Linear Regression Analysis

Simple Linear Regression

A homeowner recorded the amount of electricity in kilowatt-hours (KWH) consumed in his house on each of 21 days. He also recorded the numbers of hours his air conditioner (AC) was turned on and the numbers of times his electric clothes dryer (DRYER) was operated. His objective was to relate the KWH consumption to the AC and DRYER usage. In addition, he wanted to know how many KWH's the AC used per hour and the number of KWH's used in each run of the DRYER. Statistical **regression analysis** can serve this purpose.

Following are the data in tabular form:

kwh	ac (hours)	dryer (runs)
35	1.5	1
63	4.5	2
66	5.0	2
17	2.0	0
94	8.5	3
79	6.0	3
93	13.5	1
66	8.0	1
94	12.5	1
82	7.5	2
78	6.5	3
65	8.0	1
77	7.5	2
75	8.0	2
62	7.5	1
85	12.0	1
43	6.0	0
57	2.5	3
33	5.0	0
65	7.5	1
33	6.0	0

In regression terminology, KWH is called the **dependent** variable and AC and DRYER are called the **independent** variables. The names "dependent" and "independent" come from the notion that the amount of KWH consumption *depends* on the amount of AC hours and DRYER usage. Usually, dependent variables are denoted "*y*-variables" and independent variables are denoted "*x*-variables." The prime objective of *linear regression analysis* is to obtain an equation of the form

$$KWH = b_0 + b_1 AC + b_2 DRYER$$

that quantifies the dependency of KWH on AC and DRYER.
To get started, we shall investigate the dependency of KWH on AC, alone. Later, we shall explore the dependency of KWH on AC and DRYER simultaneously.

Figure 1 shows a plot of KWH versus AC.

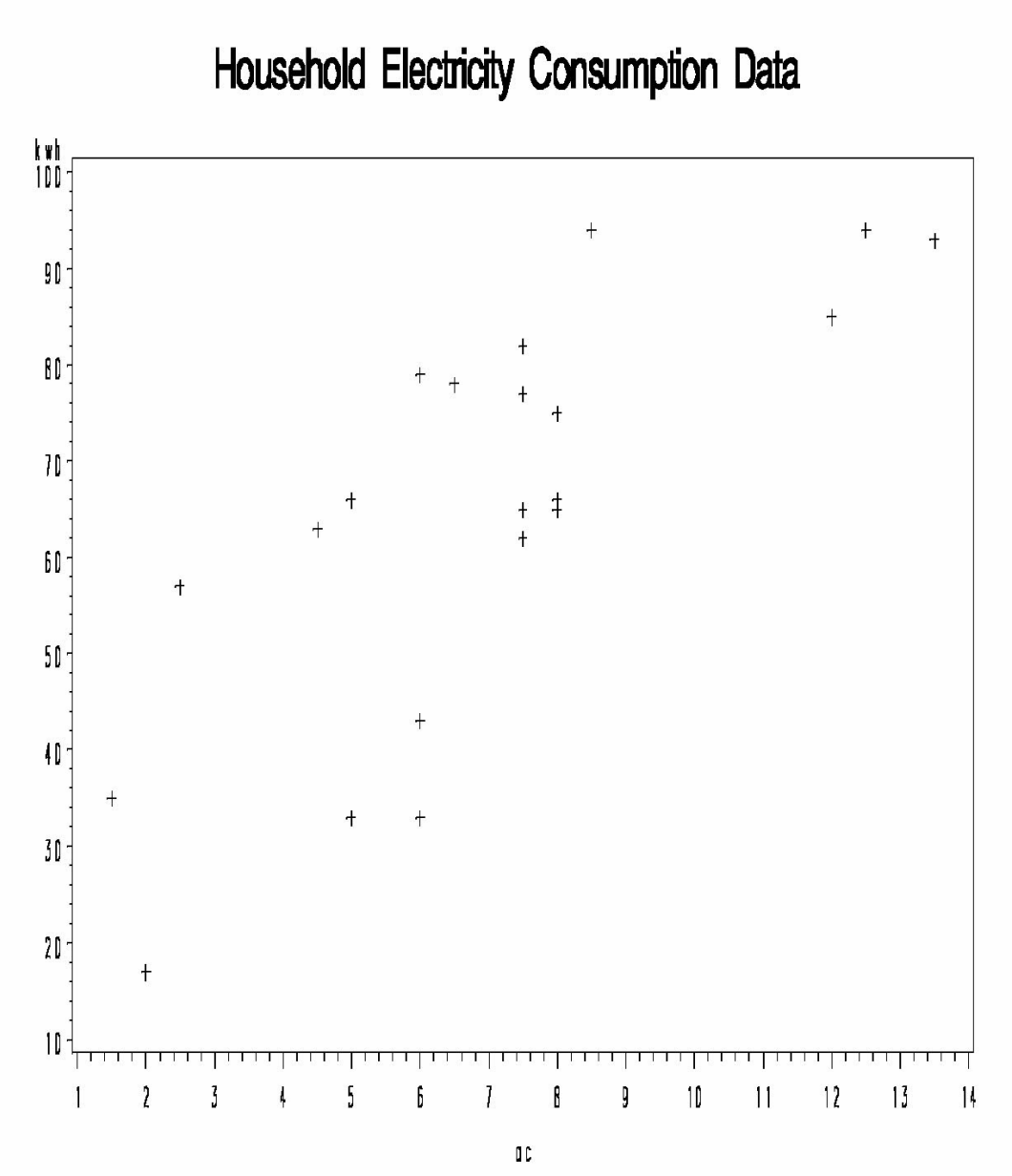

Figure 1. Kilowatt Hours versus AC Hours

Figure 1 shows that KWH increases as AC increase, as you would expect. In this application the rate of increase is more important the than the simple fact that there is an increase. We already knew that the air conditioner consumes electricity, and therefore that KWH will increase with AC. What we want to know is how much electricity the AC is using per hour; that is, the rate of consumption.

We shall obtain an equation

$$KWH = b_0 + b_1 AC$$

that will be used to quantify the rate of increase in KWH as a function of AC. This is the equation of a straight line. The coefficient b_1 is the *slope* of the straight line and it represents the rate of increase of KWH with AC. The coefficient b_0 is the *intercept* of the line. It represents the amount of KWH consumption when AC=0. The equation turns out to be

KWH = 27.85 + 5.34(AC).

This equation is plotted through the data in Figure 2.

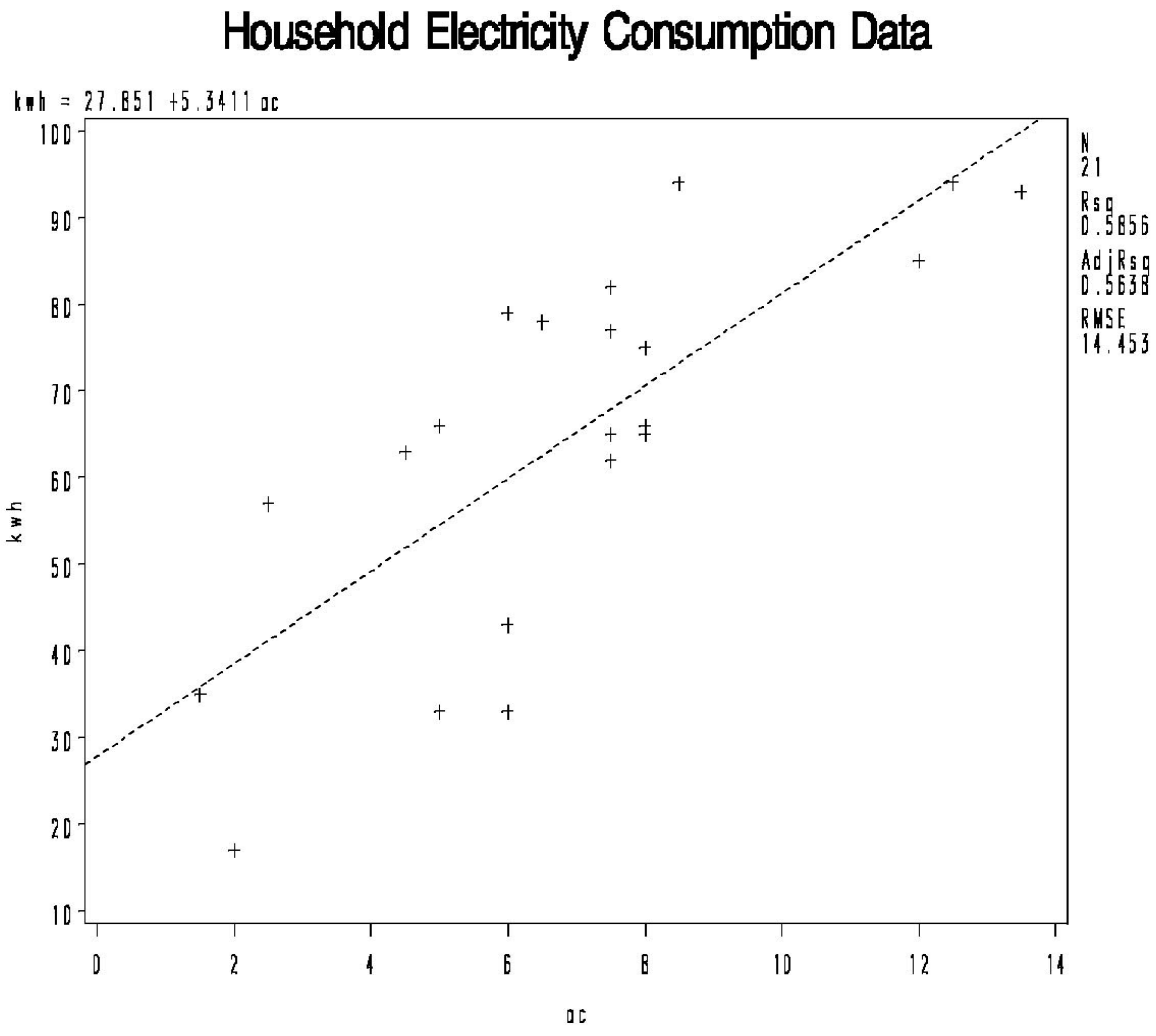

Figure 2. Regression of KWH versus AC

The number 5.34 is an estimate of the amount of electricity in KWH consumed for each hour the air conditioner is turned on. Then number 27.85 is an estimate of the amount of electricity consumed per day by all other electrical devices in the house.

Some things to think about:

1. How *precise* is 5.34 as an estimate of the *true* rate of KWH consumption by the air conditioner? Can we use 5.34 to construct a confidence interval about the true rate?

2. How *accurate* is 5.34 as an estimate of the *true* rate of KWH consumption by the air conditioner? If we did the experiment over and over again, would the estimates we obtain be clustered about the true rate? Is the *expected* value of the estimates equal to the true rate?

3. What other uses can be made of the regression equation? Can we *predict* the amount of KWH consumption on a certain day if we know the air conditioner usage was, say, AC=8 hours? What can we say about the accuracy and precision of the prediction?

4. How well does the equation fit the data? Is a linear equation appropriate for this application?

The process of obtaining the equation of the line and making inference about the coefficients is called *Linear Regression Analysis*. At the heart of linear regression analysis is a *statistical model*.

Simple Linear Regression Model

The expression "Simple Linear Regression" refers to regression applications in which there is only one independent and one dependent variable. A simple linear regression model is given by the equation

$$y = \beta_0 + \beta_1 x + \varepsilon, \qquad (1)$$

where β_0 and β_1 are unknown parameters and ε is a random variable, usually considered normally distributed.

The model equation (1) states that a value of y is equal to a linear function of x plus a random quantity ε. The parameters β_0 and β_1 are the **intercept** and **slope** of the regression line. In the electricity consumption example, y=KWH and x=AC would yield a simple linear regression model for relating KWH to AC. The parameter β_1 is the *expected* KWH's consumed per hour use of the AC, and the parameter β_0 is the *expected* combined KWH's used by all other electrical devices in the house per day. These parameters are *population* quantities, and cannot be known exactly, but we can estimate them from the data. The quantity ε is a random quantity that accounts for random deviation from expected KWH consumption. For example, suppose the AC is turned on for x = eight hours on a particular day. What is the value of y = KWH

consumption? The *expected* consumption, that is, the mean consumption in the conceptual sub-population of all similar days that the AC would be turned on for eight hours, is $E(y) = \beta_0 + \beta_1 x$. The *actual* KWH consumption for the *particular* day in question is the mean for that population plus the random quantity ε to account for deviation from the mean for that particular day. That is to say,

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon . \tag{2}$$

The mean, i.e. expected, KWH consumption for a day with known AC usage cannot be calculated because it involves the unknown parameters β_0 and β_1. If the random ε values are distributed with mean 0 and variance σ^2, then the sub-population of KWH values also has variance σ^2.

When it is necessary to write the model equation in reference to a particular observation, a subscript can be inserted on *x* and *y*. Generically, the subscript "*j*" could be used to indicate the "*j*th" observation, and the model equation would be

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j , \tag{3}$$

Fitting the Simple Linear Regression Model

In applications, the model is fitted to data using the method of least squares, giving the "prediction" equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x , \tag{4}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of β_0 and β_1 and $\hat{y}$ is a "predicted value" of *y* obtained by inserting a value of *x* into the prediction equation. The prediction equation is also useful to estimate the mean $E(y) = \beta_0 + \beta_1 x$ of the sub-population of *y* values corresponding to a given value of *x*.

The caret above the parameters is called "hat" and is used to distinguish the actual parameters from their estimates. Thus, $\hat{\beta}_1$ is called "beta-one hat." Likewise, the estimate $\hat{\sigma}^2$ of σ^2 is called "sigma-squared hat." All these parameter estimates can be computed from five summary statistics

$$\begin{aligned}
&\text{mean of } x\text{'s, } \bar{x} = (\Sigma x)/n \\
&\text{mean of } y\text{'s, } \bar{y} = (\Sigma y)/n \\
&\text{sum of squares of } x\text{'s, } S_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n \\
&\text{sum of squares of } y\text{'s, } S_{yy} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - (\Sigma y)^2/n \\
&\text{sum of products of } x\text{'s and } y\text{'s, } S_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - (\Sigma x)(\Sigma y)/n ,
\end{aligned} \tag{5}$$

where n is the total number of data points.

In the electricity consumption example n=21, and the summary statistics are:

$$\bar{x} = 145.5/21 = 6.93$$
$$\bar{y} = 1362/21 = 64.86$$
$$S_{xx} = 1204.75 - 145.5^2/21 = 1204.75 - 1008.12 = 196.64$$
$$S_{yy} = 97914 - 1362^2/21 = 97914 - 88335.4 = 9578.6$$
$$S_{xy} = 10487 - (145.5)(1362)/21 = 10487 - 9436.7 = 1050.3$$

The regression parameter estimates are:

$$\hat{\beta}_1 = S_{xy} / S_{xx} = 1050.3/196.64 = 5.34 \qquad (6)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 64.86 - (5.34)(6.93) = 27.85.$$

This gives the prediction equation

$$\text{KWH} = 27.85 + 5.34(\text{AC}).$$

Using the Prediction Equation

The prediction equation is useful for two purposes. It can be used to estimate the *mean* of the sub-population of amounts of electricity consumed on *all* days when the AC is turned on for a specified number of hours. It is also useful to predict the amount of electricity used on a *particular* day when the AC is turned on for a specified number of hours. For example, consider the conceivable days when the AC could be turned on for 10 hours. The value from the prediction equation corresponding to AC=10 is 27.85 + 5.34(10) = 81.25 kilowatt-hours. This number is an estimate of the mean KWH consumption on all days when the AC is turned on for 10 hours. Also, suppose the AC was turned on for 10 hours on a particular day of interest. The homeowner could use the prediction equation to predict the amount of electricity used on that day to be 81.25 KWH.

Accounting for Variation in Simple Linear Regression

Another aspect of regression analysis is accounting for the variation in the dependent variable as it relates to variation in the independent variable. A fundamental equation is

$$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y}). \qquad (7)$$

Equation (7) states that a deviation of a value of the dependent variable from the overall mean, $y - \bar{y}$, is equal to the sum of a deviation of the dependent variable from the predicted value,

$(y - \hat{y})$, plus the deviation of the predicted value from the overall mean, $\hat{y} - \bar{y}$. It can be shown that $\hat{y} - \bar{y} = 0$ when $x = \bar{x}$. Also, $\hat{y} - \bar{y}$ changes by an amount $\hat{\beta}_1$ for each unit of change in x. Thus, the deviation $(\hat{y} - \bar{y})$ depends directly on the independent variable x. But the deviation $(y - \hat{y})$ can be large or small, and positive or negative, for any value of x. So the deviation $(y - \hat{y})$ does not depend directly on x.

It turns out that the sums of squares of the deviations in equation (7) obey a similar equation,

$$\Sigma(y - \bar{y})^2 = \Sigma(y - \hat{y})^2 + \Sigma(\hat{y} - \bar{y})^2 . \qquad (8)$$

The sums of squared deviations have names. The left side of the equation is call the **total** sum of squares, and is denoted SS(Total)$= \Sigma(y - \bar{y})^2$. The terms on the right side are the **error** and **regression** sums of squares, SS(Regression) $= \Sigma(\hat{y} - \bar{y})^2$ and SS(Error)$= \Sigma(y - \hat{y})^2$. For brevity, we write SSR=SS(Regression), SSE=SS(Error), and SST=SS(Total), and equation (8) takes the form

$$\text{SST=SSR+SSE.} \qquad (9)$$

The total sum of squares, SST, is a measure of the total variation in the values of the dependent variable *y*. The regression sum of squares, SSR, is a measure of the variation in *y* that is attributable to variation in the independent variable *x*. Finally, the error sum of squares measures the variation in *y* that is *not* attributable to changes in *x*. Thus equation (9) shows the fundamental partitioning of the total variation into the portion *attributable* to *x* and the portion *not attributable* to *x*. The **coefficient of determination**, usually denoted $\mathbf{R^2}$, is SSR divided by SST, and thus measures the proportion of total variation in *y* that is attributable to variation in *x*,

$$R^2 = SSR/SST. \qquad (10)$$

Analysis of Variance for Simple Linear Regression

An *analysis of variance* associated with the regression is

Source of Variation	DF	SS	MS	
Regression	1	SSR	MSR	(11)
Error	n-2	SSE	MSE	
Total	n-1	SST		

The column headed DF contains **degrees of** freedom for the sums of squares. The column headed MS contains **mean squares**, which are the corresponding sums of squares divided by the degrees of freedom. The error mean square, MSE, is an estimate of σ^2, the variance of the errors, $\hat{\sigma}^2$=MSE.

Analysis of Variance Computations for Simple Linear Regression

The sums of squares in the ANOVA table can be calculated from the summary statistics as SSR= S_{xy}^2 / S_{xx}, SST = S_{yy}, and SSE = $S_{yy} - S_{xy}^2 / S_{xx}$. These computations for the KWH data are

SSR = $1050.3^2/196.64$ = 1,103,130/196.64 = 5609.7,
SST = 9578.6,

and

SSE = 9578.8 – 5609.7 = 3968.9.

For the KWH data, the analysis of variance (ANOVA) is

Source of Variation	DF	SS	MS
Regression	1	5609.7	5609.7
Error	19	3968.9	208.9
Total	20	9578.6	

The estimate of σ^2 is $\hat{\sigma}^2$ = MSE = 208.9, and the estimate of the error standard deviation is $\hat{\sigma}$ =(208.9)$^{.5}$=14.45.

Statistical Inference in Simple Linear Regression

Standard errors of the regression parameter estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ are needed in order to make statistical inference about the parameters β_1 and β_0. The variances of the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$ are

$$V(\hat{\beta}_1) = \sigma^2 / S_{xx} \tag{12}$$

and

$$V(\hat{\beta}_0) = \sigma^2 (1/n + 1/S_{xx}). \tag{13}$$

The standard errors of the parameter estimates are obtained by inserting $\hat{\sigma}^2$ = MSE in place of σ^2 and then taking the square root. Thus, the standard error of $\hat{\beta}_1$ is

$$s.e.(\hat{\beta}_1) = (MSE/S_{xx})^{.5}. \tag{14}$$

For the KWH data,

$$s.e.(\hat{\beta}_1) = (208.9/196.64)^{.5} = 1.03.$$

Tests of hypotheses and confidence intervals can be constructed using the parameter estimates and standard errors. The hypothesis H_0: $\beta_1 = \beta_{10}$, where β_{10} is a known constant, can be tested using the test statistic

$$t = (\hat{\beta}_1 - \beta_{10})/s.e.(\hat{\beta}_1). \tag{15}$$

This statistic has a *t* distribution with n-2 degrees of freedom when H_0 is true. A 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{.025}\, s.e.(\hat{\beta}_1). \tag{16}$$

In some applications it is useful to test the hypothesis H_0: $\beta_1=0$. For the KWH example, this hypothesis would *not* be of interest because there is no question that the AC uses electricity. The interesting inference is about the *amount* of electricity consumed. As already seen, the estimate is $\hat{\beta}_1=5.34$ KWH per hour. A confidence interval would give more meaningful inference about β_1 than a test of hypothesis H_0: $\beta_1=0$. With 19 degrees of freedom, $t_{.025}=2.1$. The confidence interval is

$$5.34 \pm 2.1(1.03),$$

or

$$5.34 \pm 2.16.$$

Prediction and Estimation of Sub-population Means

It is often useful to make inference about the mean of a sub-population corresponding to a given value of x, say x_0. The estimate of the subpopulation mean is obtained by inserting x_0 into the prediction equation. That is, the estimate of $E(y) = \beta_0 + \beta_1 x_0$ is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$. The standard error of $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is needed to make statistical inference about $E(y) = \beta_0 + \beta_1 x_0$. The variance of the sampling distribution is

$$V(\hat{y}) = \sigma^2 (1/n + (x_0 - \bar{x})^2 / S_{xx}). \tag{17}$$

Therefore the standard error is

$$s.e.(\hat{y}) = (MSE(1/n + (x_0 - \bar{x})^2 / S_{xx}))^{.5}. \tag{18}$$

A 95% confidence interval for the mean of the sub-population of y values corresponding to $x=x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{.025} (MSE(1/n + (x_0 - \bar{x})^2 / S_{xx}))^{.5}. \tag{19}$$

When used to *predict* a value of y, the relevant variance is the *prediction* variance

$$V(\hat{y} - y) = \sigma^2 (1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}). \tag{20}$$

The relevant standard error is

$$s.e.(\hat{y} - y) = (MSE(1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}))^{.5}. \tag{21}$$

A 95% *prediction* interval for the y value corresponding to $x=x_0$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{.025} (MSE(1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}))^{.5}. \tag{22}$$

Consider the conceivable days when the AC could be turned on for 10 hours. A 95% confidence interval for the *mean* KWH consumption on those days is

$$81.25 \pm 2.1(208.9(1/21 + (10 - 6.93)^2/196.64))^{.5},$$

or

$$81.25 \pm 9.38 = (71.87,90.63).$$

Now consider the *particular* day when the homeowner had the AC turned on for 10 hours. A 95% *prediction* interval is

$$81.25 \pm 2.1(208.9(1 + 1/21 + (10 - 6.93)^2/196.64))^{.5},$$

or

$$81.25 \pm 31.77 = (49.48,113.02).$$

SAS Program for Simple Linear Regression Analysis of KWH Data

```
options nonumber nodate;
Title1 'Household Electricity Consumption Data';
Title2 'Simple Linear Regression Analysis';
data kilowatt;
  input kwh ac dryer;
cards;
35  1.5 1
63  4.5 2
66  5.0 2
17  2.0 0
94  8.5 3
79  6.0 3
93 13.5 1
66  8.0 1
94 12.5 1
82  7.5 2
78  6.5 3
65  8.0 1
77  7.5 2
75  8.0 2
62  7.5 1
85 12.0 1
43  6.0 0
57  2.5 3
33  5.0 0
65  7.5 1
33  6.0 0
;

proc print;
run;

proc reg data=kilowatt;
  model kwh=ac;
  plot kwh*ac;
run;
```

Multiple Linear Regression Model

Multiple Linear Regression refers to regression applications in which there are more than one independent variables, $x_1, x_2, \ldots, x_k$. A multiple linear regression model with k independent variables has the equation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad (23)$$

The ε is a random variable with mean 0 and variance σ^2. A prediction equation for this model fitted to data is

$$\hat{y} = b_0 + b_1 x_1 + \ldots + b_k x_k + \varepsilon \qquad (24)$$

where $\hat{y}$ denotes the "predicted" value computed from the equation, and b_i denotes an estimate of β_i. These estimates are usually obtained by the method of **least squares**. This means finding among the set of all possible values for the parameter estimates the ones which minimize the sum of squared residuals, $\Sigma(y - \hat{y})^2$. This yields the best fitting equation in terms of minimizing the sum of squared distances of the fitted plane to the data points.

An example of a multiple linear regression with two independent variables is given by the KWH data, but now with x_1=AC and x_2=DRYER. The model equation would be

$$\text{KWH} = \beta_0 + \beta_1 \text{AC} + \beta_2 \text{DRYER} + \varepsilon.$$

Least squares parameter estimates are

$b_0 = 8.11$, $b_1 = 5.47$, and $b_2 = 13.22$.

Computation of the estimates by hand is tedious. They are ordinarily obtained using a regression computer program. Standard errors also are usually part of output from a regression program.

The prediction equation is

$$\text{KWH} = 8.11 + 5.47(\text{AC}) + 13.22(\text{DRYER}).$$

This model ascribes 5.47 KWH to hourly use of the AC and 13.22 KWH to each use of the DRYER, and 8.11 to all other electrical devices.

Compare this prediction equation with the one including only AC in the model,

$$\text{KWH} = 27.85 + 5.43(\text{AC}).$$

The intercept estimate has changed substantially from 27.85 to 8.11. This change occurs because KWH consumption due to DRYER usage is combined into the intercept estimate in the model that does not contain DRYER.

The estimate of the coefficient on AC has changed very little, from 5.34 to 5.47. This is related to the fact that AC and DRYER usage are relatively uncorrelated. In other words, use of one is not related to use of the other. Generally speaking, if AC and DRYER were positively (negatively) correlated, then the regression coefficient on AC would be reduced (increased) when DRYER was added to the model.

Compare the values of predicted KWH from the two models. Previously, AC=10 was inserted in the simple linear prediction equation to get

$$KWH = 27.85 + 5.34(10) = 81.25.$$

A value of DRYER must also be inserted into the multiple regression equation to get a predicted KWH value. Trying DRYER = 0, 1, and 2 gives

$$KWH = 8.11 + 5.47(10) + 13.22(0) = 62.81,$$
$$KWH = 8.11 + 5.47(10) + 13.22(1) = 76.03,$$
$$KWH = 8.11 + 5.47(10) + 13.22(2) = 89.25.$$

An analysis of variance for a multiple linear regression model with k independent variables fitted to a data set with n observations is

Source of Variation	DF	SS	MS	
Regression	k	SSR	MSR	(25)
Error	n-k-1	SSE	MSE	
Total	n-1	SST		

The sums of squares SSR, SSE, and SST have the same definitions in relation to the model as in simple linear regression:

$$SSR = \Sigma(\hat{y} - \overline{y})^2$$
$$SSE = \Sigma(y - \hat{y})^2 \qquad (26)$$
$$SST = \Sigma(y - \overline{y})^2$$

Also, SST=SSR+SSE. The value of SST does not change with the model. It depends only on the values of the dependent variable *y*. But SSE *decreases* as variables are added to a model, and SSR *increases* by the same amount. This amount of increase in SSR is the amount of variation due to variables in the larger model that was not accounted for by variables in the smaller model. This increase in regression sum of squares is sometimes denoted

$$SSR(\textit{added variables} \mid \textit{original variables}), \qquad (27)$$

where *original variables* represents the list of independent variables that were in the model prior to adding new variables, and *added variables* represents the list of variables that were added to

obtain the new model. The overall SSR for the new model can be partitioned into the variation attributable to the *original variables* plus the variation due to the *added variables* that is *not* due to the *original variables*,

$$\text{SSR}(\textit{all variables}) = \text{SSR}(\textit{original variables}) + \text{SSR}(\textit{added variables} \mid \textit{original variables}). \quad (28)$$

Generally speaking, larger values of the coefficient of determination R^2=SSR/SST indicate a better fitting model. The value of R^2 must necessarily increase as variables are added to the model. However, this does *not necessarily* mean that the model has actually been *improved*. The amount of increase in R^2 can be a mathematical artifact rather than a meaningful indication of an improved model. Sometimes an *adjusted* R^2 is used to overcome this shortcoming of the usual R^2. Most regression computer programs include both versions of R^2.

The analysis of variance for the two-variable model fitted to the KWH data is

Source of Variation	DF	SS	MS
Regression	2	9299.8	4649.9
Error	18	278.8	15.5
Total	20	9578.6	

Adding DRYER to the model affected a dramatic change in the value of SSR, which increased from 5609.7 to 9299.8. The value of SSE dropped accordingly from 3968.9 to 278.8. The coefficient of determination is now R^2=9299.8/9578.6=0.97. The two variables, AC and DRYER, account for 97% of the variability in KWH consumption in the house. This is up from R^2=5609.7/9578.6=0.58 for the variable AC alone.

The regression sum of squares partitioned into the amount due to AC alone plus the amount due to DRYER that was *not* attributable to AC, is

$$\text{SSR(AC and DRYER)} = \text{SSR(AC)} + \text{SSR(DRYER|AC)},$$

$$9299.8 = 5609.7 + 3690.1.$$

Thus, 3690.1 is the amount of variation due to DRYER that was not accounted for by AC.

Statistical inference about the parameters requires standard errors of the estimates. A 95% confidence interval for β_i is

$$b_i \pm t_{df,\,.025}(\text{s.e.}(b_i)), \quad (29)$$

where $t_{df,.025}$ is the critical value from a t distribution with df=n-k-1, the degrees of freedom for error.

Standard errors for parameters in the two-variable model are

$$\text{s.e.}(b_0) = 2.48$$
$$\text{s.e.}(b_1) = 0.28$$
$$\text{s.e.}(b_2) = 0.86$$

The critical value from a t distribution with df=18 is $t_{18,.025}=2.1$. Thus, a 95% confidence interval for β_1 is

$$b_1 \pm t_{18,\,.025}(\text{s.e.}(b_1)) = 5.47 \pm 2.1(0.28) = 5.47 \pm 0.59.$$

Thus, we are 95% confident that the "true" hourly KWH consumption of the AC is between 4.88 and 6.06. This is a considerably shorter interval than the interval 5.34±2.16 that was obtained from the simple linear regression model because the variance estimate (MSE) has been reduced from 208.9 to 15.5.

SAS Program for Simple Linear Regression Analysis of KWH Data

```
options nonumber nodate;
Title1 ‘Household Electricity Consumption Data’;
Title2 ‘Multiple Linear Regression Analysis’;
data kilowatt;
  input kwh ac dryer;
cards;
35  1.5 1
63  4.5 2
66  5.0 2
17  2.0 0
94  8.5 3
79  6.0 3
93 13.5 1
66  8.0 1
94 12.5 1
82  7.5 2
78  6.5 3
65  8.0 1
77  7.5 2
75  8.0 2
62  7.5 1
85 12.0 1
43  6.0 0
57  2.5 3
33  5.0 0
65  7.5 1
33  6.0 0
;

proc print;
run;

proc reg data=kilowatt;
  model kwh=ac dryer;
run;
```