11.1 - “Using the crime data set uscrime.txt, build a regression model using:

1. Stepwise regression
2. Lasso
3. Elastic net”

Please find full code for all three models attached as Hmwk 5 11.1 Code. This tracks the code from Office Hours.

Below, I provide the relevant code from the attachment for each of the three required models. As the prompt just instructs us to build the model, I don’t think there is much commentary to add. I do describe a little bit of my thinking around picking an appropriate alpha/lambda for elastic net. And the attached code does take each model a bit further, printing a plot in some cases, etc., in case helpful for context.

1. Stepwise

Relevant code is as follows:

```
model_both <- lm(Crime~., data = uscrime)
step(model_both, scope = list(lower = formula(lm(Crime~1,data = uscrime)),upper=
formula(lm(Crime~.,data=uscrime))),direction = "both")
```

2. LASSO

Relevant code is as follows:

```
model_lasso <- cv.glmnet(x=as.matrix(uscrime[,-16]),
y=as.matrix(uscrime[,16]),alpha=1,nfolds=8,nlambda=20,type.measure = "mse",family =
"gaussian", standardize = TRUE)
```

I do note that this is really an elastic net model with an “alpha” of 1, which is the same as a LASSO model.

3. Elastic net

Relevant code is as follows:

```
model_elasticnet <- cv.glmnet(x=as.matrix(uscrime[,-16]),
y=as.matrix(uscrime[,16]),alpha=.5,nfolds=8,nlambda=20,type.measure = "mse",family =
"gaussian", standardize = TRUE)
```

You can see in the attached that I played around with using a for loop to try out different values for alpha. For the most part, the minimized error from the model increased as alpha approached

1. So, the more the elastic net model weighted the LASSO restraint, the worse the minimized error got. I think this is because LASSO removes variables while Ridge Regression merely regularizes, so the more the elastic net model weights LASSO, the worse your minimized error gets, basically.

However, as I thought this through, I considered that our goal here is, in substantial part, to prevent overfitting. Chasing minimized error seems, to me, to be contrary to that goal. That is, if we wanted minimized error, we could just leave the linear regression alone without applying LASSO or Ridge Regression. But then we wouldn't have counteracted overfitting at all.

Once I thought that through, I decided to abandon testing for the lowest minimized error. I left the for loop in the attached code, but here I only reproduce an elastic net model with an alpha of .5, reflecting a 50/50 balance between regularization per Ridge Regression and variable selection per LASSO.

In an applied context, I might weigh one or the other more heavily. E.g., if the data had more variables and being able to explain it was particularly important, I might increase alpha in order to zero out more factors.

12.1 - "Describe a situation or problem from your job, everyday life, current events, etc., for which a design of experiments approach would be appropriate."

I have a friend who is single and uses Tinder. He wants to know what type of pictures get the most matches. He thinks it could be a function of time of day of the photo, the color of clothes he is wearing, whether it is an action shot or a picture posing with friends, etc. He has no idea which factors most affect his getting responses. Given the number of factors, a formal or less formal fractional factorial design experiment might be a good idea.

12.2 - "To determine the value of 10 different yes/no features to the market value of a house (large yard, solar roof, etc.), a real estate agent plans to survey 50 potential buyers, showing a fictitious house with different combinations of features. To reduce the survey size, the agent wants to show just 16 fictitious houses. Use R's FrF2 function (in the FrF2 package) to find a fractional factorial design for this experiment: what set of features should each of the 16 fictitious houses have?"

I believe this is really just one line of code:

FrF2(16,10)

The output is a grid that specifies which factors to include in each house in your survey to maximize information you can collect on each factor with only 16 runs.

13.1 - “For each of the following distributions, give an example of data that you would expect to follow this distribution (besides the examples already discussed in class).”

a. Binomial

I mentioned a friend who is single and uses Tinder in 12.1 above. Perhaps unfairly, when reviewing profiles, his habit is to just swipe right on all the profiles he sees. He then chooses from among those who respond to him as to whom he will actually follow up with. My impression is that he has to swipe right many times for each reply he gets. I think it likely does follow some rough sort of binomial distribution. Perhaps Binomial with p = .005.

b. Geometric

An example of a geometric distribution would be how many times my friend above has to swipe right before getting a response. Geometric with p = .005.

c. Poisson

The responses that he does get over time may come in as a Poisson distribution. This will not be true over longer periods of time, as people are less likely to follow up too late at night, etc., which will ruin the i.i.d. requirement. But it is likely more or less a Poisson distribution for some shorter period of time, as he swipes right.

d. Exponential

The time in between those responses could be modeled by an exponential function.

e. Weibull

Technically, the time between responses could also be modeled by a Weibull function, since an exponential distribution is a type of Weibull distribution. It is just a Weibull function where k is set to 1.

To try to find an example where k does not equal 1, consider that once he replies, the people who get turned off by him stop talking to him fairly quickly. This probably causes the distribution of people stopping talking to him after he responds to follow a Weibull distribution with a k less than 1. Most people figure out pretty quickly that he is not going to be a great fit and end discussion fast, but those that do keep replying tend to do so for a while before stopping the conversation.