PDF-XChange Click to buy NOW! www.docu-track.com

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

Assignment 1 example: multiple linear regression, and a clear confounder

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

assignment × Assignment1 × example × solution × + Add Tag

Raja Doake · a month ago

I read through Thia Kai Xin's fascinating thread with her assignment, and thought I would share mine as well, since I approached the problem a little differently and only used techniques covered in the lectures.

Here's my Analysis Text. My peer review score was 82/85, but reading various threads after the hard deadline, there are definitely some things I could have done differently. I hit the word limit on the nose, though, so it would have been tough to fit anything else in!

My code is separated in to R scripts by category: parsing/transforms for the assignment set, exploratory plots, exploratory analysis, parsing/transforms for the complete Lending Club dataset (more on that in a minute), and then final analysis/plots. Here's my code.

A short overview of my process:

Exploratory Analysis and Transformations:

I tried a few different transforms, including logs on anything related to money, but ultimately most of those distributions didn't become any more normal after doing so, so I only kept the log transform of monthly income. I decided to remove the "n/a" employment length observations (77) so that I could treat employment length as either factor or numeric during exploratory modeling. Since employment length didn't end up being significant and I was pressed for time, I didn't redo my final model with those observations added back in. I also removed the two observations with undefined (NA) debt to income ratios.

In hindsight, I should have capped FICO the way Anne Paulson (and others) did, to resolve the nonlinearity at high FICOs as the interest rate approaches the floor of ~5%.

Inferential Analysis:

I initially tried SVD to identify the most significant contributor, but since I didn't fully understand what was going on under the hood, I decided to stick to multiple linear regression. Taking FICO

PDF-XChange Click to buy NOW! www.docu-track.com

score as the most significant variable, I wrote a FOR loop that made a model with FICO score and each other variable, then stored the R-squared for each model in a matrix. In hindsight, I should have used adjusted R-squared here, but it didn't end up affecting my results this time.

Loan Length ended up adding the most to the model, so I took that as the second variable and then used a similar FOR loop to output the R-squared of each three variable model with FICO and loan length. This time, the maximum R-squared increase came from the amount requested, so I added that to the model and repeated the procedure with a fourth variable. None of the fourth variables added much to the model -- even without going to adjusted R-squared -- so I left it at three.

I wasn't surprised to see that loan length was important, since I picked up on that in my exploratory analysis (see Figure 1-1 below).

Note: After cleaning up my code to upload here, I ran it again and noticed that amount funded -- rather than amount requested -- comes out of the second iterative linear regression as the 3rd variable, albeit only barely over amount requested. I'm certain that didn't happen when I did this the first time, and I expect the difference is due to the fact that I modified the data cleanup slightly to remove some extraneous experimentation (like testing employment length as both factor and numeric). However, I wouldn't have included amount funded in my model in any case, since it's an output of the Lending Club's process, rather than an input. Also, the R-squared values reported in my analysis are the same in either case, so it doesn't affect the report.

Model Issues:

A plot of the residuals (Figure 1-2) shows a positive bias at high interest rates. Other people attempted to address this by incorporating quadratic terms into their models, but I took it as a sign that there was a confounder outside our dataset. I was pretty confident about this since none of the models I plotted (and I plotted all of them) had a significantly different residual pattern.

So, I went looking for confounders. A check of the Lending Club site revealed their current interest rate policy, which Thia Kai Xin found beforehand and used as the basis for her analysis. I kicked myself when I found the "how we set interest rates" page, but was at least pleased to see my model somewhat confirmed -- the two risk adjustment factors that Lending Club uses after their proprietary model assigns a risk class to a borrower are loan length and amount requested, with loan length being the larger of the two. Those risk adjustments aren't linear, though, so incorporating them directly into the model rather than approximating them with transforms would be handy -- although we don't know how they may have changed over time.

Lending Club also has their full loans dataset available for download. Our 2,500 loan sample is pulled from this larger dataset. I suspected that the issue date of the loan was likely to be our confounding variable, since the interest rates charged by Lending Club were very likely to have changed over time. So I downloaded the full loans dataset (~112,000 loans spanning 2007-2013) and did some exploratory analysis focused on loan issue date. The result of that can be seen in Figure 1-3 below: the maximum interest rate for both loan terms increases steadily with time. I did do a complete.cases() on the issue date/interest rate columns, leaving me with ~108,000 loans for the figure.

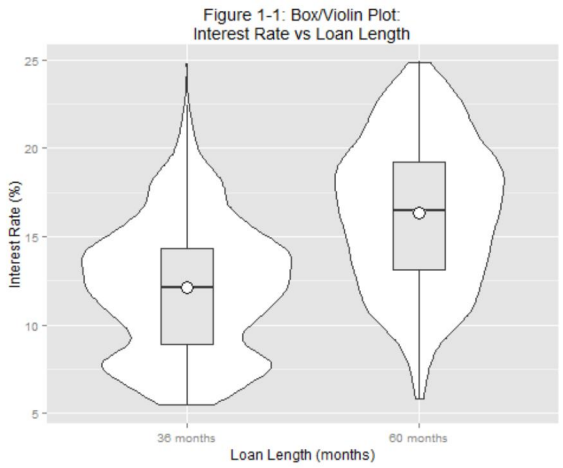
Since I was looking for a confounder that explained variance in the top interest rates, it seemed pretty likely that issue date was very significant. However, I didn't have time to repeat the whole analysis on the full dataset or figure out which issue dates were associated with our

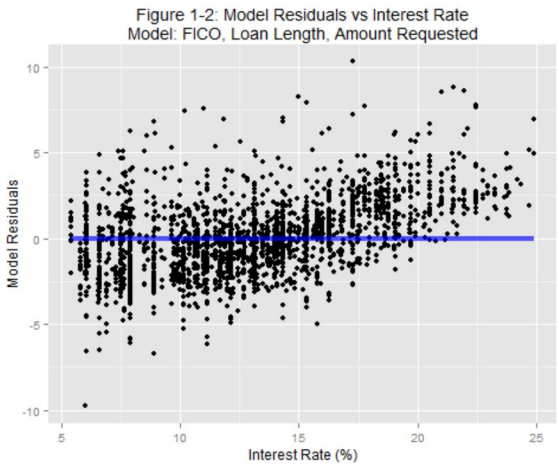
PDF-XChange Click to buy NOW! www.docu-track.com

2,500 loans, so I stated that issue date was probably the next place to look and left it at that.

Figure:

I picked up a copy of the R Graphics Cookbook before starting this class, and when I found that I couldn't get lines to display properly in the base graphics package using RStudio, I figured it would be a good time to learn ggplot2. I'm glad I did, because it's very easy to use and generates very nice plots. Being able to easily overlay the density distributions on a boxplot is pretty nice. I did have to use the gridExtra package to tile the plots; ggplot2 doesn't natively support putting multiple plots on a single graphics device.





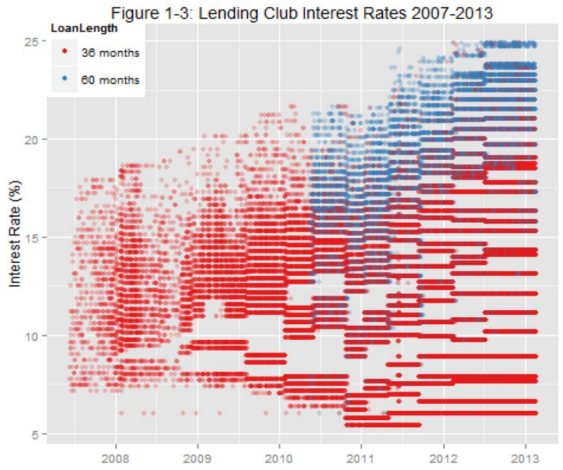


Figure 1-1: *A boxplot/violin plot of interest rates for loan terms of 36 and 60 months. The circles indicate the mean interest rate for each term, which overlap the median lines, suggesting that the interest rates are approximately normally distributed, although the density plot of the 36-month loans does bulge at roughly 7.5% interest. The 25th/75th percentiles and medians of interest rate are very noticeably higher for 60 month loans, suggesting that this is a significant factor in predicting interest rate.*

Figure 1-2: *A plot of model residuals. The residuals develop a clear positive bias as interest rate increases, suggesting an unmeasured confounding variable that influences the value of higher interest rates for at least some loans.*

Figure 1-3: *An exploratory scatterplot of the complete Lending Club funded loans dataset, showing the issue date and interest rate for 108,741 loans funded between 2007 and 2013, color-coded by loan term (36 or 60 months). The maximum interest rate for both 36 and 60 month loans increases with time, suggesting that the issue date is a significant factor in the interest rate of the loan if the rate is high.*

PDF-XChange Click to buy NOW! www.docu-track.com

18

Anonymous · a month ago

Good job, Raja! A quick question, if you finally found out the issue date is a confounder, will you include it into your model? if you include it into your model and it makes a big contribution to the variance, should we still call it a confounder or just say it is a covariate. Thanks!

0

Raja Doake · a month ago

If I had been doing this for real, rather than for an assignment, I would definitely have included loan issue date in the model. I think it would technically still be a confounder, though, since it's clearly correlated with interest rate and at least one of our dependent variables (loan length -- no 60 month loans were issued before mid-2010). It's quite possible that issue date is also correlated with some of the borrowers' traits, since the time span covers the financial crisis and I suspect that some common things happened to many borrowers' credit from 2008-2009 in particular.

If I'd been able to add issue date to the model, I definitely should have used adjusted R-squared (of course, I should have been using it already).

0

Raja Doake · a month ago

Since I can't edit the above, I'll post again. I want to clarify that I would do some additional exploratory modeling with issue date before deciding how to include it. I expect it would be significant enough to include, but I might consider making it a piecewise variable that only applies at higher interest rates, depending on what I found in the exploratory modeling.

0

Anonymous · a month ago

I got it. Thanks.

0

Soheila Dehghanzadeh · 20 days ago

Hi Friends, what is the difference between confounder and covariate? thanks.

0

PDF-XChange Click to buy NOW! www.docu-track.com

Anne Paulson COMMUNITY TA · 20 days ago

The thumbnail answer is: a covariate is one of the independent variables in your regression. A confounder is something that should have been a covariate, but wasn't.

0

Soheila Dehghanzadeh · 20 days ago

thanks Anne, could u please specify a reason that why we shoudn't consider a factor as covariate after we spotted it as a potential covariate and left it as cofounder? how do we include the cofounder in our model for prediction?

0

Soheila Dehghanzadeh · 20 days ago

well, i guess probably because we haven't it in dataset so we can't specify one of the available columns as confounder ...

0

Anne Paulson COMMUNITY TA · 20 days ago

Soheila, could you possibly use standard spelling? I find it distracting to have to translate. Thanks.

If you find the confounding variable, and you think it has an important effect (for whatever definition of important you are using) then you would put it in your model. I can't think of a reason you'd leave it out, unless you decided its effect was insignificant.

0

Anonymous · 12 days ago

This is the correct Assignment 2 using svm (support vector machine in R) :http://www.icephd.org/sites/default/files/IWAAL2012.pdf

0

Soheila Dehghanzadeh · 12 days ago

hi Raja, I added time to the model but residual pattern is still the same :-/ have you tried it with issue date? was it different for you?

0

PDF-XChange Click to buy NOW! www.docu-track.com

+ Add New Comment

 Anne Paulson COMMUNITY TA · a month ago

I like this analysis a lot. The resulting model is simple, but the analysis shows that the simple model was arrived at via a thorough process. Moreover, the explanation is crystal clear: we are led through the process, and can see why we arrive at the conclusion.

0

 Raja Doake · a month ago

Thanks. :)

It was important to me to keep the model simple, partly so it would be easy to explain, and partly because none of the more complex versions I tested had much effect on the model residuals at higher interest rates. This is actually why I'm not in favour of using a quadratic term in this model: sure, a quadratic on FICO score or some transform of it reduces the model's skew at the interest rate floor, but the interest rate floor isn't where the problem lies.

That's also how I got away without capping FICO score, although after reading your posts on the subject I would do that if I redid this work.

0

+ Add New Comment

Anonymous · a month ago

If you have to do all this work to get an 82... well there is something wrong with how these assignments are graded! Don't get me wrong: I got a decent grade for a couple of days of work. Still, as literary teachers do not request you to write valuable poetry to get an A, a data analysis assignment should have a clear objective and appropriate grades. All what is done above and beyond should be done only for its own sake. The grading criteria seemed to cope with this request, maybe they should just be loosened a little.

0

+ Add New Comment

 Diego F. Pereira-Perdomo · a month ago

PDF-XChange Click to buy NOW! www.docu-track.com

Congrats Raja!!!

Thumbs up!

0

+ Add New Comment

 Raja Doake · a month ago

I got curious about whether I could actually create a version of the model that incorporates issue date, so I took a quick look at the full loans dataset the other day and found that the ID numbers in our sample didn't directly correspond to the ones in the full set. It should still be possible to find our loans by matching the other variables, but that makes it a little more tricky and I'm busy getting started on Assignment 2!

0

Soheila Dehghanzadeh · 12 days ago

is there really a need to match full dataset to the sample dataset? can't we build model based on the full dataset? i build a linear model from the full dataset using lmX2 <- lm(bigLoans1$Interest.Rate ~ bigLoans1$ficomean + as.factor(bigLoans1$Loan.Length) + bigLoans1$Issued.Date)

plot(bigLoans1$Interest.Rate, lmX2$residuals, pch = 19, col = as.factor(bigLoans1$Loan.Length), xlab = "Interest rate", ylab = "model residual") abline(0, 0, col = "blue", lwd = 3)

and still the same residual pattern....

0

 Raja Doake · 12 days ago

Interesting. Is it any better at all?

0

A post was deleted

Soheila Dehghanzadeh · 12 days ago

nope ... even worse since the residuals tend to be either too negative or too positive ...

0

PDF-XChange Click to buy NOW! www.docu-track.com

Raja Doake · 12 days ago

Some level of transformation may be needed. For example, I think Anne was right to cap FICO score. Additionally, without some measure of the change in spread of loan rates, we don't really know if issue date affects loans with low to moderate interest rates. So you need a conditional variable where if interest rate is higher than ~17%, it's included, but otherwise it isn't.

Still, those initial results aren't promising at all. It might be necessary to repeat the looping exercise with the rest of the variables in bigLoans.

0

A post was deleted

Soheila Dehghanzadeh · 12 days ago

yes other variables should also be tested and the idea of adding conditional variable is also nice :) thanks :)

0

+ Add New Comment

Davi Souza Simon · 21 days ago

Raja,

Great work!

In order to correct for general interest rates' variation over time, what you could do is create two new variables (if you had the right data). The first one would be the base interest rate at the time (like a prime rate). The second one would be the interest over the base per each loan.

This way, your analysis would focus on the determinants of spread, instead of general interest rate.

But maybe that would not fully solve the problema, because spreads are also supposed to be variable over time, due to changes in applicants' perceived risk and general economic conditions.

What do you think about this possibility?

0

Raja Doake · 21 days ago

I think bringing in the base rate would be similar to capping FICO -- if you have the information, it's the right thing to do, but it may not help you explain substantial additional variance. The problem with the model is really with rates >15%, and we

PDF-XChange Click to buy NOW! www.docu-track.com

don't know why Lending Club's maximum rate has increased over time. That's why I suggested simply using the issue date as a proxy for it -- issue date is covariant with maximum rate.

0

+ Add New Comment

Anonymous · 21 days ago

I like your analysis. Great figures.

Where can I get bigLoansData.csv?

Thank you!

0

Raja Doake · 21 days ago

The first spreadsheet on this page is the full set of historical loan data:

https://www.lendingclub.com/info/download-data.action

0

+ Add New Comment

PDF-XChange Click to buy NOW! www.docu-track.com

Anonymous · 21 days ago

Good job. Will you be sharing the R code with the class?

0

Raja Doake · 21 days ago

My code is actually linked in my first post, but here it is again:

https://dl.dropbox.com/u/10694858/assignment1rdoake.R

0

+ Add New Comment

Wei Deng · 20 days ago

Excellent work, Raja!!

Is there a reason that you used log 10 transformation rather than the natural log?

If issue date were included, then probably it would go beyond multivariate linear regression - my two cents.

0

Raja Doake · 20 days ago

I used log10 for simplicity. If monthly income had ended up being in my model, the coefficient would have been much easier to interpret with a log10 transform: a person who makes $10,000 per month got an interest rate X% lower than a person who makes $1,000 per month. With a natural log transform interpreting the coefficient is a lot more fiddly.

For a case like this where we are explicitly reverse engineering how Lending Club sets interest rates, it's important to be able to easily explain what the model terms mean. Polynomial terms and other variable transformations make that more difficult unless you're describing a natural phenomenon where that relationship physically exists (e.g. kinetic energy varies with the square of velocity).

1

Wei Deng · 20 days ago

Great! I see. By the way did you look at interaction terms such as loan_length:amount_requested?

PDF-XChange Click to buy NOW! www.docu-track.com

PDF-XChange Click to buy NOW! www.docu-track.com

0

+ Add New Comment

 Raja Doake · 20 days ago

No, but that was just an oversight on my part! No rationale there.

1

 Wei Deng · 20 days ago

I see.

0

+ 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

PDF-XChange Click to buy NOW! www.docu-track.com

PDF-XChange Click to buy NOW! www.docu-track.com