

Introduction to Linear Models Using R

January 14, 2010

1 Some R commands for linear models

lm fit a linear model

summary summary method for **lm** object

anova ANOVA method for **lm** object(s)

plot plot data or plot residuals of **lm** object

abline add a line $y = a + bx$ to an x-y plot

update update method for **lm** object

2 The data set: whiteside

Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption.

We load the **whiteside** data, which is part of the **MASS** package. It uses 'lazy loading' so the **data** command is not needed here. We can check that it is loaded by displaying the help topic. To quickly check the format we can display the top of the file using **head** or describe it in another way using **str**.

```
> library(MASS)
> `?`(whiteside)
> str(whiteside)
```

```
'data.frame':      56 obs. of  3 variables:
 $ Insul: Factor w/ 2 levels "Before","After": 1 1 1 1 1 1 1 1 1 1 1 ...
 $ Temp : num  -0.8 -0.7 0.4 2.5 2.9 3.2 3.6 3.9 4.2 4.3 ...
 $ Gas  : num  7.2 6.9 6.4 6 5.8 5.8 5.6 4.7 5.8 5.2 ...
```

```
> head(whiteside)
```

```
   Insul Temp Gas
1 Before -0.8 7.2
2 Before -0.7 6.9
3 Before  0.4 6.4
4 Before  2.5 6.0
5 Before  2.9 5.8
6 Before  3.2 5.8
```

Descriptive statistics:

```
> by(whiteside, whiteside$Insul, FUN = summary)
```

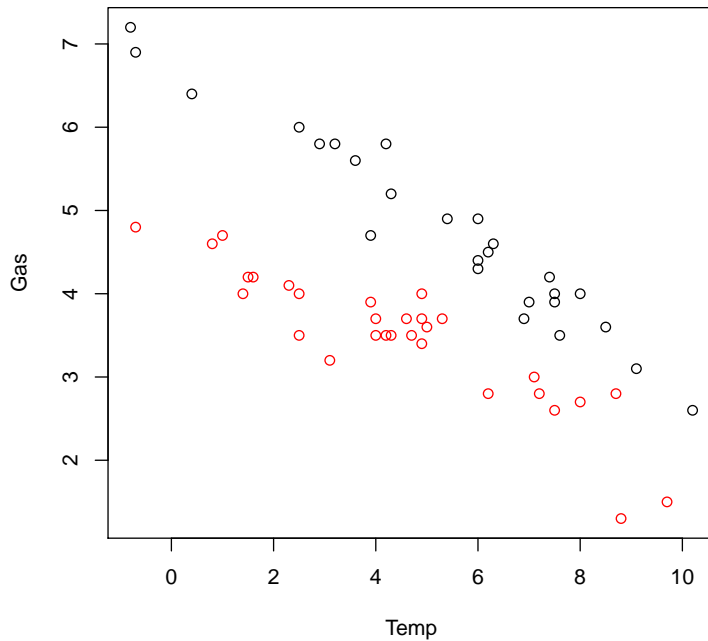
```
whiteside$Insul: Before
   Insul      Temp      Gas
Before:26  Min.   :-0.800  Min.   :2.600
After : 0   1st Qu.: 3.675  1st Qu.:3.925
          Median : 6.000  Median :4.550
          Mean   : 5.350  Mean   :4.750
          3rd Qu.: 7.475  3rd Qu.:5.750
          Max.   :10.200  Max.   :7.200
```

```
-----
whiteside$Insul: After
   Insul      Temp      Gas
Before: 0  Min.   :-0.700  Min.   :1.300
After :30  1st Qu.: 2.500  1st Qu.:3.050
          Median : 4.450  Median :3.550
          Mean   : 4.463  Mean   :3.483
          3rd Qu.: 5.975  3rd Qu.:4.000
          Max.   : 9.700  Max.   :4.800
```

3 Preliminary analysis

Display a scatterplot of the data before and after insulation.

```
> plot(Gas ~ Temp, data = whiteside, col = Insul)
```



If we did not identify the groups by color, we would perhaps consider fitting a simple linear regression model to the data.

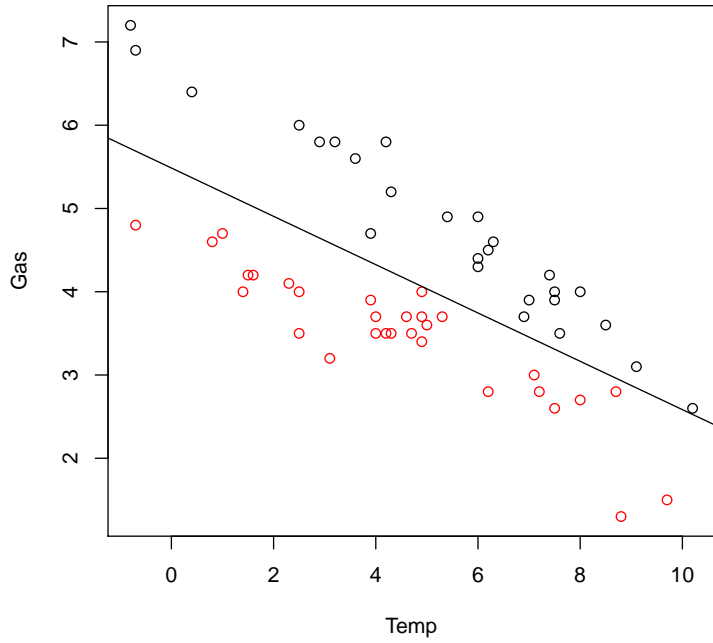
```
> fit <- lm(Gas ~ Temp, data = whiteside)
> fit
Call:
lm(formula = Gas ~ Temp, data = whiteside)
```

```
Coefficients:
(Intercept)      Temp
    5.4862      -0.2902
```

We stored the fitted model object in 'fit' and printing 'fit' displays the coefficients of the fitted model. Optionally one can use `summary` or `anova` methods to display much more information.

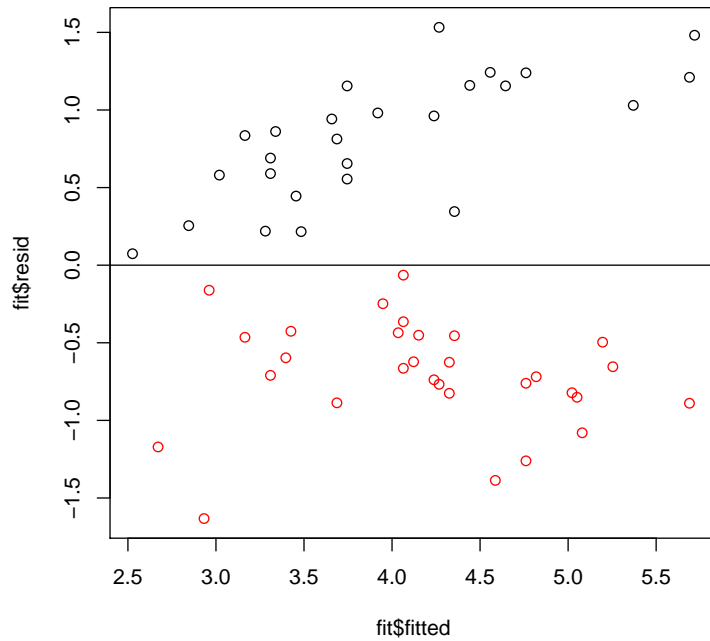
Notice that the fitted line does not fit either subset of the data.

```
> plot(Gas ~ Temp, data = whiteside, col = Insul)
> abline(fit$coef)
```



A plot of fitted values vs residuals:

```
> plot(fit$fitted, fit$resid, col = whiteside$Insul)  
> abline(h = 0)
```



4 lattice and panel display

Here is another view of before/after data, using `xyplot` in `lattice` package.

```
library(lattice)
xyplot(Gas ~ Temp | Insul, data=whiteside)
```

For an `xyplot` with fitted lines added on each panel, we first define a panel function. Here is a simple version.

```
panel <- function(x, y) {
  panel.xyplot(x, y)
  panel.lmline(x, y)
}

xyplot(Gas ~ Temp | Insul, data=whiteside, panel=panel)
```

5 Analysis of the Before data

Let us fit the model to the Before insulation data only.

```
> fitBefore <- lm(Gas ~ Temp, data = whiteside,
+ subset = (Insul == "Before"))
> fitBefore
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = (Insul == "Before"))

Coefficients:
(Intercept)      Temp
      6.8538      -0.3932
> summary(fitBefore)
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = (Insul == "Before"))

Residuals:
      Min       1Q   Median       3Q      Max
-0.62020 -0.19947  0.06068  0.16770  0.59778

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.85383    0.11842   57.88  <2e-16 ***
Temp        -0.39324    0.01959  -20.08  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2813 on 24 degrees of freedom
Multiple R-squared:  0.9438,    Adjusted R-squared:  0.9415
F-statistic: 403.1 on 1 and 24 DF,  p-value: < 2.2e-16
> anova(fitBefore)
Analysis of Variance Table

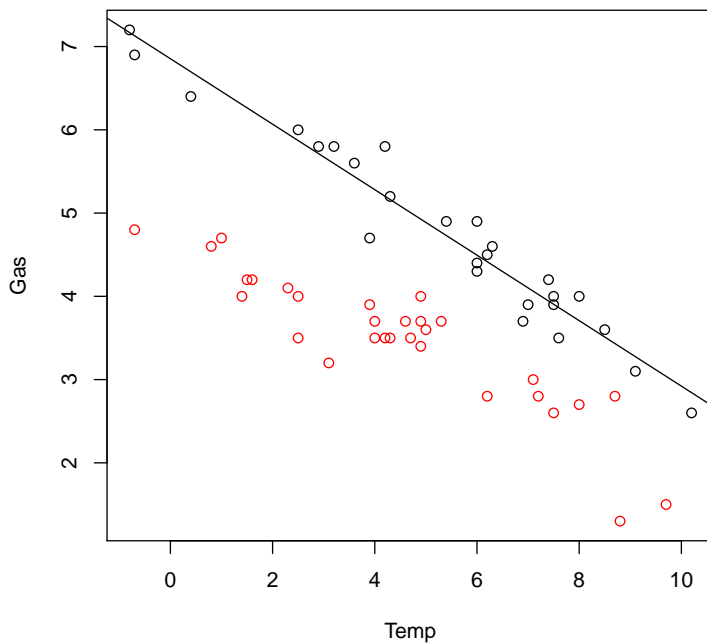
Response: Gas
          Df Sum Sq Mean Sq F value    Pr(>F)
Temp       1  31.905   31.905  403.11 < 2.2e-16 ***
Residuals 24   1.900    0.079

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Now the resulting fitted line plot should fit the Before-insulation data; we keep both subsets in the plot for comparison with our original attempt.

```
> plot(Gas ~ Temp, data = whiteside, col = Insul)
> abline(fitBefore$coef)
```



To obtain a set of residual plots automatically we can use the `plot` method for the `lm` object. (Output of commands not shown.)

```
# residual plots (default set)
plot(fitBefore)
```

6 Fit both models and compare them

An easy way to fit both models in two steps uses the `update` method.

```

> fitBefore <- lm(Gas ~ Temp, data = whiteside,
+   subset = (Insul == "Before"))
> fitAfter <- update(fitBefore, subset = (Insul ==
+   "After"))
> fitBefore
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = (Insul ==
"Before"))

Coefficients:
(Intercept)      Temp
      6.8538      -0.3932
> fitAfter
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = (Insul ==
"After"))

Coefficients:
(Intercept)      Temp
      4.7238      -0.2779

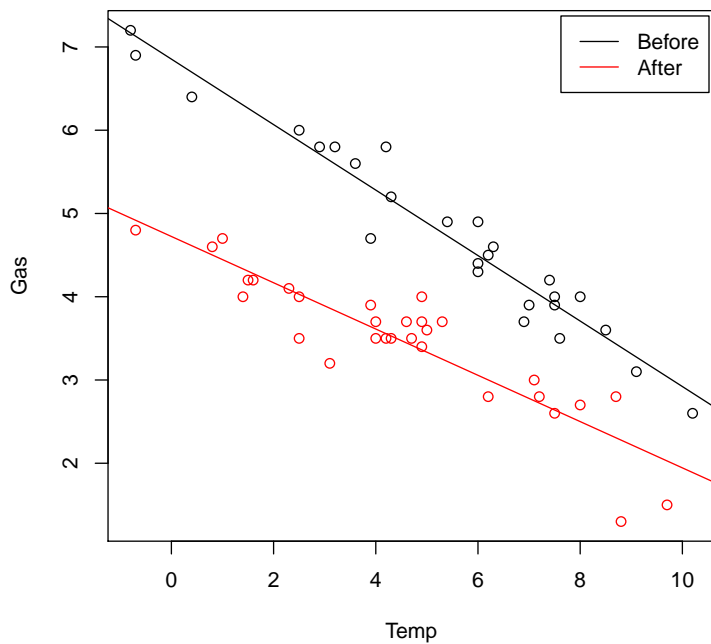
```

Next, we review our original plot of data, and add the second fitted line for the After-insulation data. (Only need to add the second `abline` if the plot is still displayed.) The legend is optional.

```

> plot(Gas ~ Temp, data = whiteside, col = Insul)
> abline(fitBefore$coef)
> abline(fitAfter$coef, col = 2)
> legend("topright", inset = 0.01, c("Before", "After"),
+   col = 1:2, lty = 1)

```



7 Model including Before-After (Insul) Terms

Let us include insulation in the model. Slopes may be different. Specify a model with insulation included, fitting with different slopes.

```
> fit2 <- lm(Gas ~ Insul * Temp, data=whiteside)
```

```
> fit2
```

Call:

```
lm(formula = Gas ~ Insul * Temp, data = whiteside)
```

Coefficients:

(Intercept)	InsulAfter	Temp	InsulAfter:Temp
6.8538	-2.1300	-0.3932	0.1153

```
> summary(fit2)$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.8538277	0.13596397	50.409146	7.997414e-46
InsulAfter	-2.1299780	0.18009172	-11.827185	2.315921e-16
Temp	-0.3932388	0.02248703	-17.487358	1.976009e-23
InsulAfter:Temp	0.1153039	0.03211212	3.590665	7.306852e-04

Note: The formula

`Gas ~ Insul*Temp`

expands to:

`Gas ~ 1 + Insul + Temp + Insul:Temp`

and the estimates are

1. intercept for Before
2. difference in intercepts
3. slope for Before
4. difference in slopes

It is clear that the slopes are different (see `InsulAfter:Temp`). The estimated difference in slopes is significantly different from 0. The term `Insul:Temp` is necessary in the model, so the additive model is not adequate.

However, let us compare (anyway) with the additive model, which assumes parallel lines (equal slopes). We can compare the models using one `anova` statement as follows.

```
> fit3 <- lm(Gas ~ Insul+Temp, data=whiteside)
> anova(fit3, fit2)
```

Analysis of Variance Table

```
Model 1: Gas ~ Insul + Temp
Model 2: Gas ~ Insul * Temp
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     53 6.7704
2     52 5.4252  1    1.3451 12.893 0.0007307 ***
```

The result of the `anova` command on the two nested models produces a different style of ANOVA table. This table also confirms that the term `Insul:Temp`, which is added in the second model, is significant.

