Microsoft: DAT209x Programming in R for Data Science

Bookmarks

- 0. Start Here
- 1. Introduction
- 2. Functions and Data Structures
- 3. Loops and Flow Control
- 4. Working with Vectors and Matrices
- 5. Reading in Data
- 6. Writing Data to Text Files
- 7. Reading Data from SQL Databases

11. Linear Models > Lab > Lab

Bookmark

Install the package **R330**, load the package and access the **wine.df** dataset with the following commands:

```
install.packages("R330")
library(R330)
data(wine.df)
```

wine.df is a data set which attempt to assess the quality of various Bordeaux vintages based upon certain variables. In **wine.df**, 400 mm rain in the preceding winter period **w.rain** is a small value, while 800mm is a big value. The data contains no evidence that rain in the preceding winter impacts on the rain in the harvest period **h.rain**. You may assume that **h.rain** does not depend on **w.rain**, when answering the questions below.

Question 1

(1/1 point)

We shall regress price on **year**, **temp**, **h.rain** and **w.rain** with the `lm()` function. Allow for an interaction between **h.rain** and **w.rain**.

What is Adjusted R-squared of the model?

- 8. Working with Data
- 9. Manipulating Data
- 10. Simulation
- 11. Linear Models

Lecture

Knowledge Checks
Quiz due Jun 27, 2016 at 23:30 UTC

Lab
Lab due Jun 27, 2016 at 23:30 UTC

- 12. Graphics in R
- Course Wrap-up

○ 0.804

● 0.7574 ✓

○ 0.7369

○ 0.6891

EXPLANATION

You can create the model and view the summary using the following code:

```
my.analysis<-lm(price~year+temp+h.rain+w.rain+h.rain:w.rain,data=wine.df)
summary(my.analysis)
```

Question 2

(1/1 point)

Reduce the model with `drop1()`, removing statistically insignificant explanatory variables from the model (if any), testing at the 5% level.

What can you say about the significance of the variables?

○ temp is not significant.

○ year is not significant.

○ Interaction between h.rain and w.rain is not significant.

● All variables are significant, the model cannot be reduced. ✓

EXPLANATION

You can run the following code to perform the task:

```
drop1(my.analysis,test="F")
```

Question 3

(1/1 point)

If the rain over the winter has been 400mm, we can assess the effect of rain over the harvest period, since the interaction term is equal to effect of the product of **h.rain** and **w.rain**.

The coefficients in the model object my.analysis are

```
coef(my.analysis)
  (Intercept)          year          temp        h.rain        w.rain h.rain:w.rain 
 1.709415e+03 -1.029350e+00  1.626221e+01  3.573569e-01  1.415583e-01 -7.210972e-04 
```

Thus, the regression coefficient to **h.rain**, when w.rain is kept fixed at 400mm, is the 4th coefficient plus 400 times the 6th coefficient:

```
coef(my.analysis)[4]+400*coef(my.analysis)[6]
    h.rain 
0.06891802 
```

This coefficient is positive, so the model predicts INCREASING prices with increasing rain during harvest.

Now, if the rain over the winter is equal to 800mm, what is the regression coefficient to **h.rain**?

● -0.2195209 ✓

○ 0.2195209

○ -285.8848

○ 285.8848

EXPLANATION

You can run the following code to calculate the regression coefficient:

```
coef(my.analysis)[4]+800*coef(my.analysis)[6]
```

In this case the model will predict DECREASING prices with increasing rain during harvest.

Question 4

(1/1 point)

Predict the price of a Bordeaux vintage in 1985 with the `predict()` function, if the temperature and precipitation in the harvest period and the preceding winter have values equal to the averages from the **wine.df** dataset, disregarding variation of the estimator.

What is the value of price predicted using the model?

○ 8.241431

● 8.341431 ✓

○ 8.441431

○ 8.541431

EXPLANATION

You can run the following code to perform the task:

```
new.data<-data.frame(year=1985,
                     temp=mean(wine.df$temp),
                     h.rain=mean(wine.df$h.rain),
                     w.rain=mean(wine.df$w.rain))
predict(my.analysis,newdata=new.data)
```

Question 5

(1/1 point)

Now regress `log(price)` (rather than price itself) on **year**, **temp**, **h.rain** and **w.rain** with the `lm()` function. Again, allow for an interaction between **h.rain** and **w.rain**.

What is Adjusted R-squared of the model?

○ 0.6891

○ 0.7574

○ 0.8001

● 0.8251 ✓

EXPLANATION

You can create the model and view the summary using the following code:

```
my.analysis<-lm(log(price)~year+temp+h.rain+w.rain+h.rain:w.rain,data=wine.df)
summary(my.analysis)
```

Question 6

(1/1 point)

Reduce the model with `drop1()`, removing statistically insignificant explanatory variables from the model (if any), testing at the 5% level.

What can you say about the significance of the variables?

○ temp is not significant.

○ year is not significant.

● Interaction between h.rain and w.rain is not significant. ✓

○ All variables are significant, the model cannot be reduced.

EXPLANATION

You can run the following code to perform the task:

```
drop1(my.analysis,test="F")
```

In this case, the p-value for the interaction terms is above 5%, rending the effect insignificant. We can therefore remove the interaction from the model, which we do with the update() function, after which we apply drop1() again on the reduced model:

```
my.analysis<-update(my.analysis,~.-h.rain:w.rain)
drop1(my.analysis,test="F")
```

Now, all factors are significant, and the model cannot be reduced further.

Question 7

(1/1 point)

What is Adjusted R-squared of the updated model?

○ 0.6891

○ 0.7574

● 0.8001 ✓

○ 0.8251

EXPLANATION

You can create a new model and exlude the interaction between h.rain and w.rain, or update the model using the solution given in the previous question. The following code will create the model and show the summary.

```
my.analysis<-lm(log(price)~year+temp+h.rain+w.rain,data=wine.df)
summary(my.analysis)
```

In particular, note that here the coefficient to h.rain is negative, while the coefficient to w.rain is positive. Since h.rain and w.rain doesn't interact, the effect of h.rain doesn't depend on w.rain. In all three cases, the model will therefore predict DECREASING prices with increasing rain during harvest, because the coefficient to h.rain is negative.

Question 8

(1/1 point)

Predict the price of a Bordeaux vintage in 1985 with the `predict()` function, if the temperature and precipitation in the harvest period and the preceding winter have values equal to the averages from the **wine.df** dataset, disregarding variation of the estimator.

What is the value of price predicted using the model?

● 15.10931 ✓

○ 15.11931

○ 15.12931

○ 15.13931

EXPLANATION

Since we have modeled log-transofrmed data, we must transform the result with the exponential.

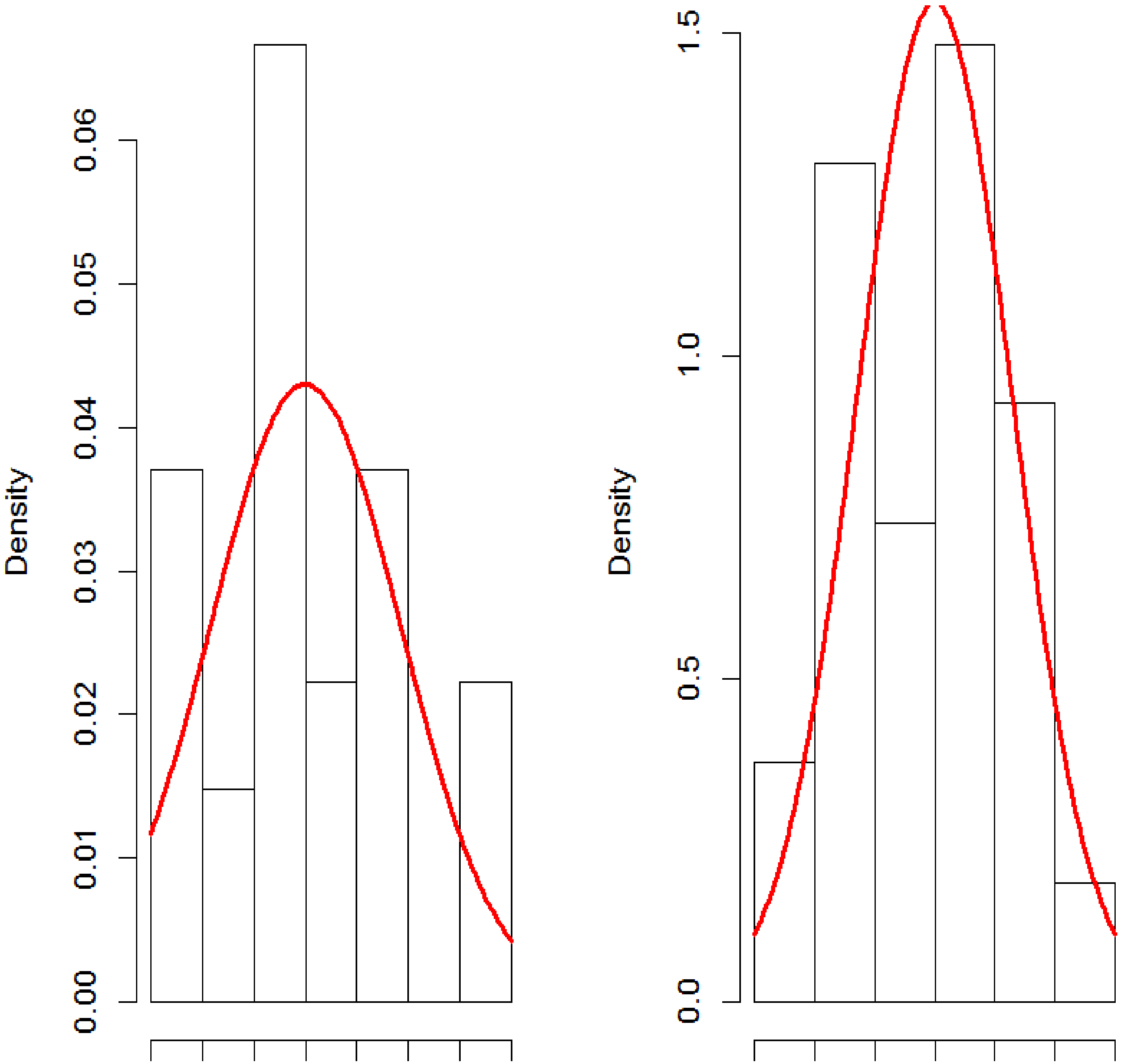
You can run the following code to perform the task:

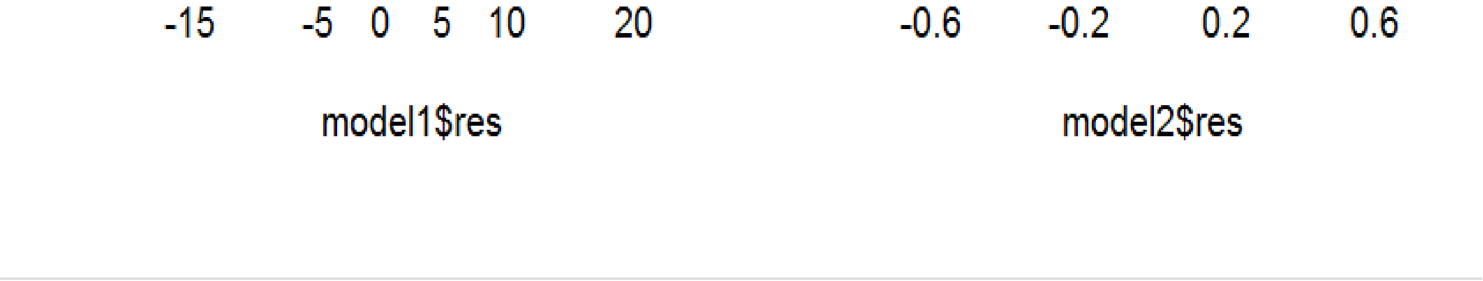
```
new.data<-data.frame(year=1985,
                     temp=mean(wine.df$temp),
                     h.rain=mean(wine.df$h.rain),
                     w.rain=mean(wine.df$w.rain))
exp(predict(my.analysis,newdata=new.data))
```

Assuming we have assigned the model in **Question 1** as **model1** and the model in **Question 7** as **model2**, we can compare the histograms of the residuals by using the following command:

```
par(mfrow=c(1,2))
g<-function(x){dnorm(x,sd=sd(model1$res))}
hist(model1$res,probability=TRUE)
curve(g,col="red",lwd=3,add=TRUE)
hist(model2$res,probability=TRUE)
g<-function(x){dnorm(x,sd=sd(model2$res))}
curve(g,col="red",lwd=3,add=TRUE)
```

Residuals from model 2 resemble normality to a much higher degree than those from model 1. Thus, we conclude that model 2 is more reliable.


Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Density
0.0
0.5
1.0
1.5


-15
-5
0
5
10
20
model1$res
-0.6
-0.2
0.2
0.6
model2$res