1. Linear Regression Analysis

Introduction

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 particular, 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:

<table>
<thead>
<tr>
<th>KWH</th>
<th>AC (hours)</th>
<th>DRYER (runs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>63</td>
<td>4.5</td>
<td>2</td>
</tr>
<tr>
<td>66</td>
<td>5.0</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>2.0</td>
<td>0</td>
</tr>
<tr>
<td>94</td>
<td>8.5</td>
<td>3</td>
</tr>
<tr>
<td>79</td>
<td>6.0</td>
<td>3</td>
</tr>
<tr>
<td>93</td>
<td>13.5</td>
<td>1</td>
</tr>
<tr>
<td>66</td>
<td>8.0</td>
<td>1</td>
</tr>
<tr>
<td>94</td>
<td>12.5</td>
<td>1</td>
</tr>
<tr>
<td>82</td>
<td>7.5</td>
<td>2</td>
</tr>
<tr>
<td>78</td>
<td>6.5</td>
<td>3</td>
</tr>
<tr>
<td>65</td>
<td>8.0</td>
<td>1</td>
</tr>
<tr>
<td>77</td>
<td>7.5</td>
<td>2</td>
</tr>
<tr>
<td>75</td>
<td>8.0</td>
<td>2</td>
</tr>
<tr>
<td>62</td>
<td>7.5</td>
<td>1</td>
</tr>
<tr>
<td>85</td>
<td>12.0</td>
<td>1</td>
</tr>
<tr>
<td>43</td>
<td>6.0</td>
<td>0</td>
</tr>
<tr>
<td>57</td>
<td>2.5</td>
<td>3</td>
</tr>
<tr>
<td>33</td>
<td>5.0</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>7.5</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>6.0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1 shows a plot of KWH versus AC.

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.

**Regression with a Single Independent Variable: Simple Linear Regression**

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. 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 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.

![Household Electricity Consumption Data](image)

**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? If we did the experiment over and over again, how much would the slopes vary from one experiment to the next? 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? In other words, is the expected value of the estimate 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

\[ y = \beta_0 + \beta_1 x + \varepsilon, \]  

where \( \beta_0 \) and \( \beta_1 \) are unknown parameters and \( \varepsilon \) 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 \( \varepsilon \). The parameters \( \beta_0 \) and \( \beta_1 \) are the intercept and slope of the regression line. In the electricity consumption example, \( y=\text{KWH} \) and \( x=\text{AC} \) would yield a simple linear regression model for relating KWH to AC. The parameter \( \beta_1 \) is the expected KWH consumed per hour use of the AC, and the parameter \( \beta_0 \) is the expected combined KWH used by all other electrical devices in the house per day. These parameters are population quantities and cannot be known exactly, but can be estimated from the data. The quantity \( \varepsilon \) 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 $E(y) = \beta_0 + \beta_1 (8)$. The actual KWH consumption for the particular day in question is the mean for that population plus the random quantity $\varepsilon$ to account for deviation from the mean for that particular day. That is to say,

$$y = E(y) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon .$$  \hfill (2)

The mean, i.e. expected, KWH consumption for a day with known AC usage cannot be calculated because it involves the unknown parameters $\beta_0$ and $\beta_1$. If the random $\varepsilon$ values are distributed with mean 0 and variance $\sigma^2$, then the sub-population of KWH values also has variance $\sigma^2$.

A subscript can be inserted on $x$ and $y$ when it is necessary to write the model equation in reference to a particular observation. Generically, the subscript “$j$” could be used to indicate the “$j$th” observation, and the model equation would be

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j ,$$  \hfill (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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$  \hfill (4)

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimates of $\beta_0$ and $\beta_1$ 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 $\sigma^2$ is called “sigma-squared hat.” All these parameter estimates can be computed from five summary statistics shown in Table 1, where $n$ is the total number of data points.
mean of $x$'s \[ \bar{x} = \frac{\Sigma x}{n} \]

mean of $y$'s \[ \bar{y} = \frac{\Sigma y}{n} \]

sum of squares of $x$'s \[ S_{xx} = \Sigma (x - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2 / n \]

sum of squares of $y$'s \[ S_{yy} = \Sigma (y - \bar{y})^2 = \Sigma y^2 - (\Sigma y)^2 / n \]

sum of products of $x$'s and $y$'s \[ S_{xy} = \Sigma (x - \bar{x})(y - \bar{y}) = \Sigma xy - (\Sigma x)(\Sigma y) / n \]

Table 1. Summary statistics for regression computations

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

\[
\begin{align*}
\bar{x} &= 145.5 / 21 = 6.93 \\
\bar{y} &= 1362 / 21 = 64.86 \\
S_{xx} &= 1204.75 - 145.5^2 / 21 = 1204.75 - 1008.12 = 196.64 \quad (5) \\
S_{yy} &= 97914 - 1362^2 / 21 = 97914 - 88335.4 = 9578.6 \\
S_{xy} &= 10487 - (145.5)(1362) / 21 = 10487 - 9436.7 = 1050.3
\end{align*}
\]

The regression parameter estimates are:

\[
\begin{align*}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = 1050.3 / 196.64 = 5.34 \quad (6) \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 64.86 - (5.34)(6.93) = 27.85.
\end{align*}
\]

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

\[ y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y}). \]  

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 \( \beta \) 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,

\[ \Sigma (y - \bar{y})^2 = \Sigma (y - \hat{y})^2 + \Sigma (\hat{y} - \bar{y})^2. \]

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

\[ SST = SSR + SSE. \]

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 \( R^2 \), is \( SSR \) divided by \( SST \), and thus measures the proportion of total variation in \( y \) that is attributable to variation in \( x \),

\[ R^2 = SSR/SST. \]

Analysis of Variance for Simple Linear Regression

The generic analysis of variance associated with the regression is shown in Table 2.
Table 2. Generic ANOVA Table for Regression Analysis

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 $\sigma^2$, the variance of the errors, $\hat{\sigma}^2 = \text{MSE}$.

**Analysis of Variance Computations for Simple Linear Regression**

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

$$\text{SSR} = \frac{1050.3^2}{196.64} = 1,103,130/196.64 = 5609.7,$$
$$\text{SST} = 9578.6,$$
$$\text{SSE} = 9578.8 - 5609.7 = 3968.9.$$  

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

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>5609.7</td>
<td>5609.7</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>3968.9</td>
<td>208.9</td>
</tr>
<tr>
<td>Total</td>
<td>20</td>
<td>9578.6</td>
<td></td>
</tr>
</tbody>
</table>

The estimate of $\sigma^2$ is $\hat{\sigma}^2 = \text{MSE} = 208.9$, and the estimate of the error standard deviation is $\hat{\sigma} = (208.9)^{\frac{1}{2}} = 14.45$.

**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 $\beta_1$ and $\beta_0$. The variances of the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$ are
\[ V(\hat{\beta}_i) = \sigma^2 / S_{xx} \]  

and  

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

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

\[ \text{s.e.}(\hat{\beta}_i) = (\text{MSE} / S_{xx})^{0.5}. \]  

For the KWH data,  

\[ \text{s.e.}(\hat{\beta}_i) = (208.9/196.64)^{0.5} = 1.03. \]

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 \( \beta_{10} \) is a known constant, can be tested using the test statistic  

\[ t = (\hat{\beta}_1 - \beta_{10}) / \text{s.e.}(\hat{\beta}_1). \] 

This statistic has a \( t \) distribution with \( n-2 \) degrees of freedom when \( H_0 \) is true. A 95% confidence interval for \( \beta_1 \) is  

\[ \hat{\beta}_1 \pm t_{0.025} \text{s.e.}(\hat{\beta}_1). \] 

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 \( \beta_1 \) than a test of hypothesis \( H_0: \beta_1 = 0 \). With 19 degrees of freedom, \( t_{0.025} = 2.1 \). 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
\[
V(\hat{y}) = \sigma^2 (1/n + (x_0 - \bar{x})^2 / S_{xx}).
\]
(16)

Therefore the standard error is
\[
s.e.(\hat{y}) = (\text{MSE}(1/n + (x_0 - \bar{x})^2 / S_{xx}))^{1/2}.
\]
(16)

A 95% confidence interval for the mean of the sub-population of \( y \) values corresponding to \( x = x_0 \) is
\[
\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{0.025} (\text{MSE}(1/n + (x_0 - \bar{x})^2 / S_{xx}))^{1/2}.
\]
(17)

When used to predict a value of \( y \), the relevant variance is the prediction variance
\[
V(\hat{y} - y) = \sigma^2 (1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}).
\]
(18)

The relevant standard error is
\[
s.e.(\hat{y} - y) = (\text{MSE}(1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}))^{1/2}.
\]
(19)

A 95% prediction interval for the \( y \) value corresponding to \( x = x_0 \) is
\[
\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{0.025} (\text{MSE}(1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}))^{1/2}.
\]
(20)

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
\[
81.25 \pm 2.1(208.9(1/21 + (10 - 6.93)^2/196.64))^{1/2},
\]
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
\[
81.25 \pm 2.1(208.9(1 + 1/21 + (10 - 6.93)^2/196.64))^{1/2},
\]
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;