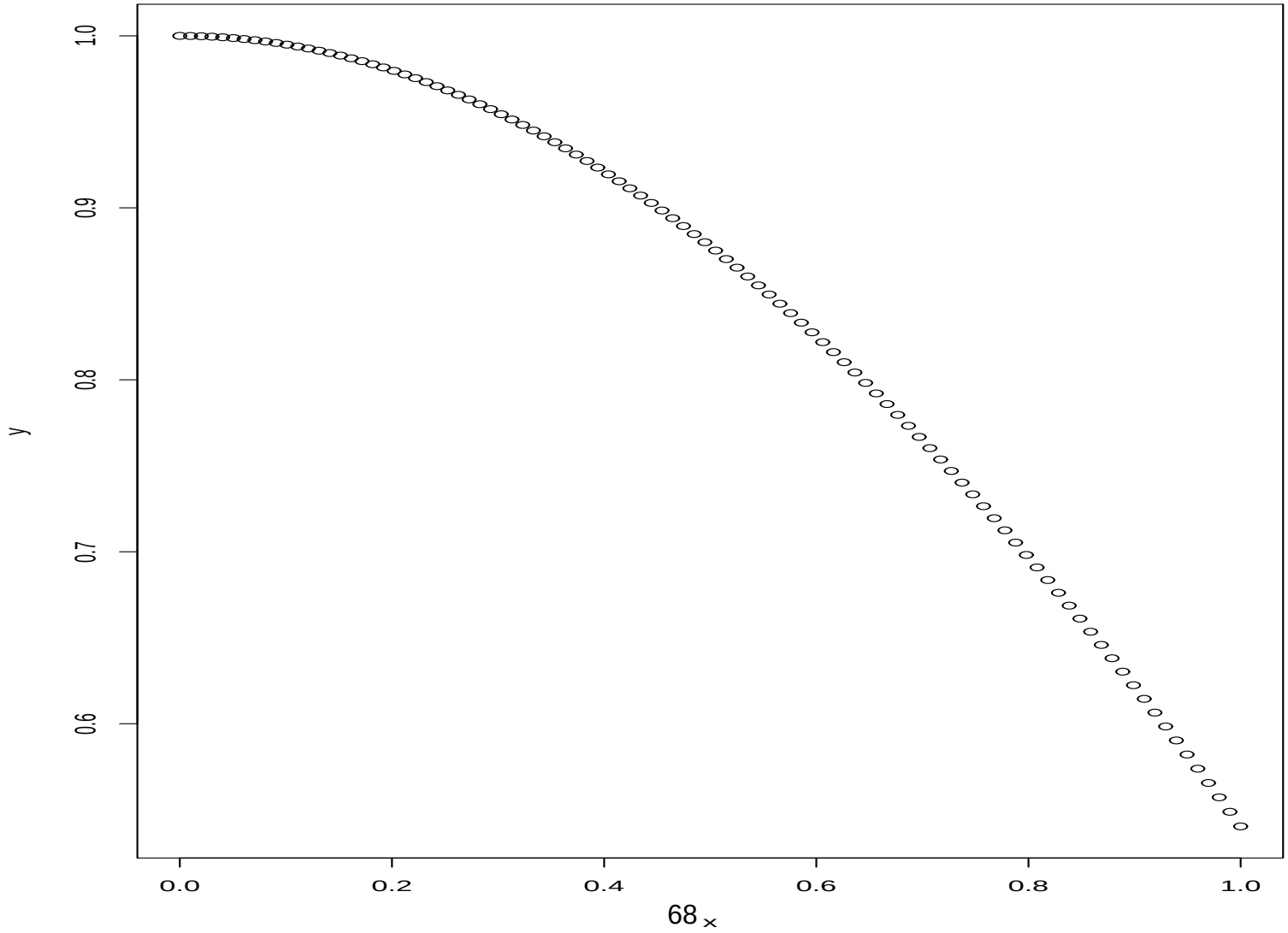
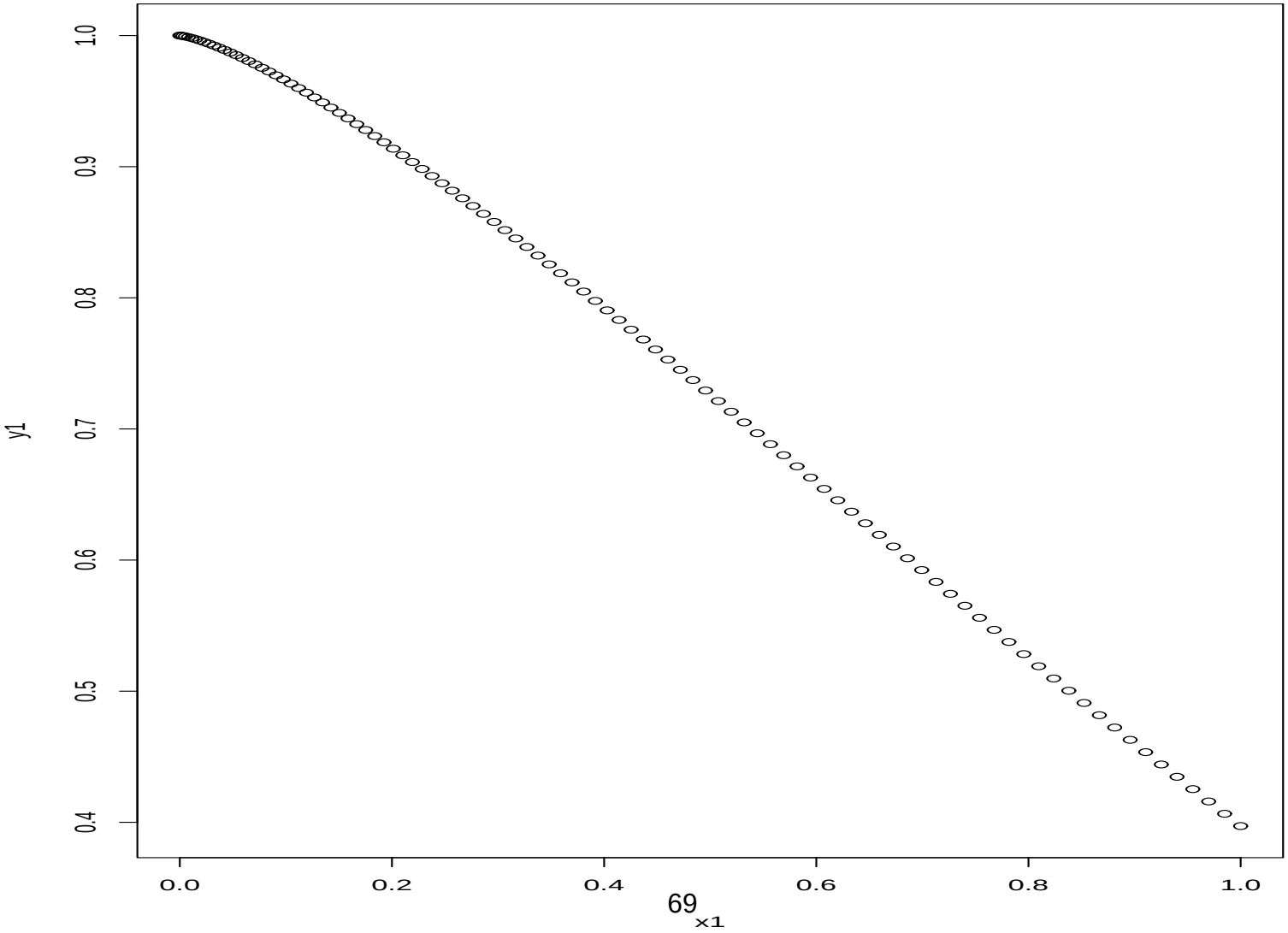


Figure 18: Plot of $x^{1.5}, y^{1.5}$. $y = \cos x$ in the first quadrant. After transformation $x^* = x^{1.5}$ and $y^* = y^{1.5}$, the curve is much straighter than the one in Figure 17



Power Transformation:

- If the residuals appear to be normal with constant variance, and the relationship is linear, then go ahead with the regression model. No transformation is needed.

- Transformation is used to deal with model violations. Commonly used transformation is the power transformation (Box-Cox transformation)

$$y^* = \begin{cases} y^\lambda & \lambda \neq 0 \\ \ln y & \text{if} \quad \lambda = 0. \end{cases}$$

$$x^* = \begin{cases} x^\lambda & \lambda \neq 0 \\ \ln x & \text{if} \quad \lambda = 0. \end{cases}$$

- If the residuals appear to be normal with constant variance, but the relationship is non-linear, try transforming the X's to make it a straight line. The transformation on Y may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.
- If the residuals are not randomly scattered around zero, but have trends. Try transforming Y .
- If you choose a transformation, you need to go back and do all the diagnostics all over again.

- Box and Cox (1964) developed a method to suggest an appropriate transformation of the response variable y, so that, the transformed y is appropriate for the simple linear regression model. The transformation are power transformation. The method selects the λ power to minimize the SSE of the regression

$$y^{\lambda} = \beta_0 + \beta_1 x + \varepsilon$$

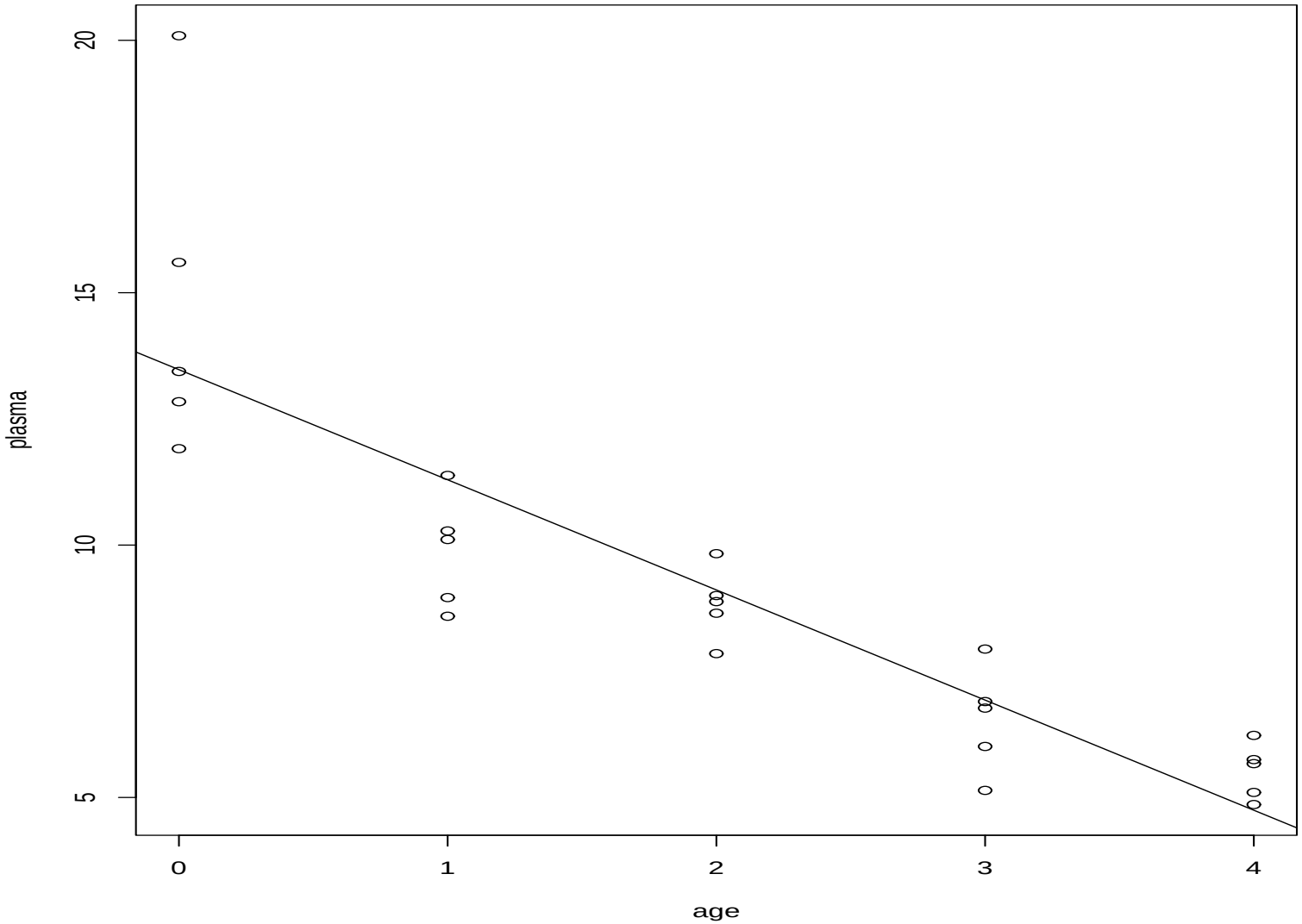
and use maximum likelihood to estimate λ. The method runs the regression for a range of transformations between -2 and +2, pick the one that minimizes $SSE(\lambda)$. Eventually you would probably suggest the same transformation by eye.

- What transformation to use? Pictures of various prototype situations are given in section 3.9. Fig. 3.13, p. 130: transformations on X for non-linear relationships. Fig. 3.15, p. 132: transformations on Y for non-constant variance (and possibly non-linear relationships)
- Not all scatter plots can be straighten by a power transformation
- Box-Cox suggests a transformation but there is no guarantee it will solve all our problems. We still have to check residuals, assumptions, etc.
- There may be a number of transformations that adequately “straighten” a scatterplot. Pick the transformation that is most interpretable (or the simplest).
- If variable ranges over several orders of magnitude, natural logs transformation usually work; often needed for economic data

- $1/Y$ often makes intuitive sense: If Y is customers per hour, $1/Y$ is hours per customer.
- Square root may make sense if you are measuring areas (square-feet, etc).
- If $Y = 0$ for some observations, cannot do $1/Y$ or $logY$; just add a constant k to all of the Ys first

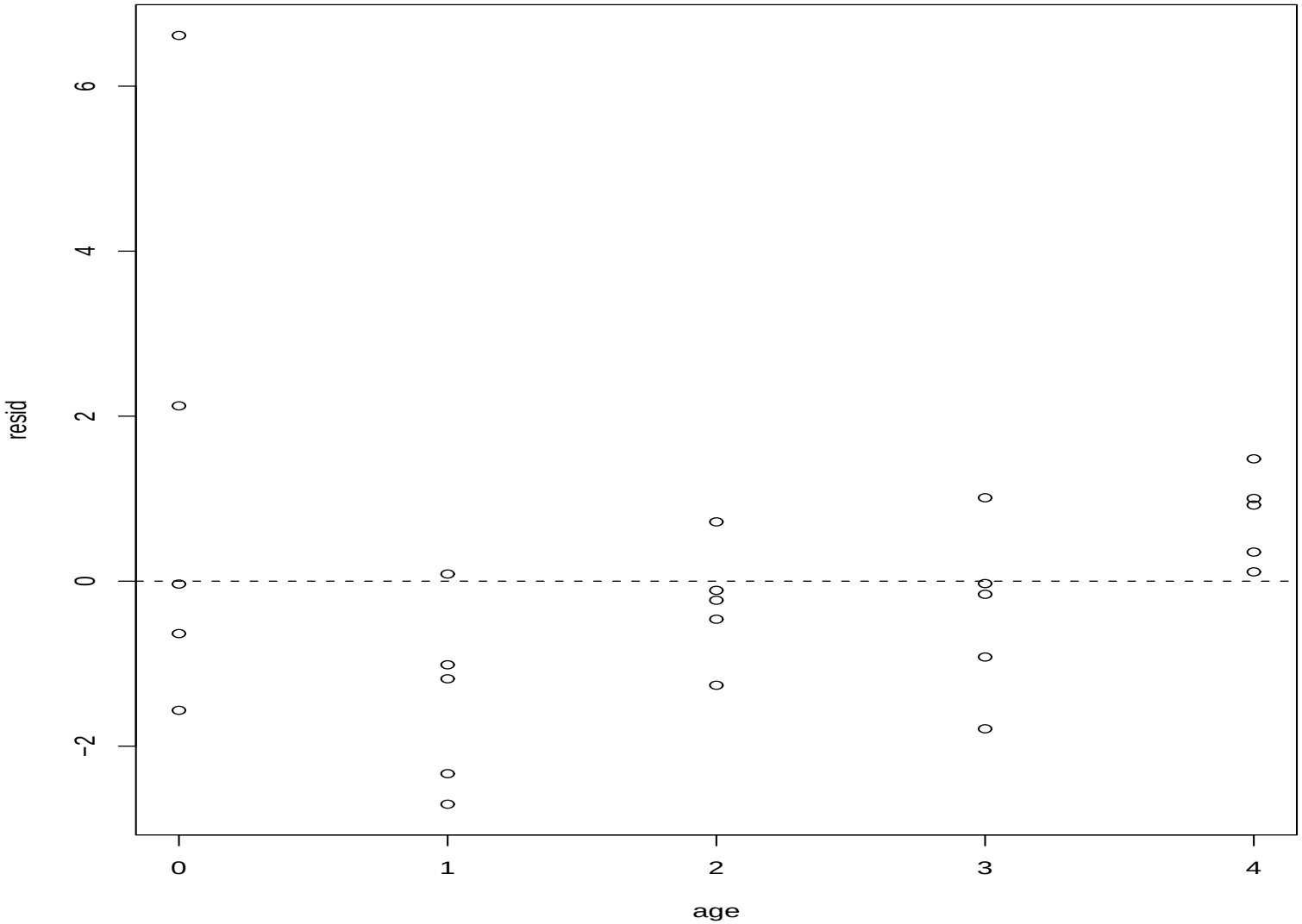
Example: Data on age and plasma level of a polyamine for a portion of 25 healthy children in a study is used for this example.

Figure 19: plot and fitted line



Notice the curvilinear regression relationship, as well as the greater variability for younger children than for old ones

Figure 20: residual vs x, check nonconstant variance, outliers



Residuals are not randomly scattered around $y = 0$. BP test gives a value of 11.1063 with p-value = 0.0008604. Reject the constant variance assumption

Figure 21: boxplot to check outliers

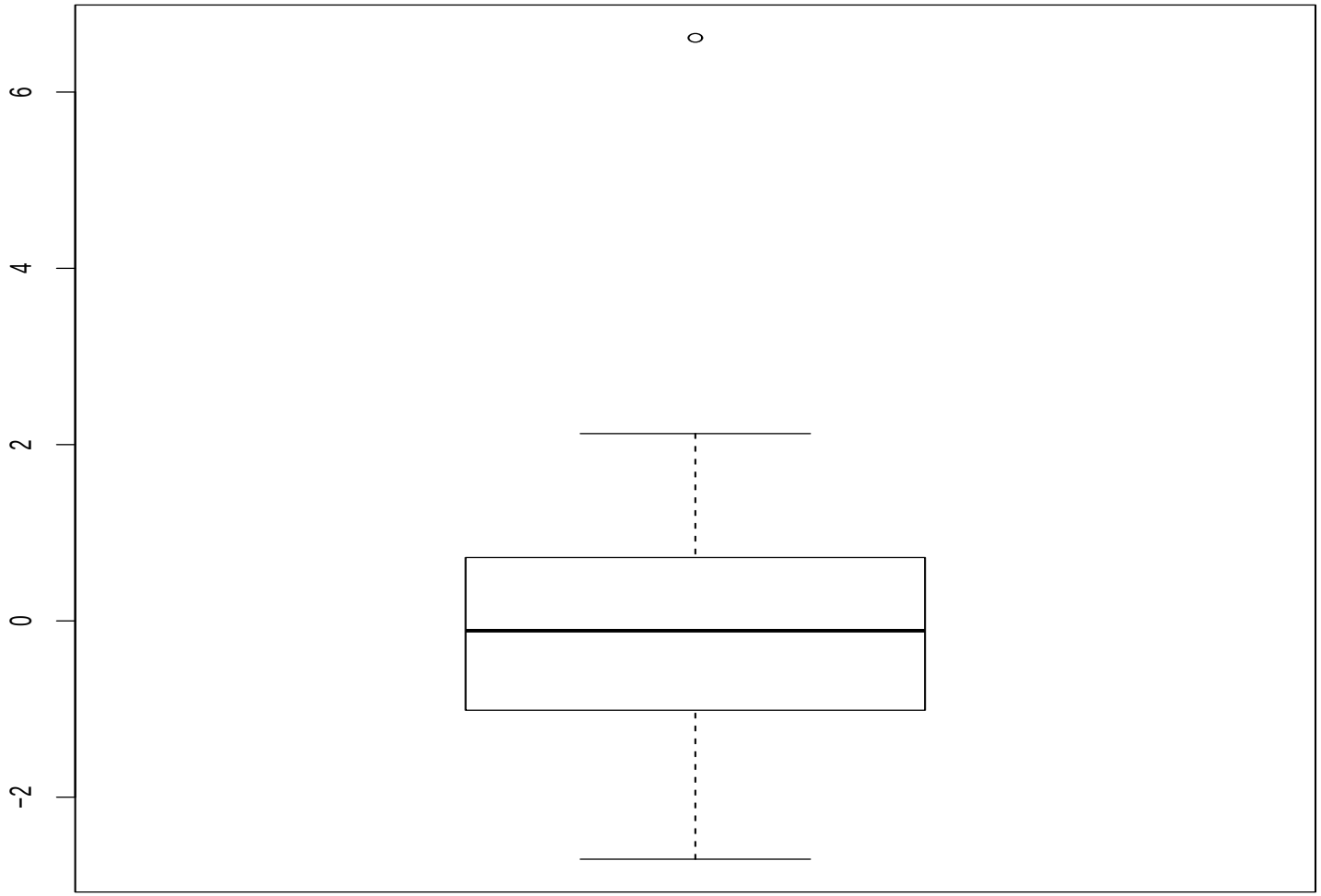
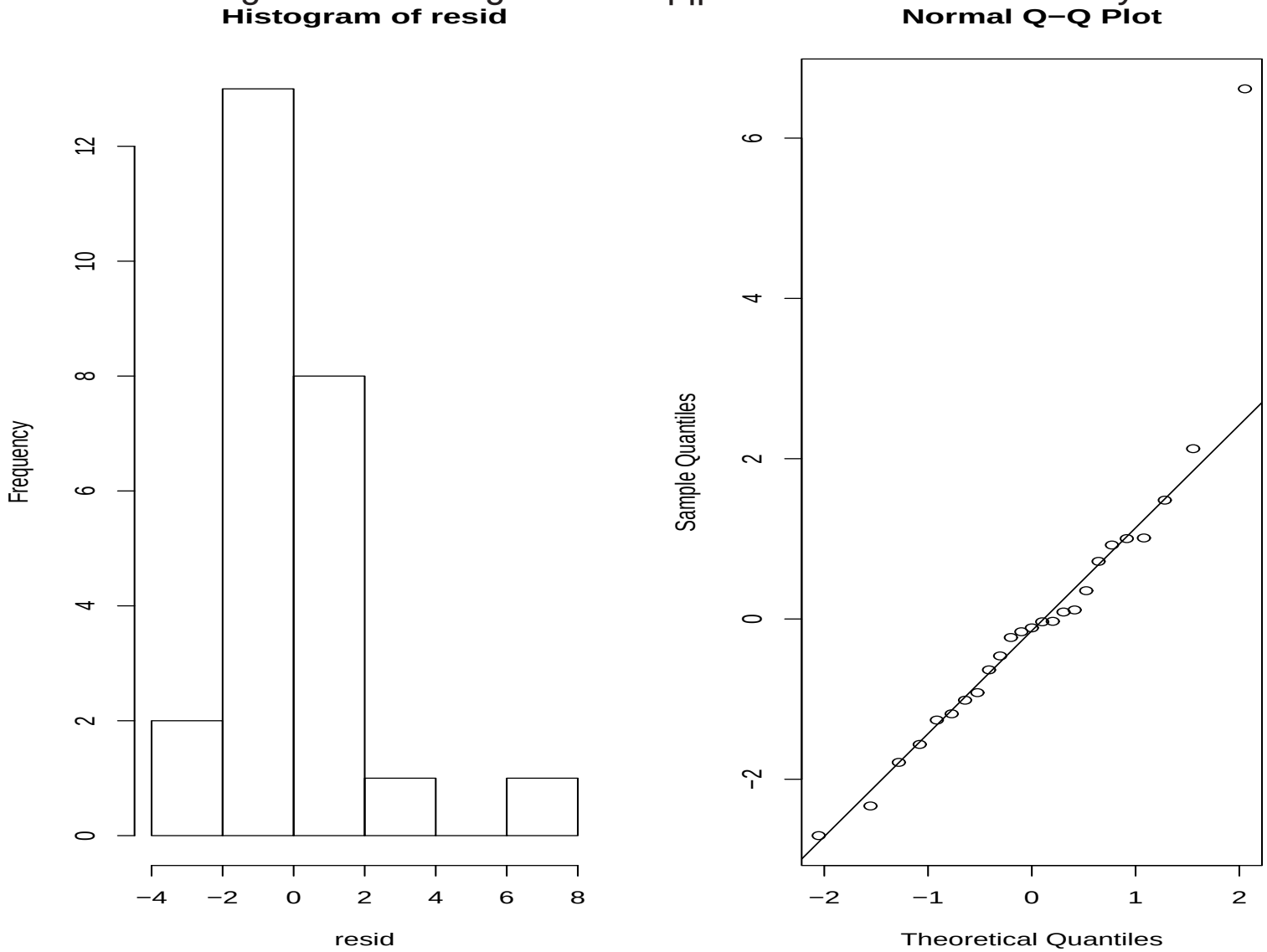


Figure 22: histrogram and qqplot to check nonnormality

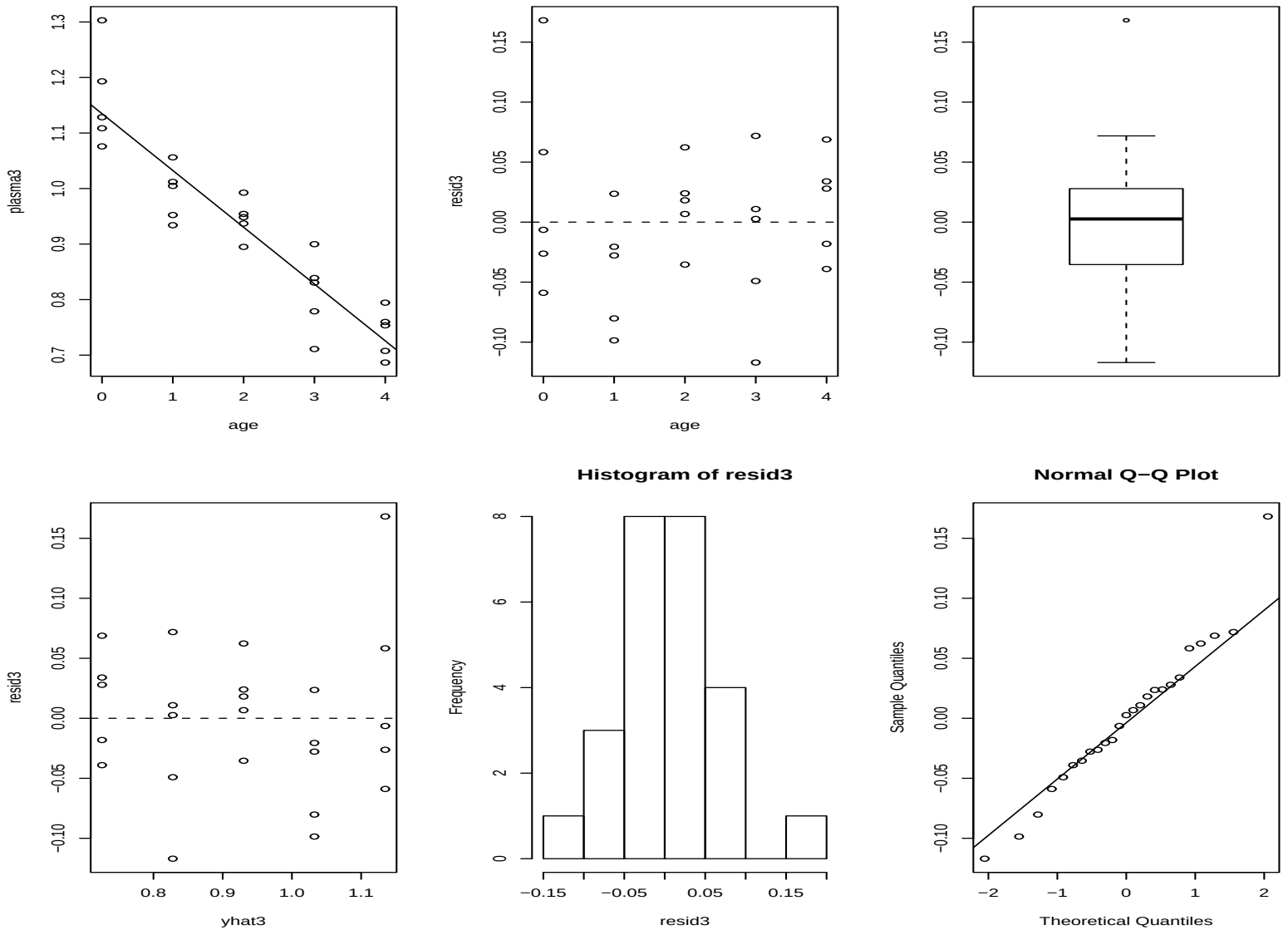


Histogram plot is skewed to right, qq plot is not satisfied. Shapiro-Wilk normality test gives a p-value of 0.001098, reject the normality assumption

Based on the prototype regression pattern, we shall try the logarithmic transformation $y' = log10(y)$

Child	Age	Plasma Level y_i	Transformed Data $log10(y)$
1	0	13.44	1.1284
2	0	12.84	1.1086
3	0	11.91	1.0759
4	0	20.09	1.3030
⋮	⋮	⋮	⋮
25	4	6.23	0.7945

Figure 23: Analysis of data after transformation $Y' = log10(Y)$



- Transformation led to a reasonably linear regression relation
- Variability at the different levels of x also has become reasonably constant
- Residual plots look reasonably well
- Breusch-Pagan test gives a value of 2.0352 with a p-value .1537, supporting the constant variance assumption. Shapiro-Wilk normality test gives a p-value of 0.6784, supporting the normality assumption.
- All of this evidence supports the appropriateness of regression model for the transformed y data
- Regression line for the transformed data $\hat{y}' = 1.135 - .1023x$
- Convert back to the original unit $\hat{y} = 10^{1.135-.1023x}$

Example 7.3.1. Hooker data (Christensen)

Forbes (1857) reported data on the relationship between atmospheric pressure and the boiling point of water that were collected in the Himalaya mountains by Joseph Hooker. Weisberg (1985, p. 28) presented a subset of 31 observations that are used as our example.

Figure 24: Hooker data with fitted regression line

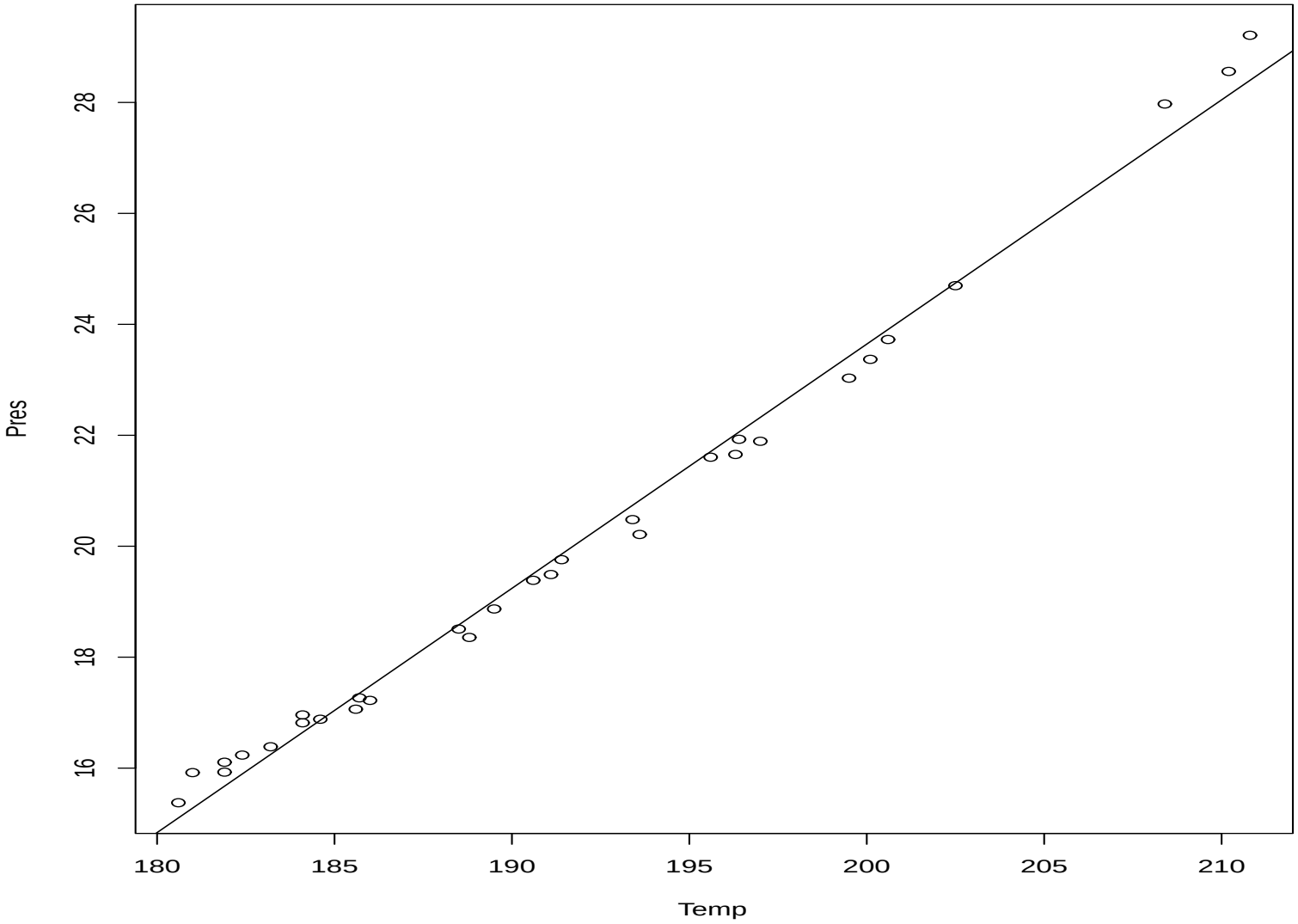


Figure 25: Hooker data: residual vs fitted values

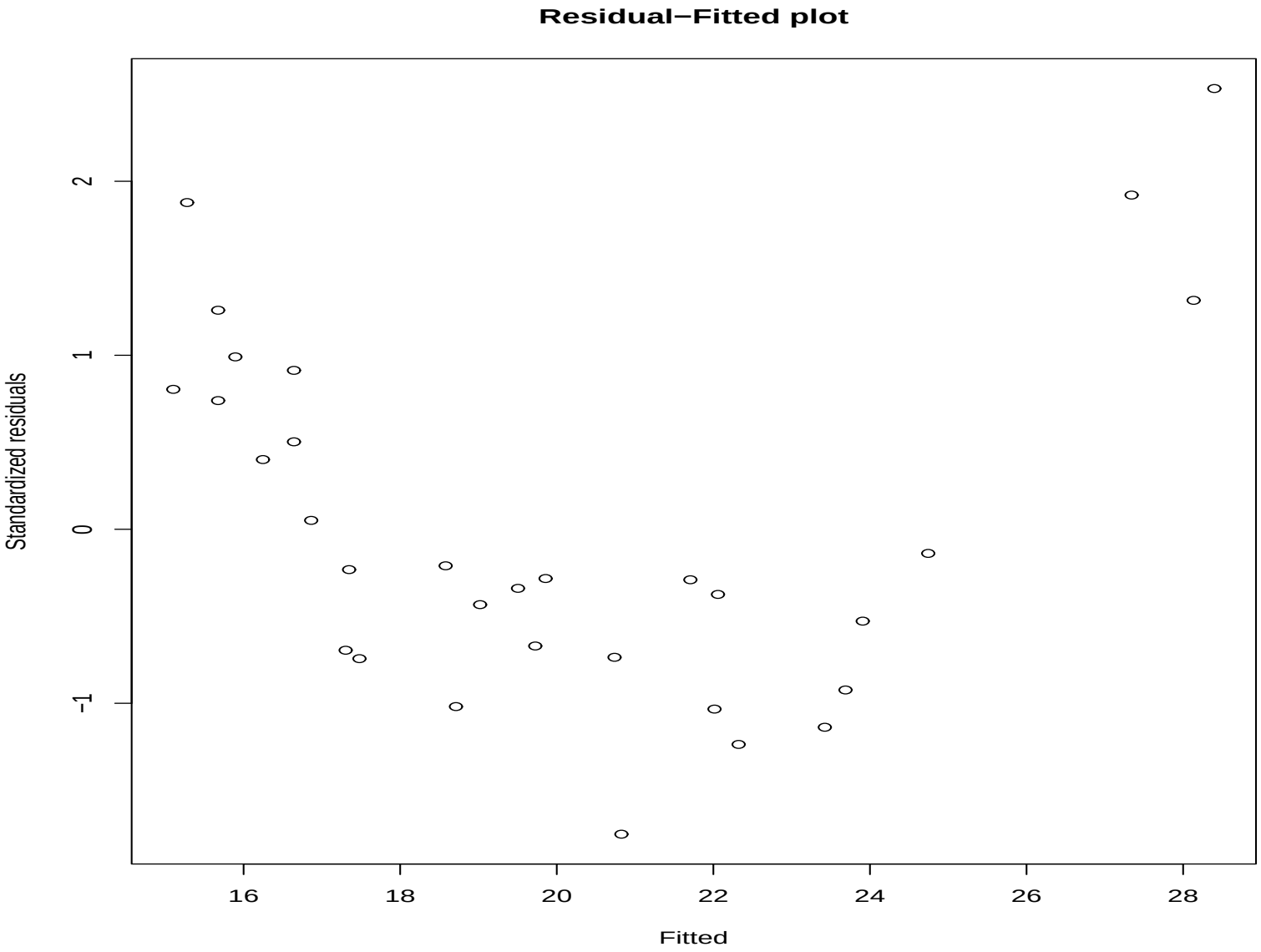
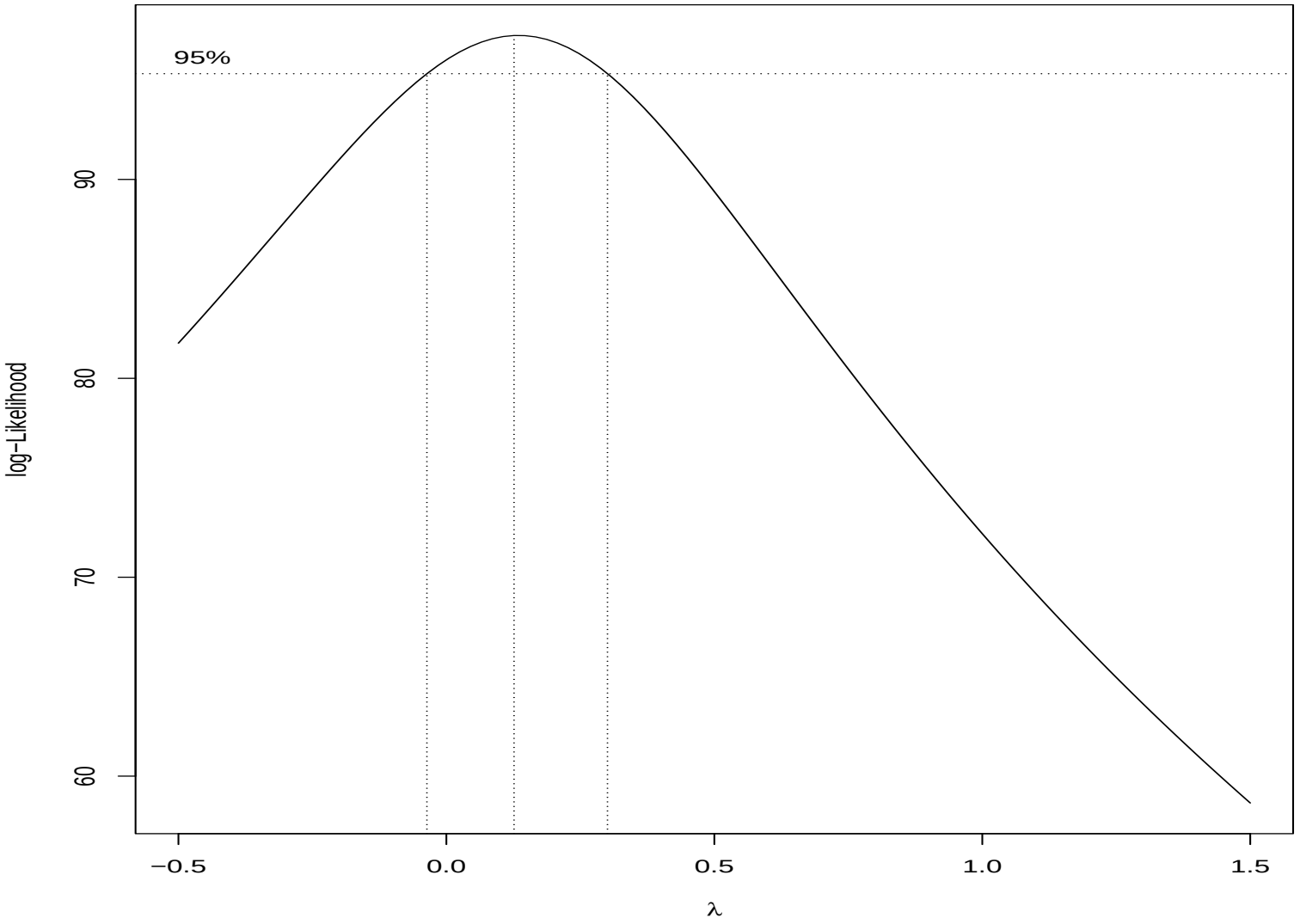


Figure 26: Box-Cox Transformation



- Loglikelihood reaches maximum when λ between (-0.01) to 0.25
- Prefer a log transformation of $\lambda = 0$ (easier for interpretation)

Figure 27: Hooked data with log transformation on Pres: Scatterplot of LPres vs Temp with fitted regression line

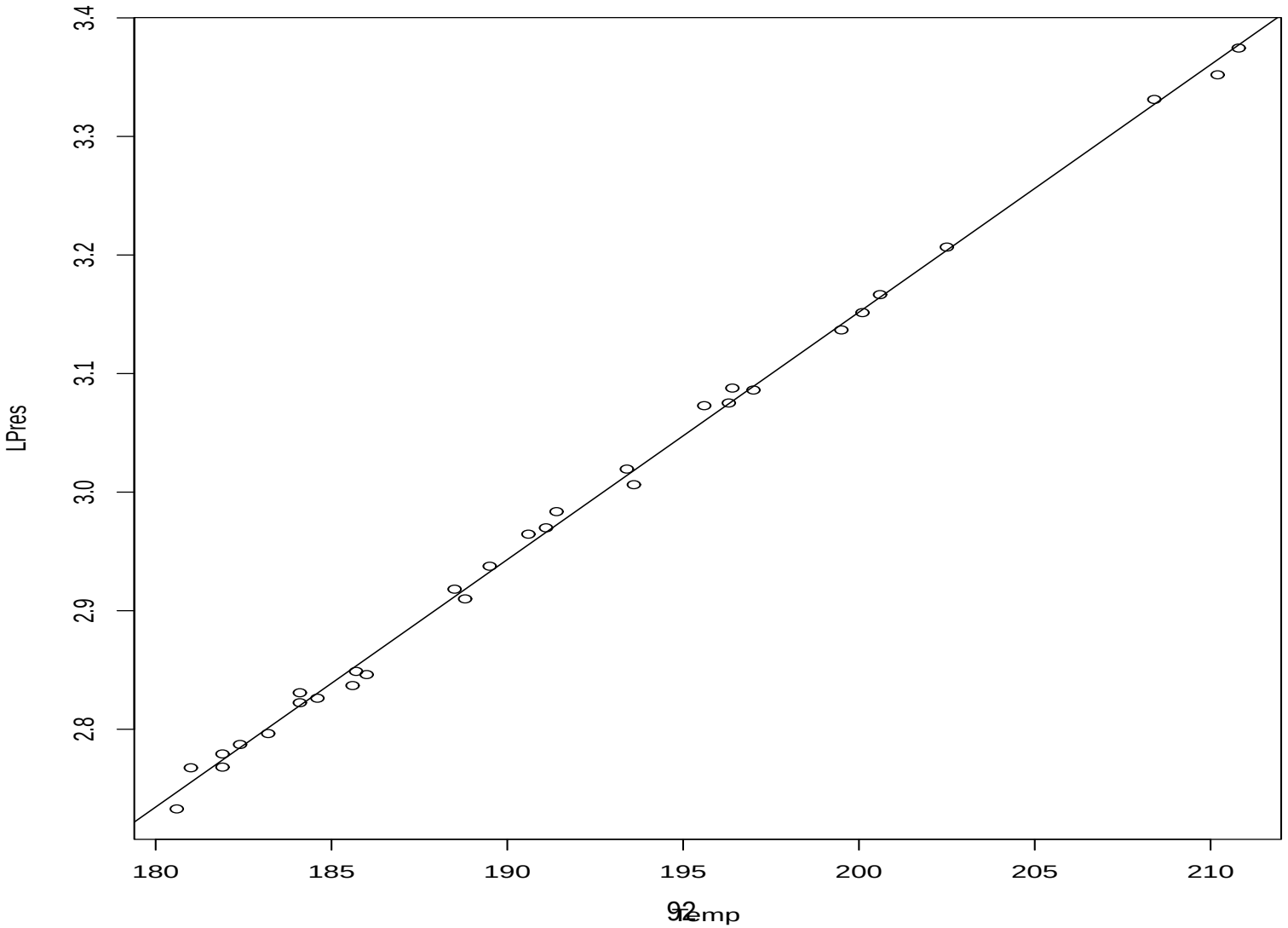


Table 2: Table of coefficient: log Hooker data

Predictor	$\hat{\beta}_k$	$SE(\hat{\beta}_k)$	t	P
Constant	-1.0221374	0.0336450	-30.38	$< 2e-16$
Temp	0.0208698	0.0001753	119.08	$< 2e-16$

Figure 28: Standardized residuals versus predicted values, logs of Hooker data

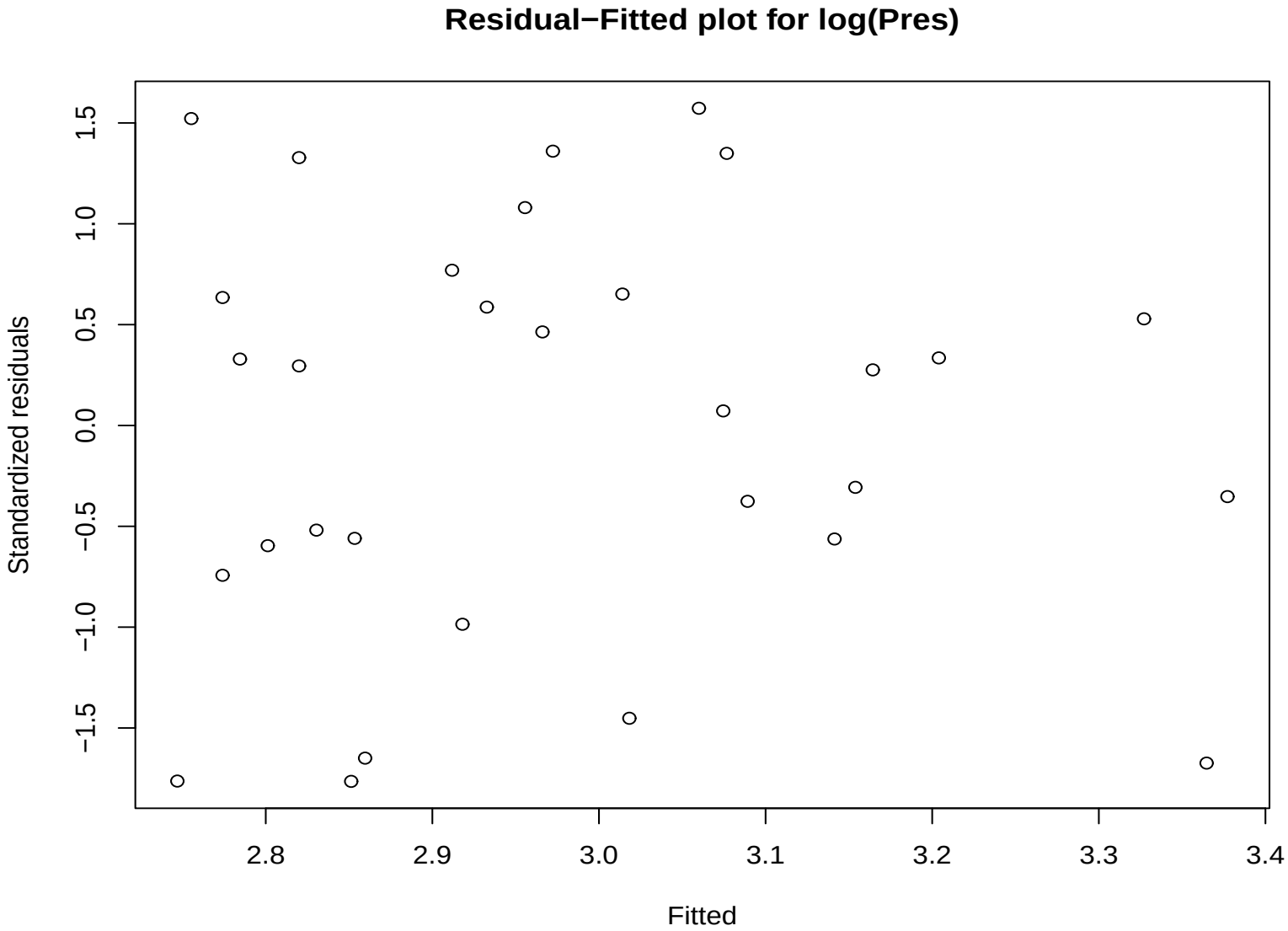
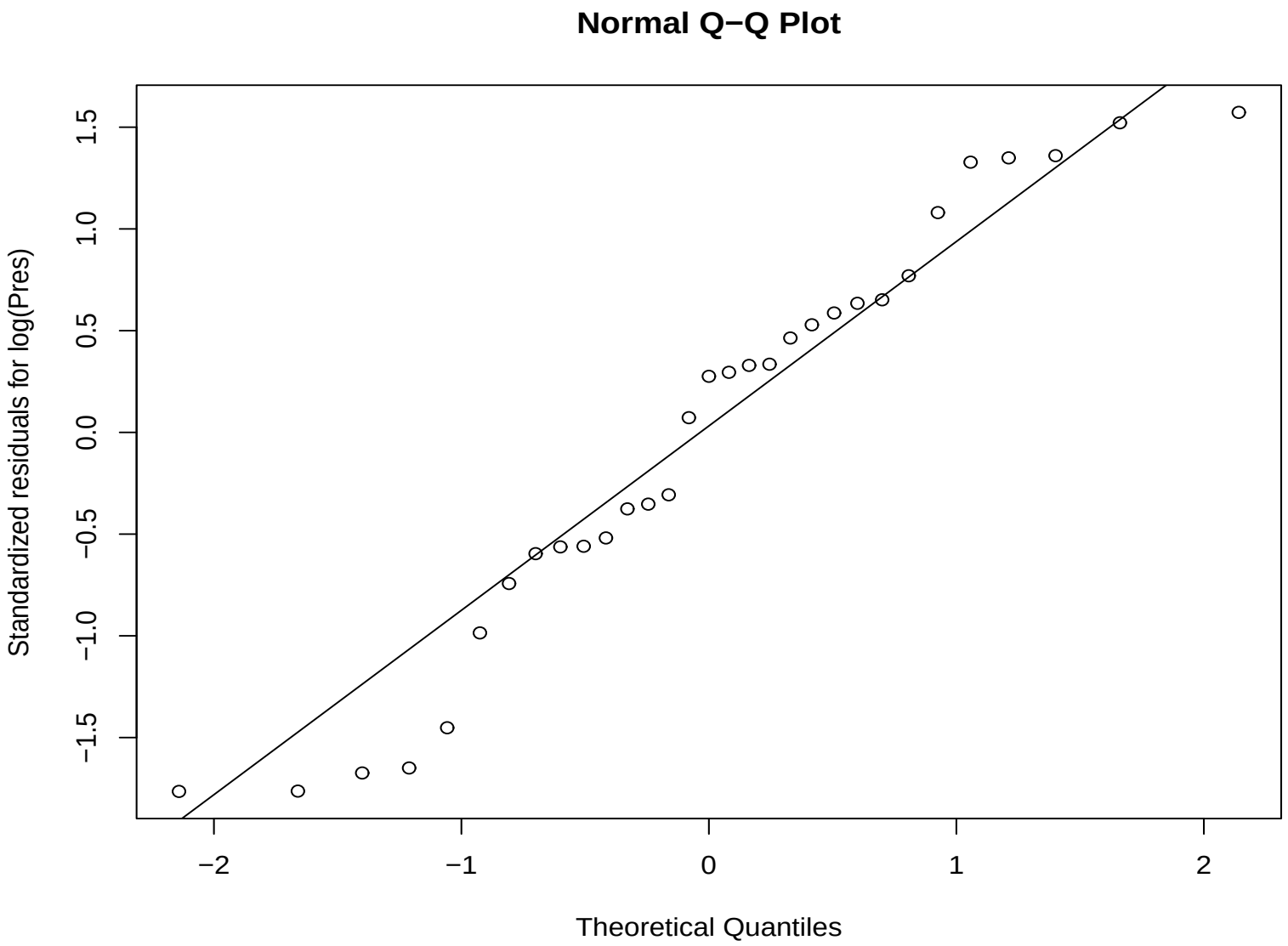


Figure 29: Normal plot for logs of Hooker data



Note:

- It may be desirable to introduce a constant into a transformation of Y, such as when Y may be negative.
- When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient. While such a transformation may stabilize the error variance, it will also change the linear relationship to a curvilinear one. A transformation on X may also be required.