Pattern regonizition lab

3.6.4

Interaction terms

T

P

F

相关性

3.6.5

非线性模型

二次

X方拟合更好

四张图a> par(mfrow=c(2,2)) > plot(lm.fit2)

poly(a, n)

a的n价 多 项 式

3.6.6

Lab: Linear Regression

Greetings!

Concept:

Summary

plot

anova

Library: MASS ISLR

In this lab, we have used the library MASS (which is a very large collection of data sets and functions) and ISLR (which includes the data sets associated with this book).

Data: Boston & Carseats

The data sets we used in the lab are Boston and Carseats.

The Carseats data is part of the ISLR library. We will attempt to predict Sales (child car seat sales) in 400 locations based on a number of predictors.

The standard linear regression model provides interpretable results and works quite well on many real-world problems. However, it makes several highly restrictive assumptions that are often violated in practice. Two of the most important assumptions state that the relationship between the predictors and response are additive and linear.

So we use R to find out the relationship vividly.

3.6.4 Interaction Terms

An interaction term, which is constructed by computing the product of X1 and X2. As a result, we can easily find out whether the effect of changes in a predictor on the response is independent of the values of the other predictors, which is defined as addictive assumption.

The Boston data is part of the MASS library. We will talk about medv (median house value), age(average age of houses), and Istat (percent of households with low socioeconomic status) and their interaction terms.

The syntax lstat:black tells R to include an interaction term between lstat and black.

The syntax lstat\*age simultaneously includes lstat, age, and the interaction term lstat×age as predictors; it is a shorthand for lstat+age+lstat:age.

3.6.5 Non-linear Transformations of the Predictors

The lm() function can also accommodate non-linear transformations of the predictors.

We could see the result is better.

3.6.6 Qualitative Predictors

The Carseats data includes qualitative predictors such as Shelveloc, an indicator of the quality of the shelving location—that is, the space within a store in which the car seat is displayed—at each location. The predictor Shelveloc takes on three possible values, Bad, Medium, and Good.

3.6.7 Writing Functions

Liking other coding languages, we could easily create our own function in R.

As we have seen, R comes with many useful functions, and still more functions are available by way of R libraries. However, we will often be interested in performing an operation for which no function is available. In this setting, we may want to write our own function. For instance, below we provide a simple function that reads in the ISLR and MASS libraries?????, called LoadLibraries(). Before we have created the function, R returns an error if we try to call it.

We now create the function. Note that the + symbols are printed by R and should not be typed in. The { symbol informs R that multiple commands are about to be input. Hitting Enter after typing { will cause R to print the + symbol. We can then input as many commands as we wish, hitting Enter after each one. Finally the } symbol informs R that no further commands will be entered. page133image329346880

Now if we type in LoadLibraries, R will tell us what is in the function.

> LoadLibraries  
function (){  
library(ISLR)  
library(MASS)  
print("The libraries have been loaded.") }

If we call the function, the libraries are loaded in and the print statement is output.

> LoadLibraries()  
[1] "The libraries have been loaded ."

s

对于glm模型summary()输出的汇总结果，如何解读是非常重要的，它直接影响得出的结论。

例如下面这样一个输出结果，该如何理解呢？

Call:

glm(formula = bl ~ I, family = gaussian,data = anaData)

Deviance Residuals:

   Min       1Q   Median      3Q      Max

-62.364 -14.278   -0.462   14.293  51.475

Coefficients:

           Estimate Std. Error t value Pr(>|t|)

(Intercept)   67.067     5.766  11.632  < 2e-16 \*\*\*

I             10.368      3.647  2.843  0.00521 \*\*

根据课程所学并查阅有关参考书箱，理解如下：

一、调用

Call:

glm(formula = bl ~ I, family = gaussian,data = anaData)

当创建模型时，以上代码表明glm是如何被调用。

二、残差统计量

Deviance Residuals:

   Min       1Q   Median   3Q      Max

-62.364  -14.278   -0.462  14.293   51.475

残差第一四分位数（1Q）和第三分位数（Q3）有大约相同的幅度，意味着有较对称的钟形分布。

三、系数

Coefficients:

           Estimate    Std. Error   t value   Pr(>|t|)

(Intercept)     67.067      5.766    11.632  < 2e-16 \*\*\*

I             10.368      3.647   2.843    0.00521 \*\*

标记为Estimate的列包含由普通最小二乘法计算出来的估计回归系数。

标记为Std. Error的列是估计的回归系数的标准误差。

从理论上说，如果一个变量的系数是0，那么该变量是无意义的，它对模型毫无贡献。然而，这里显示的系数只是估计，它们不会正好为0。因此，我们不禁会问：从统计的角度而言，真正的系数为0的可能性有多大？这是t统计量和P值的目的，在汇总中被标记为t value和Pr(>|t|)。

P值估计系数不显著的可能性，有较大P值的变量是可以从模型中移除的候选变量。

<https://www.jianshu.com/p/464bd93c3410>

**线性 Residuals vs Fitted 残差图与拟合图 若应变量与自变量线性相关，那么残差值与拟合值就没有任何系统关联，在图中可以看到有一个明显的曲线关系，这说明可能要对回归模型添加一个二次项**

**正态性 Normal Q-Q 正态QQ图 用于检验残差正态性 如果满足正态假设，那么图上的点就应该均匀的落在呈45°角的直线上（图中虚线），不然就违反了正态性的假设**

**同方差性 Scale-Location 位置尺度图 如果满足同方差性，那么图中水平线周围的点应该随机分布**

**Residuals VS Leverage 残差与杠杆图 从图形中可以鉴别出离群点，高杠杆值点和强影响点 离群点：表明拟合回归模型对其预测效果不佳 高杠杆值点：是一个异常的预测变量值的组合 强影响点：表明它对模型参数的估计产生的影响过大，非常不成比例**

1. Concept

Summary

Plot 4张图

Anova

Library data

1. Lab
2. **3.6.4 interaction term**

有相关性

Problem1 和之前的模型进行对比

1. **3.6.5 non-linear**

效果好

有图

二维

五维

Log

Anova函数

**3) 3.6.6**

Sheloc有三个非数字的值 定性分析

Problem2

R has created a ShelveLocGood dummy variable that takes on a value of 1 if the shelving location is good, and 0 otherwise. It has also created a ShelveLocMedium dummy variable that equals 1 if the shelving location is medium, and 0 otherwise. A bad shelving location corresponds to a zero for each of the two dummy variables. The fact that the coefficient forShelveLocGood in the regression output is positive indicates that a good shelving location is associated with high sales (relative to a bad location). And ShelveLocMedium has a smaller positive coefficient, indicating that a medium shelving location leads to higher sales than a bad shelving location but lower sales than a good shelving location.

4）3.6.7 有无括号

***Checking residuals is a way to discover new insights in your model and data!***

* ***So, what does having patterns in residuals mean to your research?*** It’s not just a go-or-stop sign. It tells you about your model and data. Your current model might not be the best way to understand your data if there’s so much good stuff left in the data.
* In that case, you may want to go back to your theory and hypotheses. Is it really a linear relationship between the predictors and the outcome? You may want to include a quadratic term, for example. A log transformation may better represent the phenomena that you’d like to model. Or, is there any important variable that you left out from your model? Other variables you didn’t include (e.g., age or gender) may play an important role in your model and data. Or, maybe, your data were systematically biased when collecting data. You may want to redesign data collection methods.