Math 3080 § 1.	**Truck Example:**	Name: Example
Treibergs	**Simple Regression**	Feb. 15, 2014

Simple regression fits the best linear function through observed points in the plane. Today's data was obtained in J. Yanowitz', PhD Thesis "In-Use Emission of Heavy-Duty Diesel Vehicles," Colorado School of Mines 2001 as quoted by Navidi, *Statistics for Engineers and Scientists,* 2nd ed., McGraw Hill, 2008.

It is assumed that the response variable Y, Mileage in this case, is normally distributed with a constant variance σ^2 and with a mean that depends linearly on the explanatory variable x, Weight in this case,

$$y = \beta_0 + \beta_1 x + \epsilon \qquad \text{where } \epsilon \sim N(0, \sigma^2).$$

Given the observed values $\{(x_i, y_i)\}_{i=1,\ldots,n}$, the best line is the one that minimizes the sum of squared deviations. If the proposed line is $y = a + bx$ then the ith deviation is $y_i - a - bx_i$ and the sum square of deviations is

$$f(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

This is the sum of convex quadratic functions, so is convex in (a, b) and whose minimum may by determined by setting partial derivatives to zero and solving

$$\begin{aligned} 0 &= \frac{\partial f}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) \\ 0 &= \frac{\partial f}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i \end{aligned} \tag{1}$$

The resulting system of equations is

$$\begin{aligned} na + \left(\sum_{i=1}^{n} x_i \right) b &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i \right) a + \left(\sum_{i=1}^{n} x_i^2 \right) b &= \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Using Cramer's rule we find the solution $(a, b) = (\widehat{\beta_0}, \widehat{\beta_1})$ given by

$$\widehat{\beta_1} = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix}} = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{S_{xy}}{S_{xx}}$$

where

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right) \end{aligned}$$

Note that the system is nonsingular if there are at least two distinct x_i so $S_{xx} \neq 0$. If so, we may solve (1) to get

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}.$$

The *mean response* when $x = x^*$ is the point on the fitted line

$$y^* = \widehat{\beta_0} + \widehat{\beta_1}x^*.$$

The *fitted value* is the mean response in case $x = x_i$, in other words

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1}x_i.$$

The *predicted value*, $\hat{y}$, is an estimator for the next observation when $x = x^*$. The variability comes from both the error in the next observation and the variability of the response.

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x^*.$$

As usual, we write random variables as upper case letters. For example, the coefficients of regression $\widehat{B_i}$ computed from a random sample are random variables.

Lemma 1. *$\widehat{B_0}$ and $\widehat{B_1}$ are unbiased estimators for β_0 and β_1 resp. Moreover, for fixed x^*, let the mean response $Y^* = \widehat{B_0} + \widehat{B_1}x^*$ and the predicted value be $\hat{Y} = \widehat{B_0} + \widehat{B_1}x^*$. We have*

$$E(\widehat{B_0}) = \beta_0 \qquad E(\widehat{B_1}) = \beta_1 \qquad \textit{and} \qquad E(Y^*) = E(\hat{Y}) = \beta_0 + \beta_1 x^*.$$

Proof. We have

$$\begin{aligned} E(\bar{Y}|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) \\ &= \beta_0 + \beta_1 \bar{x}. \end{aligned}$$

Now, given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$,

$$\begin{aligned} E\left(\widehat{B_1}\right) &= E\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{j=1}^n (x_j - \bar{x})^2}\right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} E\left(Y_i - \bar{Y}\right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{j=1}^n (x_j - \bar{x})^2} \left(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x}\right) \\ &= \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} \\ &= \beta_1 \end{aligned}$$

so

$$E\left(\widehat{B_0}\right) = E\left(\bar{Y} - B_1\bar{x}\right) = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$$

and

$$E\left(Y^*\right) = E\left(\widehat{B_0} + \widehat{B_1}x^*\right) = \beta_0 + \beta_1 x^*.$$

□

We compute the variances of the regression coefficients, mean response and predicted value.

Lemma 2. *For fixed x^*, the mean response and predicted value is $Y^* = \hat{Y} = \widehat{B_0} + \widehat{B_1}x^*$. The variances of $\widehat{B_0}$, $\widehat{B_1}$, Y^* and $\hat{Y}$ are given by*

$$V(\widehat{B_0}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right), \qquad V(\widehat{B_1}) = \frac{\sigma^2}{S_{xx}},$$

$$V(Y^*) = \sigma^2\left(\frac{1}{n} + \frac{(\bar{x} - x^*)^2}{S_{xx}}\right) \qquad \textit{and} \qquad V(\hat{Y}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(\bar{x} - x^*)^2}{S_{xx}}\right)$$

Proof. First,

$$\begin{aligned} V(Y_i - \bar{Y}) &= V\left(\frac{n-1}{n}Y_i - \frac{1}{n}\sum_{j\neq i} Y_j\right) \\ &= \frac{(n-1)^2}{n^2}\sigma^2 + \frac{n-1}{n^2}\sigma^2 = \frac{n-1}{n}\sigma^2 \end{aligned}$$

Note that $\sum_{i=1}^n (x_i - \bar{x}) = 0$ so

$$\begin{aligned} V(\widehat{B_1}) &= V\left(\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{j=1}^n (x_j - \bar{x})^2}\right) \\ &= \frac{1}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} V\left(\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y}\sum_{i=1}^n (x_i - \bar{x})\right) \\ &= \frac{1}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} V\left(\sum_{i=1}^n (x_i - \bar{x})Y_i\right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2 V(Y_i)}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}. \end{aligned}$$

Second, since $\mathrm{Cov}(Y_i, Y_j) = 0$ if $i \neq j$,

$$\begin{aligned} \mathrm{Cov}(\bar{Y}, \widehat{B_1}) &= \mathrm{Cov}\left(\bar{Y}, \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{k=1}^n (x_k - \bar{x})^2}\right) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})\,\mathrm{Cov}\left(\bar{Y}, Y_i - \bar{Y}\right)}{\sum_{k=1}^n (x_k - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})\,\mathrm{Cov}\left(\frac{1}{n}\sum_{j=1}^n Y_j, \frac{n-1}{n}Y_i - \frac{1}{n}\sum_{k\neq i} Y_k\right)}{\sum_{k=1}^n (x_k - \bar{x})^2} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})\left[(n-1)V(Y_i) - \sum_{j\neq i} V(Y_j)\right]}{n^2\sum_{k=1}^n (x_k - \bar{x})^2} \\ &= \sigma^2\frac{\sum_{i=1}^n (x_i - \bar{x})[(n-1) - (n-1)]}{\sum_{k=1}^n (x_k - \bar{x})^2} = 0. \end{aligned}$$

Thus

$$\begin{aligned} V(\widehat{B_0}) &= V(\bar{Y} - \widehat{B_1}\bar{x}) \\ &= V(\bar{Y}) - 2\bar{x}\,\mathrm{Cov}(\bar{Y}, \widehat{B_1}) + \bar{x}^2 V(\widehat{B_1}) \\ &= \sigma^2 \left(\frac{1}{n^2} - 0 + \frac{\bar{x}^2}{S_{xx}} \right). \end{aligned}$$

The variance of the mean response follows:

$$\begin{aligned} V(Y^*) &= V(\widehat{B}_0 + \widehat{B_1}x^*) \\ &= V(\bar{Y} + \widehat{B_1}(x^* - \bar{x})) \\ &= V(\bar{Y}) - 2(x^* - \bar{x})\,\mathrm{Cov}(\bar{Y}, \widehat{B_1}) + (x^* - \bar{x})^2 V(\widehat{B_1}) \\ &= \sigma^2 \left(\frac{1}{n^2} - 0 + \frac{(x^* - \bar{x})^2}{S_{xx}} \right). \end{aligned}$$

Finally, the error in predicting the next point Y_{n+1} at $x_{n+1} = x^*$ with Y^* is the variability of

$$Y_{n+1} - \widehat{B_0} - \widehat{B_1}x^*.$$

But Y_{n+1} is independent of $\{Y_i\}_{i=1,\dots,n}$ so

$$\begin{aligned} V(\hat{Y}) &= V(Y_{n+1} - \widehat{B}_0 - \widehat{B_1}x^*) \\ &= V(Y_{n+1}) + V(\widehat{B}_0 + \widehat{B_1}x^*) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n^2} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right). \quad \square \end{aligned}$$

The *sum square error* is

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

with corresponding degrees of freedom $n - 2$ because the parameters in the formula $(\widehat{\beta_0}, \widehat{\beta_1})$ have already used up two degrees of freedom.

Lemma 3. *The sum squared error has a shortcut formula*

$$SSE = \sum_{i=1}^{n} y_i^2 - \widehat{\beta_0} \sum_{i=1}^{n} x_i - \widehat{\beta_1} \sum_{i=1}^{n} x_i y_i.$$

Proof. Observe that

$$\sum_{i=1}^{n} \left[y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i \right] = \sum_{i=1}^{n} \left[y_i - \bar{y} - \widehat{\beta_1}(x_i - \bar{x}) \right] = n\bar{y} - n\bar{y} - \widehat{\beta_1}(n\bar{x} - n\bar{x}) = 0$$

and using this,

$$\begin{aligned}\sum_{i=1}^{n}\left[y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i\right]x_i &= \sum_{i=1}^{n}\left[y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i\right](x_i-\bar{x})\\ &= \sum_{i=1}^{n}\left[y_i-\bar{y}-\widehat{\beta_1}(x_i-\bar{x})\right](x_i-\bar{x})\\ &= \sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+\widehat{\beta_1}\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &= \sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+\frac{R_{xy}}{R_{xx}}\sum_{i=1}^{n}(x_i-\bar{x})^2\\ &= \sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})-\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})\\ &= 0.\end{aligned}$$

Then the short cut formula follows

$$\begin{aligned}SSE &= \sum_{i=1}^{n}(y_i-\widehat{y_i})^2\\ &= \sum_{i=1}^{n}(y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i)^2\\ &= \sum_{i=1}^{n}(y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i)y_i-\widehat{\beta_0}\sum_{i=1}^{n}(y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i)-\widehat{\beta_1}\sum_{i=1}^{n}(y_i-\widehat{\beta_0}-\widehat{\beta_1}x_i)x_i\\ &= \sum_{i=1}^{n}y_i^2-\widehat{\beta_0}\sum_{i=1}^{n}y_i-\widehat{\beta_1}\sum_{i=1}^{n}x_iy_i.\end{aligned}$$

□

The *mean square error* is

$$MSE=\frac{SSE}{n-2}.$$

Lemma 4. *The mean squared error is an unbiased estimator for the variance*

$$E(MSE)=\sigma^2.$$

Proof. Given $X_1=x_1, X_2=x_2,\ldots,X_n=x_n$, because

$$E(y_i-\widehat{B_0}-\widehat{B_1}x_i)=\beta_0+\beta_1x_i-\beta_0-\beta_1x_i=0,$$

the identity $E(Z^2)=V(Z)+E^2(Z)$ gives

$$\begin{aligned}E(SSE) &= E\left(\sum_{i=1}^{n}(Y_i-\widehat{Y_i})^2\right)\\ &= E\left(\sum_{i=1}^{n}(Y_i-\widehat{B_0}-\widehat{B_1}x_i)^2\right)\\ &= \sum_{i=1}^{n}E\left((Y_i-\widehat{B_0}-\widehat{B_1}x_i)^2\right)\\ &= \sum_{i=1}^{n}V(Y_i-\hat{Y}-\widehat{B_1}(x_i-\bar{x}))\end{aligned}$$

But

$$V(Y_i - \bar{Y} - \widehat{B_1}(x_i - \bar{x})) = V(Y_i - \bar{Y}) - 2(x_i - \bar{x})\operatorname{Cov}(Y_i - \bar{Y}, \widehat{B_1}) + (x_i - \bar{x})^2 V(\widehat{B_1})$$

Finally, since $\operatorname{Cov}(Y_i, Y_j) = 0$ if $i \neq j$ and $\sum_{i \neq j}(x_j - \bar{x}) = -x_i + \bar{x}$ imply

$$\begin{aligned}
\operatorname{Cov}(Y_i - \bar{Y}, \widehat{B_1}) &= \operatorname{Cov}\left(Y_i - \bar{Y}, \frac{\sum_{j=1}^n (x_j - \bar{x})(Y_j - \bar{Y})}{\sum_{k=1}^n (x_k - \bar{x})^2}\right) \\
&= \frac{(x_i - \bar{x})V(Y_i - \bar{Y}) + \sum_{j \neq i}(x_j - \bar{x})\operatorname{Cov}\left(Y_i - \bar{Y}, Y_j - \bar{Y}\right)}{\sum_{k=1}^n (x_k - \bar{x})^2} \\
&= \frac{(n-1)(x_i - \bar{x})\sigma^2}{n\sum_{k=1}^n (x_k - \bar{x})^2} + \frac{\sum_{j \neq i}(x_j - \bar{x})\operatorname{Cov}\left(\frac{n-1}{n}Y_i - \frac{1}{n}\sum_{k \neq i} Y_k, \frac{n-1}{n}Y_j - \frac{1}{n}\sum_{\ell \neq j} Y_\ell\right)}{\sum_{k=1}^n (x_k - \bar{x})^2} \\
&= \frac{(n-1)(x_i - \bar{x})\sigma^2}{n\sum_{k=1}^n (x_k - \bar{x})^2} + \frac{\sum_{j \neq i}(x_j - \bar{x})\left(-(n-1)V(Y_i) - (n-1)V(Y_j) + \sum_{k \neq i,j} V(Y_k)\right)}{n^2\sum_{k=1}^n (x_k - \bar{x})^2} \\
&= \frac{(n-1)(x_i - \bar{x})\sigma^2}{n\sum_{k=1}^n (x_k - \bar{x})^2} - \frac{\sum_{j \neq i}(x_j - \bar{x})\sigma^2}{n\sum_{k=1}^n (x_k - \bar{x})^2} \\
&= \frac{(n-1)(x_i - \bar{x})\sigma^2}{n\sum_{k=1}^n (x_k - \bar{x})^2} + \frac{(x_i - \bar{x})\sigma^2}{n\sum_{k=1}^n (x_k - \bar{x})^2} \\
&= \frac{(x_i - \bar{x})\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2}
\end{aligned}$$

Inserting yields

$$\begin{aligned}
E(SSE) &= \sum_{i=1}^n \left[V(Y_i - \bar{Y}) - 2(x_i - \bar{x})\operatorname{Cov}(Y_i - \bar{Y}, \widehat{B_1}) + (x_i - \bar{x})^2 V(\widehat{B_1})\right] \\
&= \sum_{i=1}^n \left[\frac{n-1}{n} - 2\frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\sigma^2 \\
&= [n-2]\,\sigma^2
\end{aligned}$$

□

These lemmas gives estimated standard errors on the quantities.

$$\begin{aligned}
s &= \sqrt{MSE}, \\
s_{\widehat{\beta_0}} &= s\sqrt{\frac{1}{n^2} + \frac{\bar{x}^2}{S_{xx}}}, \\
s_{\widehat{\beta_1}} &= \frac{s}{\sqrt{S_{xx}}}, \\
s_{Y^*} &= s\sqrt{\frac{1}{n^2} + \frac{(x^* - \bar{x})^2}{S_{xx}}}, \\
s_{\hat{Y}} &= s\sqrt{1 + \frac{1}{n^2} + \frac{(x^* - \bar{x})^2}{S_{xx}}}
\end{aligned}$$

We may also define the *total sum square* which adds squares of deviations from the mean, and the *regression sum square* or *residual sum square* measuring the part of variation accounted for by the model by the formulae

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \widehat{\beta_1}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{S_{xy}^2}{S_{xx}}$$

The total sum square has $n-1$ degrees of freedom and the regression sum square has one degree of freedom.

Lemma 5. *The analysis of variance identity,* also called the *sum of squares identity* holds

$$SST = SSR + SSE.$$

Proof. By the shortcut formula,

$$\begin{aligned}
SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \\
&= \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i + \widehat{\beta_0} + \widehat{\beta_1} x_i \right)^2 - n\bar{y}^2 \\
&= \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i \right)^2 + 2 \sum_{i=1}^{n} (y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)(\widehat{\beta_0} + \widehat{\beta_1} x_i) + \sum_{i=1}^{n} \left(\widehat{\beta_0} + \widehat{\beta_1} x_i \right)^2 - n\bar{y}^2 \\
&= SSE + 0 + \sum_{i=1}^{n} \left(\bar{y} + \widehat{\beta_1}(x_i - \bar{x}) \right)^2 - n\bar{y}^2 \\
&= SSE + n\bar{y}^2 + 2\widehat{\beta_1}\bar{y} \sum_{i=1}^{n} (x_i - \bar{x}) + \widehat{\beta_1}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - n\bar{y}^2 \\
&= SSE + SSR
\end{aligned}$$

□

It follows that the *coefficient of determination,*

$$r^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

which falls between $0 \le r^2 \le 1$ is the fraction of variation accounted for by the model. r is called the *residual standard error.*

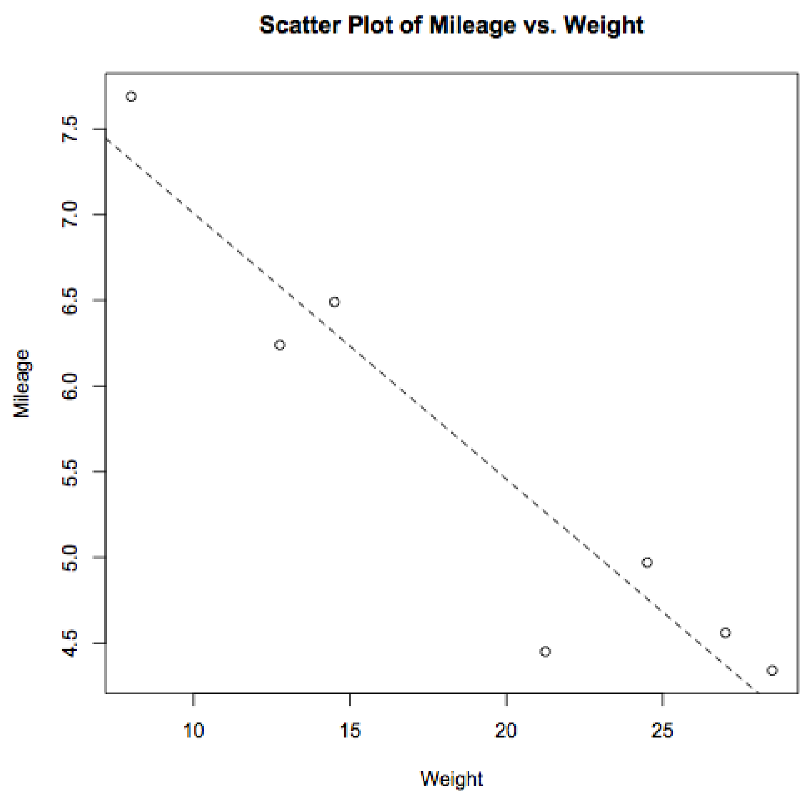
Data Set Used in this Analysis :

```
# Math 3080 - 1      Truck Data          Feb. 15, 2014
# Treibergs
#
# The following data was obtained by J. Yanowitz, PhD Thesis "In-Use
# Emission of Heavy-Duty Diesel Vehicles," Colorado School of Mines 2001 as
# quoted by Navidi, Statistics for Engineers and Scientists, 2nd ed.,
# McGraw Hill, 2008. Inertial weights (in tons) and fuel economy
# (in mi/gal) was measured for a sample of seven diesel trucks.
#
"Weight" "Mileage"
   8.00     7.69
  24.50     4.97
  27.00     4.56
  14.50     6.49
  28.50     4.34
  12.75     6.24
  21.25     4.45
```

R Session:

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.41 (5874) i386-apple-darwin9.8.0]

[History restored from /Users/andrejstreibergs/.Rapp.history]

> tt=read.table("M3082DataTruck.txt",header=T)
> attach(tt)
```

```
> ###############  RUN REGRESSION  ############################
> f1=lm(Mileage~Weight); f1
Call:
lm(formula = Mileage ~ Weight)

Coefficients:
(Intercept)       Weight
     8.5593      -0.1551
> ###############  SCATTERPLOT WITH REGRESSION LINE  ##########
> plot(tt,main="Scatter Plot of Mileage vs. Weight");
> abline(f1,lty=2)
```




```
> ################  COMPUTE ANOVA TABLE 'BY HAND"  #############
> WtBar=mean(Weight); WtBar
[1] 19.5
> MiBar=mean(Mileage); MiBar
[1] 5.534286
> SWt=sum(Weight); SWt
[1] 136.5
> SMi=sum(Mileage); SMi
[1] 38.74
> S2Wt=sum(Weight^2); S2Wt
[1] 3029.875
> S2Mi=sum(Mileage^2); S2Mi
[1] 224.3264
> SWtMi=sum(Weight*Mileage); SWtMi
[1] 698.3225
```

```
> n=length(Weight); n
[1] 7
> Sxx=S2Wt-SWt^2/n; Sxx
[1] 368.125
> Sxy=SWtMi-SWt*SMi/n; Sxy
[1] -57.1075
> b1hat=Sxy/Sxx; b1hat
[1] -0.1551307
> b0hat=MiBar-b1hat*WtBar; b0hat
[1] 8.559335
> SSE = S2Mi -b0hat*SMi -b1hat*SWtMi; SSE
[1] 1.069043
> MSE=SSE/(n-2); MSE
[1] 0.2138087
> SST=S2Mi-MiBar^2/n; SST
[1] 219.9509
> SST=S2Mi-SMi^2/n; SST
[1] 9.928171
> SSR=SST-SSE; SSR
[1] 8.859128
> MSR = MSR
> f=MSR/MSE; f
[1] 41.43484
> PVal=pf(f,1,n-2,lower.tail=F); PVal
[1] 0.001344843
> ########  PRINT THE "ANOVA BY HAND" TABLE  ###############
> at=matrix(c(1,n-2,n-1,SSR,SSE,SST,SSR,MSE,-1,f,-1,-1,PVal,-1,-1),ncol=5)
> colnames(at)=c("      DF","      SS","      MS","       f","   P-Value")
> rownames(at)=c("Weight","Error","Total")
> noquote(formatC(at,width=9,digits=7))

             DF        SS        MS         f     P-Value
Weight        1  8.859128  8.859128  41.43484 0.001344843
Error         5  1.069043 0.2138087        -1        -1
Total         6  9.928171        -1        -1        -1


> ################  COMPARE TO CANNED TABLE  ##############
> anova(f1)
Analysis of Variance Table

Response: Mileage
          Df Sum Sq Mean Sq F value   Pr(>F)
Weight     1 8.8591  8.8591  41.435 0.001345 **
Residuals  5 1.0690  0.2138
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
```

```
> ##########  RESIDUALS, COEFFICIENTS AND R2 BY HAND  ######
>
>  noquote(formatC(matrix(Mileage-b0hat-b1hat*Weight,ncol=7)))
     [,1]   [,2]   [,3]   [,4]   [,5]   [,6]    [,7]
[1,] 0.3717 0.2114 0.1892 0.1801 0.2019 -0.3414 -0.8128

> SSE
[1] 1.069043
> #########  SSE  THE HARD WAY  ############################

> sum((Mileage-b0hat-b1hat*Weight)^2)
[1] 1.069043

> SSR
[1] 8.859128

> ###############  ESTIMATE OF SIGMA SQUARED  ##############

> MSE
[1] 0.2138087

> ##### ESTIMATE OF SIGMA = RESIDUAL STANDARD ERROR  #######

> s = sqrt(MSE); s
[1] 0.4623945

> ###########  COEFFICIENT OF DETERMINATION  ###############
> R2 = SSR/SST; R2
[1] 0.8923222

> ############  ADJUSTED R^2: WEIGHT NUMBER OF COEFF.  #####
> 1-((n-1)*SSE)/((n-1-1)*SST)
[1] 0.8707867

> ##### ESTIMATE STANDARD ERRORS FOR  BETA_0, BETA_1  #######

> s0 = s*sqrt(1/n + WtBar^2/Sxx); s0
[1] 0.5013931

> s1=s/sqrt(Sxx); s1
[1] 0.02409989

> ### T-SCORES, P-VALUES ASSUMING 2-SIDED, NULL HYP. = 0  ##

> t0 = b0hat/s0; t0
[1] 17.07111

> p0 = 2*pt(abs(t0),n-2,lower.tail=F); p0
[1] 1.262291e-05

> t1 = b1hat/s1; t1
[1] -6.43699
```

```
> p1 = 2*pt(abs(t1),n-2,lower.tail=F); p1
[1] 0.001344843

> ###########  COMPARE TO CANNED SUMMARY  ##################
> summary(f1)
Call:
lm(formula = Mileage ~ Weight)

Residuals:
      1       2       3       4       5       6       7
 0.3717  0.2114  0.1892  0.1801  0.2019 -0.3414 -0.8128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.5593     0.5014  17.071 1.26e-05 ***
Weight       -0.1551     0.0241  -6.437  0.00134 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4624 on 5 degrees of freedom
Multiple R-squared: 0.8923,Adjusted R-squared: 0.8708
F-statistic: 41.43 on 1 and 5 DF,  p-value: 0.001345
```