Luke Watson

Exercise 2

958

*Part 1*

To investigate how murder arrests was related to assault arrests in 1973, I ran a simple linear regression using a state’s reported assault arrest rate compared to their murder arrest rate. The model was highly significant (*F*(1, 48) = 86.45, *p* < .001, *Multiple* R2 = .64, *Adj.* R2 = .64). Assault arrests seem to predict a more murder arrests such that with every increase of 1 assault per 100,000 that state was predicted to have a .04-point increase in murder arrests per 100,000 (*B* = .042, *SE* = .005, *p* < .001). Model assumptions were checked using the performance package in R. Visual analysis of posterior predictive check, linearity, homogeneity of variance, and influential data points revealed no substantial deviations from assumptions. The normality of the errors appeared questionable, showing deviations from normality particularly at the tails (Supp. Figure 1). Is a linear model appropriate for analyzing this data? Given the descriptive statistics used for this analysis, arrests per 100,000, it seems questionable to predict fractions of an arrest. A more appropriate analysis would account for the fact that there cannot be half of an arrest. Poisson regression would be better suited to answer the question of whether assaults predict murders in the U.S. based on the discrete count qualities of the arrest data.

*Part 2*

Expanding on the model in part 1, I wanted to see if another regression including all predictors would provide better estimates of murder rate. So, I ran a multiple regression using rape and assault arrests per 100,000 along with a states proportion of population in urban centers to predict murder arrests per 100,000. The overall model was again significant (*F*(3, 46) = 31.42, *p* < .001, *Multiple* R2 = .67, *Adj.* R2 = .65), and the increased adjusted R2 suggests this model was slightly better at fitting the data than the previous model. The individual predictor of assault was the only significant predictor of murder arrests (*B* = .04, *SE* = .006, *p* < .001), while rape (*B* = .061, *SE* = .056, *p* = .276) and urban population (*B* = -.055, *SE* = .028, *p* = .056) were not significant according to the model. I checked the assumptions of this model with the same function from the performance package. These checks indicated no serious violations of assumptions aside from normality of residuals, which appeared to deviate (in a similar way to the first model) at the tails. Like the first model, I believe this relationship would be better captured by a Poisson regression.

*Part 3*

Expanding further on the multiple regression model done previously in part 2, I wanted to investigate if there were any interactions occurring between each of the predictors. I conducted a full factorial multiple regression with assault arrests, rape arrests, and proportion of urban population predicting murder arrests. The overall model was again significant (*F*(7, 42) = 14.95, *p* < .001, *Multiple* R2 = .71, *Adj.* R2 = .67), and the increase in both multiple and adjusted R2 indicates that this model fit the data better than the previous two models (although a much more conservative increase in the adjusted R2). No new predictors showed significance in this model compared to the previous models (see Table 1 for summary). I checked assumptions using the performance package in R and found that all of the variables were severely collinear. This will affect the interpretation of each of the parameter estimates below.

**Table 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Estimate | B | SE | t | p |
| Intercept | -9.493 | 10.95 | -.87 | .391 |
| Assault | .156 | .054 | 2.14 | .039 |
| Rape | .905 | .806 | 1.12 | .268 |
| Urban Population | .081 | .18 | .45 | .653 |
| Assault x Rape | -.004 | .003 | -1.24 | .223 |
| Assault x Urban Population | -.001 | .0009 | -.96 | .342 |
| Rape x Urban Population | -.009 | .012 | -.73 | .465 |
| Assault x Rape x Urban Population | .0004 | .0005 | .89 | .379 |

*Note.* Regression weights and standard errors are predicting units of murder arrests per 100,000. Assaults and rapes were measured the same as rates per 100,000. Urban population is the proportion of that state’s population living in urban centers. Data were obtained from state level arrest reports in 1973.

Finally, to compare the three nested models, I ran a model comparison. The test revealed that no significant improvement in model fit was made between the first and second (*F*(2, 46) = 2.13, *p* = .131) as well as the second and third models (*F*(4, 42) = 1.53, *p* = .212). This is also reflected in the negligible differences in the adjusted R2 values reported with each model as well as the AIC (Model 1: 242.5, Model 2: 242.3, Model 3: 243.5).

*Part 4*

Some differences jump out right away when using the anova() command as opposed to the summary() and Anova(, type =”III”). It contains no regression weights or standard errors, only the SS, mean SS, F, and p. Along with the different reports, one of the interactions became significant (Assault x Rape interaction). It seems like it makes a big difference how SS is calculated with larger models with interactions.

*Part 5*

To address multicollinearity issues in my interaction model from part 3, I centered the predictors and then ran the regression again. Some differences that stood out to me were the change in the estimates, particularly the intercept and the main effects which all were significant or approaching significance (rape *p* = .059). These estimate values were much smaller and more in line with the second model’s values in part 2. Centering each of the predictors changes the interpretation of the intercept, which now represents the value when all of the predictors are set to zero.

Another consideration for normalizing the errors could be a transformation of the outcome variable, murder per 100,000. Using the boxcox command from the MASS library in R revealed that the ideal transformation to normalize the errors would be a square root transform of the criterion.

*Part 6*

I changed the diet factor to be effect coded and compared the results of that model to the model with dummy coded diet. The initial code didn’t result in any different results, but after I saved and changed the contrasts on a new data frame object it worked. The differences that stand out to me are the intercept and the other estimates changed in value, as well as dropping the Diet 4 estimate for the diet 1 estimate when effect coding that variable. The intercept in the first, dummy-coded model captures the estimate for the first diet level, while the intercept in the effect coded model captures the grand mean of weights for all diets.

Mean diet 4:

Chick\_model(dummy)

* Times diet2 and diet3 by 0, diet4 by 1, add to intercept
* Or use emmeans

Sum\_chick\_model(effect)

* Times diet1, diet2, and diet3 by -1, add to intercept
* Or use emmeans

A screenshot of a computer code

Description automatically generated