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

My 72/85 assignment

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

No tags yet.+ Add Tag

 Jana sedivy · 25 days ago

Hi, Since Prof Leek and the TAs have not posted sample analysis but some of the other students have, I thought I would also share my report.

I got a 72/85 which I was pretty happy with. Other posters who posted great reports here: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=2698 and here: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=2637 used techniques that were not mentioned in the lectures because they had previous experience.

I was struggling with just the basics, so I only used what was in the lectures, leaning heavily on Prof Leek's example. I posted my report here in the spirit of sharing and learning. I know this is not a perfect assignment. Please be gentle :-)

https://www.dropbox.com/s/swd7ywulyv7ehjg/FinalAnalysis.pdf

17

 ANIMESH KUMAR · 25 days ago

this is very useful. Thanks Jana

1

+ Add New Comment

Markus Wesoly · 25 days ago

well done. You concentrated on the basics and did it. The coloring for the 36 vs. 60 months is quite striking. I noticed a minor thing: tradition has it (in most areas) that predictors (i.e. FICO

score) go on the x-axis (abcissa) while the obersvations (i.e. interest rate) go on the y-axis (ordinate). Also, I would have been interested to see what tools (x.y plot? other?) you used exactly during the model building by "extensive exploration of the data".

0

+ Add New Comment

Jakub Chromec · 25 days ago

Díky, Jano!

Thanks, Jana. Nice write-up, too. Could you please provide the code as well? I know there is this security issue but Kai Xin posted his markdown file already, so I think there should be no problem if you do the same. Or, if you don't feel comfortable with it, you can just list the functions - especially those you used to color your sexy plots! :)

Thanks in advance, Jakub

0

Anonymous · 24 days ago

I was interested in what code you used for the plots as well. Perhaps oddly, I was interested in how you made the legends. I added mine at the last minute and struggled to get the fonts big enough to read without overwhelming the graph.

0

+ Add New Comment

Daniel P Thomas · 24 days ago

Jana, I think this is a very good example of an analysis which followed the template we were given and, by using what was covered in the lectures, did not overwhelm or perhaps perplex students like myself who are new to data analysis through this course.

1

+ Add New Comment

Jana sedivy · 24 days ago

Thanks for the comments everyone! Ahoj Jakub! I see you spotted my Czech name! :-)

Markus, thanks for pointing out how the x and y-axis should be used. I seem to remember learning that at some point but I had forgotten it until you pointed it out!

My "extensive exploration" mostly consisted of doing scatter plots and coloring them by all the different variables. For the numeric variables, I generally split them into quartiles like so:

```
requestedQuartile <- as.integer(cut(loans$AmountRequested, quantile(loans$AmountReque
sted, probs=0:4/4), include.lowest=TRUE))
```

I also did some box plots but didn't find much interesting. I also did some linear regression fits to fit the overall FICO/Interest rate relationship and then also the different loan lengths. I didn't include those in the report because they didn't contribute directly to the overall narrative.

For those of you that were interested in the code for the "sexy plots", here you are.

colored by loan length

```
plot(loans$InterestRate, loans$lowFICO, col=(loans$LoanLength==36)+1, pch=19, xlab="I
nterest Rate (%)", ylab="FICO score")
legend(x="topright", c("36 months", "60 months"), col=c("red", "black"), pch=c(19,19)
)
```

colored by amount requested

```
plot(loans$InterestRate, loans$lowFICO, col=requestedQuartile, pch=19, xlab="Interest
 Rate (%)", ylab="FICO score")
legend(x="topright", c("Amount Requested 1st quartile", "Amount Requested 2nd quartil
e", "Amount Requested 3rd quartile", "Amount Requested 4th quartile"), col=seq(1:4),
pch=c(19,19))
```

without adjustment residuals

```
plot(loans$InterestRate, lmAll$residuals, col=requestedQuartile, pch=19, xlab="Intere
st Rate (%)", ylab="no adjustment residuals")
legend(x="bottomright", c("Amount Requested 1st quartile", "Amount Requested 2nd quar
tile", "Amount Requested 3rd quartile", "Amount Requested 4th quartile"), col=seq(1:4
), pch=c(19,19))
```

full model residuals

```
plot(loans$InterestRate, lmAmtLen2$residuals, col=requestedQuartile, pch=19, xlab="In
terest Rate (%)", ylab="full model residuals")
legend(x="bottomright", c("Amount Requested 1st quartile", "Amount Requested 2nd quar
tile", "Amount Requested 3rd quartile", "Amount Requested 4th quartile"), col=seq(1:4
), pch=c(19,19))
```

One thing that I was wondering about - for most of the sections my self-grade was pretty compatible with my peer's grade (in fact, my self-grade was often 1 point lower!) except for the category on confounders. I thought that I had done a pretty good job describing confounders and gave myself a 4. My peers however, gave me a 0. Obviously, I'm not understanding something! If anyone did a good job explaining the confounders and can tell me what I should have done, I would appreciate some tips for next time!

0

Jakub Chromec · 24 days ago

Thanks a lot! Btw, splitting the numerical variables into quartiles for plotting is great idea since it makes the distribution pretty easy to understand. Let's see if I can spice up my plots next time :)

And yes, I was pleased to see a Czech name although I guess (judging by the way you write your surname) that you don't live in Czech Republic.

0

Jana sedivy · 24 days ago

Thanks Jakub!
You are right. My parents are Czech and some of my siblings were born there - but I was born in Canada (like many Czech ex-pats, my family left in 1969). However, a bunch of us are going there in May for a family pilgrimage - which we are all looking forward to!

0

+ Add New Comment

Anonymous · 24 days ago

Nice job! I love the coloring by quartiles.

I too got a 0 for "confounders", just because I did not mention the word "confounder" in my write-up. Looks like many reviews were done by "Ctrl-F" and not by reading.

0

Jana sedivy · 24 days ago

I don't judge the reviewers to harshly - they're just people like us right? Doing the best they can. I hardly felt qualified to evaluate other people's work - I'm sure I'm not the only one. The important thing is that I can figure out what the "right" thing to do is - so that I can do better next time :-)

0

+ Add New Comment

 Priyanka Deshmukh · 24 days ago

wow, Jana that's really a gr8 write up. u were perfect buddy, n thanx for that lovely explanation of those nice figures. I had struggled a lot for that coloring part. n now I know where I was doing it wrong. specially the coloring of loan length. thanQ very much. and one more thing even I got 0 for the confounder thing though I mentioned the confounders, explained why they r the confounders but still got the "0" cant help it :-) but I am happy with my score (59/85) I am not ashamed of it because I am a total newbie to computers its languages and data. but still I am able to make sense of data, use statistical modals, use a programing language like R. thanx to prof. Jeff n prof. Peng n u all of course! THE FORAM FRIENDS . I feel supported.

1

 Jana sedivy · 24 days ago

Thanks Priyanka! I also greatly appreciate all the help and support on the forums!

1

+ Add New Comment

 Anne Paulson COMMUNITY TA · 24 days ago

Could you explain a little what you mean by "Furthermore, an SVD analysis of the numeric variables indicated that the highest contributor to the percent variance explained accounted for less than 30% of the variance with the second contributor contributing 18%"? What SVD analysis is this?

About the issue with the confounders: Poking around the net, I notice that there seems to be more than one definition of confounding variables. In experimental design, a confounding variable is something unintended that influences your experimental results: you thought you proved that A is correlated with B, but you forgot to control for C, which is actually causing your effect. But no one set up an experiment here; we're just analyzing some data. Sometimes in data analysis, we call some variable *that we don't know about and that might be having an effect on our data* a confounding variable. So for example, some people mentioned that the loan date would be a confounding variable, because the interest rate offered to a particular borrower would be higher if interest rates generally were higher. We don't know the loan dates for our loans, but we may think that the loan dates are explaining some of the variance.

In the lectures, Prof. Leek seemed to indicate that any variable you hadn't yet accounted for in your analysis would be considered a confounding variable, even though you had that variable available to you. That is, if you set up InterestRate ~ FICO, then Amount Requested would be considered, by this definition, to be a confounding variable. But as far as I can tell, that is not a usual usage of the terminology.

And this is what may have caused the disparity in scores for the rubric item about confounders. Some people thought they dealt with the issue by mentioning variables that they put in their analysis. Other people may have thought that those were never confounding variables in the first place.

▲ 1 ▼

+ Add New Comment

 Jana sedivy · 24 days ago

Hi Anne, thanks for picking up on the SVD thing.

I did an SVD analysis on the loans data and plotted the singular variable and the % variance explained.

```
svd1=svd(scale(loansNumeric))
plot(svd1$d, xlab="column", ylab="singular value", pch=19)
plot(svd1$d^2/sum(svd1$d^2), xlab="column", ylab="percent Variance explained", pch=19
)
```

The resulting graphs

So, I'm not sure if I'm interpreting this correctly but - looking at the rightmost graph, it looks to me like the biggest contributor is only contributing less than 30% of the variance. This is quite different from the examples given in the lectures where the biggest contributor usually contributed over 90%. I took this to mean that there wasn't any single dominating variable, and that most of the variables contributed some thing. Is this a correct interpretation? I'm not fully confident that I performed the svd operation correctly either... Would love your thoughts.

About the confounding variables - that makes sense. I was thinking that I was describing the confounding variables because I had included them in the data. I didn't realize that this was not the conventional usage. I'll know for next time. I like the example of the loan date being a confounding variable - Thanks!

▲ 1 ▼

 Jana sedivy · 24 days ago

Gah! I can't figure out how to share an image on this forum. Here is the dropbox link: https://www.dropbox.com/s/jboej5qnm6ak2he/svdImage.jpg

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

0

Jana sedivy · 24 days ago

I should also mention that I did the SVD analysis only on the numeric values and did not include any of the factor variables. I suppose I probably should have done an ANOVA to look at the factor variables - but I didn't think of it at the time...

0

Anne Paulson COMMUNITY TA · 24 days ago

I see. I'm glad to bring this issue up, because I think there is some confusion.

That's a misleading graph.

It tells you what percent of variance would be explained *if you did PCA (which begins with SVD) and transformed all your variables into new variables*. It says nothing about the actual variables you have, which you are not transforming in the rest of the analysis. The percent variance explained for the actual variables you actually have will be different. I don't know whether you included FICO in that graph, but FICO itself explains about 50% of the variance.

0

Jana sedivy · 24 days ago

Hmm. OK. Still not sure I understand. I'm finding this a very challenging concept.

I was trying to follow what was done in slide 17 of the Dimension Reduction lecture (week 3) where Prof Leek used the face image to demonstrate dimension reduction.

But sounds like what you are saying is that the SVD function does a transformation on those variables - so you can't use it to go back and say "Aha! This is the most important variable!" ? Is that right?

And yet, in slides 10-12 in the Clustering Example lecture (Week 4 - with the Samsung Data) it seems like that's exactly what is being done. We find the "maximum contributor" of the right singular vector and then use that information for the clustering attempt.

I'm obviously not understanding something. I THOUGHT that you could use SVD analysis to give you some insight into which variables were the biggest contributors to the data variance, and then use that information to hone your analysis to figure out exactly how they are contributing to the outcome you are interested in. Is that not correct?

You said that FICO explains about 50% of the variance - how would you go about figuring that out?

Thanks again for your help Anne!

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

1

 Anne Paulson COMMUNITY TA · 24 days ago

You know that FICO explains about 50% of the variance, because when you do a linear regression with just FICO and nothing else, it explains about 50% of the variance.

I'll go check out the slides you mention from Week 4 and get back to you with more comments.

0

A post was deleted

 Anne Paulson COMMUNITY TA · 24 days ago

Let's call the variables in your data set base variables, and the variables that you would get if you did PCA the PCA variables. The lecture you mentioned, the Clustering Example of Week 4, showed, first, how to find a PCA variable that might distinguish two different values of a base variable. In the example, we found a PCA variable, I think the second principal component, that distinguishes walking up from walking on the flat.

Then, the lecture showed how to figure out which base variables contribute to that PCA variable, so we figured out which base variables contribute to the second principal component. Then we included that PCA variable in our clustering and got better results.

This, by the way, is a roundabout way of solving the problem. Rather than trying to go back and forth between PCA variables and base variables, if we wanted to go the PCA route, we would be more likely to transform our variable set to the PCA variables, and thereafter, just work with the PCA variables and forget all about the base variables.

0

 Anne Paulson COMMUNITY TA · 24 days ago

It's not a dumb question.

Adjusted R-squared: 0.5028

0

 Jana sedivy · 24 days ago

OK. I think I'm sort of understanding this! And I can see now where I missed the moving between PCA variables and the base variables. So, if a base variable is contributing to a PCA variable, that doesn't NECESSARILY mean that it is contributing to the overall data variance right?

I'm still not sure that I understand your last paragraph though. Wouldn't we WANT to go back to the base variable to see what is causing the effect? Or does it not matter because you are going to use the PCA variable (not the base variable) in the clustering algorithm anyways - and the clustering is what you care about (in this case).

Is there any kind of analysis that DOES let you see the contributions of the various variables to the data? Seems like that's what you want most of the time. But maybe that's not possible - and it's why these kinds of problems are tough to analyze.

P.S. Adjusted R-squared - got it. THANKS!

0

 Anne Paulson COMMUNITY TA · 24 days ago

Wouldn't we WANT to go back to the base variable to see what is causing the effect? Or does it not matter because you are going to use the PCA variable (not the base variable) in the clustering algorithm anyways - and the clustering is what you care about (in this case).

Right-- we care about the answer. We don't particularly care how we got it, in this case.

0

 Jana sedivy · 24 days ago

OK got it. Thank you SOOO much. This was really helpful!

0

Anonymous · 10 days ago

Hi Anne, Jana, thanks for illustrating this topic.I have 2 questions regarding your discussions: 1-how to get the percentage of variance explained by actual values since you said that "when you do a linear regression with just FICO and nothing else, it explains about 50% of the variance." I did an anova over this linear model and there is no result in anova outcome to show the percentage of variance:

anova(lm(a$Interest.Rate ~ a$ficoMean))

Analysis of Variance Table

Response: a$Interest.Rate

Df Sum Sq Mean Sq F value Pr(>F)

a$ficoMean 1 21937 21937 2527

<2e-16 *

Residuals 2498 21689 9

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

2-if i want to add those columns of PCA variables to my actual values and then try to predict output based on them[those PCA variables that explain the majority of variance] how can we do this?

0

Brian C. Holt · 7 days ago

Anne, this is news to me, which isn't to say that it's wrong; it's literally new to me and I'm wondering which lecture/ or slides elaborate on your post below. Jeff mentions that if you set up the PCA analysis correctly, it's basically doing an SVD. But what I can't wrap my head around is how to construct the new variables to use them in further models. For what it's worth, I understand the idea behind "base variables" and "composite variables", I just don't think there was any instruction on how to actually use the composite variables, let's say for example in a regression model. Thanks for your help, by the way; You've been a consistent, frequent, and helpful TA.

"Let's call the variables in your data set base variables, and the variables that you would get if you did PCA the PCA variables. The lecture you mentioned, the Clustering Example of Week 4, showed, first, how to find a PCA variable that might distinguish two different values of a base variable. In the example, we found a PCA variable, I think the second principal component, that distinguishes walking up from walking on the flat.

Then, the lecture showed how to figure out which base variables contribute to that PCA variable, so we figured out which base variables contribute to the second principal component. Then we included that PCA variable in our clustering and got better results.

This, by the way, is a roundabout way of solving the problem. Rather than trying to go back and forth between PCA variables and base variables, if we wanted to go the PCA route, we would be more likely to transform our variable set to the PCA variables, and thereafter, just work with the PCA variables and forget all about the base variables."

0

Anne Paulson COMMUNITY TA · 7 days ago

The lectures on PCA and SVD could have been clearer. It helps to know linear algebra. If you do, you might like this tutorial: http://www.snl.salk.edu/~shlens/pca.pdf It helped me when I was struggling with PCA and SVD.

Take your time with it :) You won't learn it in a minute.

0

Brian C. Holt · 7 days ago

Thanks Anne, for that resource.

I got about 90% of the way through this resource: http://www.ime.unicamp.br/~marianar/MI602/material%20extra/svd-regression-analysis.pdf

but got lost a bit around step # 62 (page 23) as they introduced a new concept with the var(y-hat). So close! But I'll keep pushing.

0

Brian C. Holt · 6 days ago

Thanks Anne, for that resource.

I got about 90% of the way through this resource: http://www.ime.unicamp.br/~marianar/MI602/material%20extra/svd-regression-analysis.pdf

but got lost a bit around step # 62 (page 23) as they introduced a new concept with the var(y-hat). So close! But I'll keep pushing.

0

Brian C. Holt · 6 days ago

Thanks Anne, for that resource.

I got about 90% of the way through this resource: http://www.ime.unicamp.br/~marianar/MI602/material%20extra/svd-regression-analysis.pdf

but got lost a bit around step # 62 (page 23) as they introduced a new concept with the var(y-hat). So close! But I'll keep pushing.

0

+ Add New Comment

Jana sedivy · 24 days ago

Whoah - the formatting on that previous post got completely screwed up!

let me try again:

```
Call:
lm(formula = loans$InterestRate ~ loans$lowFICO)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9898 -2.1363 -0.4565  1.8351 10.1935

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   72.838757   1.190658   61.17   <2e-16 ***
loans$lowFICO -0.084675   0.001685  -50.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.946 on 2496 degrees of freedom
Multiple R-squared: 0.503,  Adjusted R-squared: 0.5028
F-statistic:  2526 on 1 and 2496 DF,  p-value: < 2.2e-16
```

Is it the r-squared that gives the measure of variance?

▲ 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