

## Section 2: Multiple Regression

For Multiple Regression, our response variable is  $Y$ , and our explanatory or predictor variables are  $X_1, X_2$ , etc.

Our model is:

$$Y_i = \mathbf{b}_0 + \mathbf{b}_1 X_{1,i} + \mathbf{b}_2 X_{2,i} + \dots + \mathbf{e}_i$$

with error term  $\mathbf{e}_i \sim N(0, \mathbf{s}^2)$

It is helpful to think of the response as a surface in space.

Polynomial regression is a special case of multiple regression. We get more explanatory variables by taking powers of one variable. Predictor variables are numerical, but can be indicator variables (*i.e.*, categorical).

We will begin our discussion of multiple regression with a Predictor and Response variable, and also have a Categorical variable to split up the data. Then we will add additional Predictor variables and interaction terms. We will also pay attention to error structure, as we want to do some inference.

The data on home gas consumption is from OzData, and can be found at

<http://www.maths.uq.edu.au/~gks/data/general/insulgas.html>,

a good website which has data categorized by type of analysis.

The data provide the weekly gas consumption (1000 cubic feet) for a home in England in the 1960's. The average outside temperature for each week is also recorded. There are 26 weeks of data before insulation was installed and 18 weeks of data after insulating. The house thermostat was set at 20°C throughout. The categorical variable is insulation: 0 for without insulation and 1 for with insulation.

Before Insulation						After Insulation					
Temp	Gas	Temp	Gas	Temp	Gas	Temp	Gas	Temp	Gas	Temp	Gas
-0.8	7.2	4.3	5.2	7.4	4.2	-0.7	4.8	3.1	3.2		
-0.7	6.9	5.4	4.9	7.5	4.0	0.8	4.6	3.9	3.9		
0.4	6.4	6.0	4.9	7.5	3.9	1.0	4.7	4.0	3.5		
2.5	6.0	6.0	4.3	7.6	3.5	1.4	4.0	4.0	3.7		
2.9	5.8	6.0	4.4	8.0	4.0	1.5	4.2	4.2	3.5		
3.2	5.8	6.2	4.5	8.5	3.6	1.6	4.2	4.3	3.5		
3.6	5.6	6.3	4.6	9.1	3.1	2.3	4.1	4.6	3.7		
3.9	4.7	6.9	3.7	10.2	2.6	2.5	4.0	4.7	3.5		
4.2	5.8	7.0	3.9			2.5	3.5	4.9	3.4		

In order to look at inference, standard errors and tests of hypothesis, we need to have random sampling. We assume the first 26 weeks to be a random sample of all weeks before insulation, and the last 18 to be a random sample of all weeks afterward. For inference, we also need to have the appropriate error structure  $e_i \sim N(0, \mathbf{s}^2)$ . Each outside temperature has a distribution of gas consumption responses. The value we get is a random value from that distribution. Any inferences that we make are restricted to this one house.

### Building a Regression Model

In the following example, we will add variables to a model and see how it influences our ability to predict. The response is gas, specifically gas consumption in 1000's of cubic feet. The predictors are *Temperature* (average outside temperature in °C) and *Insulation* (0 before insulation, 1 after insulation). Later, we will create another variable by multiplying  $Temp \cdot Insul$ .

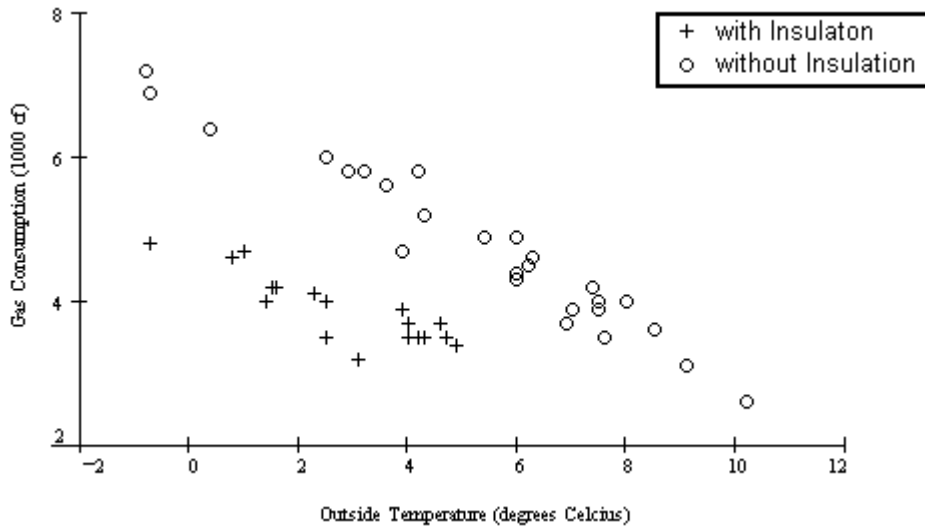


Figure 10: Plot of Gas Consumption/Temp data

The graph in Figure 10 shows two clear linear trends with a gap between them. As temperature goes up, gas consumption goes down in both situations. So our general trends are:

- As outside temperature increases, gas consumption goes down.
- Less gas is consumed after insulation is added.

The following regression analysis shows results from combining all of the data.

## Regression Analysis (Minitab output)

The regression equation is  
 Gas = 5.33 - 0.216 Temp

Predictor	Coef	StDev	T	P
Constant	5.3291	0.2427	21.96	0.000
Temp	-0.21602	0.04773	-4.53	0.000

S = 0.8533      R-Sq = 32.8%      R-Sq(adj) = 31.2%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	14.912	14.912	20.48	0.000
Error	42	30.578	0.728		
Total	43	45.490			

Note that  $R^2$  is low but the slope is highly significantly different from zero. The test for the slope asks, “Am I better off knowing the value of the explanatory variable?”  $P \approx 0$  tells us that the linear model is better than a horizontal line at  $\bar{y}$ .

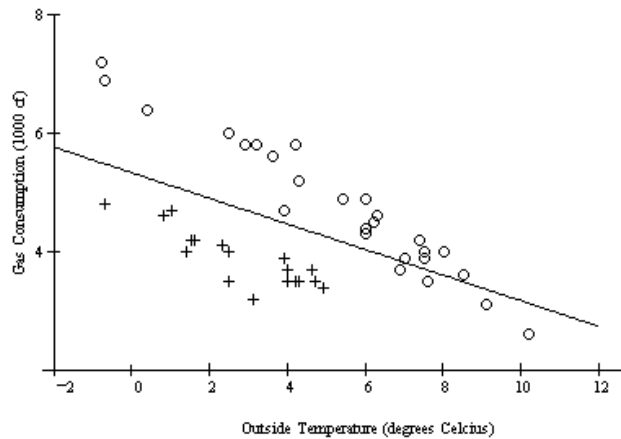


Figure 11: Simple Linear Regression

Using simple linear regression, we get the prediction equation:  $fitGas = 5.33 - 0.216Temp$

The slope is negative, as expected.  $R^2 = 32.8\%$  of the variability in gas consumption is explained by the linear relationship with average outside temperature; that's not very much. Notice from the regression output that  $t^2 = F$  ( $(-4.53)^2 = 20.48$ ) and

$$\frac{SS_{Regression}}{SS_{Total}} = \frac{14.912}{45.49} \approx 0.328 \rightarrow R^2. \text{ Also, } 1 - \frac{\left(\frac{30.578}{42}\right)}{\left(\frac{45.49}{43}\right)} = 1 - \frac{0.728}{1.0579} \approx 0.312 \rightarrow Adj R^2.$$

Simple linear regression shows that temperature is statistically significant. We have  $t = \frac{-0.21602}{0.04773} = -4.53$  with  $p = 0.00$ . Even though the slope is significantly different from 0, there is still a good deal of unexplained variability. Since only 32.8% of the variation is explained, we have 67.2% unexplained—so we'll include an indicator variable and look at the two groups. Moreover, a look at the residual plot indicates that the residuals from the before insulation data appears to have a significant negative slope and the after insulation residuals appear to be scattered around  $-1$  rather than 0.

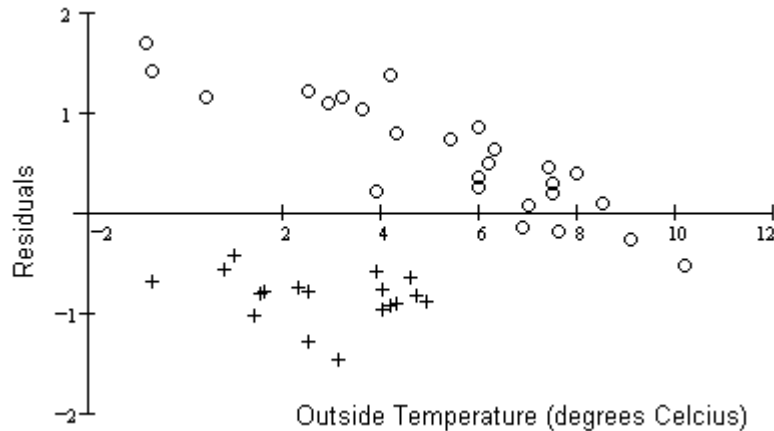


Figure 12: Residual Plot for  $fitGas = 5.33 - 0.216Temp$

### Adding a Categorical (Indicator) Variable

With the home gas consumption data, we want to know if additional variability can be explained by whether or not there was insulation in the home. So, we add the indicator variable, *Insul*, to the previous model. Note that the *Insul* variable is either 0 or 1.

The regression model is

$$fitGas = 6.72 - 0.368Temp - 1.79 Insul .$$

Using this model, we have different intercepts before and after insulation.

-Before insulation: if  $Insul = 0$ , then  $fitGas = 6.72 - 0.368Temp$

-After insulation: if  $Insul = 1$ , then  $fitGas = 4.93 - 0.368Temp$

When we interpret this model, we are assuming there is the same slope both before and after insulation was added to the home. Holding temperature constant, installing insulation reduces gas consumption, on average, 1790 cubic feet ( $-1.79 \cdot 1000 \text{ cubic feet} = -1790$ ).

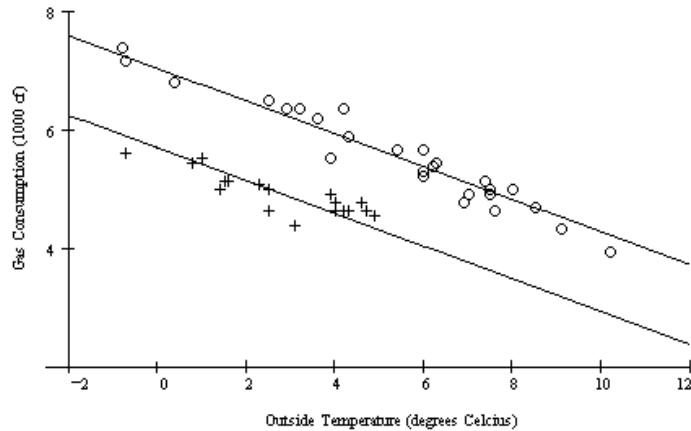


Figure 13: Two parallel lines for with and without Insulation

For the home gas consumption data, both lines appear to do a good job. Notice that the two lines are parallel. Holding all other variables constant, Gas Consumption drops, on average, 368 cubic feet for every 1° C increase in outside temperature.

## Questions of Statistical Significance

### Regression Analysis (Minitab Output)

The regression equation is  
 Gas = 6.72 - 0.368 Temp - 1.79 Insul

Predictor	Coef	StDev	T	P
Constant	6.7171	0.1169	57.48	0.000
Temp	-0.36769	0.01889	-19.47	0.000
Insul	-1.7946	0.1035	-17.33	0.000

S = 0.2993      R-Sq = 91.9%      R-Sq(adj) = 91.5%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	41.818	20.909	233.48	0.000
Error	41	3.672	0.090		
Total	43	45.490			

Source	DF	Seq SS
Temp	1	14.912
Insul	1	26.906

Remember that when we have multicollinearity, variables share information. We must be careful interpreting slope coefficients and tests of significance. The *t*-test for the variable *Insul* essentially looks at whether adding the predictor variable *Insulation* to a model with the predictor variable *Temperature* is statistically significant. Given the other variable in the model, does adding this new one contribute significantly to the models predictive ability?

We find that the variable *Temperature* is statistically significant. We also find the indicator variable *Insulation* to be statistically significant. This significance means that each variable adds predictive ability to the model.

In considering the significance of the *t*-values, we assume all other variables are part of the model. The question is, does this particular variable, when added to the model containing all other variables, add predictive ability. A significant *t*-value says the answer is "Yes".

In interpreting the Sequential Sums of Squares (Seq SS in the table), there is an implied order. By reading from the table, we see that having only *Temperature* as a predictor adds 14.912 to the sums of squares over just the average gas consumption. So  $\frac{14.912}{45.49} \approx 0.328 = R^2$ . With *Temperature* already in the model, does adding *Insulation* increase  $R^2$  significantly? From the Seq SS table, we see that 26.906 is added to the sum of squares, so  $\frac{26.906}{45.49} \approx 0.591$  is added to  $R^2$  giving a total of  $\frac{14.912 + 26.906}{45.490} \approx 0.919$ .

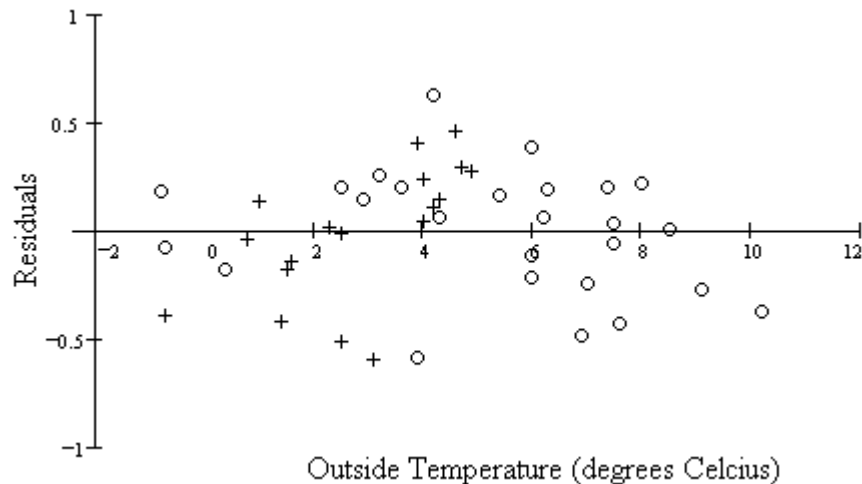


Figure 14: Plot of residuals (*Temp* and *Insul*)

If you cover up the points in the lower left-hand corner, the residuals for the crosses (+) would slope up while the O's (o) will slope down. This causes us concern that maybe the slopes should not be the same. Also note that all crosses (after insulation) are for low temperatures indicating that our sample might not be as random as we originally assumed.

### Adding an Interaction Term

Should there be a different slope after insulation than before insulation? To investigate this question, we will bring in a new interaction variable:  $Temp \cdot Insul$ . This will allow for different slopes, as we will soon see. This variable will yield 0's when there is no insulation and temperature values when there is insulation.

## Minitab output

### Regression Analysis

The regression equation is  
Gas = 6.85 - 0.393 Temp - 2.26 Insul + 0.144 Temp\*Insul

Predictor	Coef	StDev	T	P
Constant	6.8538	0.1136	60.32	0.000
Temp	-0.39324	0.01879	-20.93	0.000
Insul	-2.2632	0.1728	-13.10	0.000
Temp*Insul	0.14361	0.04455	3.22	0.003

S = 0.2699      R-Sq = 93.6%      R-Sq(adj) = 93.1%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	42.575	14.192	194.77	0.000
Error	40	2.915	0.073		
Total	43	45.490			

Source	DF	Seq SS
Temp	1	14.912
Insul	1	26.906
Temp*Insul	1	0.757

Note that  $R^2$  is now 93.6% and  $p=0.003$  for the new variable  $Temp \cdot Insul$ . Adding this new variable is still statistically significant. This  $p$ -value essentially tells us that the change in  $R^2$  value from 91.9% to 93.6% is statistically significant. When we add this variable, we get more of an increase in  $R^2$  than we would expect by chance alone, or from a variable that is not related.

Our prediction equation using this multiple regression is

$$fitGas = 6.85 - 0.393Temp - 2.26Insul + 0.144Temp \cdot Insul$$

There are now two separate lines.

Before insulation, we let  $Insul = 0$ , so  $fitGas = 6.85 - 0.393Temp$ .

After insulation, we let  $Insul = 1$ , and  $fitGas = 4.59 - 0.249Temp$ .

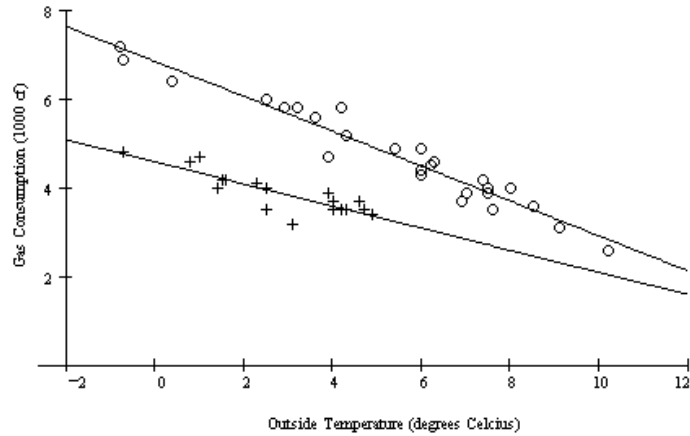


Figure 15: Home Gas Consumption Graph (Full Model)

Before insulation, there is a higher initial predicted gas consumption (6850 cubic feet at 0° C) and a steeper average decline (393 cubic feet less for every 1°C increase in average outside temperature.) After insulation, there is a lower initial predicted gas consumption (4590 cubic feet at 0°C) and a slower average decline (249 cubic feet less for every 1°C increase in average outside temperature). Obviously, gas consumption doesn't change as much when outside temperature changes if the house is insulated.

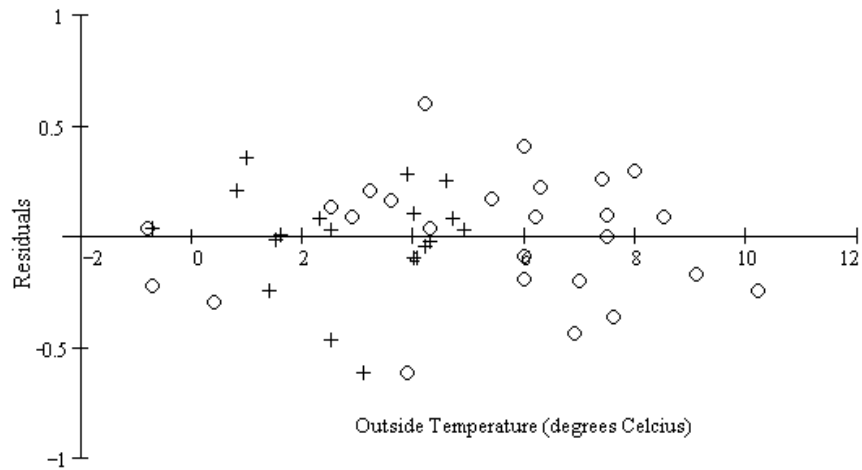


Figure 16: Plot of Residuals (for Full Model)

But are these apparent differences statistically significant?

	Coef	StDev	t	p
Predictor	6.85	0.114	60.32	0.000
Constant	-0.393	0.019	-20.93	0.000
Temp	-2.26	0.173	-13.10	0.000
Insul	0.144	0.045	3.22	0.003
Temp*Insul				

Given the other variables in the model, each individual variable is statistically significant. From earlier models, we've learned that temperature is significant, adding the indicator for insulation adds significantly, and adding the interaction term adds significantly. By adding the interaction term, we have explained more of the variability, and the increase is more than could be expected by adding an unrelated variable. Recall that every time you add a variable  $R^2$  increases, but you give up a  $df$ . You always pick up some  $SS_{Error}$  and move it up to the  $SS_{Model}$ . This essentially adds to the increased  $R^2$ .

### Insulation Only Model

One model left out of the analysis is the *Insulation* only model. If we fit a Simple Linear Regression for Gas Consumption against Insulation, we get the model

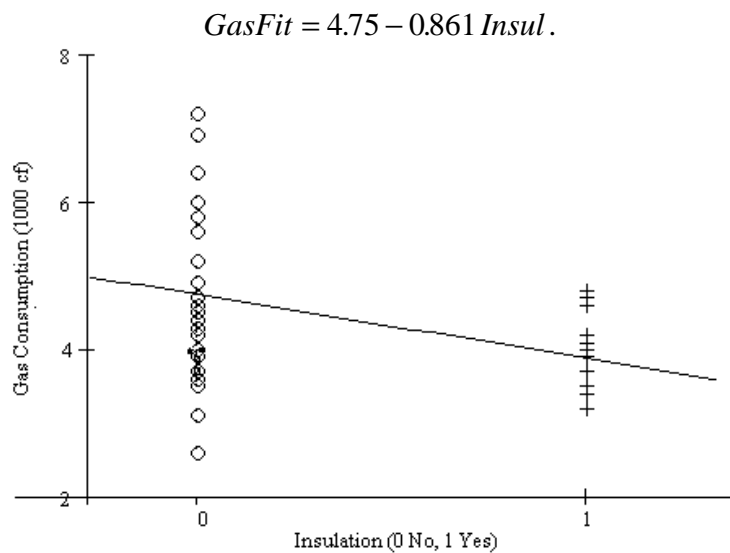


Figure 17: Insulation only model

Here  $R^2 = 17.3\%$  with  $t = -2.97$ ,  $p = 0.005$  for the test of  $H_0: \beta_1 = 0$  against  $H_a: \beta_1 \neq 0$ . This slope is significant, even though only 17.3% of the variation in Gas Consumption is explained by Insulation.

The average *Gas Consumption* without *Insulation* is 4750 cubic feet while the average *Gas Consumption* with *Insulation* is 3889 cubic feet. The fact that the slope is significant means that there is a difference in these two means. We will see this more clearly shortly.

### Change in $R^2$

Insulation only:  $R^2 = 17.3\%$

Temperature only:  $R^2 = 32.8\%$

Temperature and Insulation:  $R^2 = 91.9\%$

Temperature, Insulation and Temp\*Insul:  $R^2 = 93.6\%$

Each change is statistically significant; we know this because each variable was significant as it was added. What is the relationship between the change in  $R^2$  and the statistical significance of the  $t$ -test?

The change in  $SS_{Error}$  is what is added to  $R^2$  from one model to the next.

Our partial  $F$ -statistic is  $F = \frac{\left( \frac{SS_{Error(Reduced)} - SS_{Error(Full)}}{df_{Error(Reduced)} - df_{Error(Full)}} \right)}{MS_{Error(Full)}}$ . The difference in the degrees of freedom will be 1 if we are looking at one value at a time.

**The Full Model** (includes  $Temp$ ,  $Insul$ , and  $Temp \cdot Insul$ )

Source	df	SS
Regression	3	42.575
Error	40	2.915
Total	43	45.490

**Reduced Model** (no  $Temp \cdot Insul$ )

Source	df	SS
Regression	2	41.818
Error	41	3.672
Total	43	45.490

From the two tables above, we see that

$$F = \frac{\left( \frac{(3.672 - 2.915)}{(41 - 40)} \right)}{0.073} = \frac{0.757}{0.073} = 10.37 \quad (0.757 \text{ is the change in } SS_{Error})$$

The  $p$ -value is 0.0025. Note that this  $p$ -value for the interaction term matches the  $p$ -value associated with the  $t$ -value for the interaction term in the regression output. Also,  $t_{df} = \sqrt{F_{1,df}}$ , so we must be adding one variable at a time for the  $t$  and  $F$  to be equivalent.

For example, for  $Temp \cdot Insul$ ,  $t = \sqrt{10.37} = 3.22$ . In other words, 3.22 is the table value of  $t$  that indicates the significance of adding the variable. Note that as you toss in one variable at a time, the  $t$ -score from the computer output gives the same information as this  $F$ , so the  $t$ -score can be interpreted as whether the change in  $R^2$  is significant. The  $F$ -test is more general in that we can add several variables at a time and test the null hypothesis  $H_0$ : all new slopes = 0.

We could also think of this problem as being two separate sets of data, the Before Insulation and After Insulation. We have two approaches to consider. First, fit a separate simple linear regression to each data set. How do these separate regressions compare to the full model given here? Second, perform a two-sample t-test on the means. Is the average amount of gas used after insulation less than the mean before insulation. How does this test compare to the regression analysis?

### Fit Separate Linear Models

If we fit two separate linear equations, one to the before insulation data and another to the after insulation data, how would they compare to the full multivariate model?

Before insulation, our prediction equation is

$$\text{fitGas} = 6.85 - 0.393\text{Temp}, R^2 = 94.4\%, MS_{Error} = 0.079, \text{ and } df_{Error} = 24.$$

This is the same equation we would get if we used the multiple regression model with  $Insul = 0$ .

After insulation, our prediction equation is

$$\text{fitGas} = 4.59 - 0.250\text{Temp}, R^2 = 73.3\%, MS_{Error} = 0.063, \text{ and } df_{Error} = 16.$$

This is the same equation we would get if we set  $Insul = 1$  in the result of the multiple regression.

Note that in the multivariate model, we assumed that  $\mathbf{s}^2$  was the same for all  $x$ 's. This assumption was not necessary for the two separate models.

The predictions from these two separate models will be the same as those from the Full Model with  $Temperature$ ,  $Insulation$ , and  $Temp \cdot Insul$ . If you pool the  $MS_{Error}$  from these separate analyses (0.079 from before insulation and 0.063 from after insulation), you get the  $MS_{Error}$  for the Full Model.

$$\frac{[24(0.079) + 16(0.063)]}{40} = \frac{2.915}{40} = 0.073$$

Essentially the difference in the two separate linear models and the single multivariate model is whether or not you pool the variances. So which is the better model? Predictions will be the same from both, and  $R^2$  is generally better when the two linear models are combined. Although the predictions will be the same from either model, the prediction errors will likely be smaller with the full model. Since pooling variances gives us a better estimate of  $\mathbf{s}^2$ . It could, however, be argued that all we want is the information on insulation, since we will never remove insulation that is already present.

## 2-Sample $t$ -Test

If we perform a 2-sample  $t$ -test on the before and after insulation gas consumption, pooling the variance, we find some other interesting comparisons.

Here we have:

<u>Before Insulation</u>	<u>After Insulation</u>
$\bar{x} = 4.75$	$\bar{x} = 3.889$
$s^2 = 1.35$	$s^2 = 0.2234$
$n = 26$	$n = 18$

The pooled variance is  $s^2 = \frac{25(1.35) + 17(0.2234)}{42} = 0.8953$ , so the  $t$ -value is

$$t_{42} = \frac{4.75 - 3.889}{\sqrt{0.8953} \sqrt{\frac{1}{26} + \frac{1}{18}}} = 2.968. \text{ This is the same } t\text{-value that we got in the } \textit{Insulation}$$

only model. There, we compared the two intercepts, 4.75 if  $I = 0$  and 3.889 if  $I = 1$ . The  $p$ -value corresponding to this  $t$ -value is  $p = 0.005$ .