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

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \qquad (1)$$

The ε is a random variable with mean 0 and variance σ^2. The parameter β_1 represents the expected change in y resulting from a single unit change in x_1, *holding all other independent variables fixed.*

A prediction equation for this model fitted to data is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k \qquad (2)$$

where $\hat{y}$ denotes the "predicted" value computed from the equation, and $\hat{\beta}_i$ 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, $\sum_{i=1}^{n} (y_i - \hat{y})^2$. The least squares estimates yield the best fitting equation in terms of minimizing the sum of squared distances of the fitted plane to the data points. The interpretation of the parameter estimates is the same as the interpretation of the model parameters, except with respect to the fitted model. The parameter estimate $\hat{\beta}_1$ represents the change in $\hat{y}$ resulting from a single unit change in x_1, *holding all other independent variables fixed.*

Example of Multiple Linear Regression

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. Figure 1 shows a plot or KWH versus DRYER.

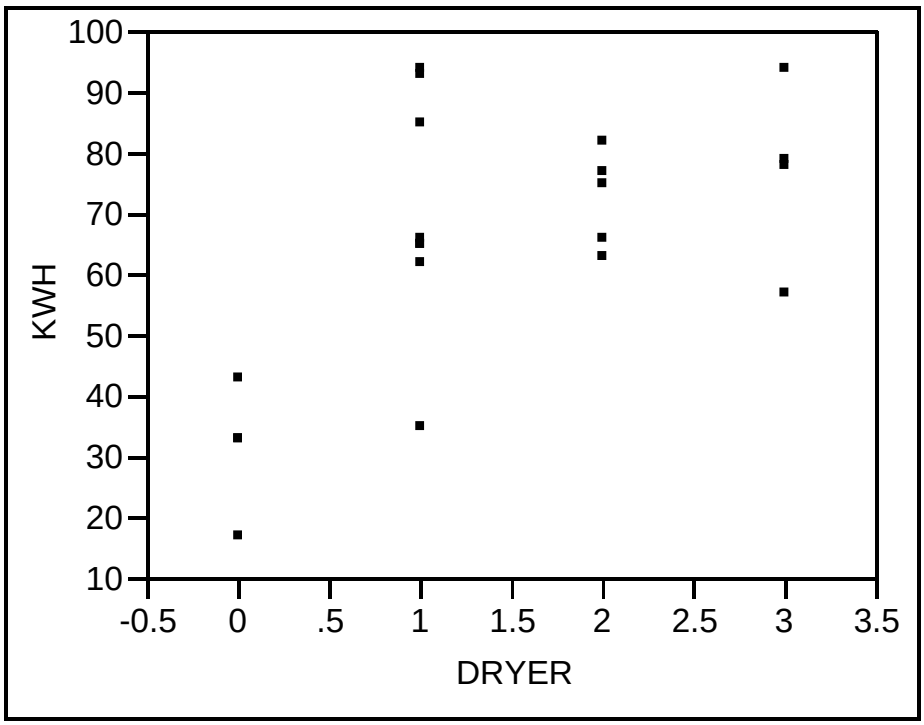

Figure 1. Plot of KWH versus DRYER.

The plot in Figure 1 clearly shows KWH increases with increasing runs of the dryer, but the plot does not take into account the variable AC. Visualizing the simultaneous effects of AC and DRYER on KWH would requite a plot in three dimensions, which is difficult to construct.

The model equation would be

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

Least squares parameter estimates are

$$\hat{\beta}_0 = 8.11, \hat{\beta}_1 = 5.47, \hat{\beta}_2 = 13.22$$

Computation of the estimates by hand is tedious, and infeasible for more than two independent variables. Estimates are ordinarily obtained using a regression computer program. Standard errors also are usually part of output from a regression program.

The prediction equation for the KWH data is

$$\text{KWH} = 8.11 + 5.47(\text{AC}) + 13.22(\text{DRYER}). \qquad (3)$$

This model ascribes 5.47 KWHs to hourly use of the AC and 13.22 KWHs to each use of the DRYER, and 8.11 to all other electrical devices, combined. Remember that $\hat{\beta}_1 = 5.47$ is an estimate of the amount of change in KWH due to a one unit increase in AC *holding DRYER constant*.

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

$$\text{KWH} = 27.85 + 5.43(\text{AC}). \qquad (4)$$

The intercept estimate has changed substantially from 27.85 to 8.11. This change occurs because KWH consumption due to DRYER usage is not accounted for in the equation. The KWH consumption due to *average* DRYER usage is combined into the intercept estimate in the model that does not contain DRYER. But the change in KWH due to a one-unit increase in DRYER usage is not explicitly shown in equation (4).

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. (See Figure 2.) 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.

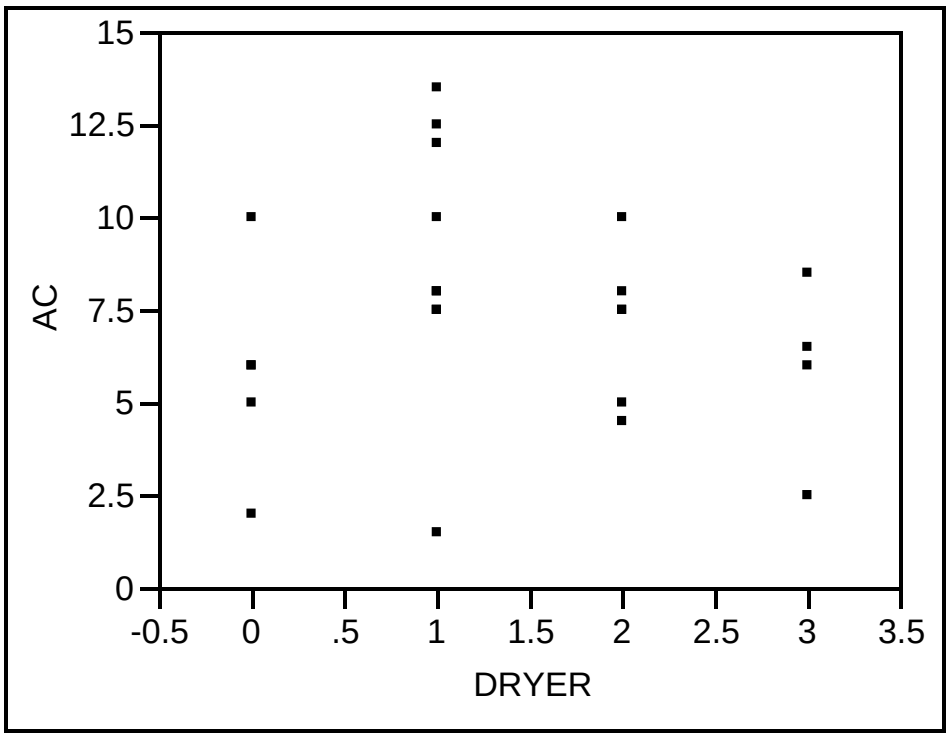

Figure 2. Plot of AC versus DRYER

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

$$\text{KWH} = 27.85 + 5.34(10) = 81.25. \tag{5}$$

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 and holding AC=10 gives

$$\begin{aligned} \text{KWH} &= 8.11 + 5.47(10) + 13.22(0) = 62.81, \\ \text{KWH} &= 8.11 + 5.47(10) + 13.22(1) = 76.03, \\ \text{KWH} &= 8.11 + 5.47(10) + 13.22(2) = 89.25. \end{aligned} \tag{6}$$

KWH consumption increases by 13.22 as DRYER goes from 0 to 1 and again from 1 to 2, *holding AC fixed at 10*.

Analysis of Variance for Multiple Regression Model

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
Error	n-k-1	SSE	MSE
Total	n-1	SSTot	

(7)

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

$$\text{SSR} = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2 \text{ , SSE} = \sum_{j=1}^{n} (y - \hat{y})^2 \text{ , SSTot} = \sum_{j=1}^{n} (y - \bar{y})^2 \qquad (8)$$

Also, SSTot=SSR+SSE. The value of SSTot 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

$$\text{SSR}(\textit{added variables} \mid \textit{original variables}), \qquad (9)$$

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*,

$$\begin{aligned}\text{SSR}(\textit{all variables}) = {} & \text{SSR}(\textit{original variables}) \\ & + \text{SSR}(\textit{added variables} \mid \textit{original variables}).\end{aligned} \qquad (10)$$

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

$$SSR(\text{AC and DRYER}) = SSR(\text{AC}) + SSR(\text{DRYER}|\text{AC}), \qquad (11)$$

$$9299.8 = 5609.7 + 3690.1.$$

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

We can expand the ANOVA table to show the breakdown of the regression SS given in equation (11):

Source of Variation	DF	SS	MS	F	P
Regression	2	9299.8	4649.9		
AC	1	5609.7	5609.7		
DRYER|AC	1	3690.1	3690.1	238.1	.0001
Error	18	278.8	15.5		
Total	20	9578.6			

The values of R^2 increases as variables are added to the model, as shown in the table:

Variables in Model	AC	AC and DRYER
R^2	.5856 = 5609.7/9578.6	.9709 = (5609.7+3690.1)/9578.6

One of the detracting features of R^2 is that it can be driven closer and closer to 1.0 by adding variables, even though the variables may have no relationship to KWH. To overcome this problem, and adjusted version is available in most computer programs.

Statistical Inference for Regression Parameters

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

$$\hat{\beta}_i \pm t_{df,.025}(\hat{\sigma}_{\hat{\beta}_i}) \tag{12}$$

where $t_{df,.025}$ is the critical value from a t distribution with df=n-k-1, the degrees of freedom for error, and $\hat{\sigma}_{\hat{\beta}_i}$ is the standard error of $\hat{\beta}_i$.

Standard errors for parameters in the two-variable model are

$$\hat{\sigma}_{\hat{\beta}_0} = 2.48,\ \hat{\sigma}_{\hat{\beta}_1} = 0.28,\ \hat{\sigma}_{\hat{\beta}_2} = 0.86. \tag{13}$$

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

$$\hat{\beta}_1 \pm t_{18,.025}\hat{\sigma}_{\hat{\beta}_1} = 5.47 \pm 2.1(0.28) = 5.47 \pm 0.59.$$

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.

It seems apparent that the model including both AC and DRYER is superior to the model containing AC alone. The value of R^2 is much higher (.9709 compared to .5856) and MS(Error) is much smaller.

You can conduct a statistical test of significance to compare the two models using the ANOVA table with the partitioned SS(Reg). The test statistic is

$$F = MS(DRYER|AC)/MS(Error) = 3690.1/15.5 = 238.1,$$

with numerator df=1 and denominator df=18. This is a huge value of F with these degrees of freedom and is significant at any reasonable level.

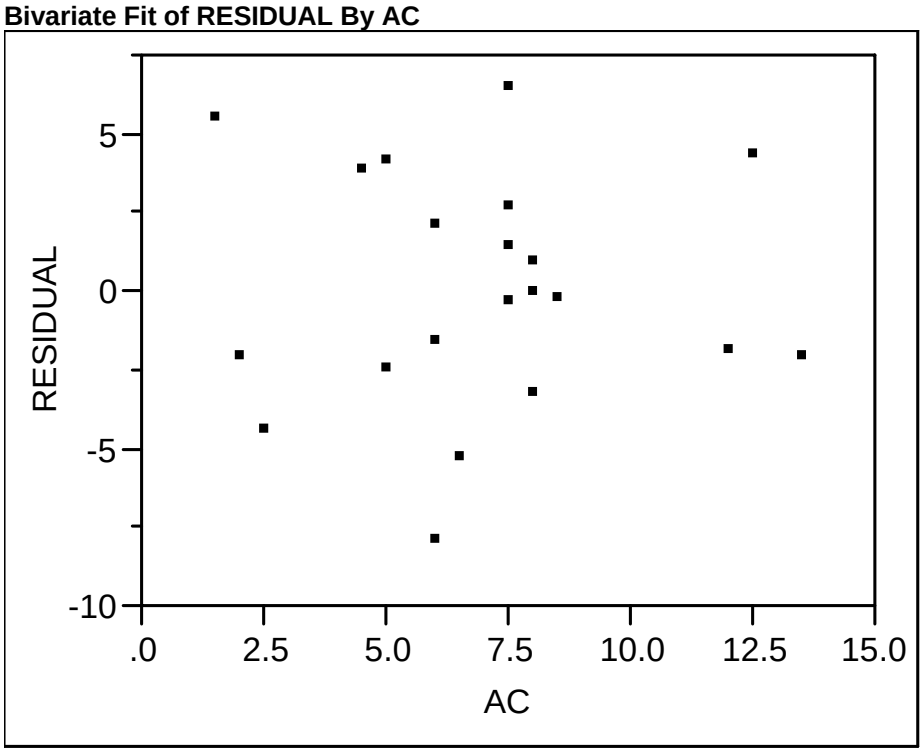

Figure 3. Plot of Residuals versus AC

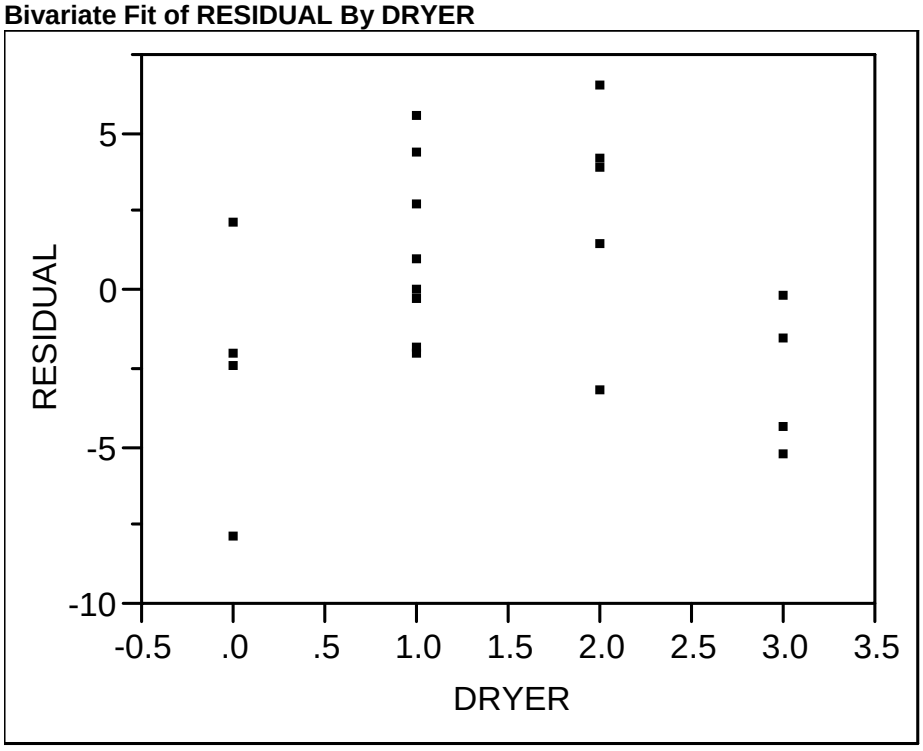

Figure 4. Plot of Residuals versus DRYER

Plots of the residual from regressing KWH on AC and DRYER in Figures 3 and 4 reveal essentially the same pattern as when KWH is regressed on AC or DRYER individually. That is because AC and DRYER are essentially uncorrelated. The curvature of the points in the residuals versus DRYER persists.

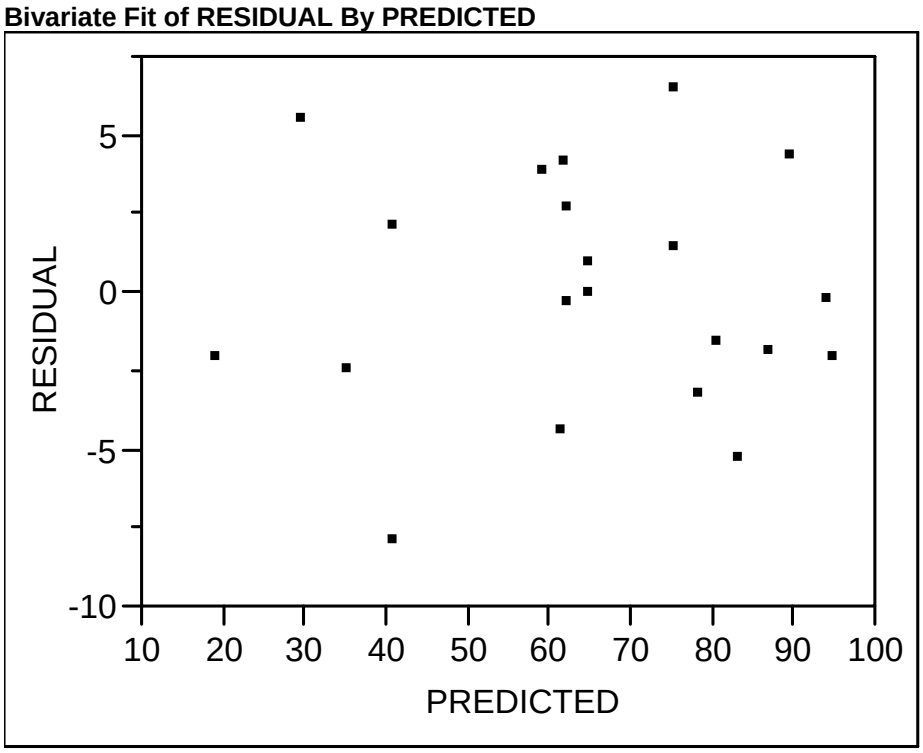

Figure 5. Plot of Residuals versus Predicted Values

Residuals from the regression of KWH on AC and DRYER plotted versus predicted values in Figure 5 shows a pattern distinctly different from the plots versus AC or DRYER.

Regression with Collinear Variables

The example on household KWH consumption utilized two independent variables that are almost uncorrelated. Thus, when DRYER was added tot the model in addition to AC, the AC regression coefficient changed very little. Also, the amount of variation attributable to DRYER is almost the same when it is the only variable in the model as when it is added to a model that already includes AC.

The following example illustrates a situation when two highly correlated variables are in a regression model: Students in a graduate statistics course recorded the spans of their left and right hands and their heights, all in inches. The objective was to develop a regression model to predict height from hand span. The variable names are HT, LSPAN, and RSPAN. Of course, LSPAN and RSPAN are highly correlated. This example illustrates the consequence of using two highly correlated variables in a multiple regression equation.

The two simple linear regression models are:

$$HT = \beta_0 + \beta_1 LSPAN + \varepsilon$$

and

$$HT = \beta_0 + \beta_2 RSPAN + \varepsilon$$

The prediction equations are:

HT=43.62+2.88*LSPAN

and

HT=41.35+3.17*RSPAN

Not surprisingly, the two equations are quite similar.

Figures 6a and 6b show HT versus LSPAN and RSPAN:

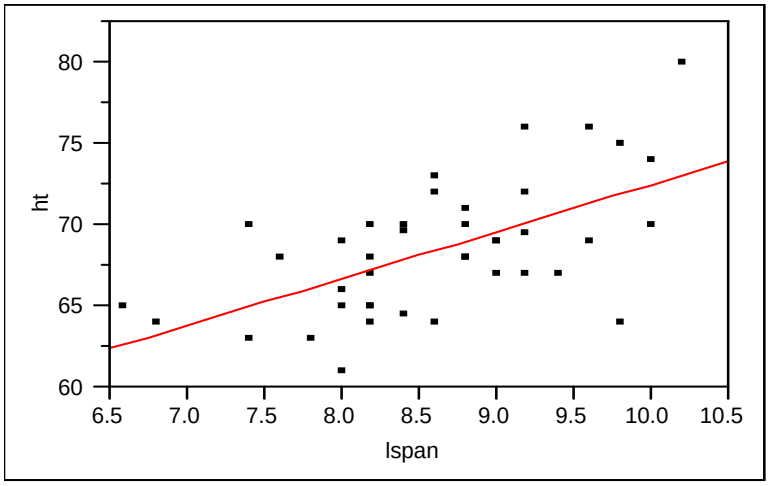

Figure 6a. Regression of HT on LSPAN

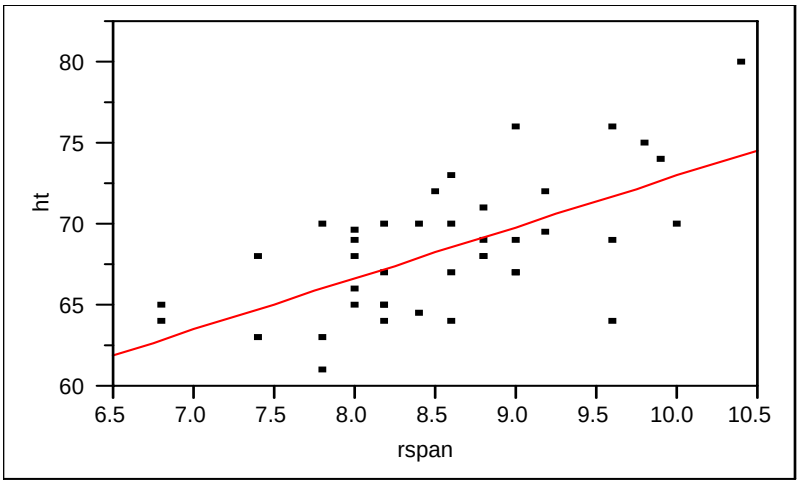

Figure 6b. Regression of HT on RSPAN

Figure 7 shows a plot of RSPAN versus LSPAN showing the high degree of collinearity.

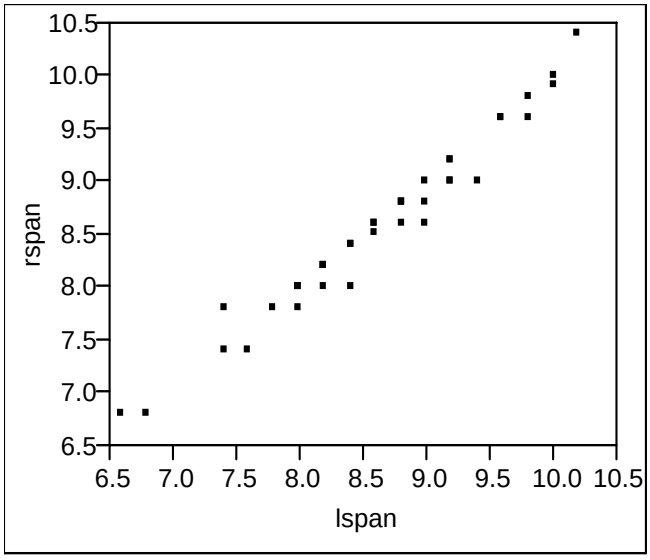

Figure 7. Plot showing collinearity between LSPAN and RSPAN

The multiple linear regression model is:

$$HT = \beta_0 + \beta_1 LSPAN + \beta_2 RSPAN + \varepsilon \qquad (14)$$

The prediction equation is

HT = 41.13 – 4.31*LSPAN + 7.53*RSPAN

At first look, this equation seems to make no sense at all. The regression coefficient on LSPAN is negative, and the coefficient on RSPAN is twice as large as the coefficient

when RSPAN was the only variable in the model. These are consequences of collinearity between LSPAN and RSPAN.

Figure 8 shows a plot of HT versus LSPAN and RSPAN:

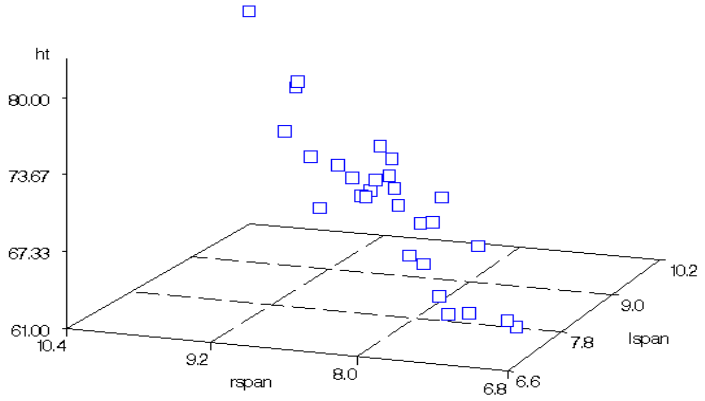

Figure 8. HT plotted versus LSPAN and RSPAN

When you observe this plot, you can imagine the instability of a plane fitted to the data due to the lack of data points for large LSPAN-small RSPAN combinations and small LSPAN-large RSPAN combinations. Such points, of course, would not be typical values of LSPAN and RSPAN.

Here is a table showing the summary statistics for three regression models usingLSPAN and RSPAN. Model 1 contains LSPAN alone, model 2 contains RSPAN alone, and model 3 contains both LSPAN and RSPAN.

Modle	Variables in Model	Intercept	lspan	rspan	R^2
1.	lspan	43.62	2.88 (se.62)	--------	.36
2.	rspan	41.34	--------	3.17 (se.61)	.41
3.	lspan& rspan	41.13	-4.31 (se 3.24)	7.53 (se 3.34)	.43

The table shows these phenomena:

1. Huge standard errors when both variables are included in the model
2. Similar intercept values for all models
3. Only very small increases in R^2 when going from model 1 to model 3 or from model 2 to model 3.

Figures 9a and 9b show the data and fitted plane on similar axes:

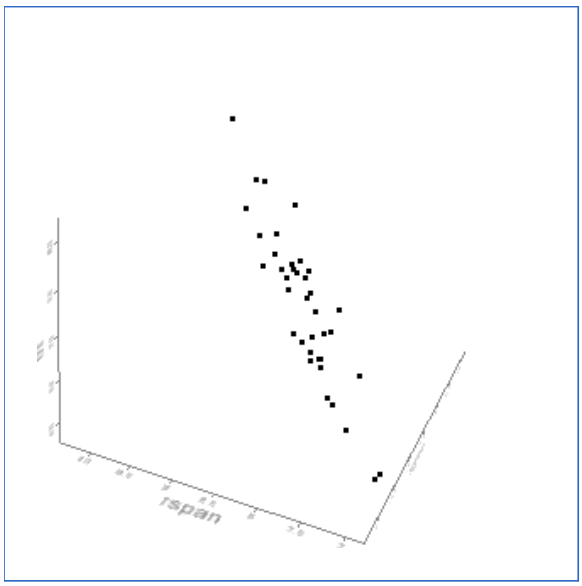

Figure 9a. HT plotted versus LSPAN and RSPAN

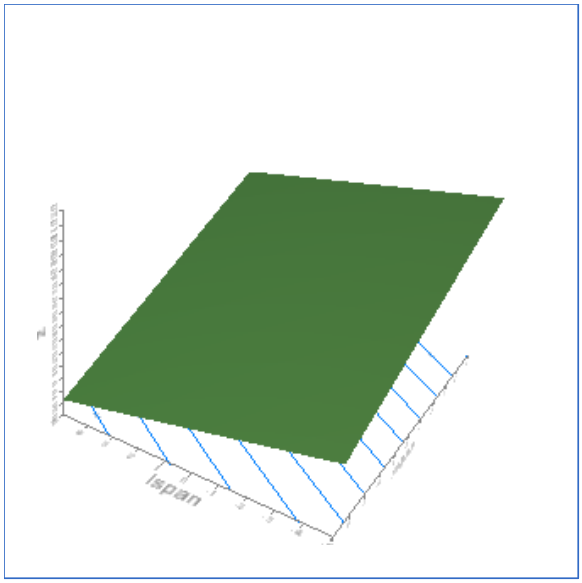

Figure 9b. Plane of predicted values plotted versus

You can see that although the slope in the LSPAN direction is negative, the plane increases in slope from the smallest values of LSPAN and RSPAN up to the largest values. This illustrates the following facts of regression with highly collinear variables:

1. Individual regression coefficients are practically meaningless.
2. Predictions over the region of observed values of the collinear independent variables are relatively stable.

SAS Program for Multiple Linear Regression Analysis of KWH Data

```
options nonumber nodate;
Title1 'Household Electricity Consumption Data';
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
.  10 0
.  10 1
.  10 2
;

proc print data=kilowatt;
run;

proc corr sscp data=kilowatt;
run;

proc sort data=kilowatt;
by ac;
run;

data acplot;
do ac=0 to 15 by .5;
kwh=.;
output;
```

```
end;
run;

proc print data=acplot;
run;

Title2 'Regression of KWH vs AC and DRYER';
run;

data acplot;
merge acplot kilowatt;
by ac;
run;

proc reg data=acplot; id ac;
  model kwh=ac/p;
  plot kwh*ac;
  output out=acplot1 p=acpred r=acresid1;
run;

proc gplot data=acplot1;
 plot acresid1*ac;
 plot acresid1*acpred;
run;

proc reg data=acplot; id ac;
  model kwh=ac/p clm cli;
  plot kwh*ac;
  output out=acplot2 p=acpred lclm=lcl uclm=ucl lcl=lpl
ucl=upl;
run;

proc print data=acplot2;
run;

proc gplot data=acplot2;
 plot kwh*ac acpred*ac lcl*ac ucl*ac / overlay;
 plot kwh*ac acpred*ac lpl*ac upl*ac / overlay;
run;

Title2 'Regression of KWH vs DRYER';
run;

proc sort data=kilowatt;
by dryer;
run;
```

```
data dryplot;
do dryer=0 to 3 by .1;
kwh=.;
output;
end;
run;

proc print data=dryplot;
run;

data dryplot;
merge dryplot kilowatt;
by dryer;
run;

proc reg data=dryplot; id ac;
  model kwh=dryer/p;
  plot kwh*dryer;
  output out=dryplot1 p=drypred r=dryresid1;
run;

proc gplot data=dryplot1;
 plot dryresid1*dryer;
 plot dryresid1*drypred;
run;

proc reg data=dryplot; id ac;
  model kwh=dryer/p clm cli;
  plot kwh*dryer;
  output out=dryplot2 p=drypred lclm=lcl uclm=ucl lcl=lpl
ucl=upl;
run;

proc print data=dryplot2;
run;

proc gplot data=dryplot2;
 plot kwh*dryer drypred*dryer lcl*dryer ucl*dryer /
overlay;
 plot kwh*dryer drypred*dryer lpl*dryer upl*dryer /
overlay;
run;
```