Body Fat Regression Analysis - New Version

jdt

10/20/2021

Contents

Introduction **1**

Linear Model **1**
Least Squares Solution – Simple Linear Model 2
Inference in Linear Regression 3
Analysis of Variance for Regression 4

Linear Regression **6**
R 6
SAS 11

Polynomial and Multiple Regression **17**
R 22
SAS 26

Introduction

In this document I am going to modify my presentation of the material. I will present a brief review of the theory (from the class notes) then present the analysis for the example using R and then SAS. In this document I will present results for linear, then polynomial, before considering a multiple regression model. In all cases the Body Fat data will be used.

Linear Model

Consider a modeling problem whereby we are interested in determining the relationship between two (or more) random variables from a single population of interest. These models are called *regression models.*

The simplest of these models, the linear model, is given by

$$\mu_y = E(Y) = \beta_0 + \beta_1 X \tag{1}$$

where the mean, μ_y, of the dependent variable of interest, Y, can be written as a linear function of a independent variable X as determine by two parameters, the slope, β_1, and y-intercept, β_0.

Least Squares Solution – Simple Linear Model

Suppose that one has two random variables X and Y for which one observes n pairs, denoted by (x_i, y_i), $i = 1, 2, \ldots, n$. The objective of the least squares problem is to determine, $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the linear equation given by $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ minimizes

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^{n} \epsilon_i^2. \tag{2}$$

where $\epsilon_i = (y_i - (\beta_0 + \beta_1 x_i))$ is called the *population residual or error* for the observation (x_i, y_i). The solution can be determined by taking the partial derivatives of $Q(\beta_0, \beta_1)$ with respect to both β_0 and β_1. That is,

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]$$

and

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i] x_i.$$

By setting these equations equal to zero, one obtains

$$\begin{aligned} n\beta_0 + \sum x_i \beta_1 &= \sum y_i \\ \sum x_i \beta_0 + \sum x_i^2 \beta_1 &= \sum x_i y_i. \end{aligned} \tag{3}$$

Equation(3) is called **the normal equations** for the linear least squares problem. From which, the unique solution is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \tag{5}$$

where

$$\begin{aligned} \bar{y} &= \sum_{i=1}^{n} y_i / n \qquad \bar{x} = \sum_{i=1}^{n} x_i / n \\ SS_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y} \\ SS_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2 = (n-1)s_x^2 \\ SS_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = (n-1)s_y^2 \end{aligned}$$

The estimated residual, $\hat{e}_i = y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i)$, is the vertical distance that the observed y_i is from the least squares line $(\hat{\beta}_0 + \hat{\beta}_1 x)$ at $x = x_i$. The *residual sum of squares* or *sum of squares due to the error* is

$$SS_E = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum \hat{e}_i^2 = SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}.$$

The predicted value (line at $x = x^*$) can be written as[1][2]

$$\hat{\mu}_{y|x^*} = \hat{y}^* = \bar{y} + \hat{\beta}_1(x^* - \bar{x})$$

Inference in Linear Regression

The procedure for computing the least squares estimates for β_0 and β_1 is a mathematical optimization problem. In order to use this method as a statistical problem one must make additional distributional assumptions concerning the response or dependent variable y. These assumptions are;

- The observed dependent data $y_1, y_2, \ldots, y_n$, are a sample from a population where $Y \sim N(\mu_{y_i}, \sigma_y^2)$. This is called the **normality assumption**.
- $E(y_i) = \mu_{y_i} = \beta_0 + \beta_1 x_i$. The expected value for y_i is a linear function of x_i. This is called the **linear assumption**.
- σ_y^2 does not depend upon the value of x_i. This is called the **homogeneity of variance assumption**.
- The data $y_1, y_2, \ldots, y_n$ are independent. This is called the **independence assumption**.

The above assumptions allow one to determine the standard errors for the statistical estimates of the population parameters of interest; β_0, β_1, and $E(y \mid x = x^*) = \mu_{y|x^*}$.

- The slope (β_1)

  $$\hat{\beta}_1 = SS_{xy}/SS_{xx}$$

  and

  $$\hat{\sigma}_{\hat{\beta}_1} = \hat{\sigma}/\sqrt{SS_{xx}}$$

  where

  $$\hat{\sigma} = \sqrt{SSE/(n-2)} = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}.$$

- The line $(\mu_{y|x^*})$

  $$\hat{\mu}_{y|x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \bar{y} + \hat{\beta}_1(x^* - \bar{x})$$

  and

  $$\hat{\sigma}_{\hat{\mu}_{y|x^*}} = \hat{\sigma}\sqrt{1/n + (x^* - \bar{x})^2/SS_{xx}}.$$

[1] Note if the estimated slope $\hat{\beta}_1$ is zero then the predicted value for every y_i is $\bar{y}$. Which indicates that the variable X was not needed in the regression model for Y.

[2] Every linear least squares regression line passes through the point $(\bar{x}, \bar{y})$.

- The y-intercept (β_0)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

and

$$\hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{1/n + \bar{x}^2/SS_{xx}}.$$

The $(1-\gamma)100\%$ Confidence intervals are of the form

$$\text{(estimate)} \pm t_{\gamma/2}(df = (n-2)) \times \text{(standard error of estimate)}.$$

where $t_{\gamma/2}(df = (n-2))$ is the critical point from a t-distribution with $df = n - 2$.

Analysis of Variance for Regression

The regression results are often presented as an analysis of variance table or ANOVA table. The basic idea is to describe how much of the variability found in the dependent variable y can be explained by the presence of the linear equation ($\beta_1 \neq 0$) versus having a line $y = \bar{y}$ ($\beta_1 = 0$).

The total adjusted (corrected for β_0) sum of squares in the dependent variable y is given by $SS_{CT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$. This corrected sum of squares can be written as $SS_{CT} = SS_M + SS_E$ where

$$SS_M = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

and

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

SS_M is the sum of squares due to the model and SS_E is the sum of squares due to the error (*Residual sum of squares*).

The above follows from

$$\begin{aligned}
\sum(y_i - \hat{y}_i)^2 &= \sum(y_i - \bar{y} + \bar{y} - \hat{y}_i)^2 \\
&= \sum(y_i - \bar{y})^2 + (\hat{y}_i - \bar{y})^2 - 2(y_i - \bar{y})(\hat{y}_i - \bar{y}) \\
&= \sum(y_i - \bar{y})^2 + \sum(\hat{y}_i - \bar{y})^2 - 2\sum(y_i - \bar{y})(\hat{y}_i - \bar{y}) \\
&= \sum(y_i - \bar{y})^2 - \sum(\hat{y}_i - \bar{y})^2
\end{aligned}$$

where

$$\begin{aligned}
-2\sum(y_i - \bar{y})(\hat{y}_i - \bar{y}) &= -2\sum(y_i - \bar{y})\hat{\beta}_1(x_i - \bar{x}) \\
&= -2\hat{\beta}_1\sum(y_i - \bar{y})(x_i - \bar{x}) \\
&= -2\hat{\beta}_1^2\sum(x_i - \bar{x})^2 \\
&= -2\sum(\hat{y}_i - \bar{y})^2.
\end{aligned}$$

If the slope of the line is nonzero then one would expect that a sizeable amount of the variability in y as specified by SS_{CT} would be attributable to SS_M. One way of measuring this is to compute $R^2 = SS_M/SS_{CT} = SS_M/(SS_M + SS_E)$. R^2 is a number between 0 and 1 which is usually expressed as a percentage. **Since SS_{CT} has been adjusted for $\hat{\beta}_0$, R^2 is the variability explained by the model relative to what can be explained by $\bar{y}$.**The closer the value is to 1 or 100% means that the amount of variability found in SS_{CT} is nearly explained by the model (or in this case having $\hat{\beta}_1$ be nonzero). On the other hand if R^2 is close to zero then very little of the variability in the data is explained by the model as opposed to just using $\hat{\mu}_y = \bar{y}$, which means that one doesn't need x in order to explain variability in y.

ANOVA Table

The analysis of variance table is given by

Source	Degrees of Freedom	Sum of Squares	Mean Square
due to $\beta_1 \mid \beta_0$	1	$SS_M = \sum(\hat{y}_i - \bar{y})^2$	$MS_M = SS_M/1$
Residual	n-2	$SS_E = \sum(y_i - \hat{y}_i)^2$	$MS_E = SS_E/(n-2)$
Corrected Total	n-1	$SS_{CT} = \sum(y_i - \bar{y})^2$	
due to β_0	1	Correction factor =$n\bar{y}^2$	
Total	n	$SS_T = \sum y_i^2$	

Linear Regression

R

```
  # clear the environment and set seed
rm(list = ls())
set.seed(123)
```

Read Body Fat Data

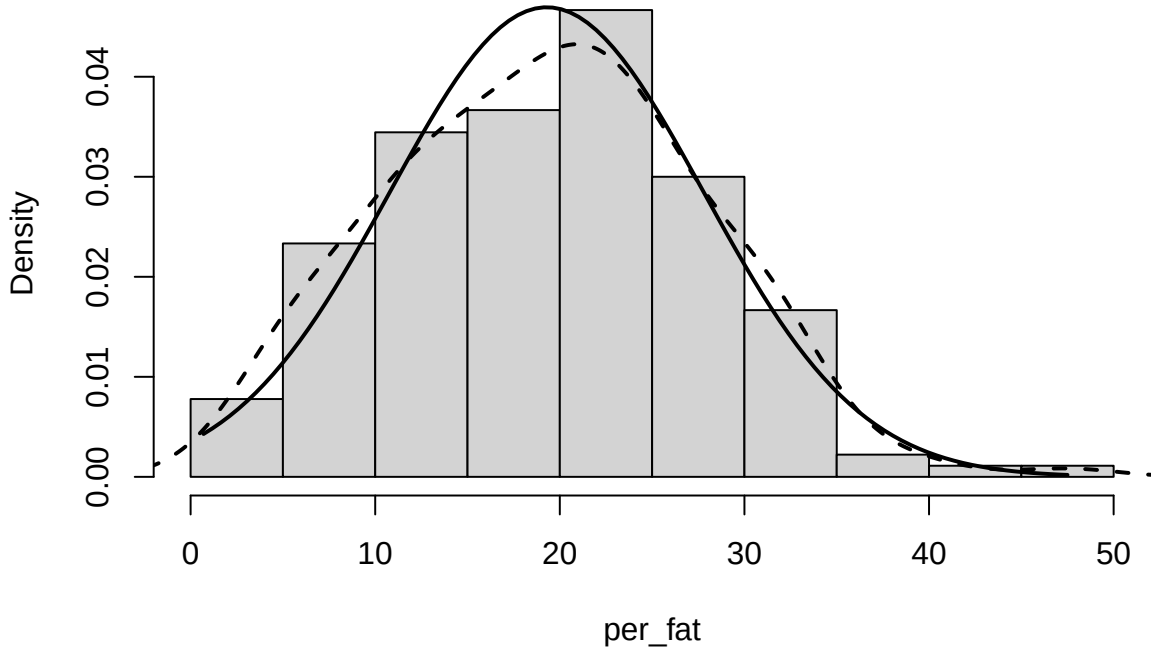
```
library(foreign)
bf = read.dbf("new_bfat.dbf")    #define bf as the bodyfat data
```

Define a few variables

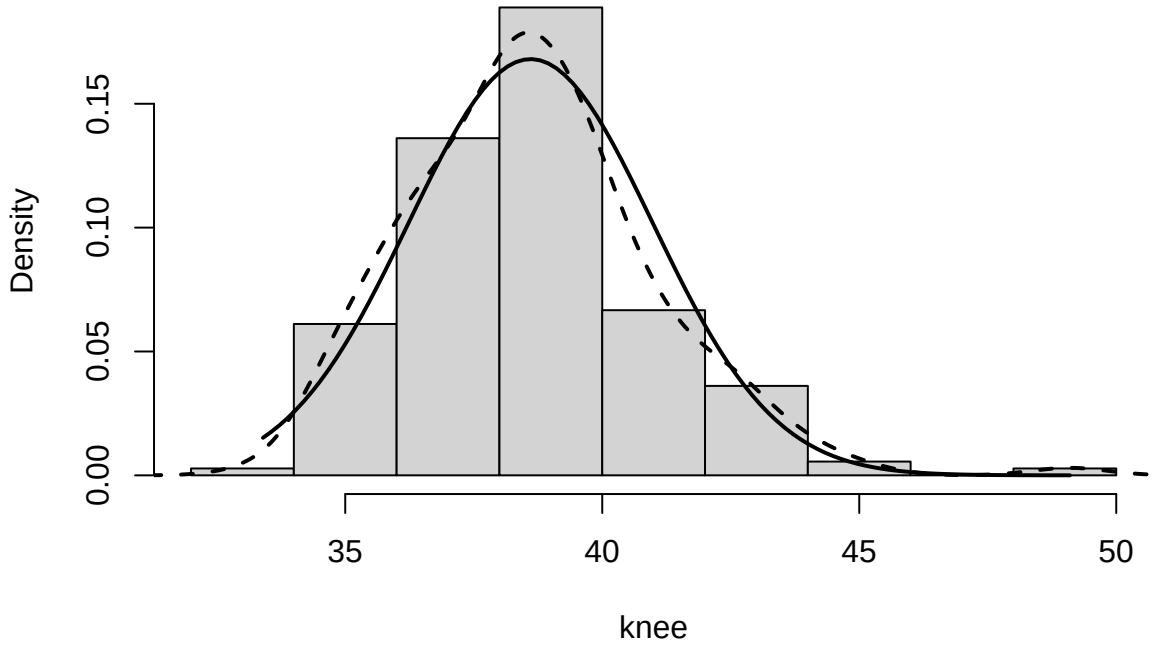
```
abdomen=bf$abdomen
thigh = bf$thigh
neck = bf$neck
per_fat = bf$per_fat
density = bf$density
age=bf$age
wt = bf$wt
ht = bf$ht
chest = bf$chest
hip = bf$hip
thigh = bf$thigh
knee = bf$knee
ankle = bf$ankle
biceps = bf$biceps
forearm = bf$forearm
wrist = bf$wrist
```

Plots of Percent Fat and Abdomen Circumference

```
with(bf, hist(per_fat, main="", freq=FALSE))
with(bf, lines(density(per_fat), main="PERCENT FAT", lty=2, lwd=2))
xvals = with(bf, seq(from=min(per_fat), to=max(per_fat), length=100))
with(bf, lines(xvals, dnorm(xvals, mean(per_fat), sd(per_fat)), lwd=2))
```




```
with(bf, hist(knee, main="", freq=FALSE))
with(bf, lines(density(knee), main="ABDOMEN", lty=2, lwd=2))
xvals = with(bf, seq(from=min(knee), to=max(knee), length=100))
with(bf, lines(xvals, dnorm(xvals, mean(knee), sd(knee)), lwd=2))
```



```r
mod1 = lm(per_fat ~ knee, data=bf)
summary(mod1)
```

```
##
## Call:
## lm(formula = per_fat ~ knee, data = bf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1863  -4.6730  -0.3861   3.9940  31.2614
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -49.7968     8.9871  -5.541 1.07e-07 ***
## knee          1.7896     0.2323   7.702 8.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.38 on 178 degrees of freedom
## Multiple R-squared:   0.25,  Adjusted R-squared:  0.2458
## F-statistic: 59.32 on 1 and 178 DF,  p-value: 8.964e-13
```

```r
covb = vcov(mod1)
coeff.mod1 = coef(mod1)

covb = vcov(mod1)
covb
```

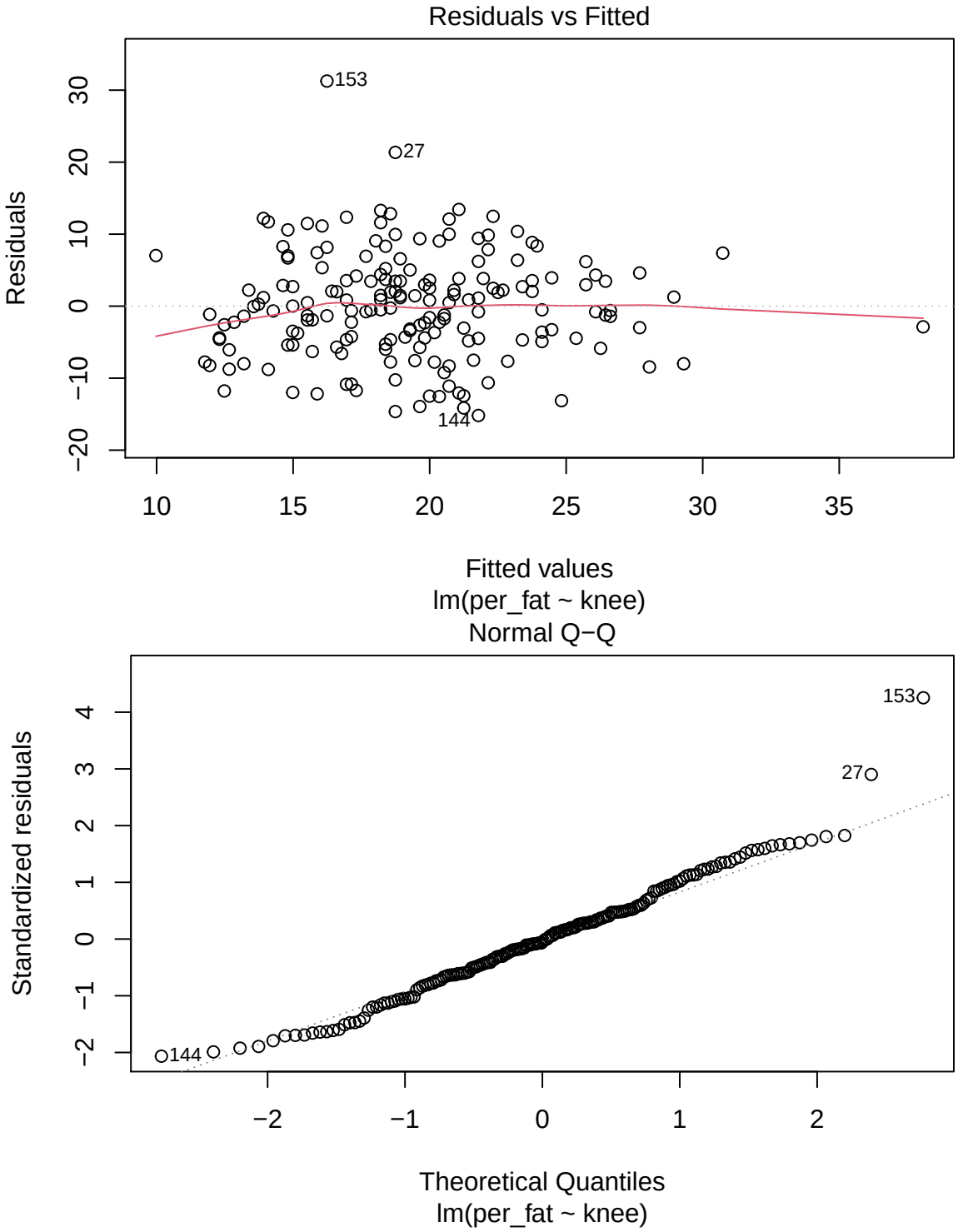
```
##             (Intercept)        knee
## (Intercept)   80.768119 -2.08423949
## knee          -2.084239  0.05398652
```

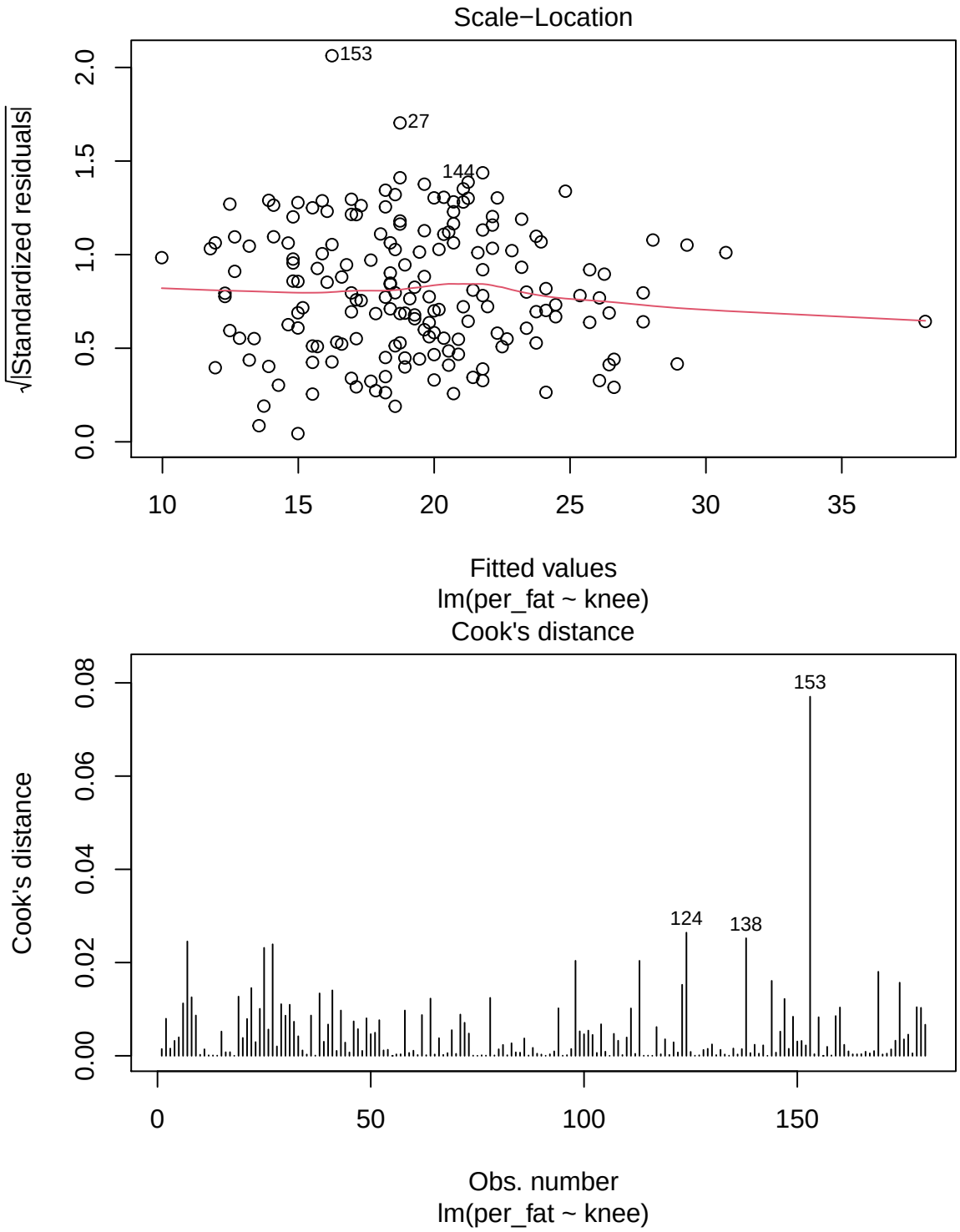
```r
pred.per_fat = predict(mod1)
res.per_fat = residuals(mod1)
summary(res.per_fat)
```

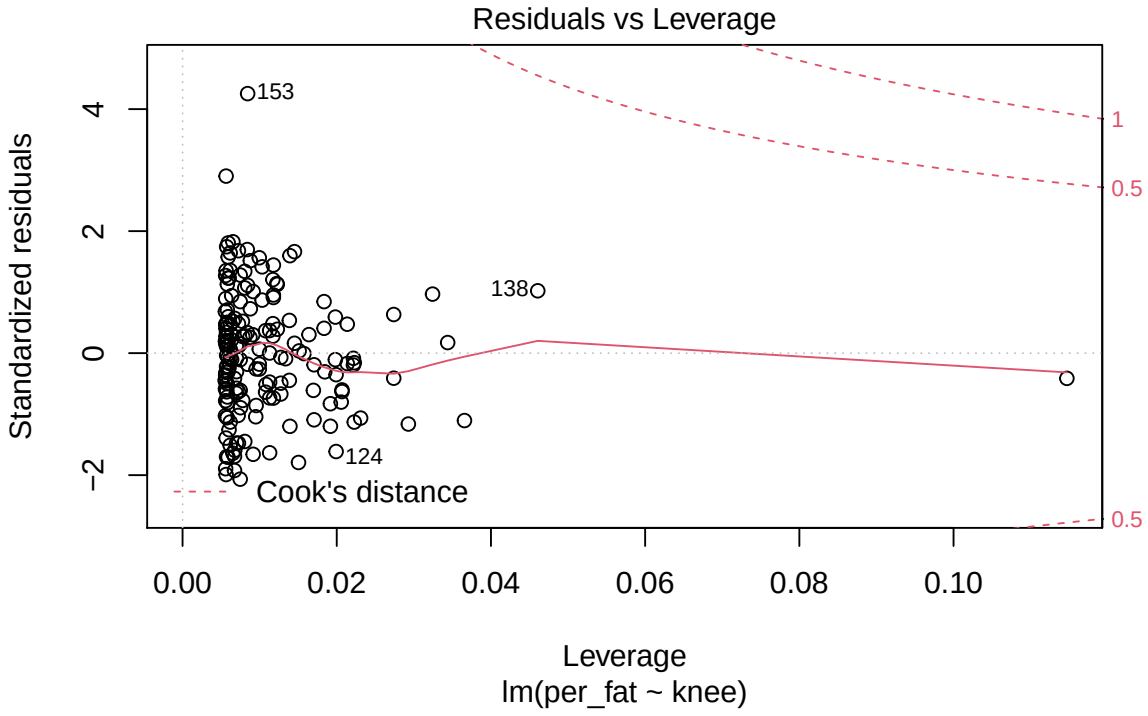
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -15.1863  -4.6730  -0.3861   0.0000   3.9940  31.2614
```

Residual Plots

```r
#par(mfrow=c(1,1))
plot(mod1, which = c(1, 2,3,4,5))
```


Residuals vs Fitted
Residuals
Fitted values
lm(per_fat ~ knee)
153
27
144
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(per_fat ~ knee)


Scale−Location
√|Standardized residuals|
153
27
144
0.0
0.5
1.0
1.5
2.0
10
15
20
25
30
35
Fitted values
lm(per_fat ~ knee)
Cook's distance
Cook's distance
153
124
138
0.00
0.02
0.04
0.06
0.08
0
50
100
150
Obs. number
lm(per_fat ~ knee)



SAS

The following code is found in the program but the output is not included here.

```
title1 'Body Fat Data';
/*
Density determined from underwater weighing
Percent body fat from Siri's (1956) equation
Age (years)
Weight (lbs)
Height (inches)
Neck circumference (cm)
Chest circumference (cm)
Abdomen 2 circumference (cm)
Hip circumference (cm)
Thigh circumference (cm)
Knee circumference (cm)
Ankle circumference (cm)
Biceps (extended) circumference (cm)
Forearm circumference (cm)
Wrist circumference (cm)
*/
```

```
data bodyfat; set ldata.bodyfat;
run;

title2 'Scatterplot of Entire Data';
proc sgscatter data=bodyfat;
matrix per_fat density age wt ht neck chest abdomen hip
/diagonal=(histogram normal);
run;

proc sgscatter data=bodyfat;
matrix per_fat density thigh knee ankle biceps forearm  wrist
/diagonal=(histogram normal);
run;


/*
Use per_fat or density as the dependent variable with a subset of the data
*/

title2 'Simple Random Sampling of size = 50';
proc surveyselect data=bodyfat
method=srs n=50 out=new_bfat seed = 12345;
run;

/*
if you wish to save this new data set of size n=50
in the SAS DATA SET folder with tag ldata in the libname command
You can do this with any data set after running the
proc surveyselect command;
*/
```

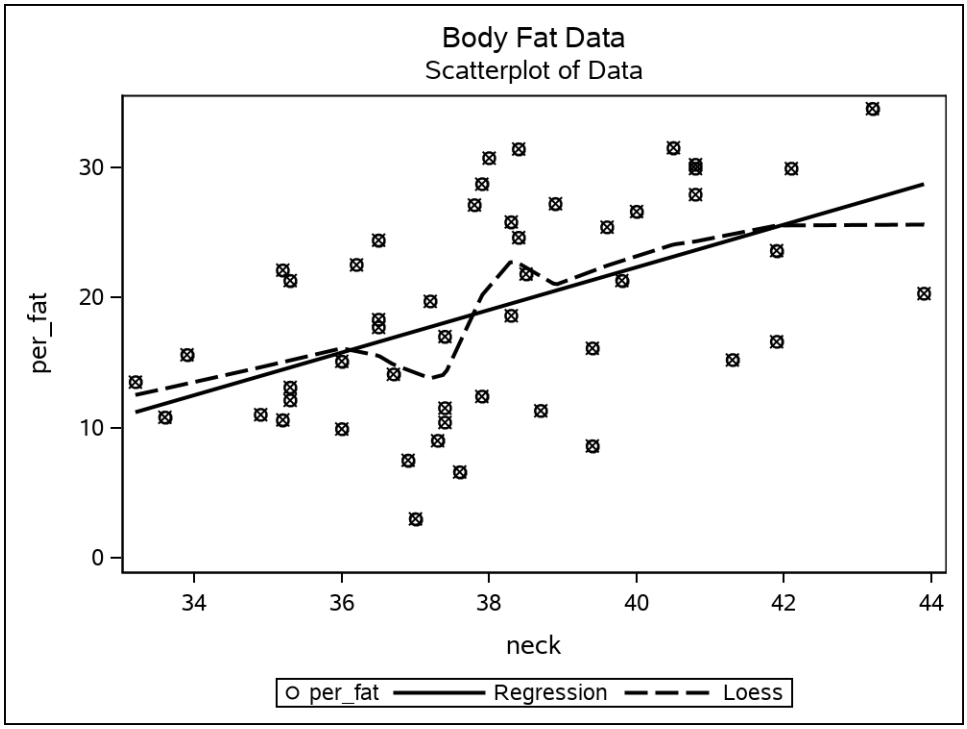
Linear Regression

```
data ldata.bfat_50; set new_bfat; run;

data new_bfat; set new_bfat;
neck2 = neck*neck; abdomen2=abdomen*abdomen; run;

title2 'Scatterplot of Data';
proc sgplot data=new_bfat;
scatter y=per_fat x=neck ;
reg y=per_fat x=neck;
loess y=per_fat x=neck;
run;

title2 'Simple Linear Model - Neck';
proc reg data=new_bfat plots = diagnostics;
model per_fat=neck;
run;
```


Body Fat Data
Scatterplot of Data
per_fat
neck
0
10
20
30
34
36
38
40
42
44
per_fat
Regression
Loess

Body Fat Data

Simple Linear Model - Neck

The REG Procedure

Model: MODEL1

Dependent Variable: per_fat

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	793.54122	793.54122	16.58	0.0002
Error	48	2297.83878	47.87164		
Corrected Total	49	3091.38000			

Root MSE	6.91893	R-Square	0.2567
Dependent Mean	19.08000	Adj R-Sq	0.2412
Coeff Var	36.26275		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	−43.15275	15.31657	−2.82	0.0070
neck	1	1.63684	0.40203	4.07	0.0002

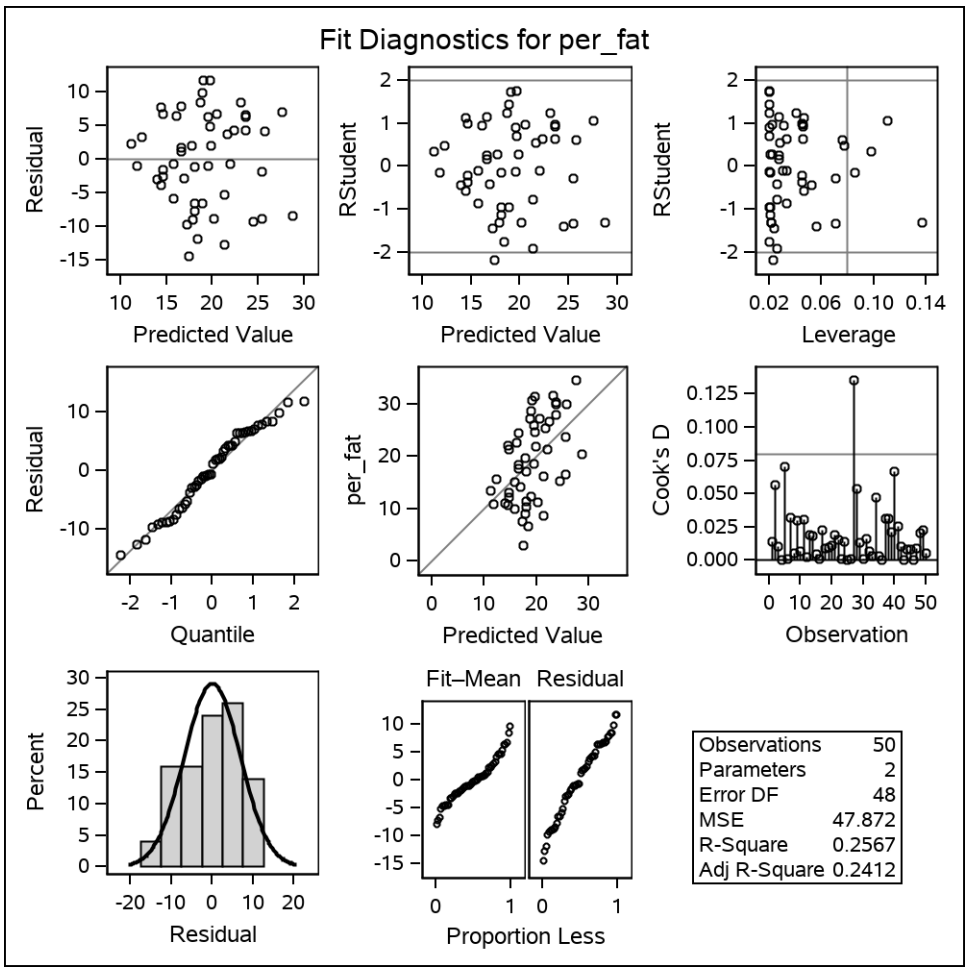
Body Fat Data

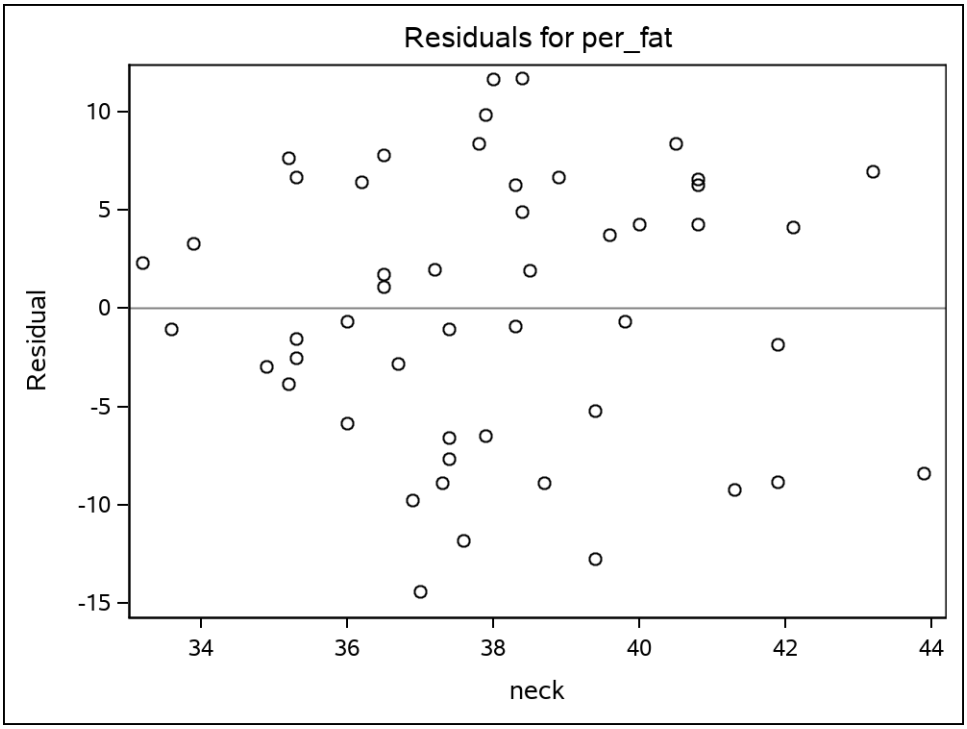
Simple Linear Model - Neck

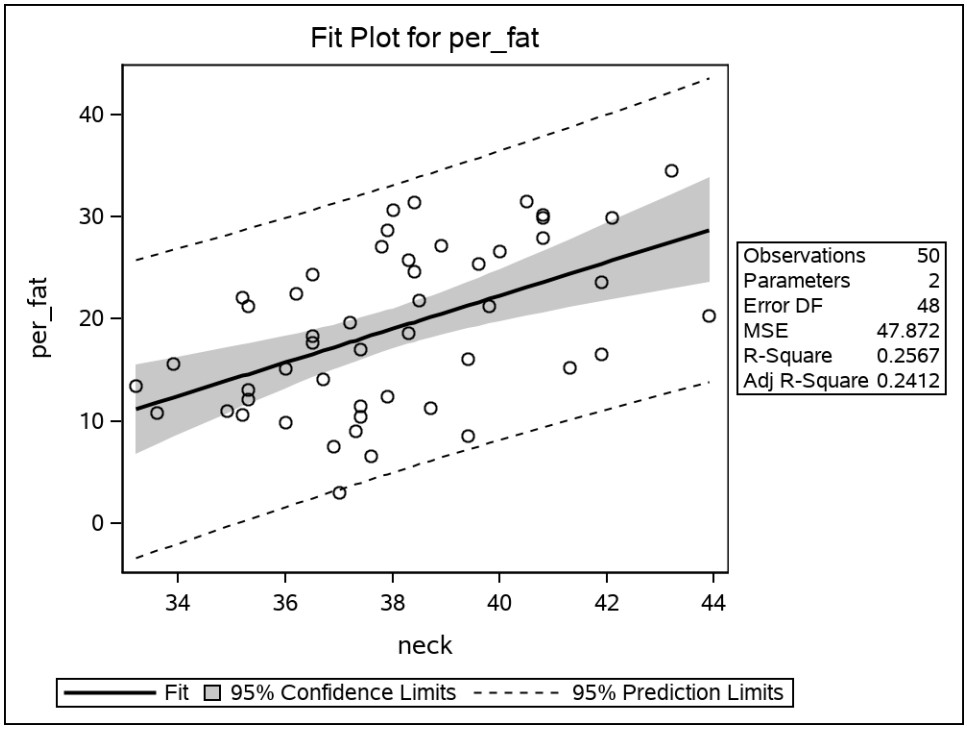
The REG Procedure

Model: MODEL1

Dependent Variable: per_fat




Residuals for per_fat
Residual
10
5
0
-5
-10
-15
34
36
38
40
42
44
neck


Fit Plot for per_fat
per_fat
40
30
20
10
0
34
36
38
40
42
44
neck
Observations 50
Parameters 2
Error DF 48
MSE 47.872
R-Square 0.2567
Adj R-Square 0.2412
Fit
95% Confidence Limits
95% Prediction Limits

Polynomial and Multiple Regression

The simple linear regression model is extended to higher order linear[3] regression models. The general linear regression model can be written as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{(p-1)i} + \epsilon_i \tag{6}$$

for $i = 1, 2, \ldots, n$ where the independent variables $x_{1i}, x_{2i}, \ldots, x_{(p-1)i}$ satisfy one of the following:

1. **Polynomial Regression** – when $x_{ji} = x_i^j, \quad i = 1, 2, \ldots, n, j = 1, 2, \ldots, (p-1)$. Equation (6) is a polynomial of degree $(p-1)$.
2. **Multiple Regression** – where each independent variable is distinct.

Equation (6) could be a combination of the above two models, however, in this chapter we will consider the model to be either polynomial or multiple regression. As in the previous chapter, the unobserved error ϵ_i is the vertical distance that the observed y_i is from the curve or surface given by $E(y_i \mid X\beta) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{(p-1)i}$. Equation (6) can be written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \ldots & x_{(p-1)1} \\ 1 & x_{12} & x_{22} & \ldots & x_{(p-1)2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{(p-1)n} \end{pmatrix} = \begin{pmatrix} \mathbf{j_n} & \mathbf{X}_* \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix},$$

where

$\mathbf{y}$ is a $n \times 1$ vector of dependent observations.

$\mathbf{X}$ is a $n \times p$ matrix of independent observations.

$\mathbf{x_j}$ is a $n \times 1$ vector for the j^{th} independent variable, $j = 1, 2, \ldots, p-1$.

$\mathbf{X}_* = \begin{pmatrix} \mathbf{x_1} & \mathbf{x_2} & \cdots & \mathbf{x_{p-1}} \end{pmatrix}$.

β is a $p \times 1$ vector of parameters.

[3] Linear means that the expected value of the dependent variable can be expressed as an additive model in terms of the independent variables, versus having multiplicative or nonlinear terms.

ϵ is a $n \times 1$ vector of unobserved errors.

As in the previous chapter the least squares estimate for β is found by minimizing

$$Q(\beta) = \epsilon'\epsilon.$$

The solution to the minimization problem satisfies

$$\frac{\partial Q(\beta)}{\partial \beta} = -2X'y + 2X'X\beta = 0.$$

The above can be written as the normal equations

$$X'X\beta = X'y.$$

If $rank[X] = p$, [X is said to be full column rank], the normal equations have an unique solution given by

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Inference

Let $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$, in which case we have

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \sim N_n(0, \sigma^2 I_n)$$

and

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \sim N_n(X\beta, \sigma^2 I_n).$$

Since, $\hat{\beta} = Ly$, $\hat{y} = Hy$, and $\hat{\epsilon} = (I - H)y$ for $L = (X'X)^{-1}X'$ and $H = X(X'X)^{-1}X'$, one has

1. $\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1})$.
   (a) $\hat{\beta}$ is an unbiased estimate of β.
   (b) $var(\hat{\beta}_i) = \sigma^2((X'X)^{-1})_{ii}$.
   (c) $cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2((X'X)^{-1})_{ij}$.
   (d) $corr(\hat{\beta}_i, \hat{\beta}_j) = ((X'X)^{-1})_{ij}/[((X'X)^{-1})_{ii}((X'X)^{-1})_{jj}]^{1/2}$.
2. $\hat{y} \sim N_n(X\beta, \sigma^2 H)$.
   (a) $var(\hat{y}_i) = \sigma^2 h_{ii}$.

(b) $cov(\hat{y}_i, \hat{y}_j) = \sigma^2 h_{ij}$, where $H = (h_{ij})$. Notice that the $\hat{y}_i's$ are not independent of one another unless each of the $h_{ij} = 0$.

(c) $corr(\hat{y}_i, \hat{y}_j) = h_{ij}/[h_{ii}h_{jj}]^{1/2}$.

3. $\hat{\epsilon} \sim N_n(0, \sigma^2(I - H))$.

(a) $var(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$.

(b) $cov(\hat{\epsilon}_i, \hat{\epsilon}_j) = -\sigma^2 h_{ij}$.

(c) $corr(\hat{\epsilon}_i, \hat{\epsilon}_j) = -h_{ij}/[(1 - h_{ii})(1 - h_{jj})]^{1/2}$.

Estimation of σ^2

The estimation of σ^2 follows from observing that

$$\begin{aligned} E(Q(\hat{\beta}) = y'(I - H)y) &= \sigma^2 tr[(I - H)] + \beta' X'(I - H)X\beta \\ &= \sigma^2 tr[(I - H)] \\ &= \sigma^2(tr[I] - tr[H]) \\ &= \sigma^2(n - p), \end{aligned}$$

where $E(y) = X\beta$ and $cov(y) = V = \sigma^2 I_n$. In which case, the least squares estimate for σ^2 is

$$\hat{\sigma}^2 = y'(I - H)y/(n - p) = SS_E/(n - p).$$

ANOVA Table

As in the previous chapter one has the analysis of variance table given by

Source	Sum of Squares	Degrees of Freedom	Mean Square
due to β	$SS(\beta) = \hat{\beta}' X' y = y' H y$	p	$MS(\beta) = SS(\beta)/p$
Residual	$SS_E = y'y - \hat{\beta}' X' y = y'(I - H)y$	n-p	$MS_E = SS_E/(n - p)$
Uncorrected Total	$y'y$	n	

Since β_0 is a free parameter, the sum of squares term is adjusted for β_0, that is

$$y' H y = y'(H - \frac{1}{n}\mathbf{j}\mathbf{j}')y + \frac{1}{n} y'\mathbf{j}\mathbf{j}' y$$

or

$$SS(\beta) = SS(\beta_0, \beta_*) = SS(\beta_* \mid \beta_0) + SS(\beta_0),$$

where $SS(\beta_0) = \frac{1}{n} y'\mathbf{j}\mathbf{j}' y = n\bar{y}^2$ is the correction factor or the sum of squares due to β_0. In which case the ANOVA table become

Source	Sum of Squares	Degrees of Freedom	Mean Square
due to $\beta_* \mid \beta_0$	$SS(\beta_* \mid \beta_0) = y'Hy - \frac{1}{n}y\mathbf{jj}'y$	p-1	$MS(\beta_* \mid \beta_0) = SS(\beta_* \mid \beta_0)/(p-1)$
Residual	$SS_E = y'y - \hat{\beta}'X'y = y'(I-H)y$	n-p	$MS_E = SS_E/(n-p)$
Corrected Total	$SS_{CT} = y'y - \frac{1}{n}y'\mathbf{jj}'y$	n-1	
due to β_0	$SS(\beta_0) = \frac{1}{n}y'\mathbf{jj}'y = n\bar{y}^2$	1	
Uncorrected Total	$y'y$	n	

where $\beta_* = (\beta_1, \beta_2, \ldots, \beta_{p-1})$

Expected Values of the Sum of Squares

As before, it follows when $V = \sigma^2 I_n$ that

1. $E(y'Hy) = \sigma^2 tr[H] + \beta'X'HX\beta = p\sigma^2 + \beta'X'X\beta$.
2. $E(y'(I-H)y) = \sigma^2 tr[(I-H)] + \beta'X'(I-H)X\beta = (n-p)\sigma^2$.

The R^2 is an indicator of how much of the variation in the data is explained by the model. It is defined as

$$R^2 = \frac{SS(\beta_* \mid \beta_0)}{SS_{CT}} = \frac{y'Hy - \frac{1}{n}y'\mathbf{jj}'y}{y'y - \frac{1}{n}y'\mathbf{jj}'y}.$$

The adjusted R^2 is the R^2 value adjusted for the number of parameters in the model and is given by

$$adjR^2 = 1 - [(n-i)/(n-p)](1-R^2)$$

where i is 1 if the model includes the y intercept (β_0), and is 0 otherwise.

Distribution of the Mean Squares

Again from the properties of the distribution of quadratic forms it can be show that

1. $SS(\beta)/\sigma^2 \sim \chi^2(df = p, \lambda = 1/2\beta'X'X\beta)$.
2. $SS_E/\sigma^2 \sim \chi^2(df = n-p)$.
3. $SS(\beta)$ and SS_E are independent.
4. $F = MS(\beta)/MS_E \sim F(df_1 = p, df_2 = n-p), \lambda = 1/2\beta'X'X\beta)$.
5. When $\beta = 0$ it follows that $SS(\beta)/\sigma^2 \sim \chi^2(df = p)$ and $F = MS(\beta)/MS_E \sim F(df_1 = p, df_2 = n-p)$.
6. $SS(\beta_* \mid \beta_0)/\sigma^2 \sim \chi^2(df = p-1, \lambda = 1/2\beta_*'X_*'X_*\beta_*)$, where

$$X_* = \begin{pmatrix} x_{11} & x_{21} & \ldots & x_{(p-1)1} \\ x_{12} & x_{21} & \ldots & x_{(p-1)1} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{2n} & \ldots & x_{(p-1)n} \end{pmatrix}$$

and

$$\beta_* = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}.$$

7. $SS(\beta_* \mid \beta_0)$ and SS_E are independent.
8. $F = MS(\beta_* \mid \beta_0)/MS_E \sim F(df_1 = p-1, df_2 = n-p, \lambda = 1/2\beta_*' X_*' X_* \beta_*)$.
9. When $\beta_* = 0$ it follows that $SS(\beta_* \mid \beta_0)/\sigma^2 \sim \chi^2(df = p-1)$ and $F = MS(\beta_* \mid \beta_0)/MS_E \sim F(df_1 = p-1, df_2 = n-p)$.
10. In the general linear regression model, one rejects the null hypothesis

$$H_0 : \beta_* = 0 \quad \text{versus} \quad H_1 : \beta_* \neq 0$$

if $F = MS(\beta_* \mid \beta_0)/MS_E > F_\alpha(p-1, n-p)$. Then the power is computed as

$$\text{Power}(\beta_*) = \Pr[W > F_\alpha(p-1, n-p)].$$

where $W \sim F(p-1, n-p, \lambda = 1/2\beta_*' X_*' X_* \beta_*)$.

The Reduction Notation

In constructing the ANOVA tables one is interested in describing the sum of squares attributed to various terms in the model. This can be done in one of two ways, sequential or partial. Suppose that one has the three term model given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i.$$

If the model is a polynomial regression model of order 2 (i.e., $x_{ji} = x_i^j$) then the model sum of squares SS_M should be written as (sequential ordering)

$$SS_M = S(\beta_0) + S(\beta_1 \mid \beta_0) + S(\beta_2 \mid \beta_0, \beta_1) + S(\beta_3 \mid \beta_0, \beta_1, \beta_2).$$

Where each term is an independent quadratic forms with non-central chi-square distributions with degrees of freedom = 1.

When the model is a multiple regression model, one is interested in determining the amount of reduction that can be attributed to each variable (partial ordering)

$$S(\beta_3 \mid \beta_0, \beta_1, \beta_2)$$
$$S(\beta_2 \mid \beta_0, \beta_1, \beta_3)$$
$$S(\beta_1 \mid \beta_0, \beta_2, \beta_3).$$

These quadratic forms are no longer independent and their sum does not equal SS_M. However, they are independent of the error sum of squares and they can be shown to be a non-central chi-square distributed with one degree of freedom.

When using SAS the sequential sum of squares is the standard output and can be specified with the Type I SS statement. The partial SS can be obtained using the Type II SS statement. The Type I SS should be used when

- Balance ANOVA model with terms in their proper sequence or order.
- Purely nested model in proper order.
- **Polynomial regression** models.

The Type II SS statement should be used when

- The model is balanced.
- Any main effects model.
- **Multiple regression** model.
- An effect not contained in any other effect (no nesting).

R

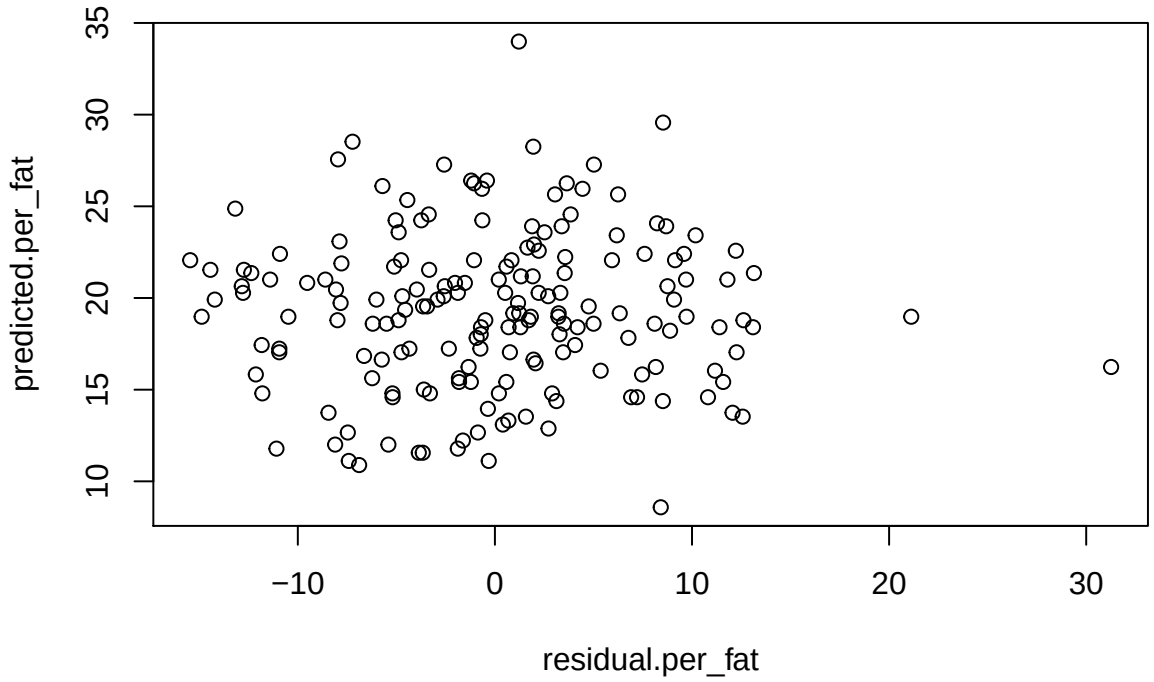
Polynominal Regression

```
knee2=knee*knee
mod2 = lm(per_fat ~ knee + knee2, data=bf)
summary(mod2)
```

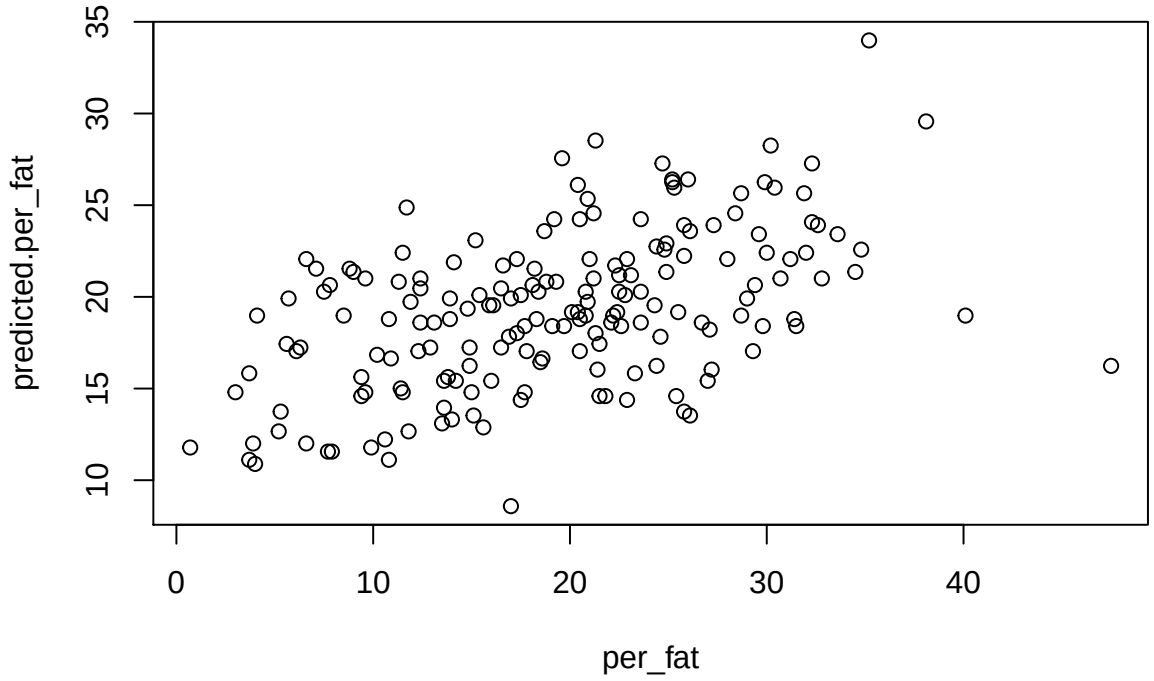
```
##
## Call:
## lm(formula = per_fat ~ knee + knee2, data = bf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4606  -4.8848  -0.0584   3.8987  31.2631
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -121.82103   89.20318  -1.366    0.174
## knee           5.45984    4.52848   1.206    0.230
## knee2         -0.04657    0.05738  -0.812    0.418
##
## Residual standard error: 7.387 on 177 degrees of freedom
## Multiple R-squared:  0.2527, Adjusted R-squared:  0.2443
## F-statistic: 29.93 on 2 and 177 DF,  p-value: 6.339e-12
```

Predicted values and Residuals

```
residual.per_fat=residuals(mod2)
predicted.per_fat=predict(mod2)
plot(residual.per_fat, predicted.per_fat)
```




```
plot(per_fat, predicted.per_fat)
```



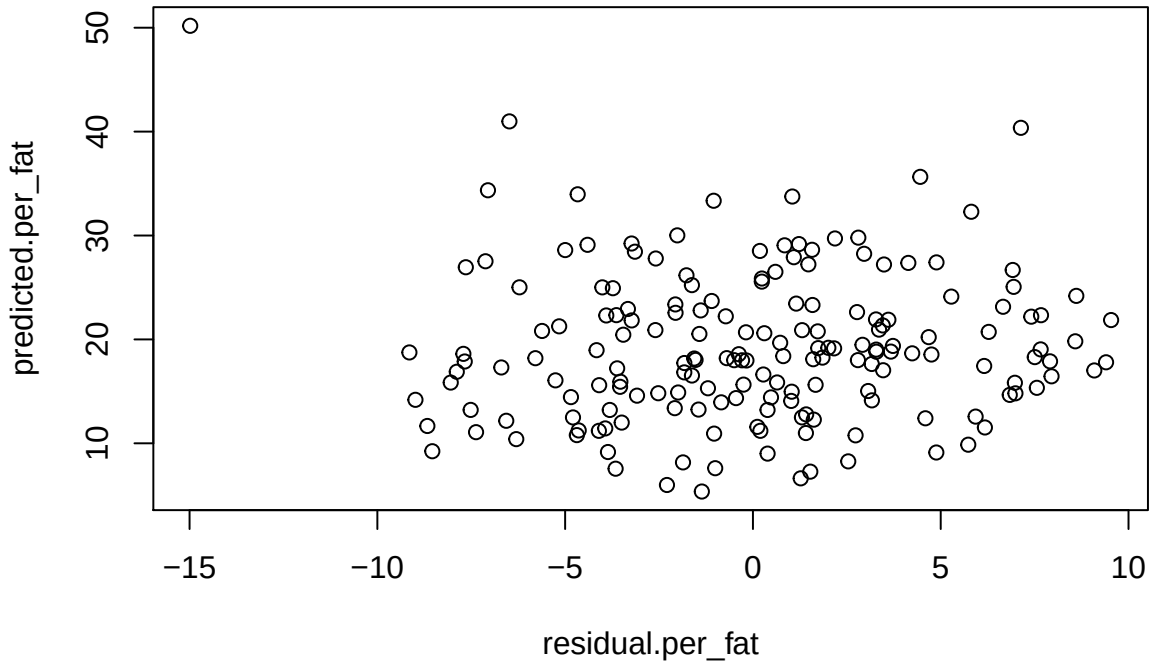
Multiple Regression

```
mod3 = lm(per_fat ~ abdomen + thigh + neck, data=bf)
summary(mod3)
```

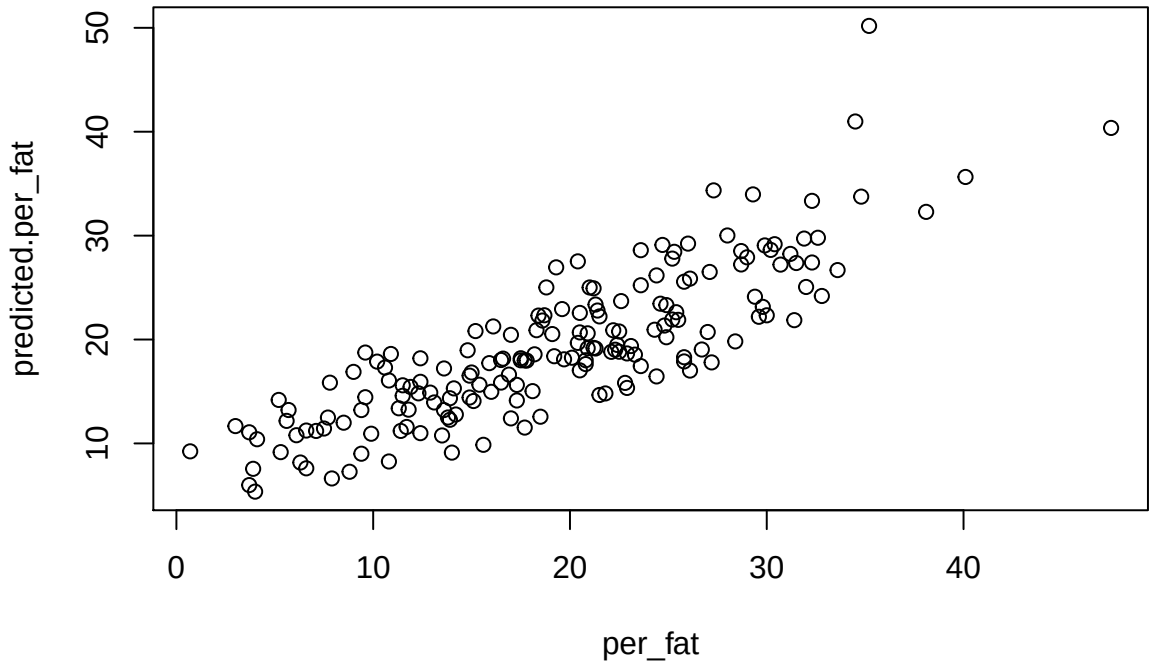
```
##
## Call:
## lm(formula = per_fat ~ abdomen + thigh + neck, data = bf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9836  -3.4626   0.2172   3.1647   9.5368
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.92951    5.81094  -2.741  0.00675 **
## abdomen       0.83165    0.05705  14.578  < 2e-16 ***
## thigh        -0.08537    0.11205  -0.762  0.44712
## neck         -0.96877    0.23741  -4.081  6.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.608 on 176 degrees of freedom
## Multiple R-squared:  0.7109, Adjusted R-squared:  0.7059
## F-statistic: 144.2 on 3 and 176 DF,  p-value: < 2.2e-16
```

Predicted values and Residuals

```
residual.per_fat=residuals(mod3)
predicted.per_fat=predict(mod3)
plot(residual.per_fat, predicted.per_fat)
```




```
plot(per_fat, predicted.per_fat)
```



SAS

```
title2 'Polynomial Regression';
proc reg data=new_bfat;*  plots = (diagnostics partial);
model per_fat=neck neck2/ press partial ss1;
run;

title2 'Multiple Regression';
proc reg data=new_bfat;*  plots = (diagnostics partial);
model per_fat=neck thigh abdomen/press partial ss1 ss2;
run;
```

Body Fat Data

Polynomial Regression

The REG Procedure

Model: MODEL1

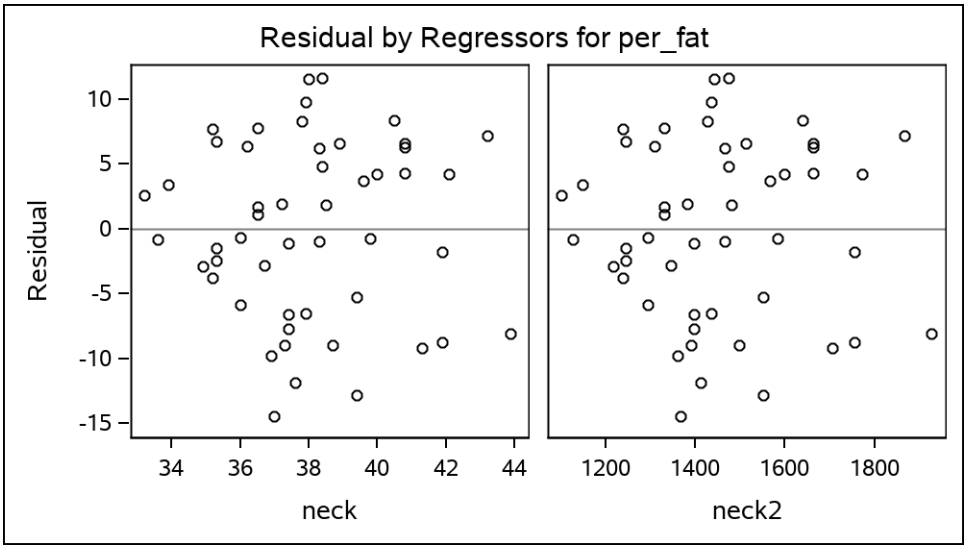
Dependent Variable: per_fat

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	793.92706	396.96353	8.12	0.0009
Error	47	2297.45294	48.88198		
Corrected Total	49	3091.38000			

Root MSE	6.99156	R-Square	0.2568
Dependent Mean	19.08000	Adj R-Sq	0.2252
Coeff Var	36.64342		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	−60.51507	196.03634	−0.31	0.7589	18202
neck	1	2.54497	10.22964	0.25	0.8046	793.54122
neck2	1	−0.01183	0.13311	−0.09	0.9296	0.38584



Body Fat Data

Multiple Regression

The REG Procedure

Model: MODEL1

Dependent Variable: per_fat

Number of Observations Read	50
Number of Observations Used	50

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2109.16734	703.05578	32.93	<.0001
Error	46	982.21266	21.35245		
Corrected Total	49	3091.38000			

Root MSE	4.62087	R-Square	0.6823
Dependent Mean	19.08000	Adj R-Sq	0.6616
Coeff Var	24.21840		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	−11.07899	11.46198	−0.97	0.3388	18202	19.94935
neck	1	−1.11408	0.45570	−2.44	0.0184	793.54122	127.62498
thigh	1	−0.14763	0.19770	−0.75	0.4590	189.73417	11.90642
abdomen	1	0.88000	0.12119	7.26	<.0001	1125.89194	1125.89194