We've made a few improvements to the forums. You can read read more on the blog.

Forums / Data Analysis Assignment 1

79/85 Assignmet. What do you think?

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

assignment × example × opinion × non-linear × quadratic × + Add Tag

Jordi Esteve · a month ago

I got 79 over 85 points. Actually, one point above my own evaluation.

i think it was important to notice that the relationship between the interest rate and the FICO was not constant. In particular, for lower FICO levels the relationship was quite strong, but it wasn't that much for high FICO scores.

One of the Assignments I reviews took that into account by splitting the data in two different sets, with an arbitrary FICO score threshold. My approach I think it is more flexlible. I introduced a quadratic FICO score term.

I understand that not taking that into account may lead to a bias in the coefficients, but I'm not sure how important that effect would be.

Opinions?

https://dl.dropbox.com/u/26651556/Data%20Assignment%201.pdf https://dl.dropbox.com/u/26651556/Figure.pdf https://dl.dropbox.com/u/26651556/Figure%20description.pdf

6

Prokhorov George · a month ago

Figures are absent, only caption descriptive. So I can't understand, why you have involved quadratic power. Why not cubic, fourth power, exponential ?

0

Anonymous · 13 days ago

IF you highlight the figure address and do right click, you can go to the figure

0

Anonymous · 7 hours ago

https://dl.dropbox.com/u/26651556/Data%20Assignment%201.pdf
https://dl.dropbox.com/u/26651556/Figure.pdf
https://dl.dropbox.com/u/26651556/Figure%20description.pdf

0

Prokhorov George · 10 minutes ago

Code, please.

0

+ Add New Comment

Jordi Esteve · a month ago

Thanks Prokhorov.

Something went wrong with the figure link. Here it is again: https://dl.dropbox.com/u/26651556/Figure.pdf

Obviously, you could introduce more regressors (cubic, fourht power, etc). However, the more you introduce, the more degrees of freedom you loose and the more likely it is that you end up overfitting. Moreover, once you see that there are no sigificant patterns left in the residuals, there is no much reason to introduce yet another degree.

In the case of exponential function, it acutally works the other way round. Maybe the logaritmic could also work. But in those cases you have to guess the relationship in advance, while the quadratic is supposed to be more flexible.

0

+ Add New Comment

Katie Cornog · a month ago

I also used a quadratic term as well as a linear term of FICO score. The important thing was to capture the shape of the relationship between FICO score and interest rate. The interest rate flattened out for higher FICO scores and it was possible to get a shallow, upward facing parabola to fit well. Until I tried this, it was very hard to tell which other variables had a true effect on the interest rate.

I thought about using two lines with a threshold, but I was unsure of how to get confidence intervals and the p-values in that case. My reason for quadratic rather than any higher order was that quadratic appeared to fit the scatter plot well by eye.

I also read a very useful reference that someone had posted on one of the boards:

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R. Cambridge: Cambridge University Press, 2008, sections 4.3.2 – 4.4.1. (A draft of this book, URL: http://www.ualberta.ca/~baayen/publications/baayenCUPstats.pdf)

▲ 2 ▼

Prokhorov George · a month ago

I still don't understand the aim of the assignment.

The first one can be a demonstration of your skill , gained in this course. (R-coding, modelling, plotting, statistics etc.) . In this case it is no matter you results . But is very important to demonstrate your ability to "label axes in plain language and large enough to read".

The second aim can be an obtaining a good result. What is the measure of this good result?

You have a set of different models (linear, multivariable, plynomial etc.). You choose\probe\test your own inuitive variant. As a result you propose your own model of Interest Rate.

Is your model close to real data?

The closer model, the better result.

So , if your model is very accurate , you can get a high mark , despite that " your axis labels are not large enough to read"

And Jeff may declare : A-mark is equal to X accuracy of a model , B-mark is equal to Y accuracy , C-mark is equal to Z accuracy.

```
So, what is the  difference beteween your model and the real data?   In tota
l .
```

May be least squares.

▲ 0 ▼

Jordi Esteve · a month ago

Hi Katie,

Thanks for the reference. It looks interesting.

Another non-linear relation that I exploited was the number of open accounts. Apparently, it is bad to many or too few open accounts (hump shaped relationship). Particularly, my analysis reported that the minimum interest rate is charged to those

borrowers with around 14 open lines of credit, while the penalization for those that only has two lines (the minimum in the sample) is more or less the same than of those with 25 lines.

Actually, I latter found this link in the forum (http://www.lendacademy.com/how-lending-club-and-prosper-set-interest-rates/) that says that the ideal number of open accounts is 6 to 21.

0

+ Add New Comment

 László Kupcsik · a month ago

While your solution looks convincing, I don't think it is fair to say that someone (like me) who used linear correlation between FICO score and interest rate did a bad job. First, the assigment was published before linear models or regression was even itroduced. And then, quadratic/nonlinear models were not discussed in the lectures during the assignment *at all*. I think when grading the assignments, we should only consider the material that was shown in the lecture material, and how well the student applied that knowledge. Of course, if you do more, it can be acknowledged, but it shouldn't be required.

3

Jordi Esteve · a month ago

Hi Laszio. I'm not sure to understand you. I didn't say that any other analysis was bad or that mine was the correct answer. I don't think there's anything such the good answer here.

The only purpose of this thread is to try to learn from others.

Maybe the misunderstandign comes from the fact that I said that "not taking that into account may lead to a bias in the coefficients". But this is not an evaluation, it is an statistic concept.

Besides, I have to acknowledge that I had some previous background on econometrics. Without this, I don't think I would had tried something like non-lineartiy.

1

+ Add New Comment

Borys Blank · a month ago

I don't think quadratic terms are justified here. An arbitrary FICO cutoff is what lenders actually do. Anyone who ever applied for a mortgage must have heard "we don't care about your FICO as long as it's above 750 (or 760, 740)". What's inetersting is why we are still seeing the association of interest with FICO for high FICO. This question is really difficult to answer without exploring causal links. Do people with high Fico request smaller amounts for shorter terms more frequently than people with low Fico? Is that causing their Fico to go up?

▲ 4 ▼

Markus Wesoly · 25 days ago

interesting to hear what this is going on in practice. The quadratic term is odd because it would eventually (at least outside the range of the fitted data) make the interest rate rise again when the FICO increases. But FICO has an upper limit not far off the range of data and as they say, all models are wrong, some are useful.

▲ 1 ▼

+ Add New Comment

Amit Dingare · 25 days ago

A quadratic term makes the model too complex and it is perhaps over-fitting the sample data.

▲ 2 ▼

+ Add New Comment

New post

Bold	*Italic*	Bullets	Numbers	Link	Image	Math		<HTML>

☐ Make this post anonymous to other students

☑ Subscribe to this thread at the same time

Add post