Chapter 11

Example: The **pressure** dataset is provided in the base version of R. The data is on the relation between temperature in degrees Celsius and vapor pressure of mercury in millimeters (of mercury).

Let's begin by looking at the header of the data and calculating the correlation.

```
data(pressure)
head(pressure)

##   temperature pressure
## 1           0   0.0002
## 2          20   0.0012
## 3          40   0.0060
## 4          60   0.0300
## 5          80   0.0900
## 6         100   0.2700

cor(pressure)

##             temperature  pressure
## temperature   1.0000000 0.7577923
## pressure      0.7577923 1.0000000
```

The correlation matrix is given in the example above. The diagonal elements of the correlation matrix shows that any variable is perfectly correlated with itself (i.e., $r_{xx} = 1$). In the off diagonal, you can see the correlation between the **temperature** and **pressure** is 0.758 which a strong, positive linear relationship.

Let's fit a simple linear regression model with the **temperature** being the predictor variable and **pressure** being the response variable.

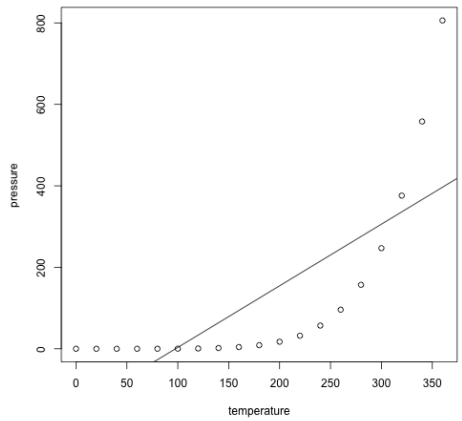
```
model = lm(pressure~temperature,data=pressure)
summary(model)

##
## Call:
## lm(formula = pressure ~ temperature, data = pressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158.08 -117.06  -32.84   72.30  409.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -147.8989    66.5529  -2.222 0.040124 *
## temperature    1.5124     0.3158   4.788 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.8 on 17 degrees of freedom
## Multiple R-squared:  0.5742,Adjusted R-squared:  0.5492
## F-statistic: 22.93 on 1 and 17 DF,  p-value: 0.000171
```

The estimated regression is $\widehat{pressure} = -147.90 + 1.51 \cdot temperature$. A one degree Celsius increase in temperate of mercury is associated with an increase of 1.51 in vapor pressure.

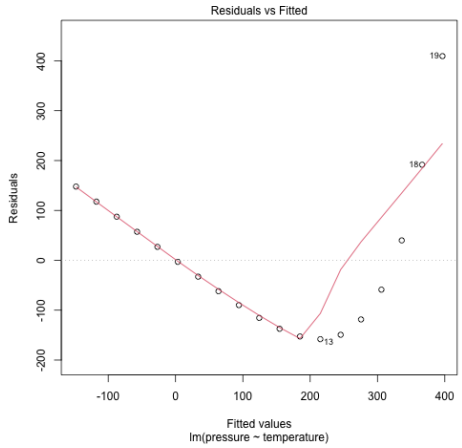
Is this model appropriate? Let's plot the data along with the line of best fit.

```
plot(pressure~temperature,data=pressure)
abline(model)
```



As seen in the above plot, our simple linear regression model is completely inadequate. We could determine that a nonlinearity exists by looking at the first residual diagnostic plot.

```
plot(model,which=1)
```



In this diagnostic plot, we see strong evidence of an unaccounted for nonlinearity as the shape of the graph is not approximately flat.

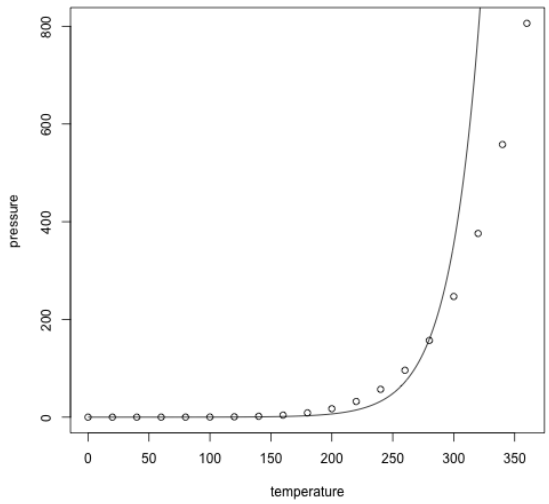
Based on the scatterplot on the previous page, there appears to be a possibly exponential relationship between temperature and pressure. Let's try log transforming pressure and refitting the linear regression model. (Note: Variable transformations are covered in Chapter 12.)

```
model2 = lm(log(pressure)~temperature,data=pressure)
summary(model2)

##
## Call:
## lm(formula = log(pressure) ~ temperature, data = pressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4491 -0.6876  0.2866  0.8716  1.1365
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.068144   0.483831  -12.54 5.10e-10 ***
## temperature  0.039792   0.002296   17.33 3.07e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 17 degrees of freedom
## Multiple R-squared:  0.9464,Adjusted R-squared:  0.9433
## F-statistic: 300.3 on 1 and 17 DF,  p-value: 3.07e-12
```

```
plot(pressure~temperature,data=pressure)

x.new = 0:350
y.new = predict(model2,newdata=list(temperature=x.new,interval="confidence"))
lines(x.new,exp(y.new))
```



This model seems to fit the data better than the untransformed simple linear regression model but is still somewhat inadequate for temperatures greater than 300C. Additionally, the residual diagnostic plot suggest our model is inadequate. A better model should be found if modeling temperatures above 300C is of interest.

```
plot(model2,which=1)
```