Stat 414 – Day 1
Review of General Linear Model

Example 1: For each scenario below, identify the response and the explanatory variables and then consider each of the LINE assumptions in the context of the study, commenting on possible problems with the assumptions.

(a) Francis Galton suspected that a son's height could be predicted using the father's height. He collected observations on heights of fathers and their firstborn sons.

Response variable = son's height
Explanatory variable = father's height
Note these are "firstborn" sons so no multiple sons from the same family, but we should still make sure there aren't any genetic links among the pairs (independence). There could be a a problem with equal variance if sons of tall fathers had much more variation in heights than sons of short fathers. Probably reasonable to assume linearity and a normal distribution of son heights at each father height.

b) Is the time spent studying predictive of success on an exam? The time spent studying for an exam, in hours, and success, measured as Pass or Fail, are recorded for randomly selected students.

Response variable = Pass or Fail
Explanatory variable = time spent studying
Binary response, so use logistic regression

(c) A researcher suspects that loud music can affect how quickly drivers react. She randomly selects drivers to drive the same stretch of road with varying levels of music volume. Stopping distances for each driver are measured along with the decibel level of the music on their car radio.

Explanatory variable = decibel level of music
Response variable = reaction time (stopping distances)
There are potential problems with the assumptions. For example, if there is threshold for the volume of music where the effect on reaction times remains the same, mean reaction times would not be a linear function of music (but rather would "level off." Another problem may occur if a few subjects at each decibel level took a really long time to react. In this case, reaction times would be right skewed and the normality assumption would be violated. Similarly, if the reaction times had a lot more variability at low decibles then at higher decibles (equal variance).

(d) Do wealthy families tend to have fewer children compared to lower income families? Annual income and family size are recorded for a random sample of families.

Explanatory variable = family income
Response variable = number of children
Number of children is likely skewed right and discrete. That alone might not be a problem, but it's likely that this is true for every value of family income, violating the normality

assumption. Poisson regression might be better (which would also allow the variability to decrease with income/smaller number of children)

(e) The yield of wheat per acre for the month of July is thought to be related to the rainfall. A researcher randomly selects acres of wheat and records the rainfall and bushels of wheat per acre.
Explanatory variable = amount of rainfall
Response variable = yield of wheat per acre
Similar to (d), could be violations of linearity if the yield increases with rainfall but then decreases with too much rainfall. The "limiting behavior" in (d) would suggest a log transformation, an increase then decrease would suggest a polynomial model. Also want to make sure these fields aren't too close together (independence).

(f) Investigators collected the weight, sex, and amount of exercise for a random sample of college students.
Response variable = weight
Explanatory variable = sex and amount of exercise
The random sampling should help with the independence assumption. Here could look at variation in weight between males and females and variation in weight with amount of exercises (equal variance assumption). Also need weight to be linearly related to amount of exercise. Note, including a categorical variable is not a problem with indicator variables.

(g) Medical researchers investigated the outcome of a particular surgery for patients with comparable stages of disease but different ages. The ten hospitals in the study had at least two surgeons performing the surgery of interest. Patients were randomly selected for each surgeon at each hospital. The surgery outcome was recorded on a scale of one to ten.
Explanatory variables = stage of disease, age
Response variable = surgery outcome (scale of 1 to 10)
The discrete nature of the response variable, may not be a problem if the conditional distributions are approximately normal. We do need to watch for independence of operations by the same surgeon and even within the same hospital (could be some hospital level contextual factors to consider).

Example 2: The **Election2004Data.xlsx** datafile contains demographic variables for the year 2000 for 3,115 U.S. counties across 49 states (no Alaska), as well as the county-level election results in 2004 (proportion voting for Bush).
(b) What are the observational units in this study?
counties

(c) Examine the histogram of the percentage of voters who voted for George W. Bush. Is this distribution approximately normal? Is it required to be? Does anything surprise you about this distribution? What about the other distributions?
The distribution has a little bit of skewness, but (a) need to consider the normality of the conditional distribution of the response in the model and (b) might be ok with a bit of

nonnormality because we have a decent sample size. Might be a little surprised at the variation in the results from county to county?

(d) Is there evidence that the Democratic Party was able to tap into the youth vote? Look at the scatterplot of PctBush vs. Pct18.24. What do you notice? What do you suggest next?
The percentage voting for Bush does seem to decrease for counties with higher percentages of younger voters. Is some evidence that could benefit from a log transformation (decreases and then levels off).

(e) Is there evidence that income and voting for the Republican candidate are positively associated? Do you want to log transform income? Are there any downsides to log transforming?
The smoother shows an increase and then a decrease. This decrease would be pretty surprising as would expect a higher percentage voting for Bush in the richer counties. The curvature is supported by the RESET test.

(f) One of the issues that year was "traditional family values." As a proxy, we have a variable, PctFamily, defined as the percentage of residents living in a household with husband, wife, and at least one child under age 18. What do you observe?
A positive association between PctFamilyand PctBush. Maybe some curvature but the linear association looks like it will explain a fair bit on its own.

(g) Find the least squares line for predicting PctBush from PctFamily and interpret the slope and intercept coefficients in context.

```
Coefficients:
             Estimate
(Intercept)  39.6189
PctFamily     1.0814
```

If PctFamily = 0 (no traditionally structured families in the county), then we predict 39.6% of the country vote for Bush. With each one percentage point increase in PctFamily, the predicted percentage voting for Bush increases by 1.08 percentage points.

(h) What do you learn from the residual plots?
Same evidence of a little bit of curvature
No real evidence of unequal variances across the fitted values
Normality assumption looks great
No highly influential counties

(i) Fit the multiple regression model without median family income (Income) and using Income as the response. Does Income appear useful to add to this model? In what form?
Edit to the R Code: I hadn't included PctFamily in both models
So we have the residuals of the model without Income – these tell us about the unexplained variation that remains in the PctBush after accounting for the set of variables. Then the second set of residuals tells us about the variability in Income that is not accounted for by the same set of variables. If these variables are related, that tells us that Income has something *unique* to contribute to the model after accounting for the first set of variables. It also tells us a

little bit about form of the association (e.g., curvature). If we fit a regression between these residuals, we find a slope coefficient of -0.2982 with $t = 10.5$. This coefficient matches the coefficient in model3. *We also see that the adjusted association is more linear than the unadjusted association.*

(j) Fit the multiple regression model with median family income (model 3), using only the complete cases. Do the validity conditions appear to be met? Multiple R-squared?
Still some evidence of curvature, but generally look pretty good.
Multiple $R^2 = 0.426$, so 42.6% of the variation in PctBush across the counties is explained by these variables. Now the coefficient of median family income would be interpreted *after adjusting for the other variables* (more on this Day 2).

(k) Examine the residuals versus each explanatory variable. What do you suggest next?
Should worry about the curvature with the ethnicity variables.

(l) Add *PctBlack*^{2} to the model. Any issues? (*Hint*: Multicollinearity? Why?)
Yes, the VIF values for PctBlack and PctBlack^2 are close to 10. This makes sense because they are the same variable and so will be related to each other. It turns out that for the values we have in these data set, these two variables are linearly related to each other.

(m) Does *centering* the variables first help? How so?
Yes, this moves the curved part of the quadratic relationship between the two variables into the middle of our x-space. The VIF values for these two variables are now below 2. The residuals vs. fitted values graph now looks gorgeous!

(n) Investigate the three suggested interactions.
Centering variables involved in interactions can also help reduce multcollinearity. Doesn't seem to be too much of a problem here. In fact, none of the interactions are significant after adjusting for the other variables.

(o) What is the best way to test "can I remove all three interactions from the model"?
A "partial F test" would compare the model with all three terms to the model with none of the terms. (The reduced model is a "subset" or is "nested" in the full model.) The numerator df would correspond to the difference in the df between the two models. The denominator df would be the df error of the full model. If the F-statistic is large, that says at least one term is useful to the model and we should use the full model rather than the reduced model.

```
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1   3094 234442
2   3097 234713 -3      -270 1.19   0.31
```

In this case, we fail to reject the null hypothesis that the three β coefficients are all zero and the reduced model is adequate.

(p) What would a significant interaction imply in this model?
The relationship between say income and PctBush differs depending on how rural or urban the county is.