Statistical Data Analysis

N. Sairam

sairam@cse.sastra.edu

School of Computing, SASTRA University, Thanjavur.

Unit-III

Multiple Linear Regression

- In straight line regression, a response variable y is regressed on a single explanatory variable x
- Multiple linear regression generalizes this methodology to allow multiple explanatory or predictor variables The
- Accurate Prediction is our focus

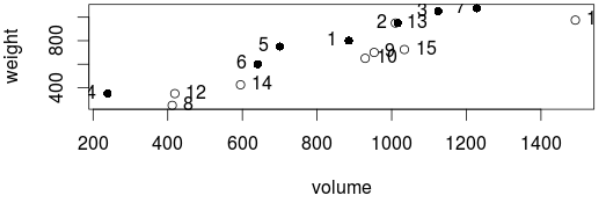
Basic Idea: Example

- Let us consider the book weight example that has two x-variables in the regression equation
- Explanatory variables are the volume of the book ignoring the covers, and the total area of the front and back covers
- weight of book $= b_0 + b_1$ x volume $+ b_2$ x area of covers

R Code

```
lot(weight   volume, data=allbacks, pch=c(16,1)
[unclass(cover)])
# unclass(cover) gives the integer codes that
 identify levels
with(allbacks, text(weight   volume, labels=paste(1:15),
pos=c(2,4)[unclass(cover)]))
```

Graph


weight
400
800
200
400
600
800
1000
1200
1400
volume
1
2
3
4
5
6
7
8
9
10
12
13
14
15

Summary of the Regression Model

```
summary(allbacks.lm <- lm(weight~volume+area, data=allbacks
Output:
Call:
lm(formula = weight ~ volume + area,
data = allbacks)

Residuals:
    Min      1Q  Median      3Q     Max
-104.06  -30.02  -15.46   16.76  212.30

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.41342   58.40247   0.384 0.707858
volume       0.70821    0.06107  11.597 7.07e-08 ***
area         0.46843    0.10195   4.595 0.000616 ***
---
```

Summary of the Regression Model

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

Residual standard error: 77.66 on 12 degrees of freedom
Multiple R-squared:  0.9285,Adjusted R-squared:  0.9166
F-statistic: 77.89 on 2 and 12 DF,  p-value: 1.339e-07
```

Analysis of the Results

- The coefficient estimates are $b_0 = 22.4$, $b_1 = 0.708$, and $b_2 = 0.468$
- Standard errors and p-values are provided for each estimate
- The p-value for the intercept suggests that it cannot be distinguished from 0
- The p-value for volume tests $b_1 = 0$, in the equation that has both volume and area as explanatory variables
- The estimate of the noise standard deviation (the residual standard error) is 77.7
- There are now 15-3 = 12 degrees of freedom for the residual

Analysis of the Results

- The null hypothesis for this test is that all coefficients (other than the intercept) are 0
- Here, we reject this hypothesis and conclude that the equation does have explanatory power
- Confidence Interval for the volume: 0.708±qt(0.975, 12)*0.0611
- Output: 0.708±2.178813*0.0611=0.575 to 0.841
- anova(allbacks.lm)
- model.matrix(allbacks.lm)

Analysis of Anova Table

- This table gives the contribution of volume after fitting the overall mean, then the contribution of area after fitting both the overall mean and volume
- The p-value for area in the anova table must agree with that in the main regression output, since both these p-values test the contribution of area after including volume in the model
- The p-values for volume will differ if there is a correlation between volume and area
- Command to compute correlation:with(allbacks, cor(volume,area))
- Here, the correlation of volume with area is 0.0015

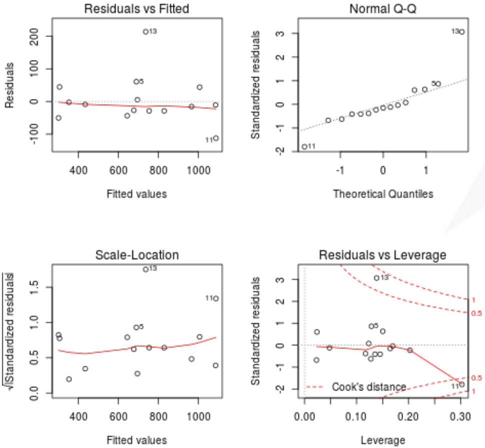
Analysis of model.matrix() Results

- Predicted values are given by multiplying the first column by b_0 (=22.4), the second by b_1 (=0.708), the third by b_2 (=0.468), and adding
- Omission of the Intercept Term:
  - allbacks.lm0 <- lm(weight -1+volume+area, data=allbacks)
  - summary(allbacks.lm0)
  - The regression coefficients now have smaller standard errors
  - The reason is that, in the model that included the intercept, there was a substantial negative correlation between the estimate of the intercept and the coefficient estimates
  - The reduction in standard error is greater for the coefficient of volume, where the correlation was -0.88, than for area, where the correlation was -0.32. Correlations between estimates can be obtained by setting corr=TRUE in the call to summary()

Diagnostic Plots

- Let us consider the following code:

```
par(mfrow=c(2,2));plot(allbacks.lm0);
dev.copy(png,'31.png');dev.off()
```



Explanation of the Diagnostic Plots

- The residual for observation 13 is large
- The observation 13 lies outside the 0.5 contour of Cook's distance, well out towards the contour for a Cook's distance of 1.
- It is a (somewhat) influential point

What happens if we omit observation 13?

- R code: allbacks.lm13 <- lm(weight -1+volume+area, data=allbacks[-13,])
- summary(allbacks.lm13)

Ouput of the above code

```
Call:
lm(formula = weight ~ -1 + volume + area,
 data = allbacks[-13, ])

Residuals:
    Min      1Q  Median      3Q     Max
-61.721 -25.291   3.429  31.244  58.856

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
volume  0.69485    0.01629   42.65 1.79e-14 ***
area    0.55390    0.05269   10.51 2.08e-07 ***
---
```

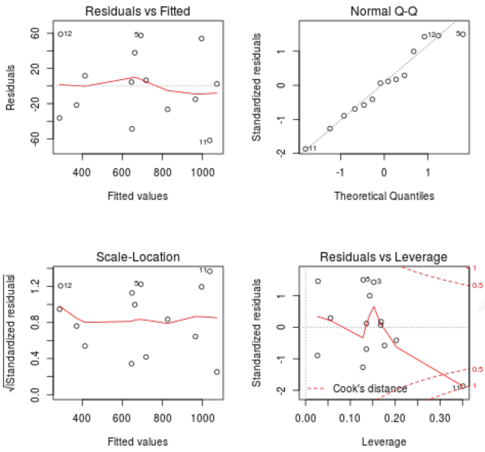
Output of the above code

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

Residual standard error: 41.02 on 12 degrees of freedom
Multiple R-squared:  0.9973,Adjusted R-squared:  0.9969
F-statistic:  2252 on 2 and 12 DF,  p-value: 3.521e-16
```

- The residual standard error is substantially smaller (41 instead of 75.1) in the absence of observation 13
- Observation 11 now has a Cooks distance that is close to 1, but does not stand out in the plot of residuals

Results


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

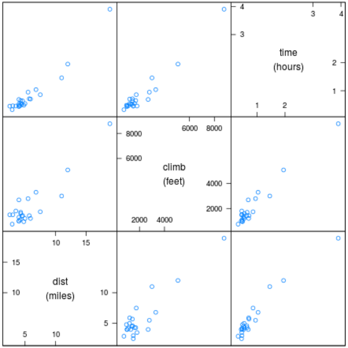
Interpretation of Model Coefficients

- To Understand the interpretation of model coefficients then it is important to fit a model whose coefficients are open to the relevant interpretations
- Different formulations of the regression model, or different models, may serve different explanatory purposes
- Predictive accuracy is in any case a consideration, and is often the main interest

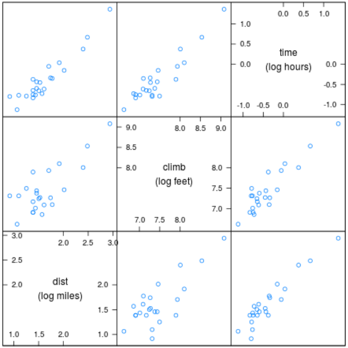
Example 1

- Let us consider the data set nihills (DAAG), that gives the distances (dist), heights climbed (climb), male record times (time), and female record times (timef), for Northern Irish hill races
- Let us begin with scatter plot matrices, both for the untransformed data and for the log transformed data
- Let us limit our attention to Male
- The diagonal panels give the x-variable names for all plots in the column above or below, and the y-variable names for all plots in the row to the left or right
- Note that the vertical axis labels alternate between the axis on the extreme left and the axis on the extreme right, and similarly for the horizontal axis labels

Scatter Plot



Scatter Plot Matrix



Scatter Plot Matrix

Investigation of Taking Logarithms

- The range of values of time is large (3.9:0.32, i.e., $>$10:1), and similarly for dist and climb. The times are bunched up towards zero, with a long tail. In such instances, use of a logarithmic transformation is likely to lead to a more symmetric distribution
- One point in particular has a time that is more than twice that of the next largest time. The values of dist and climb similarly stand out as much larger than for other points. In a regression that uses the untransformed variables, this point will have a much greater say in determining the regression equation than any other point. In the terminology, it has large leverage. Even after taking logarithms, its leverage remains large, but not quite so dominating

Investigation of Taking Logarithms

- It can be expected that time will increase more than linearly at very long times, and similarly for climb, as physiological demands on the human athlete move closer to limits of human endurance
- Such relationship as is evident between the explanatory variables (dist and climb) is more nearly linear on the logarithmic scale
- Additionally, use of a logarithmic scale may help stabilize the variance

Fitting the Equation

$$log(time) = a + b_1 log(dist) + b_2 log(climb) \quad (1)$$

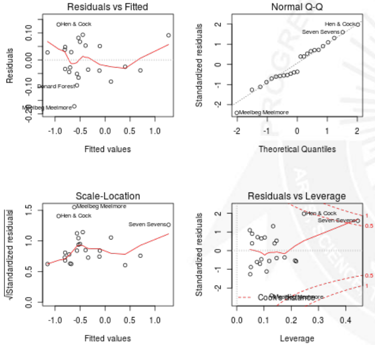
- Equivalent to Power Relationship

$$time = A(dist)^{b_1}(climb)^{b_2} \quad (2)$$

- where a = log(A)

R Code for fitting the model

```
nihills.lm<-lm(log(time)
~log(dist)
+log(climb),data=nihills)
plot(nihills.lm)
```



Summary

```
summary(nihills.lm)$coef

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -4.9611313 0.27387193 -18.11479 7.085048e-14
log(dist)    0.6813596 0.05517831  12.34832 8.186381e-11
log(climb)   0.4657575 0.04530181  10.28121 1.980592e-09
```

Interpreting the coefficients

- The estimated equation is log(time) $=$ -4.96+0.68 x log(dist)+0.47x log(climb)
- Exponentiating both sides of this equation, and noting exp(-4.96) $=$ 0.0070, gives time $=$ 0.00070 x dist$^{0.68}$ x climb$^{0.47}$
- This equation implies that for a given height of climb, the time taken is smaller for the second three miles than for the first three miles

A meaningful coefficient for logdist

- The coefficient for logdist will be more meaningful if we regress on logdist and log(climb/dist)
- R Code:

```
>lognihills <- log(nihills)
> names(lognihills) <- paste("log", names(nihills),
 sep="")
> lognihills$logGrad <- with(nihills, log(climb/dist))
> nihillsG.lm <- lm(logtime   logdist + logGrad,
 data=lognihills)
>nihillsG.lm <- lm(logtime~ logdist + logGrad,
 data=lognihills)
> summary(nihillsG.lm)$coef
```

Output for the R Code

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -4.9611313 0.27387193 -18.11479 7.085048e-14
logdist      1.1471170 0.03459867  33.15494 5.896354e-19
logGrad      0.4657575 0.04530181  10.28121 1.980592e-09
```

Analysis of the above code

- The coefficient of logdist is now, greater than 1

```
cor(lognihills$logdist,lognihills$logGrad)
[1] -0.06529222
cor(lognihills$logdist,lognihills$logclimb)
[1] 0.780067
```

- The correlation between logdist and logGradient is 0.065, negligible relative to the correlation of 0.78 between logdist and logclimb

```
nihills.lm<-lm(logtime~logdist,data=lognihills)
> summary(nihills.lm)
```

Analysis of the above code

```
summary(nihills.lm)$coeff
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -2.210125 0.14284610 -15.47207 5.910392e-13
logdist      1.123892 0.08446864  13.30543 1.060019e-11
```

- Because the correlation between logdist and logGradient is so small, the coefficient of logdist (=1.124) in the regression on logdist alone is almost identical to the coefficient of logdist (=1.147) in the regression on logdist and logGradient
- The standard error of the coefficient of logdist is smaller - 0.035 as against 0.045 - when the second explanatory variable is logGradient rather than logclimb
- Note that the predicted values do not change
- The models nihills.lm nihillsG.lm are different mathematical formulations of the same underlying model

Scatter Plot Matrix

```
library(lattice); library(DAAG)
splom(  nihills[, c("dist","climb","time")],
 cex.labels=1.2,
varnames=c("dist\n(miles)","climb\n(feet)",
 "time\n(hours)"))
## Panel B: log transformed data
splom(  log(nihills[, c("dist","climb","time")]),
 cex.labels=1.2,
varnames=c("dist\n(log miles)", "climb\n(log feet)",
"time\n(log hours)"))
```

Exercise

- In the data set cement (MASS package), examine the dependence of y (amount of heat produced) on x1, x2, x3 and x4 (which are proportions of four constituents). Begin by examining the scatterplot matrix. As the explanatory variables are proportions, do they require transformation, perhaps by taking log(x/(100-x))?

Plots that show the contribution of individual terms

- For simplicity, the discussion will assume just two explanatory variables, x_1 and x_2, with the intention of showing the contribution of each in turn to the model
- The fitting of a regression model makes it possible to write:

$$\begin{aligned} y &= b_0 + b_1 x_1 + b_2 x_2 + e \qquad (3) \\ &= \hat{y} + e \qquad (4) \end{aligned}$$

- Another way to write the model that is to be fitted is:

$$y - \bar{y} = a + b_1(x_1 - \bar{x_1}) + b_2(x_2 - \bar{x_2}) + e \qquad (5)$$

Plots that show the contribution of individual terms

- For fitting the model in this form:
  - The observations are y-$\bar{y}$, with mean zero
  - The first explanatory variable is x_1-$\bar{x_1}$, with mean zero, and the first term in the model is $b_1(x_1 - \bar{x_1})$, with mean zero
  - The second explanatory variable is x_2-$\bar{x_2}$, with mean zero, and the first term in the model is) $b_2(x_2 - \bar{x_2})$, with mean zero
- The residuals e are exactly the same as before, and have mean zero
- The fitted model can then be written:

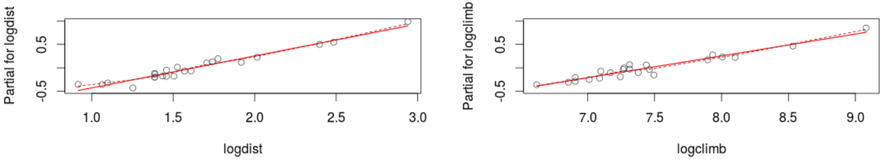
$$
\begin{aligned}
y = \bar{y} + b_1(x_1 - \bar{x_1}) + b_2(x_2 - \bar{x_2}) + e \quad (6) \\
= \bar{y} + t_1 + t_2 + e \quad (7)
\end{aligned}
$$

- Splits the response value y into three parts - an overall mean $\bar{y}$, a term that is due to x_1, a term that is due to x_2, and a residual e
- Moreover, the values of t_1 and t_2 sum,in each case, to zero

- The predict() function has an option (type="terms") that gives t_1 and t_2
- yterms <- predict(nihills.lm, type="terms")
- The first column of yterms has the values of $t_1 = b_1 (x_1 - \bar{x_1})$, while the second has the values of t_2
- Values in both these columns sum to zero

Partial Residual Plot

- The solid lines of the component plus residual plot in the figure given below show the contributions of the individual terms to the model
- The solid line in the left panel shows a plot of b_1 (x_1 -$\bar{x}_1$) against x_1 , while the solid line in the right panel shows a plot of b_2 (x_2 - $\bar{x}_2$) against x_2

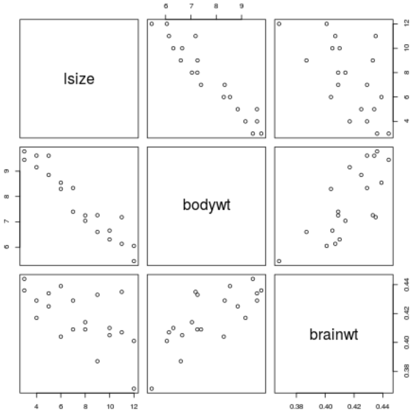


Analysis of the graph

- The lines can be obtained directly with the termplot() command
- The plotted points are the partial residuals, for the respective term
- The vector $t_1+e = \hat{y}-t_2$ holds the partial residuals for x_1 given x_2, i.e., they account for that part of the response that is not explained by the term in x_2
- The vector t_2+e holds the partial residuals for x_2 given x_1

Mouse Brain Weight Example

- The litters data frame (DAAG library) has observations on brain weight, body weight, and litter size of 20 mice



Mouse Brain Weight Example

- The explanatory variables lsize and bodywt are strongly correlated(From the graph)
- Regression of brainwt on lsize: summary(lm(brainwt lsize,data = litters))$coef

```
               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 0.447000000 0.009624762 46.442707 3.391193e-20
lsize      -0.004033333 0.001198423 -3.365534 3.444524e-03
```

Mouse Brain Weight Example

- Regression of brainwt on lsize and bodywt

```
summary(lm(brainwt~lsize+bodywt,data = litters))$coef
               Estimate  Std. Error  t value    Pr(>|t|)
(Intercept) 0.178246962 0.075322590 2.366448 0.030097278
lsize       0.006690331 0.003132075 2.136070 0.047513226
bodywt      0.024306344 0.006778653 3.585719 0.002278441
```

Interpretation of the results

- In the first regression, variation in brainwt is being explained only with lsize, regardless of bodywt
- No adjustment has been made for the fact that bodywt increases as lsize decreases: individuals having small values of lsize have brainwt values corresponding to large values of bodywt, while individuals with large values of lsize have brainwt values corresponding to low bodywt values
- In the multiple regression, the coefficient for lsize is a measure of the change in brainwt with lsize, when bodywt is held constant
- For any particular value of bodywt, brainwt increases with lsize

Multiple regression assumptions, diagnostics, and efficacy measures

- Given the explanatory variables x_1,x_2, . . . ,x_p , the assumptions are that:
  - The expectation E[y] is some linear combination of x_1,x_2, . . .,x_p :
  $$E[y] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \tag{8}$$
  - The distribution of y is normal with mean E[y] and constant variance, independently between observations

Detection of outliers

Next Class..