實習單位：國際哈佛統計科技顧問有限公司

用 R 解密數據之謎：生物醫學研究設計與資料分析

Unraveling Data Mysteries with R: Designing and Analyzing Biomedical Data

實習學生：薛沐儀、翁梓瑄　　指導老師：王彥雯 老師　　單位指導：胡賦強 老師、溫芳羽

Participated in the Daily Statistical Consulting Clinic

Our 8-week summer internship at the International-Harvard (I-H) Statistical Consulting Company in Taipei allowed us to participated in daily statistical consulting meetings with clients. The clients we met were physicians or researchers working in hospitals and biotechnology companies with the need for statistical analysis of their clinical data. Through observing the statistical consulting activities during the 8-weeks, we have developed a good understanding of the standard operating procedure (SOP) the company has established to ensure the superior quality of statistical design and analysis. A general process for performing a thorough regression analysis consists of three important steps: **(1) model selection (esp., variable selection), (2) goodness-of-fit (GOFs) assessment, and (3) regression diagnostics and remedies**, which should be applied to any statistical modeling to ensure validity and reliability of analysis results. A distinguishing feature of this company is its interactive statistical consulting service — that is, the entire statistical analysis process is carried out during their online meetings with clients. Consequently, clients get to provide their insight and knowledge about the subject matter throughout the meetings, which is a crucial part of obtaining a meaningful result, and the company explains its rationale for analysis, procedure, and results clearly and concisely as a meeting goes. This way of collaboratively working together gives both parties more clarity in their research issues and creates an efficient partnership that greatly improves the quality of statistical analysis. We had participated in roughly 4-6 hours of meetings a day during the 8-week internship. This gave us the opportunity to meet many different clients each with their interesting research topics and unique service needs. We even had the opportunity to help with R coding for for an urgent research project of a client on the last day of our internship. This first-hand work experience truly gave us a better understanding of how much work and mental focus are required to perform statistical analysis interactively during an online statistical consulting meeting.

Learned R and updated the My.Stepwise package

The company submitted its 'My.stepwise' R package to the Comprehensive R Archive Network (CRAN) in 2017, which helps R users **perform modern stepwise variable selection procedures in various types of regression analysis.** Since then, the company has not only made improvements to the existing R functions in the package, but also added new features to them, including the option of inverse probability weighting (IPW). During our internship, we also helped with the package update process. To update an R package, we need to create various files, including documentation, description, updated and checked R functions, examples, and more. These files required in the re-submission of an R package to CRAN can be prepared with the aid of the 'roxeygen2' package of R. Our task was to compile a flow chart of the steps necessary to update an R package and to figure out how to use 'roxeygen2' to create all the required files. In the end, we successfully ran the 'roxygenizing' process using an example R code file. The package update process still has more work to do, and the rest will be carried out by Dr. Fu-Chang Hu and his colleague, Ms. Fang-Yu Wen, in the near future. However, this was an interesting and unique learning experience to be part of the package update process.

Motivation of analyzing the New York Air Quality Data in R

Aside from learning though real life cases during client meetings, we have also been studying textbook material on statistical design and analysis with R. We studied and worked on the first eight chapters of the R e-book, written by Dr. Fu-Chang Hu, and passed the weekly quiz each Monday afternoon. We wanted to try and utilize this new found knowledge by completing a full data analysis using a built in data in R and the My.stepwise package. **We aimed to find important predictors of daily ozone concentrations with the available meteorological measurements by using various regression techniques.**

Methodology

- ✓ Daily air quality measurements in New York from May to September in 1973, which is freely available in R.
- ✓ Data Cleansing
  1. Set up dummy variables for every month (5 months in total)
  2. Removed missing data, NA's. As a result, 42 of 153 days were removed, 111 days remaining.
  3. Defined days of 'Weekend' according to the 1973 calendar.
- ✓ Response variable: Ozone (ppb)
- ✓ Covariates: Temperature (°F), wind speed (mph), solar radiation, month and weekend
- ✓ Data analysis: Linear Regression and Logistic Regression
- ✓ Stepwise regression: My.Stepwise R package
- ✓ Dealing with separation (or high discrimination): Chi-square Test and Firth's Bias-Reduced Logistic Regression

Results: Linear Regression Analysis

- ✓ Transformation for symmetric distribution to $Y^{0.25}$
- ✓ Solving **non-linear effects of continuous covariates by fitting their semi-parametric extension of GLMs, GAM** + discretized into categorical variables
- ✓ Multicollinearity check
- ✓ Involving interactions
- ✓ Variable selection: stepwise regression with My.Stepwise package

Linear Model: Ozone.025~Month8 + Solar.R.high + Temp.high + Solar.R.high.x.Wind.low				
	β Estimate	Std. Error	t value	p-value
Intercept	1.8844	0.0522	36.113	<2e-16
Month8	0.2299	0.0694	3.311	0.0013
Solar.R.high	0.2342	0.0663	3.535	0.0006
Temp.high	0.3825	0.0614	6.230	9.59e-09
Solar.R.high.x.Wind.low	0.4372	0.0725	6.028	2.45e-08
Multiple R-squared = 0.6703, Adjusted R-squared = 0.6578, F test p-value < 0.0001				

After adjusting for the effects of other covariates, the mean ozone level was higher in August 1973 and increased in hot temperatures, probably because of less wind and more direct sunlight. Moreover, the binary covariate 'Solar.R.high' had a positive effect on the mean ozone level, since ultraviolet (UV) rays helped to create ozone. Yet, **we had previously observed a surprisingly lower positive effect on ozone for very high values of 'Solar.R' in the GAM plot.** A possible explanation for this odd phenomenon is that solar flares and large solar storms could actually deplete upper-level ozone, as documented in an article of August 3, 2001, by the National Aeronautics and Space Administration (NASA) of the USA (*https://earthobservatory.nasa.gov/features/ProtonOzone*). We also found that there was an interactive effect of 'Solar.R.high' and 'Wind.low' (Wind.low = 1 if the wind level was below the cut-off point, meaning a lower wind level) on the mean ozone level, indicating that the effect of the wind on the ozone level was different under different levels of solar rays, and vice versa. That is, when the solar-ray level was high, low wind levels resulted in higher mean ozone levels, whereas high-speed winds might just blow ozone away.

Results: Logistic regression

- ✓ The above multiple linear regression model did not fit the continuous responses 'Ozone.025' very well → cut it as a binary response variable by the eight-hour ozone (O3) standard set by the EPA in Taiwan: **60 ppb**, where ppb = part(s) per billion
- ✓ Solving non-linear effects of continuous covariates on logit(p) (i.e., log[p/(1-p)], where p is the probability that the ozone level is higher than the threshold value of the air ozone standard) by fitting GAM analyses + discretized solar radiation levels to a dichotomous variable Solar.R.high = 0 or 1, with the cut-off value of 156.5736
- ✓ Multicollinearity check
- ✓ Checking separation and performing **Chi-square Test**

	Ozone.high		
Solar.R.high	**0**	**1**	**Total**
0	37 (100.0%)	0 (0.0%)	37 (33.3%)
1	45 (60.8%)	29 (39.2%)	74 (66.7%)
Total	82	29	111
Pearson's Chi-squared test: χ^2= 19.6281, d.f. = 1, p-value < 0.0001			

Separation (or high discrimination) occurred due to a zero count in one of the four cells, and thus the estimated odds ratio went to infinity, causing a statistical estimation problem in logistic regression analysis. The Chi-squared test of the association between these two dichotomous variables yielded statistical significance ($p < 0.05$).

- ✓ Involving interactions
- ✓ Variable selection: stepwise regression with My.Stepwise package without solar radiation (Solar.R.high)
- ✓ Performing **Firth's Bias-Reduced Logistic Regression** with solar radiation

Exact Logistic Regression Model 2: Ozone.high~ Solar.R.high.x.Temp + Wind + Solar.R.high				
	β Estimate	Std. Error	Chisq	p-value
Intercept	0.6279	2.0415	0.0819	7.7480e-01
Solar.R.high.x.Temp	0.4397	0.1276	30.9984	2.5824e-08
Wind	-0.5332	0.1824	11.2698	7.8777e-04
Solar.R.high	-32.7567	10.7112	15.6005	7.8236e-05
Wald Test = 21.5088, d.f. = 3, p-value < 0.0001				
AUC = 0.987, 95% Confidence Interval [CI] = [0.972,1]				

After adjusting for the effects of other covariates, an increase in wind speed would reduce the chance that the ozone level was higher than the threshold value of the air ozone standard, indicating that ozone levels are lower if windy and higher in stagnant air. However, the dichotomous covariate 'Solar.R.high' had a large negative effect on the 'Ozone Threshold'. This result is quite surprising since all previous linear regression models and the above logistic regression model without interactions had shown a significantly positive effect of 'Solar.R.high'. However, this seemly odd result will make clear sense if we combine the interaction term 'Solar.R.high × Temp' and the main effect term 'Solar.R.high' together in making an interpretation as shown below:

$$0.4397 \times (Solar.R.high \times Temp) + (-32.7567) \times Solar.R.high$$
$$= Solar.R.high \times (0.4397 \times Temp - 32.7567).$$

This means that when the solar ray level was high, the temperature did not have an effect on the chance that the ozone level was higher than the threshold value of the air ozone standard until 0.4397 × Temp > 32.7567. Technically speaking, **the effect of 'Temp' was nested in 'Solar.R.high'.**

Discussion

At present, we cannot provide an insightful and knowledgeable explanation for this interesting result. Understanding why or how this could happen requires the input of experts who have a professional knowledge of air quality and pollutants. This case demonstrates **the importance of having experts in the research topic work together when performing data analysis**, as their input is truly as crucial as our statistical expertise in making decisions during statistical analysis, interpreting the analysis results, and examining whether the research findings are plausible.