Chapter 15

Multiple Linear Regression

15.1 Introduction

We have considered the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where $\epsilon \sim \text{Gaussian}(0, \sigma^2)$. We now extend this basic model to include multiple independent variables $x_1, x_2, ..., x_k$. This is more realistic because Y often depends on multiple variables (not just one). Specifically, we consider models of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

We call this a multiple linear regression model.

There are now $p = k + 1$ regression parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$. In simple linear regression, $k = 1$ and $p = 2$. The regression parameters describe the population for which this model is applicable. They are unknown and are to be estimated with the observed data; i.e., based on a sample from the population.

We continue to assume that $\epsilon \sim \text{Gaussian}(0, \sigma^2)$. We also assume that the independent variables $x_1, x_2, ..., x_k$ are fixed and are measured without error.

Example 15.1. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study from the LaTrobe Valley of Victoria, Australia, specimens of cheddar cheese were analyzed for their chemical composition and were subjected to taste tests. For each specimen, the taste Y was obtained by combining the scores from several tasters. Data were collected on the following variables:

$$\begin{aligned} Y &= \text{taste score (TASTE)} \\ x_1 &= \text{concentration of acetic acid (ACETIC)} \\ x_2 &= \text{concentration of hydrogen sulfide (H2S)} \\ x_3 &= \text{concentration of lactic acid (LACTIC).} \end{aligned}$$

The variables ACETIC and H2S were measured on the log scale. The variable LACTIC has not been transformed. Table 15.1.1 contains concentrations of the chemicals in a random sample of $n = 30$ specimens of cheddar cheese and the corresponding taste scores. Researchers postulate

Specimen	TASTE	ACETIC	H2S	LACTIC	Specimen	TASTE	ACETIC	H2S	LACTIC
1	12.3	4.543	3.135	0.86	16	40.9	6.365	9.588	1.74
2	20.9	5.159	5.043	1.53	17	15.9	4.787	3.912	1.16
3	39.0	5.366	5.438	1.57	18	6.4	5.412	4.700	1.49
4	47.9	5.759	7.496	1.81	19	18.0	5.247	6.174	1.63
5	5.6	4.663	3.807	0.99	20	38.9	5.438	9.064	1.99
6	25.9	5.697	7.601	1.09	21	14.0	4.564	4.949	1.15
7	37.3	5.892	8.726	1.29	22	15.2	5.298	5.220	1.33
8	21.9	6.078	7.966	1.78	23	32.0	5.455	9.242	1.44
9	18.1	4.898	3.850	1.29	24	56.7	5.855	10.20	2.01
10	21.0	5.242	4.174	1.58	25	16.8	5.366	3.664	1.31
11	34.9	5.740	6.142	1.68	26	11.6	6.043	3.219	1.46
12	57.2	6.446	7.908	1.90	27	26.5	6.458	6.962	1.72
13	0.7	4.477	2.996	1.06	28	0.7	5.328	3.912	1.25
14	25.9	5.236	4.942	1.30	29	13.4	5.802	6.685	1.08
15	54.9	6.151	6.752	1.52	30	5.5	6.176	4.787	1.25

Table 15.1.1: Cheese data. ACETIC, H2S, and LACTIC are independent variables. The response variable is TASTE.

that each of the three variables $x_1, x_2, and x_3$ is important in describing TASTE and consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

to model this relationship.

15.2 Least Squares Estimation

Suppose we have a random sample of n individuals from a population. In a multiple linear regression application, we can envision the observed data as follows:

Individual	Y	x_1	x_2	$\cdots$	x_k
1	Y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	Y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Y_1	x_{11}	x_{12}	$\cdots$	x_{1k}

Each of the n individuals contributes a response Y and a value of each of the independent variables. The value

$$x_{ij} = \text{measurement on the jth independent variable for the ith individual}$$

for $i = 1, 2, ..., n$ and $j = 1, 2, ..., k$. For the n individuals, we write

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$.

Matrix representation: To estimate the population parameters $\beta_0, \beta_1, \beta_2, ..., \beta_k$, we again use least squares. In doing so, it is advantageous to express multiple linear regression models in terms of matrices and vectors. This streamlines notation and makes the presentation easier. Define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

With these definitions, the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$, can be expressed equivalently as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

In this representation,

- $\mathbf{Y}$ is an $n \times 1$ (random) vector of responses
- $\mathbf{X}$ is an $n \times p$ (fixed) matrix of independent variable measurements ($p = k + 1$)
- $\boldsymbol{\beta}$ is a $p \times 1$ (fixed) vector of unknown population regression parameters
- $\boldsymbol{\epsilon}$ is an $n \times 1$ (random) vector of unobserved errors.

Illustration: Here are $\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ for the cheese data in Example 15.1. Recall there are $n = 30$ individuals and $k = 3$ independent variables. The data are in Table 15.1.1.

$$\mathbf{Y} = \begin{pmatrix} 12.3 \\ 20.9 \\ \vdots \\ 5.5 \end{pmatrix}_{30\times 1}, \ \mathbf{X} = \begin{pmatrix} 1 & 4.543 & 3.135 & \cdots & 0.86 \\ 1 & 5.159 & 5.043 & \cdots & 1.53 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 6.176 & 4.787 & \cdots & 1.25 \end{pmatrix}_{30\times 4}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}_{4\times 1}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{30\times 1}.$$

The notion of least squares is the same in multiple linear regression as it was in simple linear regression. Specifically, we want to find the values of $\beta_0, \beta_1, \beta_2, ..., \beta_k$ that minimize

$$Q(\beta_0, \beta_1, \beta_2, ..., \beta_k) = \sum_{i=1}^{n} \left[Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})\right]^2.$$

First recognize that

$$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

is the inner (dot) product of the ith row of $\mathbf{X}$ and $\boldsymbol{\beta}$. Therefore,

$$Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

is the ith entry in the difference vector $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. The objective function Q is

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

the inner (dot) product of $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ with itself; i.e., the squared length of $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$.

We want to find the value of $\boldsymbol{beta}$ that minimizes $Q(\boldsymbol{beta})$. Because $Q(\boldsymbol{beta})$ is a scalar function of the $p = k + 1$ elements of $\boldsymbol{beta}$, it is possible to use calculus to determine the values of the p elements that minimize it. Formally, we can take p partial derivatives, one with respect to each of $\beta_0, \beta_1, \beta_2, ..., \beta_k$, and set these equal to zero. Using the calculus of matrices, we can write this resulting system of p equations (and p unknowns) as follows:

$$\mathbf{X}'\mathbf{X}\boldsymbol{beta} = \mathbf{X}'\mathbf{Y}.$$

These are called the **normal equations**. Provided that $\mathbf{X}'\mathbf{X}$ is full rank, the (unique) solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}_{4\times 1},$$

This is the **least squares estimator** of $\boldsymbol{\beta}$.

Technical note: For the least squares estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

to be unique, we need $\mathbf{X}$ to be of full column rank; i.e., $r(\mathbf{X}) = p = k + 1$. This will occur when there are no linear dependencies among the columns of $\mathbf{X}$. If $r(\mathbf{X}) < p$, then $\mathbf{X}'\mathbf{X}$ does not have a unique inverse, and the normal equations can not be solved uniquely. Statistical software such as R will alert you when $\mathbf{X}'\mathbf{X}$ is not full rank.

Remark: These estimators are equivalent to the maximum likelihood estimators. In OLS regression analysis, we assume that the errors are approximately Gaussian; e.g.,

$$\epsilon_i = (Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})) \sim \text{Gaussian}\left(\mu_\epsilon = 0, \sigma_\epsilon = \sigma\right),$$

where σ is estimated using the observations.

Under this assumption we take $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k)$ to be

$$\underset{(\hat{\beta}_0,\hat{\beta}_1,...,\hat{\beta}_k)\in\mathbb{R}^{(k+1)}}{\arg\max} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\sum_{i=1}^{n} \frac{(Y_i-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}))^2}{2\sigma^2}},$$

the maximum likelihood estimates under the normality assumption on ϵ_i.

We now use R to calculate the least squares estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ for the cheese data in Example 15.1:

```
> dat.cheese<-data.frame(
+    taste=c(12.3,20.9,39.0,47.9,5.6,25.9,37.3,21.9,18.1,21.0,34.9,57.2,0.7,25.9,54.9,
+    40.9,15.9,6.4,18.0,38.9,14.0,15.2,32.0,56.7,16.8,11.6,26.5,0.7,13.4,5.5),
+    acetic=c(4.543,5.159,5.366,5.759,4.663,5.697,5.892,6.078,4.898,5.242,5.740,6.446,
+    4.477,5.236,6.151,6.365,4.787,5.412,5.247,5.438,4.564,5.298,5.455,5.855,5.366,
+     6.043,6.458,5.328,5.802,6.176),
+    h2s=c(3.135,5.043,5.438,7.496,3.807,7.601,8.726,7.966,3.850,4.174,6.142,7.908,
```

```
+    2.996,4.942,6.752,9.588,3.912,4.700,6.174,9.064,4.949,5.220,9.242,
+    10.200,3.664,3.219,6.962,3.912,6.685,4.787),
+    lactic=c(0.86,1.53,1.57,1.81,0.99,1.09,1.29,1.78,1.29,1.58,1.68,1.90,1.06,1.30,
+    1.52,1.74,1.16,1.49,1.63,1.99,1.15,1.33,1.44,2.01,1.31,1.46,1.72,1.25,1.08,1.25)
+ )
> mod.cheese<-lm(taste~acetic+h2s+lactic,data=dat.cheese)
> coef(mod.cheese)
(Intercept)      acetic          h2s      lactic
-28.8767658   0.3280084   3.9117818  19.6696760
```

This output gives the value of the least squares estimate

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}_{4\times 1} = \begin{pmatrix} -28.8768 \\ 0.3280 \\ 3.9118 \\ 19.6697 \end{pmatrix},$$

Therefore, the estimated regression model based on the data is

$$\hat{Y} = -28.8768 + 0.3280x_1 + 3.9118x_2 + 19.6697x_3,$$

or, in other words,

$$\hat{\text{TASTE}} = -28.8768 + 0.3280\text{ACETIC} + 3.9118\text{H2S} + 19.6697\text{LACTIC}.$$

15.3 Estimating the error variance

In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where $\epsilon \sim \text{Gaussian}(0, \sigma^2)$, we now turn our attention to estimating σ^2, the **error variance**.

The residual sum of squares is given by

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

just as in simple linear regression. In matrix notation, we can write this as

$$\begin{aligned} SS_{res} &= (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{e}'\mathbf{e}. \end{aligned}$$

- The $n \times 1$ vector $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ contains the least squares fitted values.
- The $n \times 1$ vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ contains the least squares residuals.
- R calculates these upon request; e.g., we can ask `R` for these values as follows.

```
> fitted(mod.cheese)
1         2         3         4         5         6         7         8
1.792734 22.637150 25.036988 37.937065  7.017870 24.165298 32.563950 39.290146
9        10        11        12        13        14        15        16
13.164061 20.248520 30.077222 41.544331  5.161283 17.743291 29.451072 44.942408
17        18        19        20        21        22        23        24
10.813125 20.591607 29.057207 47.505989 14.599800 19.441193 37.389541 52.479947
25        26        27        28        29        30
12.983371 14.415142 34.307180 12.760848 20.419850 16.461809
> residuals(mod.cheese)
1           2           3           4           5           6           7
10.5072662  -1.7371496  13.9630119   9.9629354  -1.4178700   1.7347015   4.7360502
8           9          10          11          12          13          14
-17.3901465   4.9359386   0.7514803   4.8227780  15.6556686  -4.4612828   8.1567092
15          16          17          18          19          20          21
25.4489278  -4.0424079   5.0868749 -14.1916075 -11.0572071  -8.6059895  -0.5998002
22          23          24          25          26          27          28
-4.2411929  -5.3895410   4.2200534   3.8166285  -2.8151417  -7.8071802 -12.0608485
29          30
-7.0198505 -10.9618087
```

The residual mean squares

$$MS_{res} = \frac{SS_{res}}{n-p}$$

is an unbiased estimator of σ^2, that is,

$$E(MS_{res}) = \sigma^2.$$

The quantity

$$\hat{\sigma} = \sqrt{MS_{res}} = \sqrt{\frac{SS_{res}}{n-p}}$$

estimates σ and is called the **residual standard error**. This result is analogous to the simple linear regression result. The only difference is in the divisor in MS_{res}.

15.4 Analysis of variance for linear regression

The following algebraic identity arises for a linear regression model fit (simple or multiple):

$$\begin{aligned} SS_{total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \\ &= SS_{reg} + SS_{res}. \end{aligned}$$

Remark: This essentially tells us that the total variability in the response variable is partitioned into the variability described by the regression line and the rest which is left unexplained and modeled through the residual term.

This information is used to produce an analysis of variance (ANOVA) table.

Source	df	SS	MS	F
Regression	$p-1$	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{p-1}$	$F = \frac{MS_{reg}}{MS_{res}}$
Residual	$n-p$	SS_{res}	$MS_{res} = \frac{SS_{res}}{n-p}$	
Total	$n-1$	SS_{total}		

This table summarizes how the variability in the response data is partitioned.

- SS_{total} is the total sum of squares. It measures the total variation in the response data.
- SS_{reg} is the regression sum of squares. It measures the variation in the response data explained by the estimated regression model.
- SS_{res} is the residual sum of squares. It measures the variation in the response data not explained by the estimated regression model.

We also see that the degrees of freedom (df) add down.

- The degrees of freedom for SS_{total} is the divisor in the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{SS_{total}}{n-1}.$$

- The degrees of freedom for SS_{reg} is $p-1$, the number of independent variables in the model fit (recall $p = k+1 \Rightarrow k = p-1$).
- The degrees of freedom for SS_{res} is the divisor needed to create an unbiased estimator of σ^2. Recall that

$$MS_{res} = \frac{SS_{res}}{n-p}$$

is an unbiased estimator of σ^2.

The **Mean Squares** (MS) are the sums of squares divided by their degrees of freedom, and the F statistic is formed by taking the ratio of MS_{reg} and MS_{res}.

Cheese data: I used R to calculate the ANOVA table for the cheese data:

```
> (n<-nrow(dat.cheese))
[1] 30
> (p<-length(coef(mod.cheese))) #four coefficients
[1] 4
> y.bar<-mean(dat.cheese$taste)
> (SS.res<-sum(residuals(mod.cheese)^2))
[1] 2668.378
> (MS.res<-SS.res/(n-p))
[1] 102.6299
> (SS.reg<-sum((fitted(mod.cheese)-y.bar)^2))
[1] 4994.509
> (MS.reg<-SS.reg/(p-1))
```

```
[1] 1664.836
> (SS.total<-SS.reg+SS.res)
[1] 7662.887
> (F.stat<-MS.reg/MS.res)
[1] 16.22174
```

Source	df	SS	MS	F
Regression	$4-1=3$	4994.509	$MS_{reg}=1664.836$	$F=16.22174$
Residual	$30-4$=26	2668.378	$MS_{res}=102.6299$	
Total	$30-1=29$	7662.887		

Remark: Above we calculated the table's values manually. The reason we had to do that is R does something different in displaying the analysis of variance – it breaks down the regression sums of squares further. This gives us very different information than the condensed table.

```
> anova(mod.cheese) #breaks it down farther
Analysis of Variance Table

Response: taste
          Df  Sum Sq Mean Sq F value    Pr(>F)
acetic     1 2314.14 2314.14 22.5484 6.528e-05 ***
h2s        1 2147.11 2147.11 20.9209 0.0001035 ***
lactic     1  533.26  533.26  5.1959 0.0310870 *
Residuals 26 2668.38  102.63
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The convention used by R is to partition the regression sum of squares

$$SS_{reg}=4994.508$$

into sums of squares for each of the three independent variables ACETIC, H2S, and LACTIC, as they are added to the model sequentially. These are called sequential sums of squares.

Note that, after rounding,

$$\begin{aligned} SS_{reg} &= 4994.51 = 2314.14+2147.11+533.26 \\ &= SS(\text{ACETIC})+SS(\text{H2S—ACETIC})+SS(\text{LACTIC—ACETIC, H2S}). \end{aligned}$$

where the following are called **sequential sums of squares** or **Type I sums of squares**:

- SS(ACETIC) is the sum of squares added when compared to a model that includes only an intercept term.
- SS(H2S—ACETIC) is the sum of squares added when compared to a model that includes an intercept term and ACETIC.
- SS(LACTIC—ACETIC, H2S) is the sum of squares added when compared to a model that includes an intercept term, ACETIC, and H2S.

In other words, we can use the sequential sums of squares to assess the impact of adding independent variables ACETIC, H2S, and LACTIC to the model in sequence. The p-values provided by R help you assess the statistical significance of each independent variable as you add them. Small p-values suggest statistical significance.

Remark: If you change the order of the independent variables in the lm function, then you will get a different sequential sum of squares partition. For example,

```
> mod2.cheese<-lm(taste~h2s+lactic+acetic,data=dat.cheese)
> anova(mod2.cheese)
Analysis of Variance Table

Response: taste
Df Sum Sq Mean Sq F value    Pr(>F)
h2s        1 4376.8  4376.8 42.6468 6.356e-07 ***
lactic     1  617.1   617.1  6.0131   0.02123 *
acetic     1    0.6     0.6  0.0054   0.94193
Residuals 26 2668.4   102.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This table suggests that ACETIC does not add significantly to a regression model that already includes H2S and LACTIC (p-value = 0.9419). Note that the previous sequential sum of squares partition does not enable us to see this.

Remark: There are two other types of sums of squares – Type II and Type III conditional sums of squares. This type of sums of squares remove the sequential nature of the Type I sequential sums of squares in a slightly different way. These conditional sums of squares tell us if the model with a predictor of interest helps predict the response above and beyond the model with everything else but that predictor.

Type I: These are the sequential sums of squares which tell us whether or not predictors explain a sufficient amount of variance in the response in the presence of the predictors already in the model; therefore, order matters.

- mod:
  - SS(ACETIC)
  - SS(H2S|ACETIC)
  - SS(LACTIC|ACETIC, H2S)
- mod2:
  - SS(H2S)
  - SS(LACTIC|H2S)
  - SS(ACETIC|LACTIC, H2S)

Type II: These are **conditional sums of squares** where no significant interaction is assumed. An ANOVA table that uses Type II conditional sums of squares tells us whether or not predictors explain a sufficient amount of variance in the response in the presence of all the other predictors (not including interactions); therefore, order does not matter.

- mod:
  - SS(ACETIC|H2S, LACTIC)
  - SS(H2S|ACETIC, LACTIC)
  - SS(LACTIC|ACETIC, H2S)
- mod2:
  - SS(H2S|ACETIC, LACTIC)
  - SS(LACTIC|ACETIC, H2S)
  - SS(ACETIC|H2S, LACTIC)

Type III: These are **conditional sums of squares** where a significant interaction is assumed. An ANOVA table that uses Type III conditional sums of squares tells us whether or not predictors explain a sufficient amount of variance in the response in the presence of all the other predictors (including interactions); therefore, order does not matter. Suppose we added a ACETIC×LACTIC interaction into the model that was significant, the Type III conditional sums of squares would be as follows. In reality, the interaction is not significant so we would proceed with Type II conditional sums of squares.

- mod:
  - SS(ACETIC|H2S, LACTIC, ACETIC×LACTIC)
  - SS(H2S|ACETIC, LACTIC, ACETIC×LACTIC)
  - SS(LACTIC|ACETIC, H2S, ACETIC×LACTIC)
- mod2:
  - SS(H2S|ACETIC, LACTIC, ACETIC×LACTIC)
  - SS(LACTIC|ACETIC, H2S, ACETIC×LACTIC)
  - SS(ACETIC|H2S, LACTIC, ACETIC×LACTIC)

We can ask for the Type II or Type III conditional sums of squares using the "car" package for R.

```
> Anova(mod.cheese,type = 2)
Anova Table (Type II tests)

Response: taste
          Sum Sq Df F value   Pr(>F)
acetic      0.56  1  0.0054 0.941932
h2s       1007.69  1  9.8187 0.004246 **
lactic     533.26  1  5.1959 0.031087 *
Residuals 2668.38 26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This table suggests that ACETIC does not add significantly to a regression model that already includes H2S and LACTIC (p-value = 0.9419). Note that the first sequential sum of squares

partition did not enable us to see this, and this p-value is the same as seen in the first sequential sum of squares partition in which ACETIC was added to the model last.

Remark: This discussion about the different types of sums of squares is important. When working with collaborators in different fields, we might find ourselves working with people that use a different statistical software. We would be highly confused about mismatched ANOVA table outputs if `R` returns Type I sums of squares and a different program, like SPSS, returns Type II or Type III sums of squares.

15.4.1 The Overall F Test

In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

the F statistic in the condensed ANOVA table can be used to test

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_a : \text{at least one of the } \beta_j\text{'s is not zero.}$$

In other words, F tests if at least one of the independent variables $x_1, x_2, ..., x_k$ is important in describing the response Y in the population (H_0 says "no"; H_a says "yes"). If H_0 is rejected, we do not know which one or how many of the β_j' are nonzero; only that at least one is.

When H_0 is true, both MS_{reg} and MS_{res} are unbiased estimators of σ^2. Therefore, when H_0 is true,

$$F = \frac{MS_{reg}}{MS_{res}} \approx 1.$$

The sampling distribution of F when H_0 is true is

$$F = \frac{MS_{reg}}{MS_{res}} \sim F(p-1, n-p).$$

Recall that the mean of an F distribution is around 1. Therefore, values of F in the center of this distribution are consistent with H_0; large values of F (i.e., out in the right tail) are consistent with H_a; and unusually small values of F (i.e., close to zero) might indicate there is a violation of our statistical assumptions or we have fit the incorrect model.

Remark: This is precisely what we discussed when learning about the one-way ANOVA.

For the cheese data in Example 15.1, the F statistic is used to test

$$H_0 : \beta_1 = \beta_2 = \beta 3 = 0$$
$$H_a : \text{at least one of the } \beta_j\text{'s is not zero.}$$

Based on the F statistic (F = 16.22), and the corresponding p-value (p-value < 0.0001), we have strong evidence to reject H_0. See also Figure 15.4.1. We conclude that at least one of the independent variables (ACETIC, H2S, LACTIC) is important in describing TASTE in the population. In the next section, we learn how to investigate the population-level effects of each variable separately.

```
> ggplot(data=ggdat,aes(x=f,y=f1))+
+   geom_line()+
```

```
+   geom_ribbon(data=subset(ggdat,f>=qf(1-alpha,df1=p-1,df2=n-p)),aes(ymax=f1),ymin=0,
+               fill="grey",color=NA)+
+   geom_ribbon(data=subset(ggdat,f>=f.obs),aes(ymax=f1),ymin=0,
+               fill="red",color=NA,alpha=0.25)+
+   geom_point(data=ggdat.highlight,aes(x=x,y=y),color="red")+
+   geom_hline(yintercept = 0)+
+   theme_bw()+
+   xlab("f")+
+   ylab("Density")+
+   ggtitle("Regression ANOVA F Test",
+           subtitle=bquote(H[0]*":"~beta[1]==beta[2]==beta[3]==0~"versus"~
+                H[a]*": that at least one is different"))+
+   annotate("text", x=4, y=0.075,
+            label= deparse(bquote(alpha==0.05)),parse=T,size=3.5)+
+   annotate("text", x=f.obs, y=0.1, label="Observation \n P-value<0.0001",size=3.5)
```

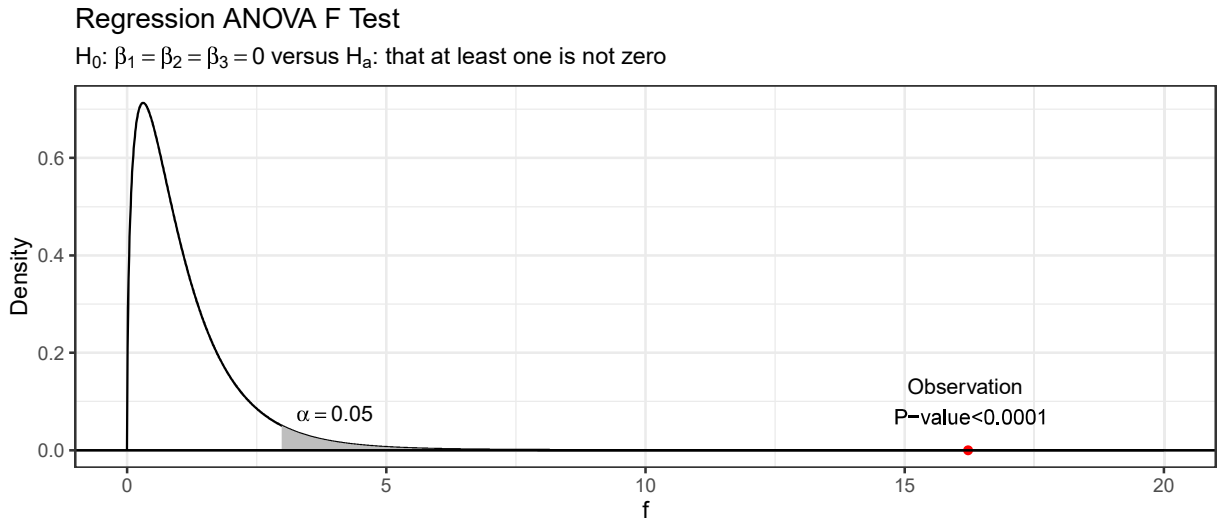

Figure 15.4.1: Cheese data: $F(3, 26)$ PDF. This is the sampling distribution of F when H_0 is true. A red point at the observation $F = 16.22$ has been added.

In the regression analysis of variance,

$$SS_{total} = SS_{reg} + SS_{res}.$$

Therefore, the proportion of the total variation in the response data explained by the estimated regression model is

$$R^2 = \frac{SS_{reg}}{SS_{total}}.$$

This statistic is called the **coefficient of determination**. By construction,

$$0 \leq R^2 \leq 1.$$

In general, the larger the R^2, the better the estimated regression model explains the variability in the response data.

For the cheese data in Example 15.1, recall the ANOVA table presented earlier:

Source	df	SS	MS	F
Regression	$4-1=3$	4994.509	$MS_{reg}=1664.836$	$F=16.22174$
Residual	$30-4$=26	2668.378	$MS_{res}=102.6299$	
Total	$30-1=29$	7662.887		

Therefore, the coefficient of determination is

$$R^2 = \frac{SS_{reg}}{SS_{total}} = \frac{4994.509}{7662.887} \approx 0.652.$$

This means that about 65.2 percent of the variability in the TASTE data is explained by the linear regression model that includes ACETIC, H2S, and LACTIC. The remaining 34.8 percent of the variability in the taste data is explained by other sources.

Warning: It is important to understand what R^2 measures and what it does not. Its value is computed under the assumption that the regression model is correct and assesses how much of the variation in the response is attributed to that relationship.

If R^2 is small, it may be that there is just a lot of random inherent variation in the data. Although the estimated regression model is reasonable, it can explain only so much of the overall variation. Alternatively, R^2 may be large (e.g., close to 1) but for an estimated model that is not appropriate for the data. A better model may exist.

15.5 Inference for Individual Regression Parameters

In the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

where $\epsilon \sim \text{Gaussian}(0, \sigma^2)$, we are interested in writing **confidence intervals** for individual regression parameters β_j.

This can help us assess the importance of using the independent variable x_j in a model that includes the other independent variables. That is, inference regarding the population parameter β_j is always conditional on the other variables being included in the model.

Under our linear regression model assumptions, a $100(1-\alpha)$ percent confidence interval for β_j, for $j = 0, 1, 2, ..., k$, is given by

$$\hat{\beta}_k \pm t_{n-p,1-\alpha/2}\sqrt{MS_{res}c_{jj}},$$

where $\hat{\beta}_j$ is the least squares estimate of β_j, MS_{res} is our estimate of the error variance σ^2 and $c_{jj} = (\mathbf{X}'\mathbf{X})^{-1}_{jj}$ is the corresponding diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix. The value $t_{n-p,1-\alpha/2}$ is the upper $\alpha/2$ quantile from the $t(n-p)$ distribution.

Note the familiar form of the interval:

$$\underbrace{\text{Point Estimate}}_{\hat{\beta}_k} \pm \underbrace{\text{Quantile}}_{t_{n-p,1-\alpha/2}} \underbrace{\text{Standard Error}}_{\sqrt{MS_{res}c_{jj}}}.$$

We interpret the interval in the same way:

"We are $100(1-\alpha)$ percent confident that the population parameter β_j is in this interval."

Of particular interest is the value $\beta_j = 0$. If the confidence interval for β_j contains "0," this suggests (at the population level) that the independent variable x_j does not significantly add to a model that contains the other independent variables. If the confidence interval for β_j does not contain "0," this suggests (at the population level) that the independent variable xj does significantly add to a model that contains the other independent variables.

We can use the `confint()` function in `R` to calculate confidence intervals for the population regression parameters:

```
> confint(mod.cheese,level=0.95)
                 2.5 %    97.5 %
(Intercept) -69.443161 11.689630
acetic       -8.839009  9.495026
h2s           1.345693  6.477870
lactic        1.932318 37.407035
```

We will ignore the intercept confidence interval, which describes $E(Y)$ when $x_1 = x_2 = x_3 = 0$, a nonsensical quantity. Here is how we interpret the other confidence intervals.

- We are 95 percent confident that β_1 (the population parameter for ACETIC) is between -8.84 and 9.50. This interval includes "0." Therefore, ACETIC does not significantly add to a model that includes H2S and LACTIC. This reaffirms what we saw in the sequential SS (Type I) when ACETIC was added last, and the conditional SS (Type II).
- We are 95 percent confident that β_2 (the population parameter for H2S) is between 1.35 and 6.48. This interval does not include "0." Therefore, H2S does significantly add to a model that includes ACETIC and LACTIC.
- We are 95 percent confident that β_3 (the population parameter for LACTIC) is between 1.93 and 37.41. This interval does not include "0." Therefore, LACTIC does significantly add to a model that includes ACETIC and H2S.

15.6 Confidence and prediction intervals for a given $x = x_0$

We would like to create $100(1-\alpha)$ percent intervals for the mean $E(Y|x_0)$ and for the new value $Y^*(x_0)$. As in simple linear regression, the former is called a confidence interval (because it is for a mean response) and the latter is called a prediction interval (because it is for a new random variable).

Suppose we are interested estimating $E(Y|x_0)$ and predicting a new $Y^*(x_0)$ when ACETIC = 5.5, H2S = 6.0, and LACTIC = 1.4, so that

$$x_0 = \begin{pmatrix} 5.5 \\ 6.0 \\ 1.4 \end{pmatrix}.$$

We use `R` to compute the prediction and confidence intervals as follows.

```
> newData<-data.frame(acetic=5.5,h2s=6.0,lactic=1.4)
> predict(mod.cheese,newData,level=0.95,interval="confidence")
fit      lwr      upr
1 23.93552 20.04506 27.82597
> predict(mod.cheese,newData,level=0.95,interval="prediction")
fit      lwr      upr
1 23.93552 2.751379 45.11966
```

Note that the point estimate/prediction is

$$\begin{aligned}\hat{Y}(x_0) &= \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \hat{\beta}_3 x_{30} \\ &= -28.877 + 0.328(5.5) + 3.912(6.0) + 19.670(1.4) \\ &\approx 23.936.\end{aligned}$$

A 95 percent confidence interval for $E(Y|x_0)$ is (20.05, 27.83). When ACETIC = 5.5, H2S = 6.0, and LACTIC = 1.4, we are 95 percent confident that the population mean taste rating is between 20.05 and 27.83.

A 95 percent prediction interval for $Y^*(x_0)$, is (2.75, 45.12). When ACETIC = 5.5, H2S = 6.0, and LACTIC = 1.4, we are 95 percent confident that the taste rating for a new specimen will be between 2.75 and 45.12.

15.7 Model diagnostics (residual analysis)

We now discuss diagnostic techniques for linear regression (simple and multiple). The term "diagnostics" refers to the process of "checking the model assumptions." This is an important exercise because if the model assumptions are violated, then our analysis and all subsequent interpretations could be compromised.

Recall: We first recall the model assumptions on the error terms in the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$$

for $i = 1, 2, ..., n$. Specifically, we have made the following assumptions:

- $E(\epsilon_i) = 0$, for $i = 1, 2, ..., n$
- $var(\epsilon_i) = \sigma^2$, for $i = 1, 2, ..., n$, that is, the variance is constant
- the random variables ϵ_i are independent
- the random variables ϵ_i are Gaussian distributed.

In checking our model assumptions, we first have to deal with the obvious problem; namely, the error terms ϵ_i in the model are never observed. However, after fitting the model, we can calculate the residuals

$$e_i = Y_i - \hat{Y}_i,$$

where the ith fitted value

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}.$$

We can think of the residuals $e_1, e_2, ..., e_n$ as "proxies" for the error terms $\epsilon_1, \epsilon_2, ..., \epsilon_n$. Therefore, we can use the residuals to check our model assumptions instead.

To check the Gaussian assumption for the errors in linear regression, we can examine several graphs. In Figures 15.7.2 15.7.3, and 15.7.4 we see little evidence that the residuals are Gaussian distributed with mean 0 and constant variance. The histogram of the residuals is tightly fit by the Gaussian PDF, the empirical distribution is tightly fit by the Gaussian CDF, and the qq plot doesn't show any severe departures from normality because the plotted points follow a straight line (approximately). These figures are created using the following `R` code.

```
> ggdat.gaussian<-data.frame(x=seq(-25,35,0.01),
+                            f=dnorm(seq(-25,35,0.01),
+                                    #ei should have mean zero
+                                    mean=0,
+                                    #ei should have common variance
+                                    sd=summary(mod.cheese)$sigma))
> ggplot(data=ggdat,aes(x=e))+
+   geom_histogram(aes(y=..density..),bins=15,
+                  fill="lightblue",color="black")+
+   geom_density(aes(color="Loess Density Estimate"),size=1)+
+   geom_line(data=ggdat.gaussian,aes(x=x,y=f,color="Under Gaussian Assumption"),
+        linetype="dashed",size=1)+
+   theme_bw()+
+   xlab("Residual")+
+   ylab("Density")+
+   labs(color = "")
```

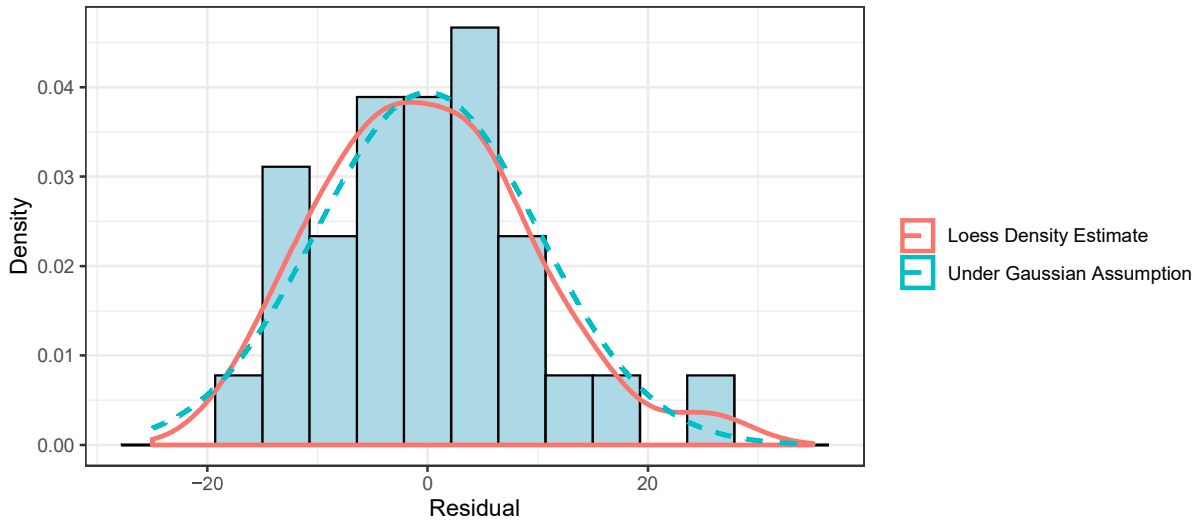

Figure 15.7.2: Cheese data: A histogram of the residuals from the multiple linear regression model. A loess density estimate (red) and the assumed Gaussian distribution with the estimated constant variance σ^2 (red).

```
> e.cdf.func<-ecdf(residuals(mod.cheese))
> e.cdf<-e.cdf.func(sort(residuals(mod.cheese)))
> ggdat<-data.frame(e=sort(residuals(mod.cheese)),
```

```
+                    e.cdf=e.cdf)
> ggdat.gaussian<-data.frame(x=seq(-25,35,0.01),
+                            CDF=pnorm(seq(-25,35,0.01),
+                                      #ei should have mean zero
+                                      mean=0,
+                                      #ei should have common variance
+                                      sd=summary(mod.cheese)$sigma))
> ggplot(data=ggdat,aes(x=e))+
+   geom_step(aes(y=e.cdf,color="Empirical CDF"))+
+   geom_line(data=ggdat.gaussian,aes(x=x,y=CDF,color="Under Gaussian Assumption"),
+        linetype="dashed",size=1)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Residual")+
+   ylab("Cumulative Density")+
+   labs(color = "")
```

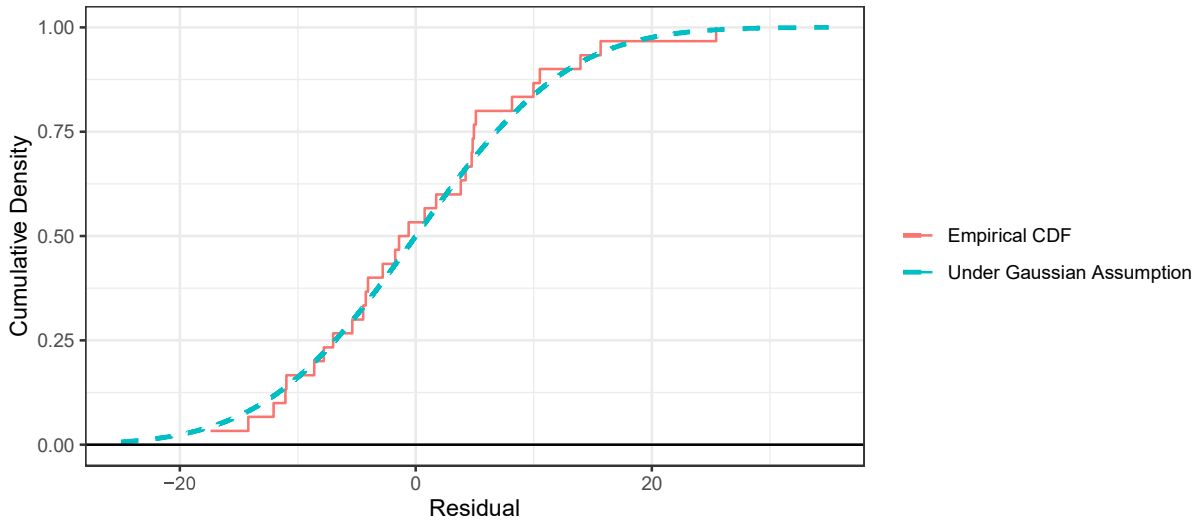

Figure 15.7.3: Cheese data: An empirical CDF of the residuals from the multiple linear regression model (red) and the assumed Gaussian distribution with the estimated constant variance σ^2 (red).

```
> library("qqplotr")
> ggplot(data=ggdat,aes(sample=scale(e)))+ #standardize e
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")
```

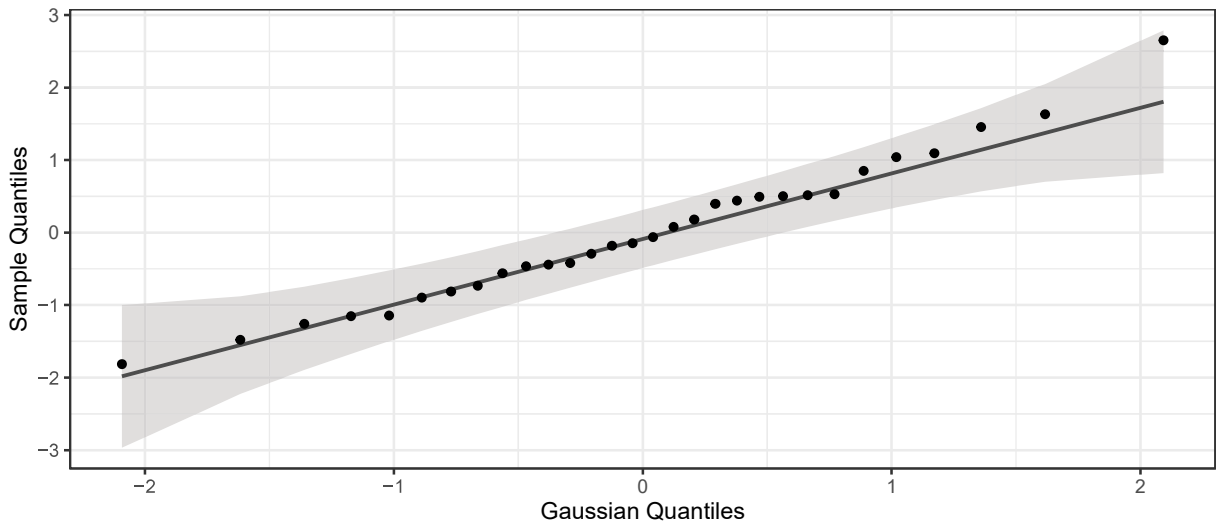

Figure 15.7.4: Cheese data: A Gaussian qq plot of the residuals from the multiple linear regression model.

If the normality assumption is violated in a linear regression analysis, this could affect population level inferences for regression parameters β_j and confidence/prediction intervals. Mild departures are generally not a problem unless the sample size is very small. Substantial departures from normality should raise concern.

A **residual plot** is a scatterplot of the residuals e_i (on the vertical axis) versus the predicted values $\hat{Y}_i$ (on the horizontal axis). A residual plot can be very useful in detecting the following violations:

- misspecifying the true regression function; i.e., a violation of the $E(\epsilon_i) = 0$ assumption
- non-constant variance (heteroscedasticity); i.e., a violation of the $var(\epsilon_i) = \sigma^2$ assumption
- correlated observations over time; i.e., a violation of the assumption that the ϵ_i's are independent random variables.

Mathematical arguments show that if all of the linear regression model assumptions hold, then the residuals and fitted values are independent. Therefore, if the residual plot appears to be random in appearance with no noticeable patterns (i.e., the plot looks like a random scatter of points), this suggests there are no model inadequacies.

On the other hand, if there are structural (non-random) patterns in the residual plot, this suggests that the model is inadequate in some way. Furthermore, the residual plot often reveals what type of model violation is occurring.

The residual plot in Figure 15.7.5 does not suggest any obvious model inadequacies. It looks completely random in appearance, the residuals are balanced around zero, and there is a roughly even vertical-spread of points going across the x axis.

```
> ggdat<-data.frame(x=fitted(mod.cheese),
+                   y=residuals(mod.cheese))
> ggplot(data=ggdat,aes(x=x,y=y))+
+   geom_point(shape=1)+
```

```
+   geom_hline(yintercept = 0,color="red",
+        linetype="dashed",size=1)+
+   xlab(bquote("Fitted Values"~(hat(Y))))+
+   ylab("Residual")+
+   theme_bw()
```

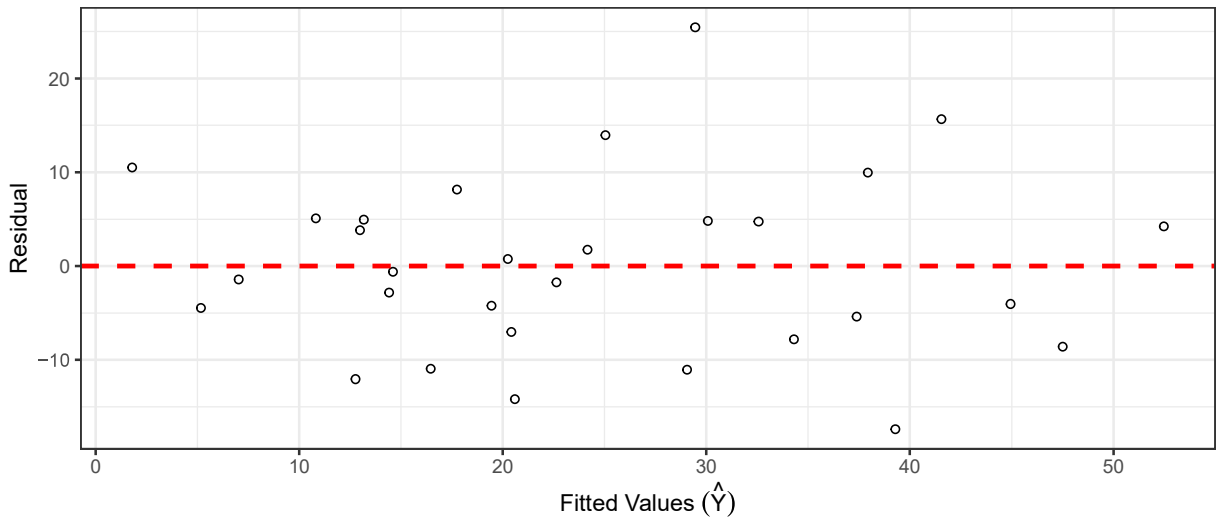

Figure 15.7.5: Cheese data: Residual plot for the multiple linear regression model fit. A horizontal line at $e_i = 0$ has been added.

We now look at two new regression examples. We use these examples to illustrate model violations that are commonly seen in practice. We also discuss remedies to handle these violations.

Example 15.2. An electric company is interested in describing the relationship between the following two variables:

$$Y = \text{peak hour electricity demand (measured in kWh)}$$
$$x = \text{total monthly energy usage (measured in kWh).}$$

This is important for planning purposes because the generating system must be large enough to meet the maximum demand imposed by customers. Engineers consider the simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

to describe the relationship. A random sample of $n = 53$ customers is obtained to estimate the model.

```
> dat.electricity<-data.frame(
+   Monthly.Usage=c(679,292,1012,493,582,1156,997,2189,1097,2078,1818,1700,747,
+                   2030,1643,414,354,1276,745,435,540,874,1543,1029,710,1434,
+                   837,1748,1381,1428,1255,1777,370,2316,1130,463,770,724,808,
+                   790,783,406,1242,658,1746,468,1114,413,1787,3560,1495,2221,
+                   1526),
+   Peak.Demand=c(0.79,0.44,0.56,0.79,2.70,3.64,4.73,9.50,5.34,6.85,5.84,5.21,
+                 3.25,4.43,3.16,0.50,0.17,1.88,0.77,1.39,0.56,1.56,5.28,0.64,
```

```
+                 4.00,0.31,4.20,4.88,3.48,7.58,2.63,4.99,0.59,8.19,4.49,0.51,
+                 1.74,4.10,3.94,0.96,3.29,0.44,3.24,2.14,5.71,0.64,1.90,0.51,
+                 8.33,14.94,5.11,3.85,3.93))
> mod.electricity<-lm(Peak.Demand~Monthly.Usage,data=dat)
> summary(mod.electricity)

Call:
lm(formula = Peak.Demand ~ Monthly.Usage, data = dat)

Residuals:
Min      1Q  Median      3Q     Max
-4.1343 -0.8218 -0.1875  1.1654  3.1578

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.8373128  0.4402394  -1.902   0.0628 .
Monthly.Usage  0.0036831  0.0003329  11.065 3.66e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.572 on 51 degrees of freedom
Multiple R-squared:  0.7059,        Adjusted R-squared:  0.7002
F-statistic: 122.4 on 1 and 51 DF,  p-value: 3.657e-15
```

The linear regression model is plotted in Figure 15.7.6.

```
> ggplot(dat.electricity, aes(x=Monthly.Usage, y=Peak.Demand)) +
+   geom_point(shape=1)+
+   geom_smooth(alpha=0.25,color="black",method="lm")+
+   theme_bw()+
+   xlab("Monthly Usage (kWh)")+
+   ylab("Peak Demand (kWh)")
```

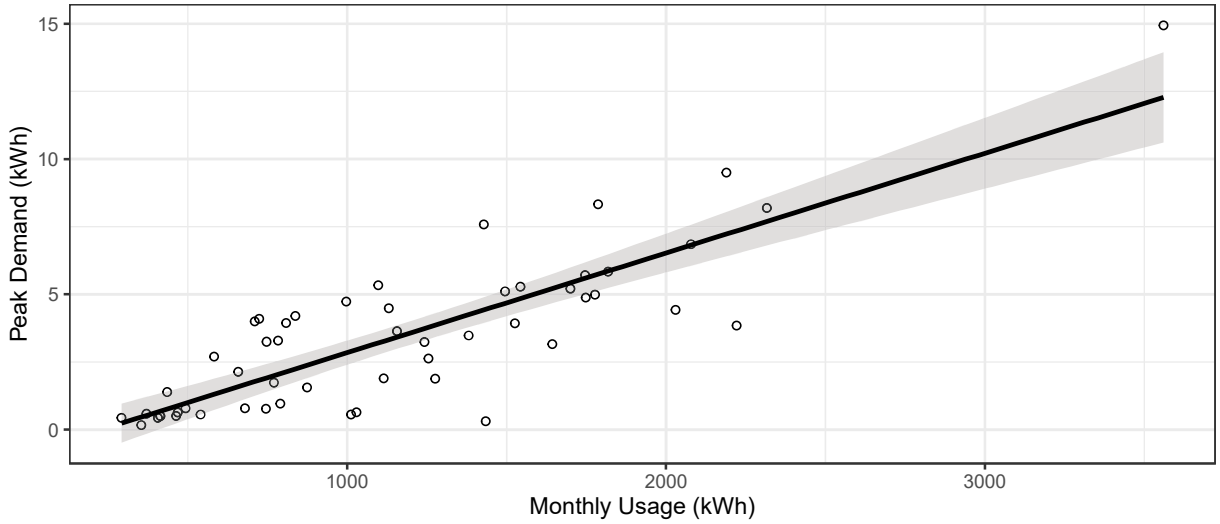

Figure 15.7.6: Electricity data: Scatterplot of peak demand (Y , measured in kWh) versus monthly usage (x, measured in kWh) with estimated simple linear regression line superimposed.

There is a clear problem with non-constant variance here. Note how the residual plot "fans out" like a megaphone. This violation may have been missed by looking at the scatterplot (Figure 15.7.6), the histogram, or the ecdf plot, but the residual plot highlights it.

```
> ####  Histogram
> ggdat<-data.frame(e=residuals(mod.electricity))
> ggdat.gaussian<-data.frame(x=seq(-5,5,0.01),
+                            f=dnorm(seq(-5,5,0.01),
+                                    #ei should have mean zero
+                                    mean=0,
+                                    #ei should have common variance
+                                    sd=summary(mod.electricity)$sigma))
> g.hist<-ggplot(data=ggdat,aes(x=e))+
+   geom_histogram(aes(y=..density..),bins=15,
+                  fill="lightblue",color="black")+
+   geom_density(aes(color="Loess Density Estimate"),size=1)+
+   geom_line(data=ggdat.gaussian,aes(x=x,y=f,color="Under Gaussian Assumption"),
+        linetype="dashed",size=1)+
+   theme_bw()+
+   xlab("Residual")+
+   ylab("Density")+
+   labs(color = "")+
+   theme(legend.position = "bottom")
> ####  ECDF
> e.cdf.func<-ecdf(residuals(mod.electricity))
> e.cdf<-e.cdf.func(sort(residuals(mod.electricity)))
> ggdat<-data.frame(e=sort(residuals(mod.electricity)),
+                   e.cdf=e.cdf)
> ggdat.gaussian<-data.frame(x=seq(-5,5,0.01),
+                            CDF=pnorm(seq(-5,5,0.01),
+                                      #ei should have mean zero
```

```
+                                     mean=0,
+                                     #ei should have common variance
+                                     sd=summary(mod.electricity)$sigma))
> g.ecdf<-ggplot(data=ggdat,aes(x=e))+
+   geom_step(aes(y=e.cdf,color="Empirical CDF"))+
+   geom_line(data=ggdat.gaussian,aes(x=x,y=CDF,color="Under Gaussian Assumption"),
+        linetype="dashed",size=1)+
+   geom_hline(yintercept=0)+
+   theme_bw()+
+   xlab("Residual")+
+   ylab("Cumulative Density")+
+   labs(color = "")+
+   theme(legend.position = "bottom")
> ####  QQplot of the residuals
> library("qqplotr")
> g.qq<-ggplot(data=ggdat,aes(sample=scale(e)))+ #standardize e
+   stat_qq_band(alpha=0.25) +
+   stat_qq_line() +
+   stat_qq_point() +
+   theme_bw()+
+   xlab("Gaussian Quantiles")+
+   ylab("Sample Quantiles")
> ####  Residual Plot
> ggdat<-data.frame(x=fitted(mod.electricity),
+                   y=residuals(mod.electricity))
> g.res<-ggplot(data=ggdat,aes(x=x,y=y))+
+   geom_point(shape=1)+
+   geom_hline(yintercept = 0,color="red",
+        linetype="dashed",size=1)+
+   xlab(bquote("Fitted Values"~(hat(Y))))+
+   ylab("Residual")+
+   theme_bw()
> #### Combine plots
> grid.arrange(g.hist,g.ecdf,g.qq,g.res)
```

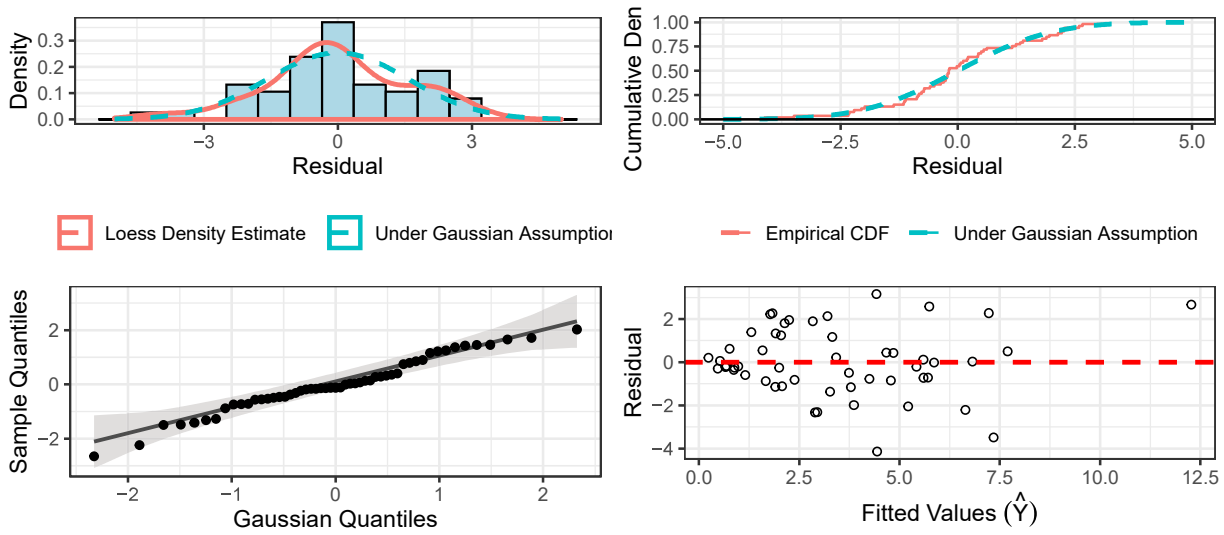

Figure 15.7.7: Electricity data: Histogram of residuals with the assumed Gaussian PDF superimposed**(top left)**, empirical CDF plot with the assumed Gaussian CDF superimposed **(top right)**, qq plot **(bottom left)**, and a residual plot **(bottom right)** for the simple linear regression model fit.

Remedy: When faced with a non-constant variance violation in regression, a common remedy is to transform the response variable Y. Common transformations are the logarithmic ($\log(Y)$), or the Box-Cox or Tukey Ladder transformation $(Y+\delta)^{\lambda}$ where λ and δ are constants; for example, $\lambda = 0.50$ and $\delta = 0$ corresponds to a square root transformation ($\sqrt{Y}$). The constant λ is the power-transformation and δ is the shift amount, which is added when there are zero or negative Y observations.

In order to select the "best" power transformation, we can optimize the likelihood of the regression model over λ. That is,

$$\arg\max_{\lambda}\left[\arg\max_{(\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k)\in\mathbb{R}^{(k+1)}}\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{\sum_{i=1}^{n}\frac{(Y_i^{\lambda}-(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_kx_{ik}))^2}{2\sigma^2}}\right],$$

the maximum likelihood estimates under the normality assumption on ϵ_i. We can complete this in `R` using the `boxcox()` function from the "MASS" library, which also plots the log-likelihood; see Figure 15.7.8.

```
> library(MASS)
> bc.transform<-boxcox(mod)
> (lambda <- bc.transform$x[which.max(bc.transform$y)])
[1] 0.5454545
```

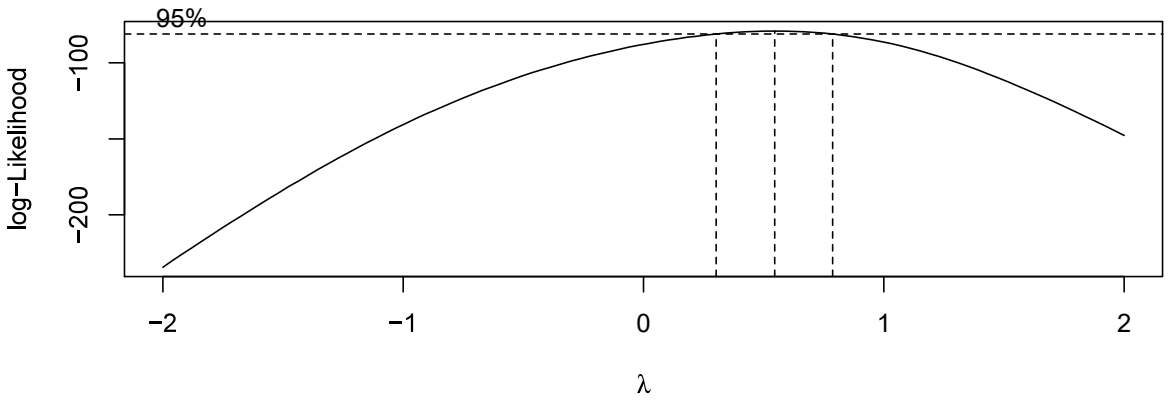

Figure 15.7.8: Electricity data: The log-likelihood function over power transformation values, λ, the dotted lines indicate a 95% confidence interval about the maximum observed value of *lambda*.

This outcome tells us that the Y^{λ} transformation is "best" at $\lambda = 0.\overline{54}$. For simplicity, we take $\lambda = 0.50$ – the square root transformation.

We consider the model

$$\sqrt{Y} = \beta_0 + \beta_1 x + \epsilon,$$

where

$$\sqrt{Y} = \text{peak hour electricity demand (measured in } \sqrt{\text{kWh}})$$
$$x = \text{total monthly energy usage (measured in kWh).}$$

Fitting this model in R gives the least squares estimates

```
> dat$sqrt.Peak.Demand<-sqrt(dat$Peak.Demand)
> mod.electricity.transformed<-lm(sqrt.Peak.Demand~Monthly.Usage,data=dat)
> summary(mod.electricity.transformed)

Call:
lm(formula = sqrt.Peak.Demand ~ Monthly.Usage, data = dat)

Residuals:
Min       1Q   Median       3Q      Max
-1.39055 -0.30440 -0.03748  0.25507  0.81157

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.808e-01  1.295e-01   4.484 4.17e-05 ***
Monthly.Usage 9.529e-04  9.793e-05   9.731 3.24e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4626 on 51 degrees of freedom
```

```
Multiple R-squared:  0.6499,        Adjusted R-squared:  0.6431
F-statistic: 94.69 on 1 and 51 DF,  p-value: 3.244e-13
```

Therefore, the estimated model on the transformed scale is

$$\sqrt{Y} = 0.5808 + 0.0009529x,$$

or, in other words,

$$\widehat{\sqrt{\text{Peak demand}}} = 0.5808 + 0.0009529 \text{ Monthly usage},$$

this regression model is plotted in Figure 15.7.9.

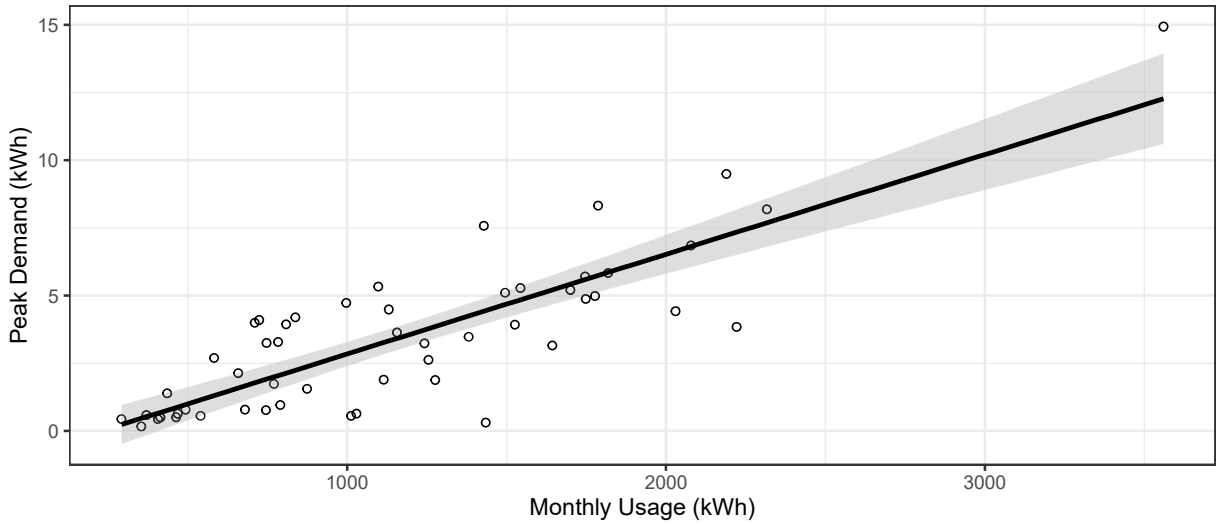

Figure 15.7.9: Electricity data: Scatterplot of peak demand (Y , measured in kWh) versus monthly usage (x, measured in kWh) with estimated transformed linear regression line superimposed..

Note that applying the transformation did help to reduce the non-constant variance problem considerably; see Figure 15.7.10. The noticeable "fanning out" shape that we saw in the residual plot previously (i.e., based on the untransformed response Y) is now largely absent. The qq plot for normality (using the residuals from the transformed model fit) reveals some potential mild departures, but nothing that was serious.

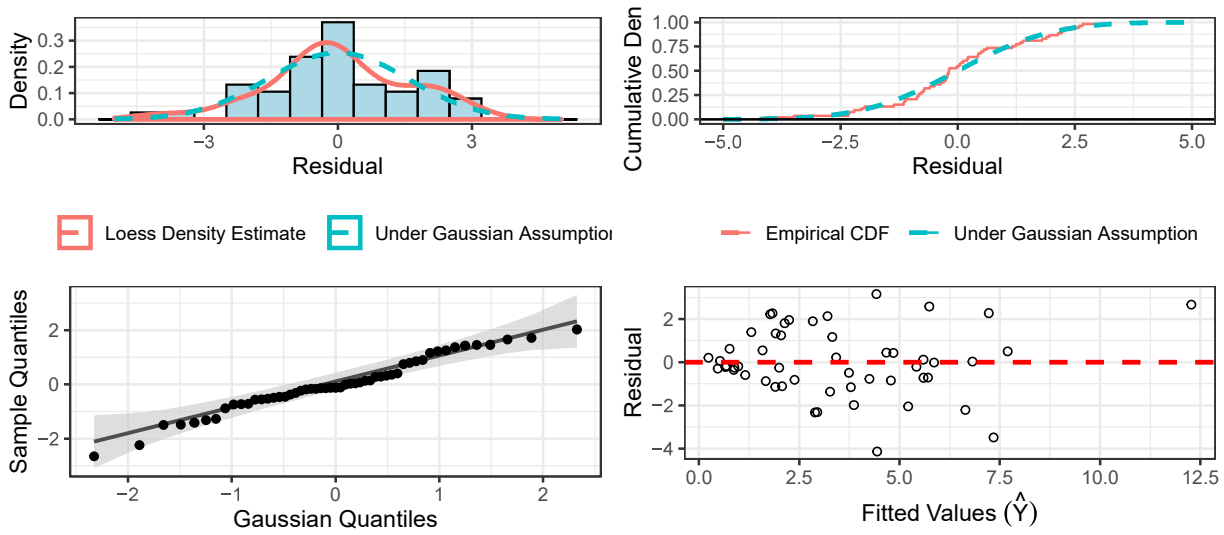

Figure 15.7.10: Electricity data: Histogram of residuals with the assumed Gaussian PDF superimposed**(top left)**, empirical CDF plot with the assumed Gaussian CDF superimposed **(top right)**, qq plot **(bottom left)**, and a residual plot **(bottom right)** for the transformed linear regression model fit.

Let's proceed with inference for β_1 to determine if the linear relationship is significant for the population. To do this, we will calculate a confidence interval for β_1 in R.

```
> confint(mod.electricity.transformed,level=0.95)
2.5 %      97.5 %
(Intercept)   0.3208043932 0.840857384
Monthly.Usage 0.0007563267 0.001149532
```

We are 95 percent confident that the population regression parameter β_1 (in the transformed model) is between 0.000756 and 0.001150.

Note that this interval does not include "0" and includes only positive values. This suggests that peak demand (on the square root scale) and monthly usage are positively related in the population. Specifically, for every one-unit increase in x (monthly usage measured in kWh), we are 95 percent confident that the mean peak demand will increase between 0.000756 and 0.001150 $\sqrt{\text{kWh}}$.

Remark: A more advanced remedy is to use a model fitting technique known as weighted least squares; this involves weighting certain observations more/less depending on their level of variability. We will discuss this in detail later.

The advantage of using a transformation is that you can still use least squares without weighting. However, all inferences will pertain to the population model with the transformed response, not the response Y itself. This can sometimes complicate how the results are interpreted.

Example 15.3. An engineer is investigating the use of a windmill to generate electricity. She has collected data on

$$Y = \text{direct current (DC) output}$$
$$x = \text{wind velocity (measured in mph)}.$$

Data for $n = 25$ observation pairs are shown in Figure 15.7.11. The engineer initially assumes a simple linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon$$

to describe the relationship and fits this model in R.

```
> dat.windmill<-data.frame(
+   wind.velocity=c(5,6,3.4,2.7,10,9.7,9.55,3.05,8.15,6.2,2.9,6.35,4.6,5.8,
+                   7.4,3.6,7.85,8.8,7,5.45,9.1,10.2,4.1,3.95,2.45),
+   DC.output = c(1.582,1.822,1.057,0.5,2.236,2.386,2.294,0.558,2.166,1.866,
+                 0.653,1.93,1.562,1.737,2.088,1.137,2.179,2.112,1.8,1.501,
+                 2.303,2.31,1.194,1.144,0.123))
> mod.windmill<-lm(DC.output~wind.velocity,data=dat.windmill)
> summary(mod.windmill)

Call:
lm(formula = DC.output ~ wind.velocity, data = dat.windmill)

Residuals:
Min       1Q   Median       3Q      Max
-0.59869 -0.14099  0.06059  0.17262  0.32184

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.13088    0.12599   1.039     0.31
wind.velocity  0.24115    0.01905  12.659 7.55e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2361 on 23 degrees of freedom
Multiple R-squared:  0.8745,        Adjusted R-squared:  0.869
F-statistic: 160.3 on 1 and 23 DF,  p-value: 7.546e-12
```

The linear regression model is plotted in Figure 15.7.11.

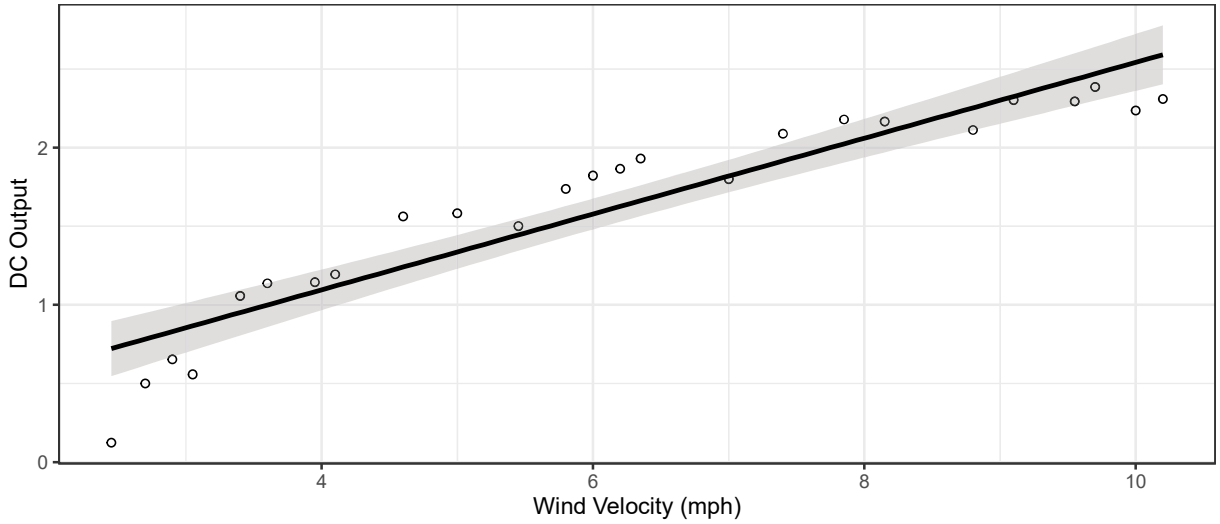

Figure 15.7.11: Windmill data: Scatterplot of DC Output (Y) versus wind velocity (x, measured in mph) with estimated simple linear regression line superimposed.

There is a clear quadratic relationship between DC output and wind velocity. The residual plot in Figure 15.7.12 from the simple linear regression model fit shows a pronounced quadratic pattern. It is easy to see why this is happening – a simple linear regression model is inappropriate here (it does not explain quadratic relationships).

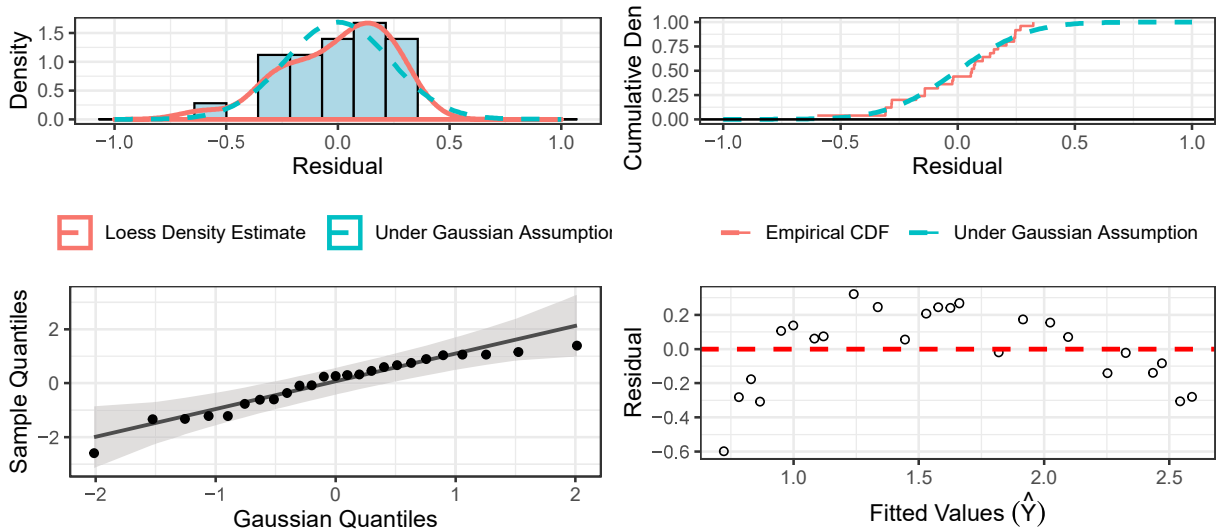

Figure 15.7.12: Windmill data: Histogram of residuals with the assumed Gaussian PDF superimposed**(top left)**, empirical CDF plot with the assumed Gaussian CDF superimposed **(top right)**, qq plot **(bottom left)**, and a residual plot **(bottom right)** for the transformed linear regression model fit.

Remark: We can use `R` to calculate the coefficient of determination from the simple linear regression model fit.

```
> summary(mod.windmill)$r.squared
[1] 0.8744932
```

A novice data analyst (especially one that doesn't even bother to graph the data) might think that because this is "pretty large," the model we have fit is a "good model." However, it is easy to see from Figure 15.7.11 that a simple linear regression model is not the best model for the data. Even though 0.874 is in fact "pretty large," its value refers specifically to a model that is inappropriate.

Remedy: Fit a multiple linear regression model with two independent variables: wind velocity x and its square x^2. The model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

is called a quadratic regression model. It is straightforward to fit a quadratic regression model in R. We simply regress Y on both x and x^2.

```
> dat.windmill$wind.velocity.sq<-dat.windmill$wind.velocity^2
> mod.windmill.transformed<-lm(DC.output~wind.velocity+wind.velocity.sq,data=dat.windmill)
> summary(mod.windmill.transformed)

Call:
lm(formula = DC.output ~ wind.velocity + wind.velocity.sq, data = dat.windmill)

Residuals:
Min       1Q   Median       3Q      Max
-0.26347 -0.02537  0.01264  0.03908  0.19903

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.155898   0.174650  -6.618 1.18e-06 ***
wind.velocity     0.722936   0.061425  11.769 5.77e-11 ***
wind.velocity.sq -0.038121   0.004797  -7.947 6.59e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1227 on 22 degrees of freedom
Multiple R-squared:  0.9676,        Adjusted R-squared:  0.9646
F-statistic: 328.3 on 2 and 22 DF,  p-value: < 2.2e-16
```

The estimated quadratic regression model is

$$\hat{Y} = -1.15590 + 0.72294x - 0.03812x^2$$

or, in other words,

$$\widehat{DCoutput} = -1.15590 + 0.72294 \text{ Wind velocity} - 0.03812 \text{ Wind velocity}^2.$$

Note that the residual plot from the quadratic model fit, shown in Figure 15.7.13, now looks much more random. The quadratic trend has disappeared (because the model now incorporates it).

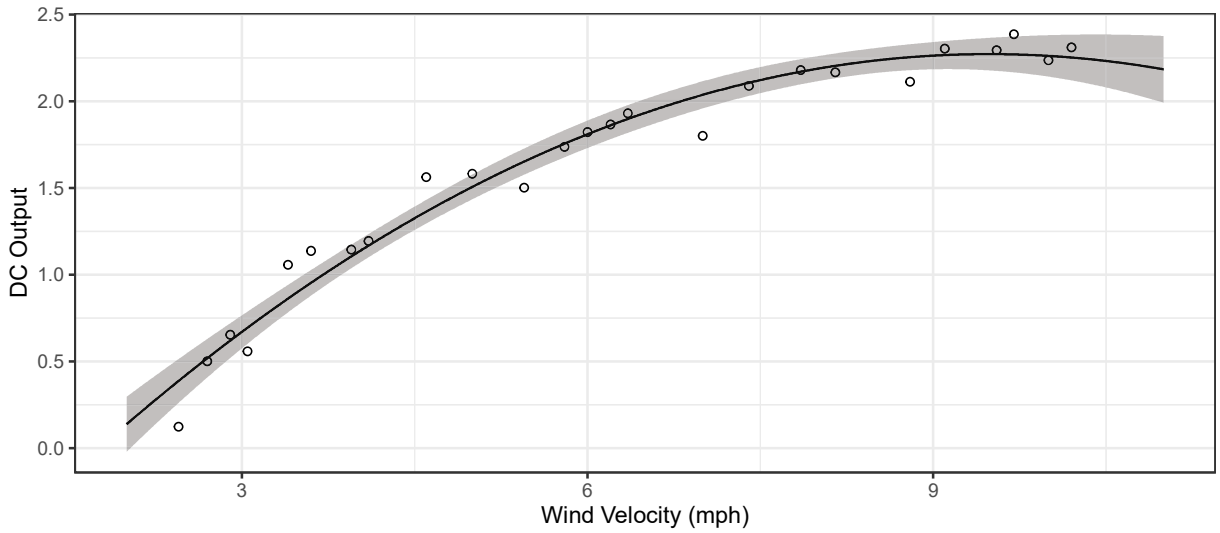

Figure 15.7.13: Windmill data: Scatterplot of DC output Y versus wind velocity (x, measured in mph) with least squares quadratic regression curve superimposed.

Figure 15.7.14 shows the diagnostic residual plots for the quadratic model fit. There's a new problem, because now it looks like the normality assumption (for the quadratic model) is violated. Interestingly, this was not a problem with the simple linear regression model. It appears that fitting the quadratic regression model fixed one problem (i.e., selecting a better regression function) but created another (normality violation).

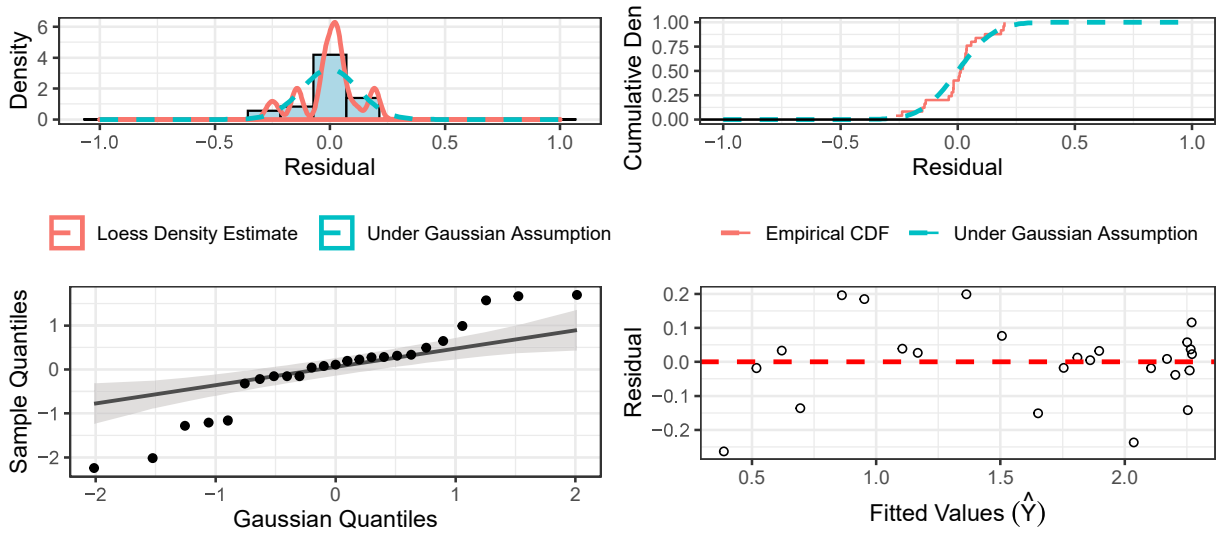

Figure 15.7.14: Windmill data: Histogram of residuals with the assumed Gaussian PDF superimposed**(top left)**, empirical CDF plot with the assumed Gaussian CDF superimposed **(top right)**, qq plot **(bottom left)**, and a residual plot **(bottom right)** for the least squares quadratic regression fit.

Confidence interval: To see if the quadratic effect between DC output and wind velocity is significant, we can write a confidence interval for β_{11}, the population parameter in the quadratic regression model that describes the quadratic effect.

```
> confint(mod.windmill.transformed,level=0.95)
2.5 %      97.5 %
(Intercept)      -1.51810023 -0.79369625
wind.velocity     0.59554751  0.85032429
wind.velocity.sq -0.04806859 -0.02817318
```

We are 95 percent confident that the population regression parameter β_{11} (in the quadratic model) is between -0.0481 and -0.0282. Note that this interval does not include "0" and includes only negative values. This suggests that quadratic effect between DC output and wind velocity is significant in the population.

Remark: We can use `R` to calculate the coefficient of determination from the quadratic regression model fit.

```
> summary(mod.windmill.transformed)$r.squared
[1] 0.9675771
```

This means that about 96.8 percent of the variability in the DC output data is explained by the estimated model that includes both wind velocity and (wind velocity)^2. The remaining 3.2 percent is not explained by the estimated model. This is an improvement over the largely meaningless $R^2 \approx 0.875$ calculated from the simple linear regression.