Statistics 688 Consulting

Report for Rietta Wagoner and Nicolas Lopez-Galvez

Nathan Hattersley
Renee Zhang
Elmira Torabzadehkhorasani

authored by
Nathan Hattersley

Clients: Rietta Wagoner and Nicolas Lopez-Galvez
Meeting Date: October 31, 2017
Contact:
nathanhattersley@email.arizona.edu
reneezhang@email.arizona.edu
elmira@statlab.bio5.org

0 Executive Summary

We discussed various approaches to analysis of data collected regarding the heat stress and pesticide exposure of farm workers in northern Mexico. We provide suggestions and code for visual data exploration as well as details of analysis, including model selection and improving model fit. The order of the sections mirrors the order of questions on the PowerPoint document you gave us. The first section concerns diagnosing heat stress, the replaceability of core body temperature measurements with a linear combination of other measures, and the prediction of urine-specific gravity with heat measures. The second section concerns the correlation of pesticide measures, comparison of pesticide measures across seasons, and the correlation of heat and pesticide measures. The final section is a short bit of editorializing on your plans for future study.

1 Heat Statistical Questions

Diagnosing Heat Stress and Substituting Heat Measures

The main issue at hand is how to predict heat stress, and whether it can be diagnosed indirectly with a combination of hand temperature and WBGT measurements. From perfunctory exploration on Wikipedia, it seems that the condition you are describing is hyperthermia, which is categorized simply as core body temperature exceeding some threshold (either 99.5 or 100.9°F). You said you had performed smoothing to eliminate the effect of anamolous readings, since there did seem to be a handful of anamolies. This is an appropriate measure to take. A short digression on smoothing follows.

There are numerous smoothing techniques commonly employed, three of which are moving average, loess, and splines. Moving average calculates the average in a small neighborhood of points, whereas loess and splines compute a regression equation on a small neighborhood of points. You can also view moving averages as local regression where the regression estimate is just the mean in that neighborhood of points—this might lend a visual explanation to the idea that moving averages are more robust to outliers. For that reason, we suggest using a moving average to smooth core body temperature measurements.

Back to the diagnosis of heat stress. Heat stress should be directly diagnosed with smoothed core body temperature. If you want to add a variable to your data frame in R that reflects a binary yes/no diagnosis of heat stress, use the *within* function:

```
THRESHOLD <- 99.5
new.df <- within(old.df,{
    STRESSED <- CORE > THRESHOLD
})
```

This assumes your data frame is called old.df, with a column named CORE that contains core body temperature measurements. The data frame new.df will be a carbon copy of old.df, except now there will be a column called STRESSED that takes on values of TRUE or FALSE based on whether or not the worker was heat stressed.

Given the infeasible price of a core temperature thermometer, we would like to replace CORE with a combination of HAND and WBGT. The most analytically tractable way to undertake this is to first predict CORE with HAND and WBGT (regressing only on the points for which you have readings for all three measures). Then, use your predicted values of CORE to diagnose predicted heat stress and compare those predictions to your true diagnoses given by CORE. In short, you want to use numerical variables in your regression and dichotomize afterward. The next sub-section details our suggestions about regression, but here is an example of post-regression dichotomization:

```
## predict the core body temperature
CORE_prediction <- lm(CORE ~ HAND + WBGT, data=my.df)$fitted.values

## add these predictions to your data frame
my.df <- cbind(my.df,COREpred=CORE_prediction)

## diagnose heat stress
THRESHOLD <- 99.5
my.df <- within(my.df, {
    STRESSED    <- CORE > THRESHOLD
    STRESSEDpred <- COREpred > THRESHOLD
})

## get the rate of true/false negatives/positives from your analysis
## the with function just allows you to reference columns of my.df more succinctly
with(my.df, c(
    TruePos = mean(STRESSED & STRESSEDpred),
    TrueNeg = mean(!STRESSED & !STRESSEDpred),
    FalsePos = mean(!STRESSED & STRESSEDpred),
    FalseNeg = mean(STRESSED & !STRESSEDpred)
))
```

Regression on Heat Measurements

As a first step in examining the relationship between the three heat measurements, we suggest visual examination. Two good ways to do that are to examine the scatterplot and correlation matrices. Simply call *pairs* or *cor* on a data.frame object in R. This will give a graphical and numerical way to cross-tabulate the relationships between the variables. Look for linear patterns or high correlations, and you can also view transformations of the variables if you think the relationship between two variables is non-linear. (Note: transformations like the squaring a measurement tend to work best when you center the data, which can be done with the *scale* function in R.) Then, once you see some relationships you would like to explore, employ regression to quantify them.

Given there are a lot of possible regression equations to predict core body temperature, it would be pertinent to consider some kind of variable selection method. The idea behind variable selection is to find the combination of predictor variables that gives the best-fitting model, while also penalizing models that have too many terms. There are myriad options (Adjusted R^2, Mallow's C, PRESS), but two common methods are AIC and BIC. These two are calculated in basically the same way, except BIC has a heavier penalty on additional terms. Perform model selection using the *step* function in R:

```
n <- nrow(my.df)
step(lm(CORE ~ ., data=my.df), direction="both", k=log(n))
```

The model formula $CORE \sim .$ says to predict CORE temperature using every other column in my.df, so be sure to add transformations of HAND and WBGT to my.df if you want to predict CORE using a non-linear function of HAND and WBGT. Setting k=log(n) performs BIC, while for AIC, set k=2.

Predicting Urine Specific Gravity

As for which heat measurement best predicts urine specific gravity, we can think about this problem as a simpler case of the variable selection outlined above. Instead of looking at every possible combination of our heat measurements, we will just consider the regression equations that have one predictor variable. Instead of using *step*, you can do this analysis manually rather easily. Compare two models using adjusted R^2. The higher the adjusted R^2 value, the better the predictor. Assuming URINE is the variable you want to predict, and CORE is the predictor variable:

```
adjR2 <- summary(lm(URINE ~ CORE, data=my.df))$adj.r.squared
```

2 Pesticide Statistical Questions

In terms of examining the correlation between pesticide measurements, we suggest the same visual approach as above as a first step. Visually examining the data will help inform your choice to use either the Spearman or Pearson correlation coefficient. Pearson tests linear relationships between two variables, and also assumes that the data are normally distributed. (You can test normality with the *ks.test* function in R.) Spearman tests monotonic relationships between data, without the normal assumption. There is no "better" statistic to use, rather your visual data analysis should inform the choice of correlation coefficient. You can obtain confidence intervals and perform significance testing on your correlation coefficient with the *cor.test* function in R.

As to whether pesticide levels vary by season, at first blush an ANOVA analysis should be used. ANOVA extends the t-test to compare more than two groups. More on performing ANOVA analysis can be found here. ANOVA assumes normality of the data and equal variance among the groups. If either of those assumptions fails, a quick remedy is to try the ANOVA analysis on the log of the data. If that analysis again violates the assumptions, then the non-parametric extension of ANOVA analysis is the Kruskal-Wallis Rank-Sum test. You noted in your PowerPoint that you had tried a Wilcoxon Rank-Sum test. Wilcoxon is an extension of the t-test, however, and should only be used to test two groups against one another. The Kruskal-Wallis test can be performed with the *kruskal.test* function in R. Keep in mind that there is a trade-off you make for the relaxed assumptions of a non-parametric test: Non-parametric tests have less power and less ability to detect differences between groups.

In order to examine the relationship between heat and pesticide, we recommend the same approach as outlined in predicting urine-specific gravity. Starting with a visual analysis, look for relationships that are worth exploring, and from there proceed with regression in order to quantify the relationships. Remember to check the assumptions of normality and homogeneity of variance. For an exhaustive list of regression assumptions and how to check them, see here.

3 Approach to Upcoming Research

In general, we don't see any issues with how you are going to approach the four points outlined in this slide. The only item that we discussed at any length was the selection of a control group against which to compare the pesticide levels of the farm workers. We don't have a "statistical" answer to that question—you will have to think about what different populations might exhibit in terms of pesticide exposure in order to find the group that best serves for comparison. Some possibilities floated in discussion amongst ourselves were citizens of a nearby town—so that you test the farm workers against people who eat the same food but don't work on a farm—or other farm workers, perhaps on a different farm. The validity of your analysis here will rest upon the qualitative differences you observe between the worker and control groups.

NOTE: If you have any questions about this report, please feel free to contact us. Though the semester may be concluded, we are still available to ensure that you understand all the ideas herein presented.