**Confounders ->**

**Collinearity -> ask yourself which one actually tells us the story**

**\*run an Anova between Wealth and Inequality included**

**\*for confounder, don’t be overly concerned with the pvalue (10% rule of thumb)**

**-meet 4 steps of linearity**

**-if log outcome, make sure it’s interpreted on the non log scale - look up how to explain the log transformation**

**-use F-test for comparing models**

**-kick something that is collinear if a high VIF; or if related because only looking at EFFECT**

**Clinically/policy relevant vs. statistical relevance “if focusing on crime right, maybe don’t initially worry about this”**

**Maybe discuss overfitting but don’t throw out significant data points**

**Add to methods:**

**-justification of log**

**-F-test for taking out wealth**

**-testing the interaction of South and NW: While this interaction was statistically significant in the model, the adjusted R square only increase from .717 to .769. This could be an interesting variable if evaluating crime rates across regions, however, due to model complexity we chose to remove it from the final model.**

**-confounding test: NW (tested, no a 10% change for prob or time)**

**BST 210** (https://github.com/karlamejia/BST-210/)

**Homework 4 / First Group Project**

Due Tuesday, October 10, 2017 (due to Columbus Day), 2:00 PM, submitted online

For this data analysis group project, you should work with 1 to 3 colleagues (groups of size 2 to 4, no singletons). Groups will submit one solution and group members will receive the same grade.

Criminologists, sociologists, politicians, citizen groups, and others are interested in the effect of punishment regimes on crime rates. You will be **studying data on crime rates (# of offenses per 100,000 population in 1960 in 47 states of the United States) to assess the effects of punishment regimes circa 1960**. You will likely also wish to adjust for certain demographic and disparity measures, as appropriate. See the other file for more details on the data set.

Use basic statistics and linear regression and your knowledge of model building and model selection, confounding and effect modification, and regression diagnostics to develop and assess an appropriate model (or models) to address this research question. Use logic, common sense, and good statistical methods to do this.

Write up your findings by developing:

**· A title**

* Effect of punishment regimes on crime rate
* Regression analysis of …..

**· A 250 word abstract matching a standard journal style (e.g., Background, Methods, Results, Conclusions)**

* We are studying …. - objective
* Methods: linear regression piecewise linear, stepwise, log of crime
  + Mention both models (with and without log)
* Results: beta coefficients and p-value and conclusion

**· One to two paragraphs describing your statistical methods**

Discuss more details than scientific papers on your thought process through checking for models and choosing

Methods:

1. **Exploratory data analysis:**
   * Histogram -
     + Checking normality of the variables (exploratory analysis)
     + Evaluated Crime and noticed a right skew - may need to considering transforming this variable. Especially because dealing with a rate, which normally has a Poisson distribution.
   * Correlation analysis
     + Collinearity - if two variables, either decided to combine them or pick one
       - wealth/income inequality - yet different measurements so kept separate
       - If 2 variables are highly correlated then look at it (consider if they’re measuring different things, if they are measuring different things, then we include it in consideration for final model)
       - If they’re collinear and making the same thing, then we choose one or combine them
       - How do you combine variables with high correlation? (Wealth and ineq highly correlated but measuring different things)
2. **Univariate analysis** (one covariate vs outcome)
   * Make Table 2 - with crime and log crime
   * Loess of each variable and outcome for log and nonlog
     + Crime not logged splines:
       - LF (1 cutoff, use linear instead of cubic for easier interpretation) - 0.575
       - NW - 10
       - time - 25
     + Crime logged
       - LF (0.575)
       - NW - 10
       - Time -25
   * Data manipulation = Added new variables
     + Did an average for police expenditures
     + U2/U1 unemployment rate ratio
   * After adding new variables, went back and included them in Table 2, correlation, and loess curves. Ultimately determined..
3. **Model selection**
   * Stepwise selection - with all variables - required both Prob and Crime - default in SAS (0.15 out) (0,05 in)
   * Describe how you choose splines, cutoffs, create new variables, by observing graph, we picked it as our cutoff

**· Up to five tables and/or figures (and you might only need fewer, also no multi-panel figures)**

1. Lf splines nonlog crime
   1. Show loess you applied spline
2. Residual vs predicted value plots (for log and nonlog)
   1. Residual plots of showing log crime vs crime (based on final model) - to argue why log is better
3. Table 1
4. Final model table
5. Table 2

**· One to three paragraphs summarizing your results (including referring to your tables and figures)**

* Additive vs multiplicative scale- interpretation
* Select variable with p-value less than 0.05 and drop if more than 0.1 (changes after adding in new variable)
* Keep time in model no matter what
* Include So to eliminate heteroscedasticity - different variance
* Splines: LF (1 cutoff linear instead of cubic) - 0.575, NW, time - 25
* Try log y - AIC check residuals, residual plots, which one is better ---> interpretation of beta coefficients is different i
  + Reduces heteroscedasity of outcome
  + Rate is poisson distribution
  + Connection between log y and poisson
  + We assume residuals are normally distributed so we can’t say it’s a poisson distribution
* If NW = 1 and rate = 2, if I know beta = 1. When Nw = 2 then
  + Better

If 2 variables are highly correlated in final model → use one variable or combine variables into 1

* Choose po1 or po2
* Average of po1 and po2 - try this variable
* Interpretation: convert wealth to $1000 (consider for other variables too)
* 2nd part of linear

All of this should be appropriate for submission as part of a manuscript. Use appropriate number of decimal places, labeling of variables, footnotes, nice layouts of abstracts and tables, and a good writing style, like you were submitting your work to a journal. Look at a few articles in journals for ideas. Higher grades will be assigned to strong analyses, perhaps with some interesting or challenging methods or findings, but also answering the research question well, having a professional appearance and writing style, and having nice tables, figures, and appropriate statistical methods and summarization of results. Focus your summarization of results to only include the most important information, not all the technical details, as you would see in a manuscript. Excessive less relevant material will be considered a negative.

You will submit your write-up online, through Canvas. You may also attach at the end a limited number of computer outputs (with comments) to help justify your approach and findings, or possibly to include some details not included in your summarizing of results text. However, the grading of this assignment will focus on your findings as described above, and only a limited number of additional pages should be submitted.

Outline

**Exposure of interest:** Punishment regimes

**Outcome**: Crime rates

**Potential confounders**: demographics and disparity measures

**Methods**

* Splines/GAM
* Forward
* Backward
* Stepwise
* Compare R2, Adjusted R2, Root MSE, AIC, and BIC
* F-test to compare models
* Do models meet 4 assumptions (Residual analysis)
* What is our objective? What information is most important for us to know?
* Figures: GGPlot
* Table 1: overview of data
  + Make it easily readable and understandable for readers
* Table 2: regression models estimates and p-values
  + Only need to include output from final model to be discussed in results

*Abstract Example from Lab*

**BACKGROUND:** Neurodevelopmental disability is the most common complication for survivors of surgery for congenital heart disease (CHD).

**METHODS:** We analyzed individual participant data from studies of children evaluated with the Bayley Scales of Infant Development, second edition, after cardiac surgery between 1996 and 2009. The primary outcome was Psychomotor Development Index (PDI), and the secondary outcome was Mental Development Index (MDI).

**RESULTS:** Among 1770 subjects from 22 institutions, assessed at age 14.5 ± 3.7 months, PDIs and MDIs (77.6 ± 18.8 and 88.2 ± 16.7, respectively) were lower than normative means (each *P*< .001). Later calendar year of birth was associated with an increased proportion of high-risk infants (complexity of CHD and prevalence of genetic/extracardiac anomalies). After adjustment for center and type of CHD, later year of birth was not significantly associated with better PDI or MDI. Risk factors for lower PDI were lower birth weight, white race, and presence of a genetic/extracardiac anomaly (all *P* ≤ .01). After adjustment for these factors, PDIs improved over time (0.39 points/year, 95% confidence interval 0.01 to 0.78; *P* = .045). Risk factors for lower MDI were lower birth weight, male gender, less maternal education, and presence of a genetic/extracardiac anomaly (all *P* < .001). After adjustment for these factors, MDIs improved over time (0.38 points/year, 95% confidence interval 0.05 to 0.71; *P* = .02).

**CONCLUSIONS:** Early neurodevelopmental outcomes for survivors of cardiac surgery in infancy have improved modestly over time, but only after adjustment for innate patient risk factors. As more high-risk CHD infants undergo cardiac surgery and survive, a growing population will require significant societal resources.

*Abstract Example from Lab*

**STATISTICAL METHODS:** Group differences by cardiac class were assessed by using x2 tests for categorical variables and analysis of variance or Kruskal–Wallis tests for continuous variables. Comparisons of PDI and MDI with normative means were made using 1-sample t tests. Primary analyses examined the relationships of PDI and MDI from BSID-II with year of birth. Pearson correlations were used to examine these relationships in the full cohort as well as in homogeneous diagnostic subgroups (ie, TGA, TOF, and HLHS). The types of CHD included in clinical studies changed over time. Therefore, linear regression analyses using continuous year of birth and adjusting for center and cardiac class served as foundation models for predicting PDI and MDI. Candidate predictors in multivariable analyses were preoperative measures and patient factors, including birth weight, gestational age, race, ethnicity (Hispanic versus not Hispanic), gender, maternal education, genetic or extracardiac anomalies, prenatal diagnosis, preoperative mechanical ventilation, and neonatal status (age at ﬁrst surgery #30 days). Factors relating to operative management or postoperative course were not included. Predictors were screened to identify associations with a neurodevelopmental score at the P , .25 level after adjusting for center, cardiac class, and year of birth. Predictors meeting this criterion were included in stepwise backward analysis in which P , .05 served as the criterion for retention into the full model. Center, cardiac class, and year of birth were retained in all models regardless of P value. Standardized mean scores present the predicted PDIs and MDIs adjusting for center, cardiac class, and other statistically signiﬁcant predictors at the mean value of the covariates. Analyses were performed using SAS 9.3 (SAS Institute, Cary, NC).

\*Use Table 2 to get example for the table portion

<http://pediatrics.aappublications.org/content/135/5/816>

Research:

<http://www.statsci.org/data/general/uscrime.html>

<http://www.icpsr.umich.edu/icpsrweb/NACJD/studies/7716> :

“For each of the index crimes, there are two sanction variables included: the probability of prison commitment and the average time served by those sentenced (severity of punishment). There are 11 socioeconomic variables, including family income, income distribution, unemployment rate for urban males in the age groups 14-24 and 35-39, labor force participation rate, educational level, percentage of young males in population, percentage of non-white young males living in the population, percentage of population living in Standard Metropolitan Statistical Areas, sex ratio, and place of occurrence. The data also include per capita police expenditures for 1959 and 1960.”

<https://www.theatlantic.com/politics/archive/2016/06/crime-imprisonment-rates/486014/>

<https://www.nap.edu/read/18613/chapter/6#106>

Potential confounders/effect modifiers to test (based on above research re crime in the 1950-60s, p107):

So indicator variable for a southern state

NW percentage of nonwhites in the population

<http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=5880&context=jclc>

“Dichotomizing.this data, Gibbs used Chi-square tests…” the study looked at severity (time) and certainty (prob) of

“They also tested the data for departures from linearity. When no significant departures were found, they concluded that the logarithmic transformation models presented by Gray and Martin did not result in a significant increase in prediction and, thus, the models must be ruled out on the grounds of parsimony….”

**Outline**

* Fixed variables = Prob and Time
* Table 1 (SAS)- summary statistics
* Test M and MpF for interaction independently or with other confounders?
* Effect Modification:
  + Test Southern state (So)
* Potential confounders:
  + NW percentage of nonwhites in the population
* Consider grouping Ed and Wealth (categorize) → ANOVA, effect modification
  + Look up in literature what categories are generally used
  + Correlation of different wealth variables and see if highly correlated (decide which one to use) - categorize
* Socioeconomic Variables
  + Ed\*
  + LF
  + U1
  + U2
  + Wealth
  + Ineq\*
* Consider using Po1-Po2 to see how change in per capita expenditure may affect crime
* Univariate analysis: histogram of each
* Bivariate: (outcome with each predictor, predictor with predictors)
  + Correlations of different variables
  + Plot bivariate and check for associations
* Splines- GAM (bivariate)
* Test for collinearity
* Models from base tests:
  + Base model: Crime ~ Prob + Time
  + Forward Selection: (formula = Crime ~ Po1 + Ineq + Ed + M + Prob + U2, data = crime)
  + Backward selection: (formula = Crime ~ M + Ed + Po1 + MpF + U1 + U2 + Ineq + Prob, data = crime)
  + Stepwise: (formula = Crime ~ Po1 + Ineq + Ed + M + Prob + U2, data = crime)
  + Subset: Crime ~ M + Ed + Po1 + MpF + U1 + U2 + Ineq + Prob
  + Stepwise Selection: Crime ~ Time + Prob + Po1 + Ineq + Ed + M + U2 + U1 + MpF
* Questions for David
  + So we know some variables are probably highly correlated (is this the same as collinearity?) What’s the best way to test for this, so we aren’t including duplicate variables
    - Assess some using pearson, basic statistics (scatter plots), histograms and summaries; covariates do not need to be normally distributed (it could be binary); is it skewed, bimodal, high outliers;
    - Parallel models using 1 model (with 1, other, both, combination, interaction, show something)
  + When do we test for effect modification?
    - Test some a priori, test after seeing main effects model
    - Depends on research question of interest (can try at beginning or later on) - do main effects model first and then explore later
    - P-value in model with or without interactions may be different - take in consideration
  + M (how can we use this?)
  + MpF (how can we use this?) -> could we categorize into ‘more males than females’ (1) and ‘fewer males than females (0)’
  + Do we need to do univariate, bivariate analysis (diagnostics of each variable)?
    - Associations pearson correlation
    - How do punishment rates affect crime rates?
    - Approximate normality is good enough
    - Do with and without that point (doesn’t change point, don’t need to discuss), loess curve in bivariate but mostly in final model
  + Categorizing wealth and education variables? Analyze with groups
    - Consider categorizing for better interpretation
  + David said:
    - ”Based on the literature, we want to adjust for x, y, and z… “ and then we can go “fishing” from there on out
    - You might have a basic model and then a more intense model
    - “Testing all two-way interactions would be stupid”
    - Basically, we just want to see if the terms are associated with crime rate…
    - David and the TAs don’t wanna see literally everything we tried
      * You should include stuff that you think is important/significant
    - Is it the punishment that affects the crime rates or is it something else that affects the crime rates?
    - It sounds like we need a hypothesis before we start “fishing” for possible variates
    - May need to look at the police expenditures as part of ‘punishment regimes’
    - Maybe look at Cook’s, etc for the final model; may have more than one model; approx normality is okay
    - When there’s a linear covariate, we should assess linearity
      * Don’t use all the methods, only use the method that logically makes sense (i.e., do not go wild)
      * Sometimes categories are better than GAMs and Splines bc they are so hard to explain
    - Sounds like he doesn’t think the MtF ratio even matters (and we don’t either lol)
    - If not significant- don’t worry about interactions, splines of it, just mention in test that it didn’t appreciably change outcome
      * For Time: could test it but see not significant; don’t worry about interactions or splines with it
  + Greyson said:
    - You can use stepwise (or other stuff) to let variables in or kick them out
    - If we use stepwise, we should use thresholds .1 to kick them out and .05 to keep them in
    - If you know there are collinear models, you only include the ones that are important

**Game Plan**

* + Rose
    - Table 1
    - Run univariate (histograms) in SAS
    - Run bivariate (get coefficients and p-value) models in SAS
    - Univariate analysis: histogram, check for normality
  + Hillary
    - Plot each predictor ; Lowess
  + Karla

Kaitlyn

* Identify outliers and see if they change the fit. If they change the fit, we should report a model with the outlier(s) and a model without the outlier(s). You should probably not be tossing out anything in this dataset since there are only 47 observations.
* Think about confounding (factors that meet definition of confounding) and effect modification (factors that modify the relationship between punishment regimes and crime). There is a statistical test for effect modification.
* Don’t worry about using stepwise procedures. Focus on adjusting for confounders.
* When interpreting the log(Crime) model, make sure you interpret it in the non-log scale!
* Our justification for using the log should be based on
  + Funky-looking residuals
  + NOT THE HISTOGRAM LOOKING NON-NORMAL (bc we are not concerned with the outcome being normally distributed, we care about how the OUTCOME RELATES TO PREDICTOR)
  + If you have a lot of clustering in small crime rate values to make it meet the linearity assumption
* AIC and BIC are nice to look at, but they are not formal tests. We should consider using F test to formally test! Keep in mind that an Adjusted R2 of .7 is quite high.
* You can pick different models that are acceptable but tell different stories or you can pick one model and explain why
* We should DEFINITELY kick out collinear things, but in this class we should focus on keeping statistically significant things while also keeping in mind that we are adjusting for important confounders.