Evaluating the Predictors of Chronic Kidney Disease

Identification Number 8806
Biostatistics 699 – Project 1
01.29.2008

Appendix A contains supplemental results
Appendix B contains analysis code

Abstract:

With the ever-increasing prevalence of obesity and diabetes in the United States population, the number of people living with chronic kidney disease (CKD) is higher than ever. Despite the growing healthcare burden of CKD, its causes and predictors are poorly understood. The current study aims to discover and evaluate novel predictors of CKD.

Data was taken from the National Health and Nutrition Survey (NHANES), which is a continuous, cross-sectional survey design. The outcome of interest is of course CKD, which is measured in stages of severity ranging from 0 (no disease) to 6 (on dialysis). Subjects are assigned to a stage category via the glomerular filtration rate (GFR), which is typically determined by the net filtration pressure, the permeability of surrounding corpuscular membranes and the surface area available for filtration. In our case, we use an equation that incorporates race, gender and urine creatinine levels to estimate GFR.

Two types of models were used in analysis: a continuous outcome model, and a continuation ratio model.

Age, gender, diabetes status, smoking, blood urea nitrogen level, globulin level, and uric acid level were all associated with CKD. Various interactions were also tested and found to be associated under the CR model, but not with the continuous outcome model. An additional finding was that women have a higher predicted probability of advancement from a no-disease state to a disease state. This research may prove valuable to clinicians in identifying and treating high-risk patients before stage advancement. Predicted probabilities should be particularly useful is assessing patient risk.

Introduction:

With the ever-increasing prevalence of obesity and diabetes in the United States population, the number of people living with chronic kidney disease (CKD) is higher than ever. Despite the growing healthcare burden of CKD, its causes and predictors are poorly understood. The current study aims to discover and evaluate novel predictors of CKD.

Methods:

Data was taken from the National Health and Nutrition Survey (NHANES), which is a longitudinal, cross-sectional survey designed to measure the health of children and adults living in the United States. It contains both a self-reported portion, and a series of laboratory and examination measures. For the present analysis I used data from both sections including variables related to diabetes status, alcohol and tobacco use, and cardiovascular health, as well as lab and examination measures.

Outcomes and Variables of Interest:

The outcome of interest is of course CKD, which is measured in stages of severity ranging from 0 (no disease) to 6 (on dialysis). Subjects are assigned to a stage category via the glomerular filtration rate (GFR), which is typically determined by the net filtration pressure, the permeability of surrounding corpuscular membranes and the surface area available for filtration[1]. In this case, I used an equation that incorporates race, gender and urine creatinine levels to estimate GFR.

[1] Anything remotely biological comes from the text in the references.

Variables of interest from the survey include cancer status, diabetes status, alcoholism, and smoking. Demographic variables such as age, sex, marital status, income, and race were also considered to control for confounding.

Three additional variables were generated using the information from other variables in the set. The first generated variable is a more sensitive smoking measure, which is 1 for current smokers, 0 for subjects who never smoked and $1/(currentAge - ageQuitSmoking)$ for subjects who used to smoke. If the quitting age was equal to the current age the variable was coded as 1. This variable is meant to capture any lingering effects of smoking, which are assumed to dissipate over time.

The other two variables created for analysis were scales meant to roughly measure general health and heart health. To generate these scales, I summed the responses of a variety of 'yes/no' questions asking subjects to report if they had ever been diagnosed with particular disease, for example, diabetes, asthma, cancer, or emphysema (See appendix for complete list). The summation yielded a variable, which I call the general health scale that ranged from 0 (no diagnoses) to 16 (diagnosed with everything). Similarly, a heart health scale was generated by summing only the questions relating to heart diseases, for example, hypertension, heart attack, and angina. The heart health scale ranges from 0 to 5, however no subject had a value greater than 3.

Models and Model Selection:

Two primary models were employed in this analysis. The first was simply least squares regression for continuous outcomes using sampling weights and clustering to adjust for underrepresented individuals (e.g. Mexican-Americans, young people). The outcome for the continuous models was GFR rather than CKD stage. This type of modeling is preferable for several reasons. First, the CKD stage is defined almost exclusively with GFR level anyhow, and information can only be *lost* by categorizing an inherently continuous variable. Second, there are only 33 subjects in stages 4 through 6 – far too few to make valid inferences for those stages, but using a continuous outcome allows us to ignore this problem, and the ensuing problem of binning the subjects, entirely.

The primary drawback of using a continuous outcome in this context is that physicians do not necessarily evaluate and diagnose CKD using small gradations in GFR level, in which case a stage interpretation is more appropriate. To produce a more clinically relevant model I used a continuation ratio framework. Continuation ratio (CR) models are useful when an outcome variable is ordinal and represents

an event that moves from one level to the next, for instance, from stage 0 to stage 1. In contrast to a multinomial model with an ordinal outcome, CR allows us to model the probability of advancement from one stage to the next, which is extremely valuable to clinicians.

The outcome for the CR models was CKD stage, however the categories were collapsed due to so few subjects residing in the advanced stages. Instead of seven stages levels, I made three: 0 for no disease (stage 0), 1 for stages 1 and 2, and 2 for stages 3 through 6.

Since the predictors of CKD are not well understood, I applied a best-subsets variable selection algorithm to inform the final model. In short, best-subsets selection fits most possible models subject to certain constraints such as forcing covariates in or out, or a cap on the maximum number of covariates in the model. The procedure typically produces some sort of score or ranking with which to judge the models. I used the adjusted R^2 as the selection criteria.

All analysis was done in the statistical program R version 2.5-1 using Mac OS X version 10.4.11. The packages used include 'survey,' 'hexbin,' 'leaps,' and 'design.'

Results:

Below, table 1 displays the sample sizes and means for demographic variables and relevant predictors broken down by the collapsed CKD stage.

Table 1. Summary Statistics and Sample Sizes for Demographic and Predictor Variables

			Binned CKD Level			
			No Disease	Stages 1 & 2	Stages > 2	Total
Demographics						
	Age	MEAN, *SD*	47.9yr, *18yr*	59yr, *18.6yr*	74yr, *10yr*	51.4yr, *19.2yr*
		N	3342	387	375	4104
	Sex	%	52% Male	53% Male	41% Male	51% Male
		N	3342	387	375	4104
	Income	MEAN, SD	40K/yr, *12.5K/ yr*	30K/yr, *12.5K/yr*	30K/yr, *12.5K/yr*	40K/yr, *12.5K/ yr*
		N	3140	358	349	3847
	Marrital Status	%	62% Married	58% Married	48% Married	60% Married
		N	3340	387	375	4102
Predictors						
	Diabetes Status	%	9% Have Diab.	30% Have Diab.	30% Have Diab.	12% Have Diab.
		N	3341	387	375	4103
	General Health	MEAN, *SD*	1.38, *1.51*	2.13, *1.95*	3.32, *2.20*	1.61, *1.72*
		N	3226	368	346	3940
	Heart Health	MEAN, *SD*	0.08, *0.34*	0.23, *0.60*	0.44, *0.76*	0.12, *0.44*
		N	3318	381	365	4064
	Cancer Status	%	8% Diagnosed	11% Diagnosed	22% Diagnosed	10% Diagnosed
		N	3337	383	374	4094
	Smoking Weight	MEAN, *SD*	0.28, *0.43*	0.30, *0.43*	0.13, *0.30*	0.27, *0.42*
		N	3267	381	373	4021

Table 1. Summary statistics and sample sizes for demographic and relevant explanatory variables. Units indicated where needed.

Several predictors stand out as potentially informative in predicting CKD. In particular, diabetes status, cancer status, heart health scale, and general health scale which all increase with CKD stage,

meaning we observe more diabetes, more cancer, poorer heart health and poorer general health as CKD stage increases.

Additionally, the demographic means suggest that age needs to be controlled for in the model because it increases sharply with CKD stage. That is, older subjects are much more likely to have CKD than younger subjects – an unsurprising result.

The results of the best-subsets model selection implementation are displayed below in figure 1. The procedure was done twice, once for lab measures only and a second time for both lab measures and all other measures of interest. Any type of variable selection procedure is subject to bias and can produce results that, while statistically significant, have no biological relevance. Therefore, the final models contain variables not selected by the procedure, but that are well known to be associated with CKD (diabetes in particular).

Despite the shortcomings of the variable selection, it is very useful for selecting laboratory measures of which we have no *a priori* information on what should be included.

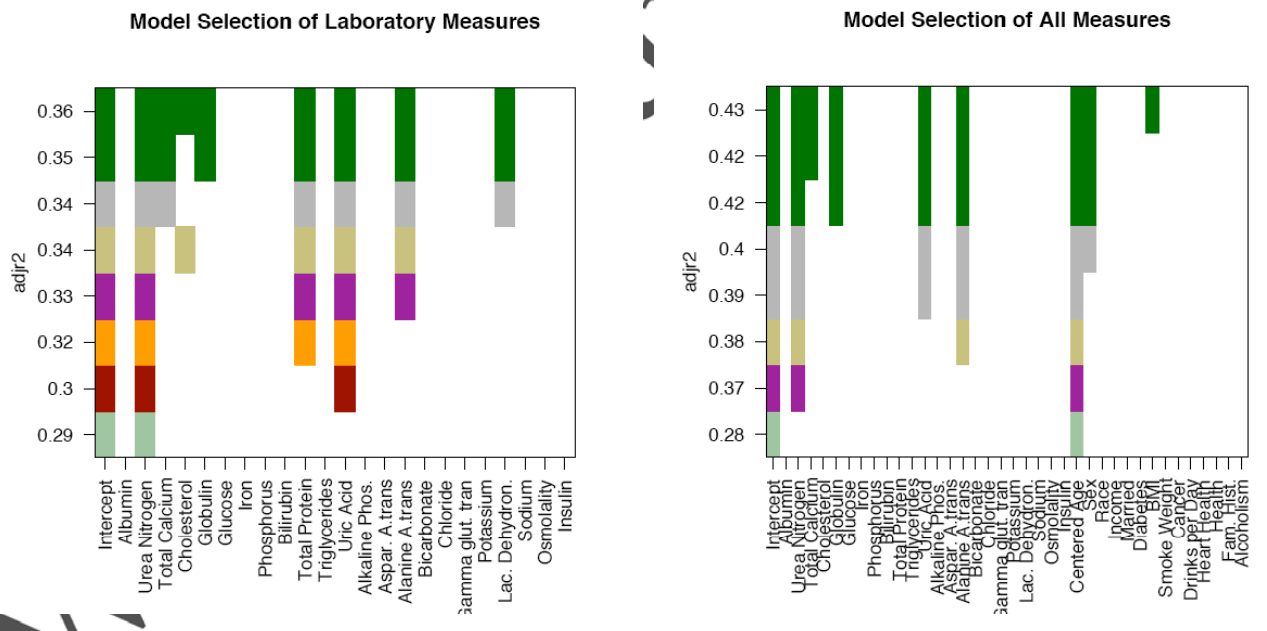


Figure 1. Results of Model selection procedure. The x-axis is a list of covariates considered, and the y-axis is the adjusted R^2. Each tick-mark on the y-axis is a different model containing the covariates corresponding to the colored regions. The difference in colors suggests marked changes in adjusted R^2.

Model output, including coefficient estimates, standard errors and the corresponding significance levels for both the continuous outcome models and CR models are located in appendix tables 2 and 3 respectively.

The final model for the continuous outcome included centered age in years, sex, diabetes status, the heart health scale, blood urea nitrogen, total calcium, globulin, uric acid and alanine aminotranferase (units reported in tables). Age, being female, having poor heart health, high levels of blood urea nitrogen, high levels of total calcium and high levels of uric acid were all associated with lower GFR and thus with CKD. Alanine aminotransferase is a measure of liver function and its elevation suggests liver dysfunction. Uric acid is a waste product collected from blood and its elevation can suggest many diseases including diabetes, and perhaps CKD. Elevated blood urea nitrogen is a known risk-factor for CKD. This model has an un-weighted, adjusted R^2 of .48 suggesting that it explains approximately 48% of the variation in the data. The directionality of diabetes is troubling, because it is well known from previous literature that diabetes is a strong predictor CKD. In this case, there is probably selection bias because individuals with diabetes *and* CKD are likely too ill to participate in the study, so individuals with diabetes in the earlier CKD stages are overrepresented, which causes the coefficient estimate to be positive and appear protective.

The CR model results are similar to the results for the continuous outcome, however CR models have an alternate interpretation via odds ratios. The odds ratios can be found below in table 2. Age and all laboratory measures were centered by the mean, so the baseline is 0.

Condition	Odds Ratio	95% CI	p-value
Female	2.06	[1.69, 2.52]	<.0001
Diabetes, CKD Stage = 0	9.62	[4.03, 22.96]	<.0001
Diabetes, CKD Stage > 0	2.49	[1.69, 3.68]	<.0001
0.5 increase in smoke weight, CKD Stage = 0	2.17	[1.42, 3.31]	0.0003
0.5 increase in smoke weight, CKD Stage > 0	1.33	[1.11, 1.60]	0.0017
Current Smoker, CKD Stage = 0	4.71	[2.01, 11.008]	0.0003
Current Smoke, CKD Stage > 0	1.79	[1.24, 2.58]	0.0017
5 Year Increase In Age	1.24	[1.21, 1.27]	<.0001
5 Unit Increase In Blood Urea Nitrogen	5.39	[3.43, 8.48]	<.0001
10 Unit Increase In Globulin, CKD Stage = 0	1.90	[1.59, 2.27]	<.0001
10 Unit Increase In Globulin, CKD Stage > 0	6.99	[4.53, 10.80]	<.0001
10 Unit Increase In Uric Acid	1.04	[1.02, 1.06]	<.0001
Advancement in Risk Stage, all interactions = 0	3.57	[2.67, 4.78]	<.0001
Advancement in Risk Stage, With Diabetes	0.92	[0.61, 1.39]	0.72

Table 2. Odds ratios, 95% confidence intervals and p-values for various contrasts or conditions in the CR final model. Age and lab measures are centered by their means and set to zero for comparisons.

Again there appears to be a protective effect for diabetes (OR goes from 9.62 to 2.49 with stage advancement). Smoking also appears to have a protective effect, but is likely subject to the same type of selection bias. The odds of advancement for a 5-year increase in age increase 124%, which recapitulates the directionality of the findings of the continuous outcome model. Uric acid is interesting because it does not markedly deviate from the null odds ratio.

It is also possible to model the probability of advancement with the CR models using predicted probabilities. Figure 2 shows the change in predicted probability of advancement over age for CKD stage equal to zero, and CKD stage greater than zero.

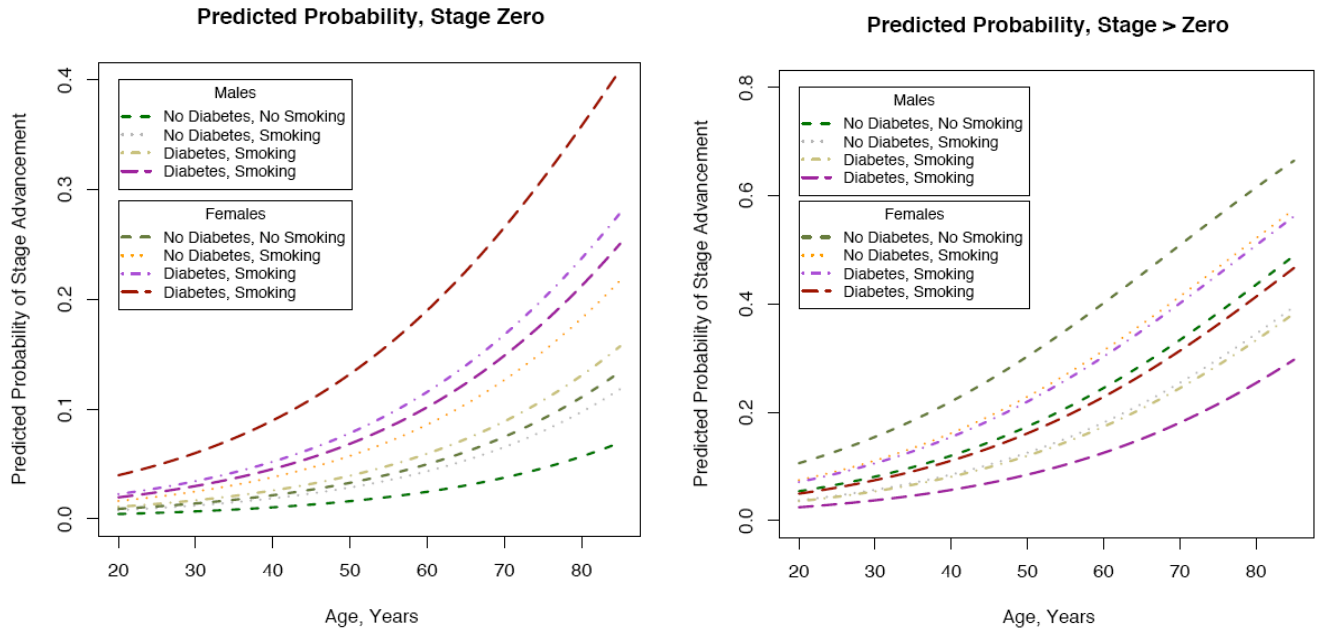


Figure 2. Plots of predicted probabilities of advancement over age. The left panel contains predicted probabilities for individuals with no disease (CKD stage = 0), and the right panel contains predicted probabilities for individuals with CKD stage > 0. Predicted probabilities are higher for women than men, in both panels. Age and lab measures are centered by their means and set to zero for comparisons. **Note: the y-axis varies between panels.**

It is interesting to note that the predicted probabilities are higher for women than for men, and particularly for women who smoke and have diabetes in the left panel.

Discussion:

The current study has many shortcomings the most obvious of which is the loss in sensitivity due to combining the CKD stages into three rather than seven groups. It is not clear that the CKD severity changes uniformly from stage to stage, and if it does not the combining of stages is not as simple as was treated in this analysis.

The issue of selection bias is also inescapable in this analysis. It is clear that individuals in later stages of CKD are not contributing to NHANES at the same rate as healthier people. The effect of this is to make covariates like diabetes and smoking appear to have protective effects. For future study, verification of effect sizes could be done using clinical data, which would likely be subject to bias in the opposite direction (very sick people overrepresented). The combination of these approaches would hopefully yield a fuller picture of the true effect sizes.

The continuous outcome model yields more sensitivity to predict small changes in GFR and explore covariates whose effect on GFR may be subtle, but it has the obvious drawback of being complicated to interpret from a clinical point-of-view. Therefore the CR models may be more clinically relevant. The predicted probabilities could be of particular interest to a clinician since they allow for immediate application in predicting CKD advancement.

Future research might expand analysis to more years of the NHANES set. Also, exploration of the gender difference could be very valuable in treating and ultimately preventing chronic kidney disease.

References:

Widmaier Eric P., Hershel Raff, and Kevin T. Strang. "Vander's Human Physiology: The Mechanisms of Body Function." 2006. McGraw-Hill Companies: New York, NY.

Appendix A:

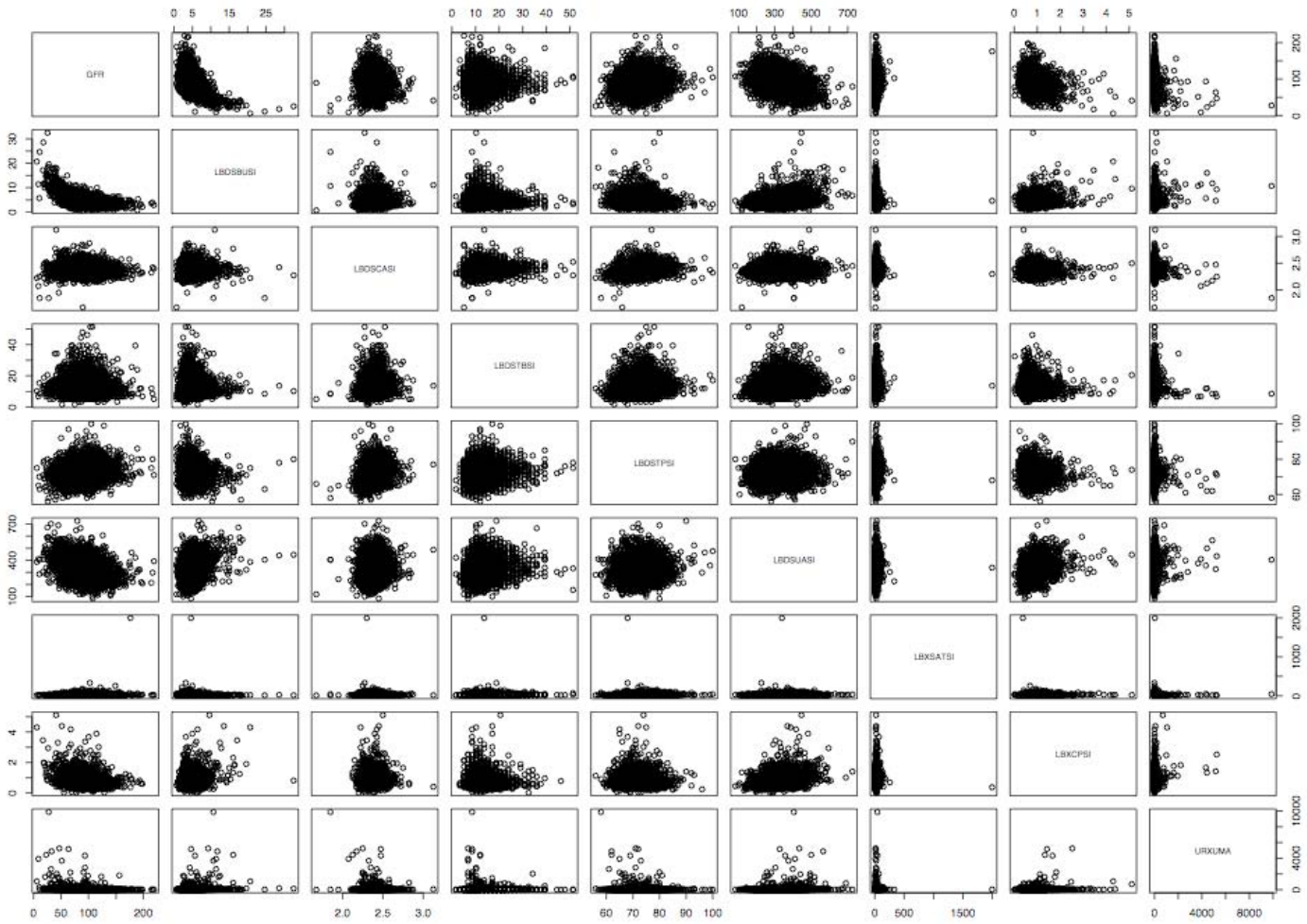
Appendix Table 1.

Table 1. Continuous, univariate Regression Models

Variable	Continuous Regression Models												
Intercept	124.823***	93.957***	95.851***	92.364***	86.398***	92.830***	95.218***	85.535***	93.147***	89.325***	92.934***	97.541***	92.459***
	(0.833)	*(0.932)*	*(1.181)*	*(1.523)*	*(0.971)*	*(0.944)*	*(1.993)*	*(0.790)*	*(0.783)*	*(1.073)*	*(0.693)*	*(0.672)*	*(1.283)*
Age	-0.709***	---	---	---	---	---	---	---	---	---	---	---	---
	(0.017)	---	---	---	---	---	---	---	---	---	---	---	---
Sex		-1.421	---	---	---	---	---	---	---	---	---	---	---
		(0.706)	---	---	---	---	---	---	---	---	---	---	---
Race			-1.331**	---	---	---	---	---	---	---	---	---	---
			(0.363)	---	---	---	---	---	---	---	---	---	---
Income				-0.084	---	---	---	---	---	---	---	---	---
				(0.149)	---	---	---	---	---	---	---	---	---
Marital Status					2.697***	---	---	---	---	---	---	---	---
					(0.287)	---	---	---	---	---	---	---	---
Diabetes Status						-7.571**	---	---	---	---	---	---	---
						(2.480)	---	---	---	---	---	---	---
BMI							-0.122	---	---	---	---	---	---
							(0.072)	---	---	---	---	---	---
Smoke Weight								11.697***	---	---	---	---	---
								(0.958)	---	---	---	---	---
Cancer Status									-15.340***	---	---	---	---
									(1.175)	---	---	---	---
Alcohol										1.556***	---	---	---
										(0.208)	---	---	---
Heart Health											-11.472***	---	---
											(0.716)	---	---
General Health												-3.750***	---
												(0.194)	---
Family History													-0.110
													(0.369)

Significance Codes: '***' 0.001, '**' 0.01, '*' 0.05

Results from univariate regressions with continuous outcomes.



Scatter plot matrix of GFR versus each lab measure.

Appendix Table 2. Output from continuous outcome models. The final model is denoted by the green shading. Age and laboratory measures are centered by their means.

Variable	Continuous Outcome Models		
	MODEL 1	***MODEL 2***	***MODEL 3***
Intercept	143.16**	99.02***	147.93***
	(10.55)	*(0.97)*	*(8.75)*
Centered Age	-0.5***	-0.54***	-0.52***
Years	(0.01)	*(0.01)*	(0.01)
Sex	-7.39*	-6.85***	-6.79***
Male = 1, Female = 2	(0.77)	(0.55)	(0.60)
Diabetes Status	4.37*	3.33**	3.20**
No = 0, Yes = 1	(1.11)	(0.79)	(0.87)
Smoke Weight	1.88	---	---
Never = 0, Current = 1	(0.94)	---	---
Cancer Status	-2.67	---	-2.19
No = 0, Yes = 1	(1.63)	---	(1.36)
Alcohol	-0.53	---	---
5 Drinks per Day	(0.80)	---	---
Heart Health	-1.71	-1.07	-0.99
Range 0 - 3	(0.73)	(0.55)	(0.54)
General Health	0.42	---	---
Range 0 -16	(0.20)	---	---
Blood Urea Nitrogen	-3.08**	-3.12***	-3.11***
mmol/L	(0.21)	(0.23)	(0.22)
Total Calcium	-13.41	-14.66*	-14.7**
mmol/L	(4.76)	(4.04)	(4.22)
Globulin	0.61**	0.61***	0.61***
g/L	(0.06)	(0.07)	(0.07)
Uric Acid	-0.05**	-0.05***	-0.05***
umol/L	(0.006)	(0.004)	(0.004)
Alanine Aminotrans	0.05	0.05*	0.05**
U/L	(0.01)	(0.01)	(0.01)
N	3092	4058	4050
AIC	27040	35770	35690
Unweight R2	0.48	0.48	0.48

Significance Codes: '***' 0.001, '**' 0.01, '*' 0.05

Appendix Table 3. Output from continuous outcome models. The final model is denoted by the green shading. Age and laboratory measures are centered by their means.

Variable	Continuous Ratio Models			
	MODEL 1	***MODEL 2***	***MODEL 3***	***MODEL 4***
Intercept	3.96	3.91**	3.88***	4.67***
	(0.38)	*(0.29)*	*(0.27)*	*(0.37)*
Centered Age	-0.04*	-0.04***	-0.04***	-0.04***
Years	*(0.004)*	*(0.002)*	*(0.002)*	*(0.002)*
Sex	-0.60	-0.71***	-0.77***	-0.72**
Male = 1, Female = 2	*(0.18)*	*(0.12)*	*(0.10)*	*(0.10)*
Diabetes Status	-0.61	-0.56**	-0.56**	-2.26**
No = 0, Yes = 1	*(0.17)*	*(0.13)*	*(0.15)*	*(0.44)*
Smoke Weight	-0.53	---	-0.41	-1.55*
Never = 0, Current = 1	*(0.19)*	---	*(0.18)*	*(0.43)*
Cancer Status	0.30	---	---	---
No = 0, Yes = 1	*(0.19)*	---	---	---
Alcohol	0.19	---	---	---
5 Drinks per Day	*(0.16)*	---	---	---
Heart Health	0.006	-0.16	---	---
Range 0 - 3	*(0.13)*	*(0.10)*	---	---
General Health	-0.05	---	---	---
Range 0 -16	*(0.04)*	---	---	---
Blood Urea Nitrogen	-0.32	-0.31***	-0.33***	-0.33**
mmol/L	*(0.05)*	*(0.04)*	*(0.04)*	*(0.04)*
Total Calcium	-0.01	0.11	---	---
mmol/L	*(0.69)*	*(0.57)*	---	---
Globulin	-0.04	-0.03*	-0.03**	-0.19***
g/L	*(0.008)*	*(0.008)*	*(0.008)*	*(0.02)*
Uric Acid	-0.004	-0.004**	-0.004**	-0.004**
umol/L	*(0.001)*	*(0.001)*	*(0.0009)*	*(0.001)*
Alanine Aminotrans	0.002	0.001	---	---
U/L	*(0.003)*	*(0.003)*	---	---
Diab*Risk Set	---	---	---	1.34**
	---	---	---	*(0.29)*
Smoke*Risk Set	---	---	---	0.96
	---	---	---	*(0.39)*
Globulin*Risk Set	---	---	---	0.13**
	---	---	---	*(0.01)*
Risk Set	-0.67	-0.73***	-0.72***	-1.27***
	(0.13)	*(0.09)*	*(0.09)*	*(0.14)*
N	3635	4803	4768	4768
AIC	30	22	18	24

Significance Codes: '***' 0.001, '**' 0.01, '*' 0.05

Appendix B:

```
########################
#BIOS 699 - Project One#
#Data Preparation      #
########################

#Set directory, read in data
setwd('/Users/***/Desktop/Class/BIOSTAT699/Project 1/NHANES')
mergedData <- read.delim('ALL_DATA.txt', header = TRUE, sep = '\t')
smokeData <- read.delim('Smoke.txt', header = TRUE, sep = '\t')
mergedData <- merge(mergedData, smokeData, by = 'SEQN')

###################################
##Generating variables for analysis##
###################################
#Make binary outcome
mergedData$binCKD <- ifelse(mergedData$STAGE == 0, 0, 1)

#Combine stages
mergedData$STAGE1 <- ifelse(mergedData$STAGE == 0, 0, 0)
mergedData$STAGE2 <- ifelse(mergedData$STAGE == 1 | mergedData$STAGE == 2, 1, mergedData$STAGE1)
mergedData$STAGECOL <- ifelse(mergedData$STAGE == 3 | mergedData$STAGE == 4 | mergedData$STAGE == 5 |
mergedData$STAGE == 6, 2, mergedData$STAGE2)
#> table(mergedData$STAGECOL)
#
#   0    1    2
#3342  387  375

#Family history scale
#Use five yes/no questions about family history of disease and summed yes responses.
#Diseases are Alzheimer's, asthma, osteoporosis, hypertension/stroke, and angina, respectively
#Scale ranges from 0 - 5, with 0 = no fam hist and 5 = all 5 in fam hist
#First remove '9' entries
mergedData$MCQ250B <- ifelse(mergedData$MCQ250B == 9, NA, mergedData$MCQ250B)
mergedData$MCQ250C <- ifelse(mergedData$MCQ250C == 9, NA, mergedData$MCQ250C)
mergedData$MCQ250E <- ifelse(mergedData$MCQ250E == 9, NA, mergedData$MCQ250E)
mergedData$MCQ250F <- ifelse(mergedData$MCQ250F == 9, NA, mergedData$MCQ250F)
mergedData$MCQ250G <- ifelse(mergedData$MCQ250G == 9, NA, mergedData$MCQ250G)
mergedData$famHistScale <- (mergedData$MCQ250B + mergedData$MCQ250C + mergedData$MCQ250E + mergedData$MCQ250F +
mergedData$MCQ250G) - 5

#Heart Health Scale, 0 = no heart disease, 3 = bad heart disease
#Sum up heart disease diagnosis questions
mergedData$MCQ160B <- ifelse(mergedData$MCQ160B == 9, NA, mergedData$MCQ160B)
mergedData$MCQ160C <- ifelse(mergedData$MCQ160C == 9, NA, mergedData$MCQ160C)
mergedData$MCQ160D <- ifelse(mergedData$MCQ160D == 9, NA, mergedData$MCQ160D)
mergedData$MCQ160B <- ifelse(mergedData$MCQ160B == 2, 0, mergedData$MCQ160B)
mergedData$MCQ160C <- ifelse(mergedData$MCQ160C == 2, 0, mergedData$MCQ160C)
mergedData$MCQ160D <- ifelse(mergedData$MCQ160D == 2, 0, mergedData$MCQ160D)
mergedData$heartHealth <- mergedData$MCQ160B + mergedData$MCQ160C + mergedData$MCQ160D

#Overall Health Scale, 0 = No diagnoses to 16
#Variables include weak kidneys, asthma, blood transfusion, trouble seeing,
#cancer, arthritis, overweight, CHF, coronary heart disease, angina, heart attack,
#stroke, emphysema, chronic bronchitis, liver condition, and thyroid problem.
mergedData$KIQ022 <- ifelse(mergedData$KIQ022 == 9, NA, mergedData$KIQ022)
mergedData$MCQ010 <- ifelse(mergedData$MCQ010 == 9, NA, mergedData$MCQ010)
mergedData$MCQ092 <- ifelse(mergedData$MCQ092 == 9, NA, mergedData$MCQ092)
mergedData$MCQ140 <- ifelse(mergedData$MCQ140 == 9, NA, mergedData$MCQ140)
mergedData$MCQ220 <- ifelse(mergedData$MCQ220 == 9, NA, mergedData$MCQ220)
mergedData$MCQ160A <- ifelse(mergedData$MCQ160A == 9, NA, mergedData$MCQ160A)
mergedData$MCQ160J <- ifelse(mergedData$MCQ160J == 9, NA, mergedData$MCQ160J)
mergedData$MCQ160E <- ifelse(mergedData$MCQ160E == 9 | mergedData$MCQ160E == 7, NA, mergedData$MCQ160E)
mergedData$MCQ160F <- ifelse(mergedData$MCQ160F == 9, NA, mergedData$MCQ160F)
```

```
mergedData$MCQ160G <- ifelse(mergedData$MCQ160G == 9, NA, mergedData$MCQ160G)
mergedData$MCQ160K <- ifelse(mergedData$MCQ160K == 9, NA, mergedData$MCQ160K)
mergedData$MCQ160L <- ifelse(mergedData$MCQ160L == 9, NA, mergedData$MCQ160L)
mergedData$MCQ160M <- ifelse(mergedData$MCQ160M == 9, NA, mergedData$MCQ160M)
#make 0 = no, 1 = yes.
mergedData$KIQ022 <- ifelse(mergedData$KIQ022 == 2, 0, mergedData$KIQ022)
mergedData$MCQ010 <- ifelse(mergedData$MCQ010 == 2, 0, mergedData$MCQ010)
mergedData$MCQ092 <- ifelse(mergedData$MCQ092 == 2, 0, mergedData$MCQ092)
mergedData$MCQ140 <- ifelse(mergedData$MCQ140 == 2, 0, mergedData$MCQ140)
mergedData$MCQ220 <- ifelse(mergedData$MCQ220 == 2, 0, mergedData$MCQ220)
mergedData$MCQ160A <- ifelse(mergedData$MCQ160A == 2, 0, mergedData$MCQ160A)
mergedData$MCQ160J <- ifelse(mergedData$MCQ160J == 2, 0, mergedData$MCQ160J)
mergedData$MCQ160E <- ifelse(mergedData$MCQ160E == 2, 0, mergedData$MCQ160E)
mergedData$MCQ160F <- ifelse(mergedData$MCQ160F == 2, 0, mergedData$MCQ160F)
mergedData$MCQ160G <- ifelse(mergedData$MCQ160G == 2, 0, mergedData$MCQ160G)
mergedData$MCQ160K <- ifelse(mergedData$MCQ160K == 2, 0, mergedData$MCQ160K)
mergedData$MCQ160L <- ifelse(mergedData$MCQ160L == 2, 0, mergedData$MCQ160L)
mergedData$MCQ160M <- ifelse(mergedData$MCQ160M == 2, 0, mergedData$MCQ160M)
#Scale
mergedData$healthScale <- mergedData$KIQ022 + mergedData$MCQ010 + mergedData$MCQ092 + mergedData$MCQ140 +
mergedData$MCQ220 + mergedData$MCQ160A + mergedData$MCQ160J + mergedData$MCQ160E + mergedData$MCQ160F +
mergedData$MCQ160G + mergedData$MCQ160K + mergedData$MCQ160L + mergedData$MCQ160M + mergedData$MCQ160B +
mergedData$MCQ160C + mergedData$MCQ160D

#Make cohabiting the same as being married
mergedData$DMDMARTL <- ifelse(mergedData$DMDMARTL == 6, 1, mergedData$DMDMARTL)

#Remove '7' and '9' characters
mergedData$INDHHINC <- ifelse(mergedData$INDHHINC == 77 | mergedData$INDHHINC == 99, NA, mergedData$INDHHINC)
mergedData$MCQ220 <- ifelse(mergedData$MCQ220 == 9, NA, mergedData$MCQ220)
mergedData$ALQ130 <- ifelse(mergedData$ALQ130 > 20, NA, mergedData$ALQ130)

#Center Age
mergedData$centeredAge <- mergedData$RIDAGEYR - mean(mergedData$RIDAGEYR)

#Center relevant lab measures so that making pred prob is easier
mergedData$centeredLBDSBUSI <- mergedData$LBDSBUSI - mean(na.omit(mergedData$LBDSBUSI))
mergedData$centeredLBDSCASI <- mergedData$LBDSCASI - mean(na.omit(mergedData$LBDSCASI))
mergedData$centeredLBDSGBSI <- mergedData$LBDSGBSI - mean(na.omit(mergedData$LBDSGBSI))
mergedData$centeredLBDSUASI <- mergedData$LBDSUASI - mean(na.omit(mergedData$LBDSUASI))
mergedData$centeredLBXSATSI <- mergedData$LBXSATSI - mean(na.omit(mergedData$LBXSATSI))

write.table(mergedData, 'ckdData.txt', sep = '\t')

######################
#BIOS 699 - Project 1 #
#Regression Models    #
######################

#Read in edited file
setwd('/Users/ers13/Desktop/Class/BIOSTAT699/Project 1/NHANES')
data <- read.delim('ckdData.txt', header = TRUE, sep = '\t')

require(survey)
require(hexbin)
require(leaps)
require(Design)

attach(data)

#NHANES uses a continuous cross-sectional survey design
#Stratified, multistage probability sample
#Dichotomous outcome variable = High_stage. 0 = no CKD, 1 = CKD.
#Scale outcome variable = SCALE
#Continuous outcome variable =
```

```
nhanes <- svydesign(ids = SDMVPSU, data = data, weights = WTMEC2YR, strata = SDMVSTRA, nest = TRUE)

#MODELS#

#Univariate, binomial models
unilogit1 <- svyglm(binCKD ~ RIDAGEYR, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit1)
unilogit2 <- svyglm(binCKD ~ RIAGENDR, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit2)
unilogit3 <- svyglm(binCKD ~ RIDRETH1, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit3)
unilogit4 <- svyglm(binCKD ~ INDHHINC, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit4)
unilogit5 <- svyglm(binCKD ~ DMDMARTL, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit5)
unilogit6 <- svyglm(binCKD ~ Diab, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit6)
unilogit7 <- svyglm(binCKD ~ BMXBMI, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit7)
unilogit8 <- svyglm(binCKD ~ smokeWeight, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit8)
unilogit9 <- svyglm(binCKD ~ MCQ220, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit9)
unilogit10 <- svyglm(binCKD ~ ALQ130, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit10)
unilogit11 <- svyglm(binCKD ~ heartHealth, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit11)
unilogit12 <- svyglm(binCKD ~ healthScale, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit12)
unilogit13 <- svyglm(binCKD ~ famHistScale, design = nhanes, family = binomial(link = 'logit'))
summary(unilogit13)

#Multivariate, binomial models
#Full model
logit1 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + RIDRETH1 + INDHHINC + DMDMARTL + Diab + BMXBMI + smokeWeight +
MCQ220 + ALQ130 + heartHealth + healthScale + famHistScale, design = nhanes, family = binomial(link = 'logit'))
summary(logit1)

#Reduced models
#Remove race because income probably accounts for it
logit2 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + INDHHINC + DMDMARTL + Diab + BMXBMI + smokeWeight + MCQ220 +
ALQ130 + heartHealth + healthScale + famHistScale, design = nhanes, family = binomial(link = 'logit'))
summary(logit2)

#Remove heart health because general health includes it
logit3 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + INDHHINC + DMDMARTL + Diab + BMXBMI + smokeWeight + MCQ220 +
ALQ130 + healthScale + famHistScale, design = nhanes, family = binomial(link = 'logit'))
summary(logit3)

#Remove family history, because general health probably includes it
logit4 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + INDHHINC + DMDMARTL + Diab + BMXBMI + smokeWeight + MCQ220 +
ALQ130 + healthScale, design = nhanes, family = binomial(link = 'logit'))
summary(logit4)

#Remove alcoholism
logit5 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + INDHHINC + DMDMARTL + Diab + BMXBMI + smokeWeight + MCQ220 +
healthScale, design = nhanes, family = binomial(link = 'logit'))
summary(logit5)

#Remove BMI because it's probably covered by diabetes status
logit6 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + INDHHINC + DMDMARTL + Diab + smokeWeight + MCQ220 + healthScale,
design = nhanes, family = binomial(link = 'logit'))
summary(logit6)
```

```
#Remove diabetes and cancer because it's included in the health measure
logit7 <- svyglm(binCKD ~ RIDAGEYR + RIAGENDR + INDHHINC + DMDMARTL + smokeWeight + healthScale, design =
nhanes, family = binomial(link = 'logit'))
summary(logit7)

#Continuous outcome models

#Univariate, continuous models
uni1 <- svyglm(GFR ~ RIDAGEYR, design = nhanes)
summary(uni1)
uni2 <- svyglm(GFR ~ RIAGENDR, design = nhanes)
summary(uni2)
uni3 <- svyglm(GFR ~ RIDRETH1, design = nhanes)
summary(uni3)
uni4 <- svyglm(GFR ~ INDHHINC, design = nhanes)
summary(uni4)
uni5 <- svyglm(GFR ~ DMDMARTL, design = nhanes)
summary(uni5)
uni6 <- svyglm(GFR ~ Diab, design = nhanes)
summary(uni6)
uni7 <- svyglm(GFR ~ BMXBMI, design = nhanes)
summary(uni7)
uni8 <- svyglm(GFR ~ smokeWeight, design = nhanes)
summary(uni8)
uni9 <- svyglm(GFR ~ MCQ220, design = nhanes)
summary(uni9)
uni10 <- svyglm(GFR ~ ALQ130, design = nhanes)
summary(uni10)
uni11 <- svyglm(GFR ~ heartHealth, design = nhanes)
summary(uni11)
uni12 <- svyglm(GFR ~ healthScale, design = nhanes)
summary(uni12)
uni13 <- svyglm(GFR ~ famHistScale, design = nhanes)
summary(uni13)

#########Lab Measures############
#Variable names
labLabels <- c('Intercept', 'Albumin', 'Urea Nitrogen', 'Total Calcium', 'Cholesterol', 'Globulin', 'Glucose',
'Iron', 'Phosphorus', 'Bilirubin', 'Total Protein', 'Triglycerides', 'Uric Acid', 'Alkaline Phos.', 'Aspar.
A.trans', 'Alanine A.trans', 'Bicarbonate', 'Chloride', 'Gamma glut. tran', 'Potassium', 'Lac. Dehydron.',
'Sodium', 'Osmolality', 'Insulin')

#Group all lab measures together for model selection.
labMeasures <- as.data.frame(cbind(GFR, LBDSALSI, LBDSBUSI, LBDSCASI, LBDSCHSI, LBDSGBSI, LBDSGLSI, LBDSIRSI,
LBDSPHSI, LBDSTBSI, LBDSTPSI, LBDSTRSI, LBDSUASI, LBXSAPSI, LBXSASSI, LBXSATSI, LBXSC3SI, LBXSCLSI, LBXSGTSI,
LBXSKSI, LBXSLDSI, LBXSNASI, LBXSOSSI, LBDINSI))

#Best subset selection and graph
labModels <- regsubsets(GFR ~., data = labMeasures)
pdf('labMeasureModelSelect.pdf')
plot.regsubsets(labModels, scale = 'adjr2', col = colors, main = 'Model Selection of Laboratory Measures', xlab
= 'Lab Measure', ylab = 'Adjusted R Squared', labels = labLabels)
dev.off()

#Further reduce sample to pare down measures
smallerLabMeasures <- as.data.frame(cbind(GFR, LBDSCHSI, LBDSGBSI, LBDSTPSI, LBXSBU, LBXSC3SI, LBXSCLSI, LBXSIR,
LBDINSI, LBXTC))

#Best subset selection and graph
smallerLabModels <- regsubsets(GFR ~., data = smallerLabMeasures)
plot.regsubsets(smallerLabModels, scale = 'adjr2', col=gray(seq(0, 0.9, length = 9)), main = 'Model Selection of
Laboratory Measures', xlab = 'Lab Measure', ylab = 'Adjusted R Squared', labels = c('Intercept', 'Cholesterol',
'Globulin', 'Total Protein', 'Blood Urea Nitrogen', 'Bicarbonate', 'Chloride', 'Iron, Refrig.', 'Insulin',
'Total Cholesterol'))
```

```
##########Models selection including lab and survey############
#Use all selected lab measures
#Use all survey measures

colors = c('darkgreen', 'darkgrey', 'darkkhaki', 'darkmagenta', 'darkolivegreen', 'darkorange', 'darkorchid',
'darkred', 'darkseagreen')

designMatrixLabels <- c('Intercept', 'Albumin', 'Urea Nitrogen', 'Total Calcium', 'Cholesterol', 'Globulin',
'Glucose', 'Iron', 'Phosphorus', 'Bilirubin', 'Total Protein', 'Triglycerides', 'Uric Acid', 'Alkaline Phos.',
'Aspar. A.trans', 'Alanine A.trans', 'Bicarbonate', 'Chloride', 'Gamma glut. tran', 'Potassium', 'Lac.
Dehydron.', 'Sodium', 'Osmolality', 'Insulin', 'Centered Age', 'Sex', 'Race', 'Income', 'Married', 'Diabetes',
'BMI', 'Smoke Weight', 'Cancer', 'Drinks per Day', 'Heart Health', 'Health', 'Fam. Hist.', 'Alcoholism')

designMatrix <- as.data.frame(cbind(GFR, LBDSALSI, LBDSBUSI, LBDSCASI, LBDSCHSI, LBDSGBSI, LBDSGLSI, LBDSIRSI,
LBDSPHSI, LBDSTBSI, LBDSTPSI, LBDSTRSI, LBDSUASI, LBXSAPSI, LBXSASSI, LBXSATSI, LBXSC3SI, LBXSCLSI, LBXSGTSI,
LBXSKSI, LBXSLDSI, LBXSNASI, LBXSOSSI, LBDINSI, centeredAge, RIAGENDR, RIDRETH1, INDHHINC, DMDMARTL, diagDIB,
BMXBMI, smokeWeight, MCQ220, ALQ130, heartHealth, healthScale, famHistScale, drink5everyday))

modelSelect <- regsubsets(GFR ~., data = designMatrix, really.big = TRUE)
summary.regsubsets(modelSelect)

pdf('allMeasureModelSelect.pdf')
plot.regsubsets(modelSelect, scale = 'adjr2', col = colors, main = 'Model Selection of All Measures', ylab =
'Adjusted R Squared', labels = designMatrixLabels)
dev.off()

##Final continuous models##
#LBDSGBSI could be included
finalModel1 <- lm(GFR ~ LBDSBUSI + LBDSCASI + LBDSGBSI + LBDSUASI + LBXSATSI + centeredAge + RIAGENDR + diagDIB
+ heartHealth)
finalSvyModel1 <- svyglm(GFR ~ LBDSBUSI + LBDSCASI + LBDSGBSI + LBDSUASI + LBXSATSI + centeredAge + RIAGENDR +
diagDIB + heartHealth, design = nhanes)
summary(finalModel1)
summary(finalSvyModel1)

finalModel2 <- lm(GFR ~ LBDSBUSI + LBDSCASI + LBDSGBSI + LBDSUASI + LBXSATSI + centeredAge + RIAGENDR + diagDIB
+ heartHealth + smokeWeight + MCQ220 + drink5everyday + healthScale)
finalSvyModel2 <- svyglm(GFR ~ LBDSBUSI + LBDSCASI + LBDSGBSI + LBDSUASI + LBXSATSI + centeredAge + RIAGENDR +
diagDIB + heartHealth + smokeWeight + MCQ220 + drink5everyday + healthScale, design = nhanes)
summary(finalModel2)
summary(finalSvyModel2)

finalModel3 <- lm(GFR ~ LBDSBUSI + LBDSCASI + LBDSGBSI + LBDSUASI + LBXSATSI + centeredAge + RIAGENDR + diagDIB
+ heartHealth + MCQ220)
finalSvyModel3 <- svyglm(GFR ~ LBDSBUSI + LBDSCASI + LBDSGBSI + LBDSUASI + LBXSATSI + centeredAge + RIAGENDR +
diagDIB + heartHealth + MCQ220, design = nhanes)
summary(finalModel3)
summary(finalSvyModel3)
#Did not find any significant interactions

#####Continuation Ratio Models######

crModelSetup <- cr.setup(STAGECOL)
crOutcome <- crModelSetup$y
crRiskSet <- crModelSetup$cohort

#Format variables to be the same length as the outcome.

crSEQN <- SEQN[crModelSetup$subs]
crMCQ220 <- MCQ220[crModelSetup$subs]
crDrink5Everyday <- drink5everyday[crModelSetup$subs]
crHealthScale <- healthScale[crModelSetup$subs]
crSDMVPSU <- SDMVPSU[crModelSetup$subs]
crWTMEC2YR <- WTMEC2YR[crModelSetup$subs]
crSDMVSTRA <- SDMVSTRA[crModelSetup$subs]
```

```
crCenteredAge <- centeredAge[crModelSetup$subs]
crRIDRETH1 <- RIDRETH1[crModelSetup$subs]
crSmokeWeight <- smokeWeight[crModelSetup$subs]
crRIAGENDR <- RIAGENDR[crModelSetup$subs]
crLBDSALSI <- LBDSALSI[crModelSetup$subs]
crLBDSBUSI <- LBDSBUSI[crModelSetup$subs]
crLBDSCASI <- LBDSCASI[crModelSetup$subs]
crLBDSCHSI <- LBDSCHSI[crModelSetup$subs]
crLBDSGBSI <- LBDSGBSI[crModelSetup$subs]
crLBDSGLSI <- LBDSGLSI[crModelSetup$subs]
crLBDSIRSI <- LBDSIRSI[crModelSetup$subs]
crLBDSTPSI <- LBDSTPSI[crModelSetup$subs]
crLBDSTRSI <- LBDSTRSI[crModelSetup$subs]
crLBDSUASI <- LBDSUASI[crModelSetup$subs]
crLBXSAPSI <- LBXSAPSI[crModelSetup$subs]
crLBXSASSI <- LBXSASSI[crModelSetup$subs]
crLBXSATSI <- LBXSATSI[crModelSetup$subs]
crLBXSC3SI <- LBXSC3SI[crModelSetup$subs]
crLBXSCA <- LBXSCA[crModelSetup$subs]
crLBXSCLSI <- LBXSCLSI[crModelSetup$subs]
crLBXSGTSI <- LBXSGTSI[crModelSetup$subs]
crLBXSKSI <- LBXSKSI[crModelSetup$subs]
crLBXSLDSI <- LBXSLDSI[crModelSetup$subs]
crLBXSNASI <- LBXSNASI[crModelSetup$subs]
crLBXSOSSI <- LBXSOSSI[crModelSetup$subs]
crLBDINSI <- LBDINSI[crModelSetup$subs]
crURXUMA <- URXUMA[crModelSetup$subs]
crLBDHDDSI <- LBDHDDSI[crModelSetup$subs]
crLBDTCSI <- LBDTCSI[crModelSetup$subs]
crLBXHDD <- LBXHDD[crModelSetup$subs]
crLBXTC <- LBXTC[crModelSetup$subs]
crBMXBMI <- BMXBMI[crModelSetup$subs]
crHeartHealth <- heartHealth[crModelSetup$subs]
crDiagDIB <- diagDIB[crModelSetup$subs]
#Centered Labs for PP plots
crCenteredLBDSBUSI <- centeredLBDSBUSI[crModelSetup$subs]
crCenteredLBDSCASI <- centeredLBDSCASI[crModelSetup$subs]
crCenteredLBDSGBSI <- centeredLBDSGBSI[crModelSetup$subs]
crCenteredLBDSUASI <- centeredLBDSUASI[crModelSetup$subs]
crCenteredLBXSATSI <- centeredLBXSATSI[crModelSetup$subs]
crSex <- ifelse(crRIAGENDR == 1, 1, 0)

crData <- as.data.frame(cbind(crSEQN, crOutcome, crSDMVPSU, crWTMEC2YR, crSDMVSTRA, crCenteredAge, crRIDRETH1,
crSmokeWeight, crLBDSALSI, crLBDSBUSI, crLBDSCASI, crLBDSCHSI, crLBDSGBSI, crLBDSGLSI, crLBDSIRSI, crLBDSTPSI,
crLBDSTRSI, crLBDSUASI, crLBXSAPSI, crLBXSASSI, crLBXSATSI, crLBXSC3SI, crLBXSCA, crLBXSCLSI, crLBXSGTSI,
crLBXSKSI, crLBXSLDSI, crLBXSNASI, crLBXSOSSI, crLBDINSI, crURXUMA, crLBDHDDSI, crLBDTCSI, crLBXHDD, crLBXTC,
crBMXBMI, crRIAGENDR, crRiskSet, crHeartHealth, crDiagDIB, crCenteredLBDSBUSI, crCenteredLBDSCASI,
crCenteredLBDSGBSI, crCenteredLBDSUASI, crCenteredLBXSATSI, crMCQ220, crDrink5Everyday, crHealthScale, crSex))

write.table(crData, 'crData.txt', sep = '\t')

crNhanes <- svydesign(ids = crSDMVPSU, data = crData, weights = crWTMEC2YR, strata = crSDMVSTRA, nest = TRUE)

####MODELS DERIVED FROM CONTINUOUS MODELS####

crFullModel <- svyglm(crOutcome ~ crCenteredLBDSBUSI + crCenteredLBDSCASI + crCenteredLBDSGBSI +
crCenteredLBDSUASI + crCenteredLBXSATSI + crCenteredAge + crRIAGENDR + crDiagDIB + crHeartHealth + crSmokeWeight
+ crMCQ220 + crDrink5Everyday + crHealthScale + crRiskSet, design = crNhanes, family = binomial(link = 'logit'),
data = crData)
summary(crFullModel)

crModel1 <- svyglm(crOutcome ~ crCenteredLBDSBUSI + crCenteredLBDSCASI + crCenteredLBDSGBSI + crCenteredLBDSUASI
+ crCenteredLBXSATSI + crCenteredAge + crRIAGENDR + crDiagDIB + crHeartHealth + crRiskSet, design = crNhanes,
family = binomial(link = 'logit'), data = crData)
summary(crModel1)
```

```r
#Remove insignificant variables
crModel2 <- svyglm(crOutcome ~ crCenteredLBDSBUSI + crCenteredLBDSGBSI + crCenteredLBDSUASI + crCenteredAge +
crRIAGENDR + crDiagDIB + crRiskSet, design = crNhanes, family = binomial(link = 'logit'), data = crData)
summary(crModel2)

#Add crSmokeWeight
crModel3 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet, design = crNhanes, family = binomial(link = 'logit'), data = crData)
summary(crModel3)

#Interaction CRMs
#Age, Not significant, p = 0.175894
crModel4 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crCenteredAge, design = crNhanes, family = binomial(link = 'logit'), data
= crData)
summary(crModel4)

#Gender, Not significant, p = 0.076251
crModel5 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crRIAGENDR, design = crNhanes, family = binomial(link = 'logit'), data =
crData)
summary(crModel5)

#smokeWeight, Significant, p = 0.036637
crModel6 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crSmokeWeight, design = crNhanes, family = binomial(link = 'logit'), data
= crData)
summary(crModel6)

#diagDIB, Significant, p = 0.003244
crModel7 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crDiagDIB, design = crNhanes, family = binomial(link = 'logit'), data =
crData)
summary(crModel7)

#LBDSUASI, Not Significant, p = 0.615712
crModel8 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crLBDSUASI, design = crNhanes, family = binomial(link = 'logit'), data =
crData)
summary(crModel8)

#LBDSGBSI, Significant, p = 0.00027
crModel9 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crLBDSGBSI, design = crNhanes, family = binomial(link = 'logit'), data =
crData)
summary(crModel9)

#LBDSBUSI, Significant, p = 0.015115
crModel10 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crLBDSBUSI, design = crNhanes, family = binomial(link = 'logit'), data =
crData)
summary(crModel10)

#diagDIB and smokeWeight, Significant, pDiagDIB = 0.005001, pSmokeWeight = 0.054489
crModel11 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crDiagDIB + crRiskSet:crSmokeWeight, design = crNhanes, family =
binomial(link = 'logit'), data = crData)
summary(crModel11)

#**diagDIB and smokeWeight and LBDSGBSI, Significant, pDiagDIB = 0.009799, pSmokeWeight = 0.070784, pLBDSGBSI =
0.002150
crModel12 <- svyglm(crOutcome ~ crCenteredLBDSBUSI + crCenteredLBDSGBSI + crCenteredLBDSUASI + crCenteredAge +
crRIAGENDR + crDiagDIB + crSmokeWeight + crRiskSet + crRiskSet:crDiagDIB + crRiskSet:crSmokeWeight +
crRiskSet:crCenteredLBDSGBSI, design = crNhanes, family = binomial(link = 'logit'), data = crData)
```

```
summary(crModel12)

#diagDIB, smokeWeight, LBDSGBSI and LBDSBUSI, Significant, pDiagDIB = 0.016919, pSmokeWeight = 0.157665,
pLBDSGBSI = 0.010077, #pLBDSBUSI = 0.079584
crModel13 <- svyglm(crOutcome ~ crLBDSBUSI + crLBDSGBSI + crLBDSUASI + crCenteredAge + crRIAGENDR + crDiagDIB +
crSmokeWeight + crRiskSet + crRiskSet:crDiagDIB + crRiskSet:crSmokeWeight + crRiskSet:crLBDSGBSI +
crRiskSet:crLBDSBUSI, design = crNhanes, family = binomial(link = 'logit'), data = crData)
summary(crModel13)

#Predicted Probability graphs, No Disease
#Initialize Variables
intercept = 4.673352; age = -0.043331; sex = -0.727316; diab = -2.264181; smokeWeight = -1.550302; diabRisk =
1.349766; smokeRisk = 0.966577; riskSet = -1.275103;
ageSeq = seq(from = 20, to = 85, by = 1)

legend1 = c('No Diabetes, No Smoking', 'No Diabetes, Smoking', 'Diabetes, Smoking', 'Diabetes, Smoking')

xlim <- range(ageSeq)
ylim  <- range(seq(0, .4, .01))
pdf('ppNoDisease.pdf')
plot(NA, xlim = xlim, ylim = ylim, xlab = 'Age, Years', ylab = 'Predicted Probability of Stage Advancement',
main = 'Predicted Probability, Stage Zero')
#MALE
ppm01 = 1 / (1 + exp((intercept + sex + age * (ageSeq - 53.7496917))))
ppm02 = 1 / (1 + exp((intercept + sex + smokeWeight + smokeRisk + age * (ageSeq - 53.7496917))))
ppm03 = 1 / (1 + exp((intercept + sex + diab + diabRisk + age * (ageSeq - 53.7496917))))
ppm04 = 1 / (1 + exp((intercept + sex + diab + diabRisk + smokeWeight + smokeRisk + age * (ageSeq -
53.7496917))))
lines(ppm01 ~ ageSeq, col = 'darkgreen', lty = 2, lwd = 2)
lines(ppm02 ~ ageSeq, col = 'darkgrey', lty = 3, lwd = 2)
lines(ppm03 ~ ageSeq, col = 'darkkhaki', lty = 4, lwd = 2)
lines(ppm04 ~ ageSeq, col = 'darkmagenta', lty = 5, lwd = 2)
legend(20, 0.40, legend = legend1, lwd = 2, lty = c(2, 3, 4, 5), col = c('darkgreen', 'darkgrey', 'darkkhaki',
'darkmagenta'), title = 'Males', cex = 0.8, merge = TRUE)
#FEMALE
ppf01 = 1 / (1 + exp((intercept + 2 * sex + age * (ageSeq - 53.7496917))))
ppf02 = 1 / (1 + exp((intercept + 2 * sex + smokeWeight + smokeRisk + age * (ageSeq - 53.7496917))))
ppf03 = 1 / (1 + exp((intercept + 2 * sex + diab + diabRisk + age * (ageSeq - 53.7496917))))
ppf04 = 1 / (1 + exp((intercept + 2 * sex + diab + diabRisk + smokeWeight + smokeRisk + age * (ageSeq -
53.7496917))))
lines(ppf01 ~ ageSeq, col = 'darkolivegreen', lty = 2, lwd = 2)
lines(ppf02 ~ ageSeq, col = 'darkorange', lty = 3, lwd = 2)
lines(ppf03 ~ ageSeq, col = 'darkorchid', lty = 4, lwd = 2)
lines(ppf04 ~ ageSeq, col = 'darkred', lty = 5, lwd = 2)
legend(20, 0.29, legend = legend1, lwd = 2, lty = c(2, 3, 4, 5), col = c('darkolivegreen', 'darkorange',
'darkorchid', 'darkred'), title = 'Females', cex = 0.8, merge = TRUE)
dev.off()

#Predicted Probability graphs, Stage > 0
#Initialize Variables
intercept = 4.673352; age = -0.043331; sex = -0.727316; diab = -2.264181; smokeWeight = -1.550302; diabRisk =
1.349766; smokeRisk = 0.966577; riskSet = -1.275103;
ageSeq = seq(from = 20, to = 85, by = 1)

legend1 = c('No Diabetes, No Smoking', 'No Diabetes, Smoking', 'Diabetes, Smoking', 'Diabetes, Smoking')

xlim <- range(ageSeq)
ylim  <- range(seq(0, 0.8, .01))
pdf('ppDisease.pdf')
plot(NA, xlim = xlim, ylim = ylim, xlab = 'Age, Years', ylab = 'Predicted Probability of Stage Advancement',
main = 'Predicted Probability, Stage > Zero')
#MALE
ppm01 = 1 / (1 + exp((intercept + sex + 2 * riskSet + age * (ageSeq - 53.7496917))))
ppm02 = 1 / (1 + exp((intercept + sex + 2 * riskSet + smokeWeight + 2 * smokeRisk + age * (ageSeq -
53.7496917))))
```

```
ppm03 = 1 / (1 + exp((intercept + sex + 2 * riskSet + diab + 2 * diabRisk + age * (ageSeq - 53.7496917))))
ppm04 = 1 / (1 + exp((intercept + sex + 2 * riskSet + diab + 2 * diabRisk + smokeWeight + 2 * smokeRisk + age *
(ageSeq - 53.7496917))))
lines(ppm01 ~ ageSeq, col = 'darkgreen', lty = 2, lwd = 2)
lines(ppm02 ~ ageSeq, col = 'darkgrey', lty = 3, lwd = 2)
lines(ppm03 ~ ageSeq, col = 'darkkhaki', lty = 4, lwd = 2)
lines(ppm04 ~ ageSeq, col = 'darkmagenta', lty = 5, lwd = 2)
legend(20, 0.80, legend = legend1, lwd = 2, lty = c(2, 3, 4, 5), col = c('darkgreen', 'darkgrey', 'darkkhaki',
'darkmagenta'), title = 'Males', cex = 0.8, merge = TRUE)
#FEMALE
ppf01 = 1 / (1 + exp((intercept + 2 * riskSet + 2 * sex + age * (ageSeq - 53.7496917))))
ppf02 = 1 / (1 + exp((intercept + 2 * riskSet + 2 * sex + smokeWeight + 2 * smokeRisk + age * (ageSeq -
53.7496917))))
ppf03 = 1 / (1 + exp((intercept + 2 * riskSet + 2 * sex + diab + 2 * diabRisk + age * (ageSeq - 53.7496917))))
ppf04 = 1 / (1 + exp((intercept + 2 * riskSet + 2 * sex + diab + 2 * diabRisk + smokeWeight + 2 * smokeRisk +
age * (ageSeq - 53.7496917))))
lines(ppf01 ~ ageSeq, col = 'darkolivegreen', lty = 2, lwd = 2)
lines(ppf02 ~ ageSeq, col = 'darkorange', lty = 3, lwd = 2)
lines(ppf03 ~ ageSeq, col = 'darkorchid', lty = 4, lwd = 2)
lines(ppf04 ~ ageSeq, col = 'darkred', lty = 5, lwd = 2)
legend(20, 0.59, legend = legend1, lwd = 2, lty = c(2, 3, 4, 5), col = c('darkolivegreen', 'darkorange',
'darkorchid', 'darkred'), title = 'Females', cex = 0.8, merge = TRUE)
dev.off()
```