Midterm 1, Part 2 - With Resources

S&DS 361

2023-02-21

For this part of the midterm, you can use R and R Studio, class notes, Google, ChatGPT, and the internet in general (e.g Stack Overflow). You may not use Ed Discussion or any other messaging/discussion forum. You may not communicate directly with any other human being. Of these resources, I'm guessing R and R Studio will be useful and the others will not be that useful.

Academic Salaries

This question involves the `academic.salaries.title.gender.rds` data set posted on Canvas under `Files/exams`. Please download that now. If you have downloaded all of the data sets in the `Files/data` folder, then you already have the file.

Build a linear regression model using the `academic.salaries.title.gender.rds` data to find the expected `salary` for a professor at a public university based on other information known about that professor. Only consider the columns `salary` (in dollars), `group.title` (Assistant Professor, Associate Professor, etc), `male` (1=male, 0=female), `score` (from US News and World Report rankings), `region`, and `state`, but feel free to transform any of those columns if you think it would help. Try a few models. For each model you try, handwrite the outcome, predictors, and Adjusted R^2 in a table like this:

```
   Outcome       Predictors AdjustedR2
 myoutcome          mypred1      0.500
 myoutcome mypred1, mypred2      0.550
```

Probably a good idea to take log transform of salary. Otherwise residuals will likely be far from normal.

Put that table here:

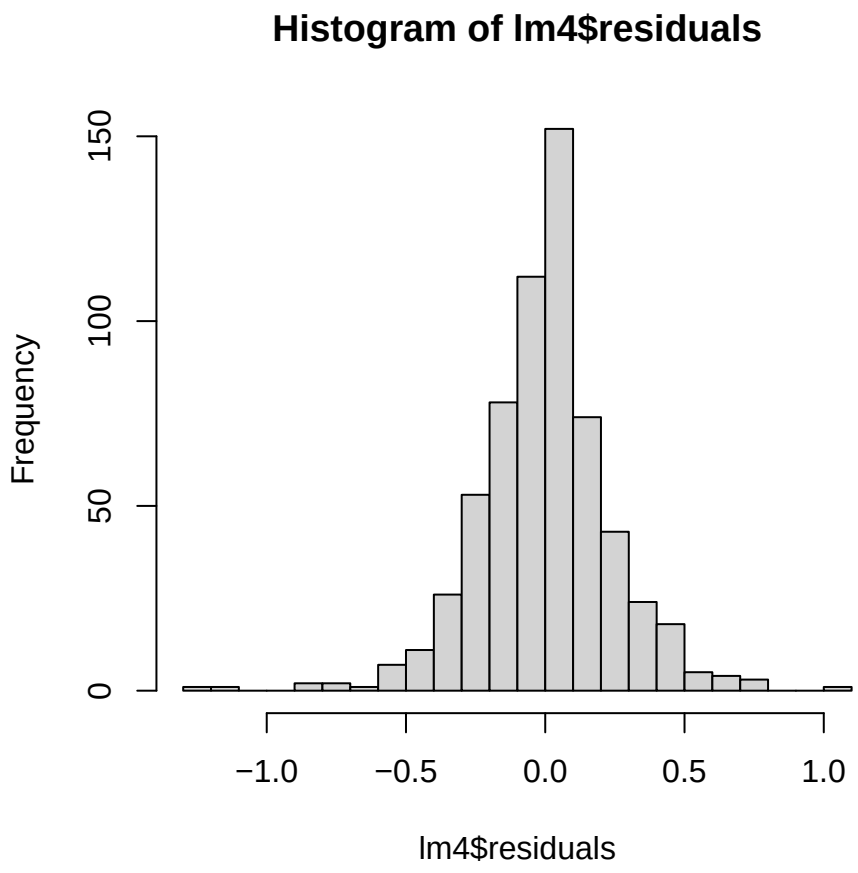
Outcome	Predictors	AdjustedR2
log(salary)	title	0.428
log(salary)	title, male	0.427
log(salary)	title, region	0.547
log(salary)	title, state	0.637
log(salary)	title, state, region	0.637
log(salary)	title, state, score	0.636

Which model would you choose as the best model? Give a short explanation (1-2ish sentences).

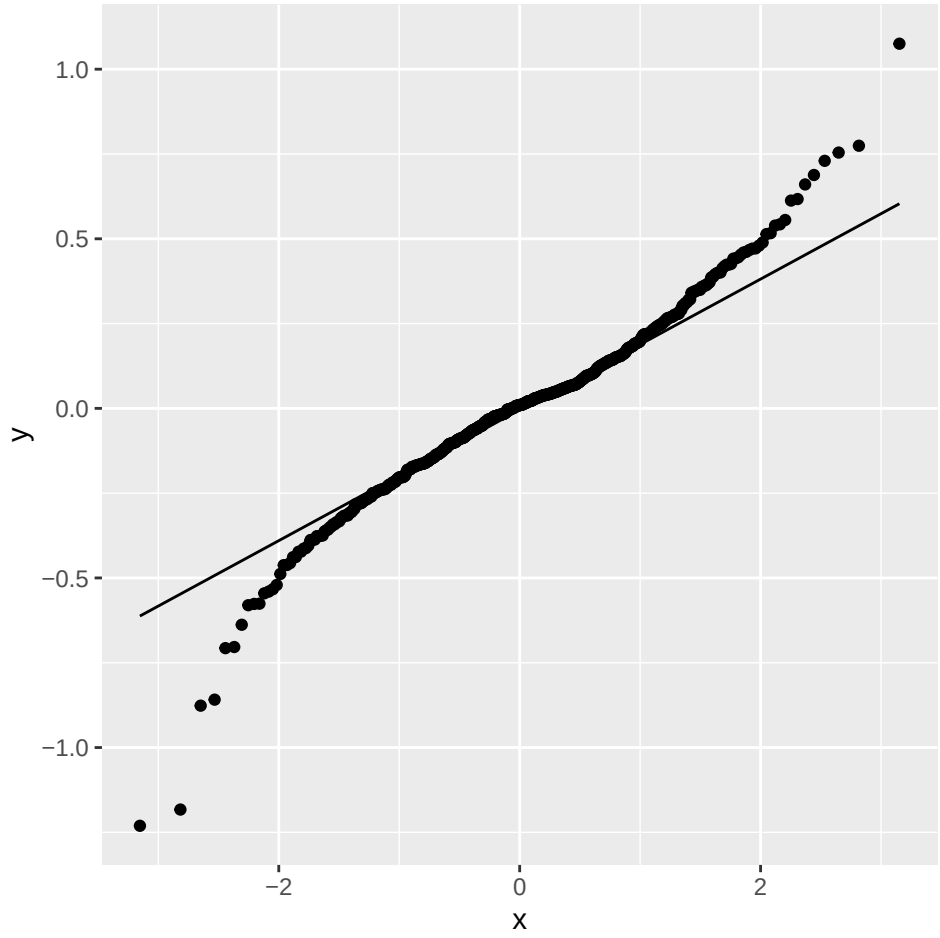
I would choose the model with `title` and `state`. The Adjusted R^2 is the highest, and those predictors are significant.

I'll check the residuals

```
hist(lm4$residuals, breaks=30)
```

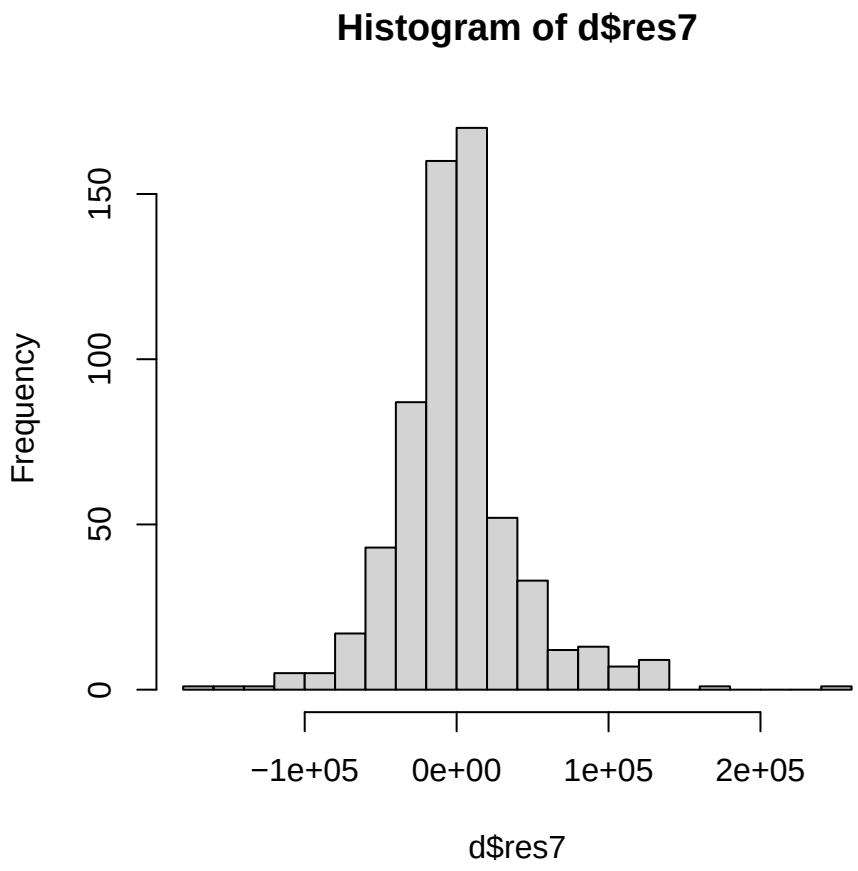



```
d$pred = predict(lm4, newdata=d)
d$res  = log(d$salary) - d$pred
ggplot(d, aes(sample=res))+
  geom_qq()+
  geom_qq_line()
```

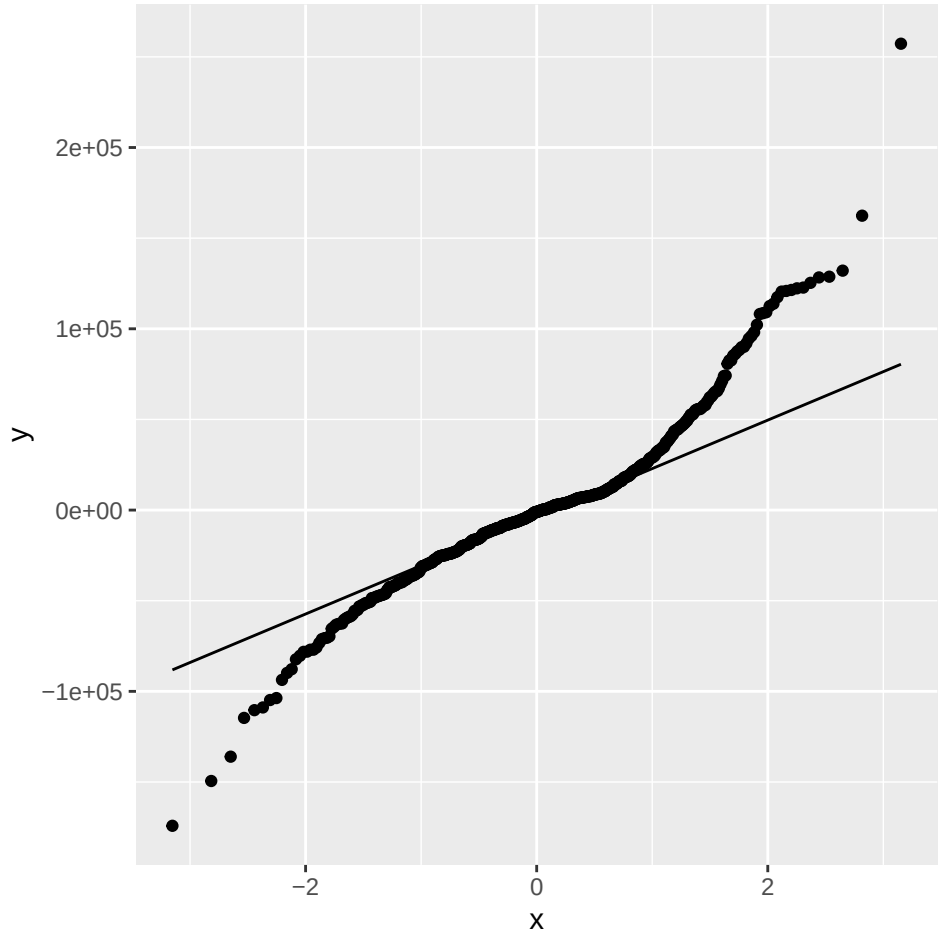


Hmm it looks not quite normal because of those tails, but it's pretty good. And I think it is better than using `salary` instead of `log(salary)`. Let's try `salary` to be sure.

```
lm7 = lm(salary ~ title + state, data=d);# summary(lm7)
d$pred7 = predict(lm7, newdata=d)
d$res7  = d$salary - d$pred7
hist(d$res7, breaks=30)
```




```
ggplot(d, aes(sample=res7))+
  geom_qq()+
  geom_qq_line()
```



That looks less symmetric, and slightly worse for the right tail, so yeah I'll go with `log(salary)`.