Appendix B: Annotated R Code and Output

Sofi Sinozich & Rachael Jackson

12/8/2020

Our full code, including initial data wrangling and brief exploratory code is available on [GitHub](https://github.com/sofisinozich/SURV617_Final_Project) but we present the key modeling code here.

# Variable selection with glmnet (LASSO)

We tested three different variations, one unweighted, one weighted using the observation weights in glmnet and one with weights as a linear predictor. We use lambda.1se rather than lambda.min to achieve more shrinkage given our suspicion that some of the variables may not be important. In all cases is used for LASSO regression (rather than elastic-net or ridge).

library(glmnet)  
library(glmnetUtils)  
  
health\_lasso <- cv.glmnet(health\_status2 ~  
 race+ gender + scale(age) + region + time\_of\_day +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(weight) + scale(height)+   
 scale(rbc\_count) + scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin) +  
 examiner\_id\_factor,  
 data=full\_data,  
 family= binomial(link = "logit"))  
  
# Test with observation weights  
health\_lasso\_weighted <- cv.glmnet(health\_status2 ~  
 race+ gender + scale(age) + region + time\_of\_day +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(weight) + scale(height)+   
 scale(rbc\_count) + scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin) +  
 examiner\_id\_factor,  
 weights = WTPFHX6,  
 data=full\_data,  
 family= binomial(link = "logit"))  
  
# Test with weights as a predictor instead  
health\_lasso\_weightaspredictor <- cv.glmnet(health\_status2 ~  
 scale(WTPFHX6) +  
 race+ gender + scale(age) + region + time\_of\_day +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(weight) + scale(height)+   
 scale(rbc\_count) + scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin) +  
 examiner\_id\_factor,  
 data=full\_data,  
 family= binomial(link = "logit"))  
  
coef(health\_lasso,s="lambda.1se")

## 40 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 8.312967e-02  
## raceWhite .   
## raceBlack -2.868103e-01  
## raceHispanic .   
## raceOther .   
## genderFemale -7.822484e-02  
## genderMale 1.795093e-15  
## scale(age) -8.816821e-01  
## regionMidwest .   
## regionNortheast .   
## regionSouth .   
## regionWest 2.435347e-01  
## time\_of\_dayAfternoon .   
## time\_of\_dayEvening .   
## time\_of\_dayMorning .   
## scale(systolic\_bp) -7.255826e-01  
## scale(diastolic\_bp) -1.670305e-01  
## locomotion -1.312071e+00  
## paralysis -1.040775e+00  
## impairment -3.572276e-01  
## infection -8.654168e-01  
## scale(bmi) -7.107897e-01  
## scale(weight) .   
## scale(height) 1.033835e-01  
## scale(rbc\_count) .   
## scale(wbc\_count) -1.253656e-01  
## scale(plt\_count) -4.016050e-02  
## scale(hemoglobin) 6.290144e-02  
## examiner\_id\_factor3007 -1.020072e+00  
## examiner\_id\_factor3001 1.182264e+00  
## examiner\_id\_factor3002 -5.323954e+00  
## examiner\_id\_factor3003 .   
## examiner\_id\_factor3004 .   
## examiner\_id\_factor3005 -3.098443e+00  
## examiner\_id\_factor3006 1.464337e+00  
## examiner\_id\_factor3008 5.974641e-02  
## examiner\_id\_factor3009 2.651085e-01  
## examiner\_id\_factor3010 .   
## examiner\_id\_factor3011 1.140956e+00  
## examiner\_id\_factor3012 .

coef(health\_lasso\_weighted,s="lambda.1se")

## 40 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 2.702116e-01  
## raceWhite .   
## raceBlack -2.864264e-01  
## raceHispanic .   
## raceOther .   
## genderFemale -1.248026e-01  
## genderMale 6.655703e-15  
## scale(age) -8.219633e-01  
## regionMidwest .   
## regionNortheast .   
## regionSouth .   
## regionWest 2.519985e-01  
## time\_of\_dayAfternoon -4.573487e-02  
## time\_of\_dayEvening .   
## time\_of\_dayMorning .   
## scale(systolic\_bp) -7.301871e-01  
## scale(diastolic\_bp) -1.539329e-01  
## locomotion -1.256039e+00  
## paralysis -8.517151e-01  
## impairment .   
## infection -1.008989e+00  
## scale(bmi) -7.660273e-01  
## scale(weight) .   
## scale(height) 1.059385e-01  
## scale(rbc\_count) .   
## scale(wbc\_count) -1.545446e-01  
## scale(plt\_count) -4.820895e-03  
## scale(hemoglobin) .   
## examiner\_id\_factor3007 -1.094242e+00  
## examiner\_id\_factor3001 7.455160e-01  
## examiner\_id\_factor3002 -5.043741e+00  
## examiner\_id\_factor3003 .   
## examiner\_id\_factor3004 .   
## examiner\_id\_factor3005 -3.629661e+00  
## examiner\_id\_factor3006 1.117377e+00  
## examiner\_id\_factor3008 -4.306518e-02  
## examiner\_id\_factor3009 1.190106e-01  
## examiner\_id\_factor3010 .   
## examiner\_id\_factor3011 1.032964e+00  
## examiner\_id\_factor3012 1.350073e-01

coef(health\_lasso\_weightaspredictor,s="lambda.1se")

## 41 x 1 sparse Matrix of class "dgCMatrix"  
## 1  
## (Intercept) 7.979564e-02  
## scale(WTPFHX6) .   
## raceWhite .   
## raceBlack -2.780908e-01  
## raceHispanic .   
## raceOther .   
## genderFemale -7.418928e-02  
## genderMale 2.056435e-15  
## scale(age) -8.731892e-01  
## regionMidwest .   
## regionNortheast .   
## regionSouth .   
## regionWest 2.350444e-01  
## time\_of\_dayAfternoon .   
## time\_of\_dayEvening .   
## time\_of\_dayMorning .   
## scale(systolic\_bp) -7.213075e-01  
## scale(diastolic\_bp) -1.596485e-01  
## locomotion -1.280293e+00  
## paralysis -9.515009e-01  
## impairment -2.793479e-01  
## infection -8.348979e-01  
## scale(bmi) -7.033813e-01  
## scale(weight) .   
## scale(height) 1.003016e-01  
## scale(rbc\_count) .   
## scale(wbc\_count) -1.208644e-01  
## scale(plt\_count) -3.791377e-02  
## scale(hemoglobin) 6.090479e-02  
## examiner\_id\_factor3007 -1.008702e+00  
## examiner\_id\_factor3001 1.157728e+00  
## examiner\_id\_factor3002 -5.229638e+00  
## examiner\_id\_factor3003 .   
## examiner\_id\_factor3004 .   
## examiner\_id\_factor3005 -3.052785e+00  
## examiner\_id\_factor3006 1.418461e+00  
## examiner\_id\_factor3008 5.095997e-02  
## examiner\_id\_factor3009 2.515295e-01  
## examiner\_id\_factor3010 .   
## examiner\_id\_factor3011 1.119024e+00  
## examiner\_id\_factor3012 .

# Evaluate the impact of examiner ID further

Even though the results of the regularization seem to suggest that we keep it, we did additional checks to gather additional evidence for an examiner-specific effect. We also double-check the usefulness of weights as a linear predictor based on the LASSO output.

We also adjusted two of our categorical variables to include only dummies for one of the categories based on the previous output, replacing race with black and region with west. We also use unscaled age for better interpretability.

# Base model - using the selected variables but NO examiner effect  
health\_model <- glm(health\_status2 ~  
 black + gender + age + west +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(height) +  
 scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin),  
 data=full\_data,  
 family = binomial(link = "logit"))  
  
# Base model limited to those that have examiner IDs  
health\_model\_lim <- glm(health\_status2 ~  
 black + gender + age + west +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(height) +  
 scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin),  
 data=subset(full\_data,!is.na(examiner\_id\_factor)),  
 family = binomial(link = "logit"))  
  
# Base model WITH examiner  
health\_model\_examiner <- glm(health\_status2 ~  
 black + gender + age + west +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(height) +  
 scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin) +  
 examiner\_id\_factor,  
 data=full\_data,  
 family = binomial(link = "logit"))  
  
# Double-check whether weights would be useful as predictors  
health\_model\_examiner\_weightaspredictor <- glm(health\_status2 ~  
 scale(WTPFHX6) +  
 black + gender + age + west +  
 scale(systolic\_bp) + scale(diastolic\_bp) +  
 locomotion + paralysis + impairment + infection +  
 scale(bmi) + scale(height) +  
 scale(wbc\_count) + scale(plt\_count) + scale(hemoglobin) +  
 examiner\_id\_factor,  
 data=full\_data,  
 family = binomial(link = "logit"))  
  
AIC(health\_model,health\_model\_lim,health\_model\_examiner,health\_model\_examiner\_weightaspredictor)

## Warning in AIC.default(health\_model, health\_model\_lim, health\_model\_examiner, :  
## models are not all fitted to the same number of observations

## df AIC  
## health\_model 16 15844.69  
## health\_model\_lim 16 15551.36  
## health\_model\_examiner 27 10649.60  
## health\_model\_examiner\_weightaspredictor 28 10651.35

BIC(health\_model,health\_model\_lim,health\_model\_examiner,health\_model\_examiner\_weightaspredictor)

## Warning in BIC.default(health\_model, health\_model\_lim, health\_model\_examiner, :  
## models are not all fitted to the same number of observations

## df BIC  
## health\_model 16 15967.24  
## health\_model\_lim 16 15673.58  
## health\_model\_examiner 27 10855.85  
## health\_model\_examiner\_weightaspredictor 28 10865.23

# Note that warning is tripped by the health\_model (health\_model\_lim and all others have same n)  
  
summary(health\_model\_examiner)

##   
## Call:  
## glm(formula = health\_status2 ~ black + gender + age + west +   
## scale(systolic\_bp) + scale(diastolic\_bp) + locomotion + paralysis +   
## impairment + infection + scale(bmi) + scale(height) + scale(wbc\_count) +   
## scale(plt\_count) + scale(hemoglobin) + examiner\_id\_factor,   
## family = binomial(link = "logit"), data = full\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0775 -0.4948 -0.0664 0.5152 3.7968   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.176264 0.099760 11.791 < 2e-16 \*\*\*  
## black -0.392087 0.061410 -6.385 1.72e-10 \*\*\*  
## genderMale 0.134344 0.078399 1.714 0.08661 .   
## age -0.050228 0.001777 -28.265 < 2e-16 \*\*\*  
## west 0.339757 0.061570 5.518 3.43e-08 \*\*\*  
## scale(systolic\_bp) -0.784110 0.044033 -17.807 < 2e-16 \*\*\*  
## scale(diastolic\_bp) -0.258453 0.033723 -7.664 1.80e-14 \*\*\*  
## locomotion -1.720242 0.169417 -10.154 < 2e-16 \*\*\*  
## paralysis -2.621998 0.671491 -3.905 9.43e-05 \*\*\*  
## impairment -1.471287 0.619688 -2.374 0.01759 \*   
## infection -1.209171 0.184963 -6.537 6.26e-11 \*\*\*  
## scale(bmi) -0.814427 0.029992 -27.155 < 2e-16 \*\*\*  
## scale(height) 0.138674 0.035060 3.955 7.64e-05 \*\*\*  
## scale(wbc\_count) -0.184620 0.028338 -6.515 7.28e-11 \*\*\*  
## scale(plt\_count) -0.065005 0.027118 -2.397 0.01652 \*   
## scale(hemoglobin) 0.086083 0.033466 2.572 0.01010 \*   
## examiner\_id\_factor3001 2.634557 0.112338 23.452 < 2e-16 \*\*\*  
## examiner\_id\_factor3002 -5.721645 0.310128 -18.449 < 2e-16 \*\*\*  
## examiner\_id\_factor3003 0.880568 0.273447 3.220 0.00128 \*\*   
## examiner\_id\_factor3004 1.137123 0.077881 14.601 < 2e-16 \*\*\*  
## examiner\_id\_factor3005 -2.507459 0.177204 -14.150 < 2e-16 \*\*\*  
## examiner\_id\_factor3006 3.177049 0.229830 13.823 < 2e-16 \*\*\*  
## examiner\_id\_factor3008 1.320059 0.092276 14.306 < 2e-16 \*\*\*  
## examiner\_id\_factor3009 1.579365 0.096483 16.369 < 2e-16 \*\*\*  
## examiner\_id\_factor3010 1.170671 0.271719 4.308 1.64e-05 \*\*\*  
## examiner\_id\_factor3011 2.566520 0.101771 25.219 < 2e-16 \*\*\*  
## examiner\_id\_factor3012 1.213634 0.089527 13.556 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 20570 on 15347 degrees of freedom  
## Residual deviance: 10596 on 15321 degrees of freedom  
## (2404 observations deleted due to missingness)  
## AIC: 10650  
##   
## Number of Fisher Scoring iterations: 8

# Interactions

We wanted to evaluate the usefulness of some interactions related to our research questions. We conclude that adding a interaction between the examiner and gender is worthwhile.

health\_model\_examiner\_intgender <- update(health\_model\_examiner,.~.+examiner\_id\_factor:gender)  
health\_model\_examiner\_intblack <- update(health\_model\_examiner,.~.+examiner\_id\_factor:black)  
health\_model\_examiner\_intblackgender <- update(health\_model\_examiner,.~.+examiner\_id\_factor:gender:black)  
  
AIC(health\_model\_examiner,health\_model\_examiner\_intgender,health\_model\_examiner\_intblack,health\_model\_examiner\_intblackgender)

## df AIC  
## health\_model\_examiner 27 10649.60  
## health\_model\_examiner\_intgender 38 10605.23  
## health\_model\_examiner\_intblack 38 10643.83  
## health\_model\_examiner\_intblackgender 50 10606.42

# Gender interaction seems to be making the most difference  
  
# Likelihood ratio tests  
(health\_model\_lim$deviance - health\_model\_examiner$deviance)>qchisq(.95,11)

## [1] TRUE

1-pchisq((health\_model\_lim$deviance - health\_model\_examiner$deviance),11)

## [1] 0

# This is greater, so we conclude that we should keep the examiner IDs  
  
(health\_model\_examiner$deviance - health\_model\_examiner\_intgender$deviance)>qchisq(.95,11)

## [1] TRUE

1-pchisq((health\_model\_examiner$deviance - health\_model\_examiner\_intgender$deviance),11)

## [1] 5.935421e-10

# This is greater, so we conclude that we should keep the gender interaction  
  
(health\_model\_examiner\_intgender$deviance - health\_model\_examiner\_intblackgender$deviance) > qchisq(.95,12)

## [1] TRUE

1-pchisq((health\_model\_examiner\_intgender$deviance - health\_model\_examiner\_intblackgender$deviance) ,12)

## [1] 0.02939283

# This is greater, but really not by much  
  
summary(health\_model\_examiner\_intgender)

##   
## Call:  
## glm(formula = health\_status2 ~ black + gender + age + west +   
## scale(systolic\_bp) + scale(diastolic\_bp) + locomotion + paralysis +   
## impairment + infection + scale(bmi) + scale(height) + scale(wbc\_count) +   
## scale(plt\_count) + scale(hemoglobin) + examiner\_id\_factor +   
## gender:examiner\_id\_factor, family = binomial(link = "logit"),   
## data = full\_data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.0930 -0.4881 -0.0644 0.5086 3.7457   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.223880 0.104710 11.688 < 2e-16 \*\*\*  
## black -0.392227 0.061762 -6.351 2.14e-10 \*\*\*  
## genderMale 0.024992 0.099278 0.252 0.80125   
## age -0.050214 0.001782 -28.174 < 2e-16 \*\*\*  
## west 0.339032 0.061730 5.492 3.97e-08 \*\*\*  
## scale(systolic\_bp) -0.794899 0.044317 -17.936 < 2e-16 \*\*\*  
## scale(diastolic\_bp) -0.258253 0.033830 -7.634 2.28e-14 \*\*\*  
## locomotion -1.706243 0.169612 -10.060 < 2e-16 \*\*\*  
## paralysis -2.622283 0.664344 -3.947 7.91e-05 \*\*\*  
## impairment -1.453394 0.621061 -2.340 0.01927 \*   
## infection -1.173097 0.186569 -6.288 3.22e-10 \*\*\*  
## scale(bmi) -0.823824 0.030201 -27.278 < 2e-16 \*\*\*  
## scale(height) 0.134666 0.035190 3.827 0.00013 \*\*\*  
## scale(wbc\_count) -0.186753 0.028514 -6.549 5.77e-11 \*\*\*  
## scale(plt\_count) -0.069635 0.027248 -2.556 0.01060 \*   
## scale(hemoglobin) 0.083600 0.033692 2.481 0.01309 \*   
## examiner\_id\_factor3001 2.606049 0.154190 16.902 < 2e-16 \*\*\*  
## examiner\_id\_factor3002 -6.090203 0.509078 -11.963 < 2e-16 \*\*\*  
## examiner\_id\_factor3003 0.711985 0.385737 1.846 0.06492 .   
## examiner\_id\_factor3004 0.805119 0.110498 7.286 3.19e-13 \*\*\*  
## examiner\_id\_factor3005 -3.615841 0.329840 -10.962 < 2e-16 \*\*\*  
## examiner\_id\_factor3006 3.240175 0.309972 10.453 < 2e-16 \*\*\*  
## examiner\_id\_factor3008 1.354903 0.136127 9.953 < 2e-16 \*\*\*  
## examiner\_id\_factor3009 1.312826 0.140835 9.322 < 2e-16 \*\*\*  
## examiner\_id\_factor3010 1.027129 0.380704 2.698 0.00698 \*\*   
## examiner\_id\_factor3011 2.683206 0.138511 19.372 < 2e-16 \*\*\*  
## examiner\_id\_factor3012 1.372432 0.116512 11.779 < 2e-16 \*\*\*  
## genderMale:examiner\_id\_factor3001 0.072577 0.214797 0.338 0.73545   
## genderMale:examiner\_id\_factor3002 0.640335 0.638972 1.002 0.31628   
## genderMale:examiner\_id\_factor3003 0.326494 0.543351 0.601 0.54791   
## genderMale:examiner\_id\_factor3004 0.646586 0.152645 4.236 2.28e-05 \*\*\*  
## genderMale:examiner\_id\_factor3005 1.936193 0.394728 4.905 9.34e-07 \*\*\*  
## genderMale:examiner\_id\_factor3006 -0.130131 0.452267 -0.288 0.77355   
## genderMale:examiner\_id\_factor3008 -0.036570 0.180929 -0.202 0.83982   
## genderMale:examiner\_id\_factor3009 0.506800 0.190224 2.664 0.00772 \*\*   
## genderMale:examiner\_id\_factor3010 0.301063 0.542487 0.555 0.57892   
## genderMale:examiner\_id\_factor3011 -0.235115 0.193119 -1.217 0.22343   
## genderMale:examiner\_id\_factor3012 -0.424428 0.178248 -2.381 0.01726 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 20570 on 15347 degrees of freedom  
## Residual deviance: 10529 on 15310 degrees of freedom  
## (2404 observations deleted due to missingness)  
## AIC: 10605  
##   
## Number of Fisher Scoring iterations: 8

# Final equation

$$
P( \operatorname{health\\_status2} = \operatorname{1} ) = invlogit(\beta\_{0} + \beta\_{1}(\operatorname{black}) + \beta\_{2}(\operatorname{gender}\_{\operatorname{Male}}) + \beta\_{3}(\operatorname{age}) + \beta\_{4}(\operatorname{west}) + \\ \beta\_{5}(\operatorname{scale(systolic\\_bp)}) + \beta\_{6}(\operatorname{scale(diastolic\\_bp)}) + \beta\_{7}(\operatorname{locomotion}) + \beta\_{8}(\operatorname{paralysis}) + \beta\_{9}(\operatorname{impairment}) + \beta\_{10}(\operatorname{infection}) +\\ \beta\_{11}(\operatorname{scale(bmi)}) + \beta\_{12}(\operatorname{scale(height)}) + \beta\_{13}(\operatorname{scale(wbc\\_count)}) + \beta\_{14}(\operatorname{scale(plt\\_count)}) + \beta\_{15}(\operatorname{scale(hemoglobin)}) +\\ \beta\_{16}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3001}}) + \beta\_{17}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3002}}) + \beta\_{18}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3003}}) + \beta\_{19}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3004}}) +\\ \beta\_{20}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3005}}) + \beta\_{21}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3006}}) + \beta\_{22}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3008}}) + \beta\_{23}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3009}}) +\\ \beta\_{24}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3010}}) + \beta\_{25}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3011}}) + \beta\_{26}(\operatorname{examiner\\_id\\_factor}\_{\operatorname{3012}}) +\\ \beta\_{27}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3001}}) + \beta\_{28}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3002}}) +\\ \beta\_{29}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3003}}) + \beta\_{30}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3004}}) +\\ \beta\_{31}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3005}}) + \beta\_{32}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3006}}) +\\ \beta\_{33}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3008}}) + \beta\_{34}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3009}}) +\\ \beta\_{35}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3010}}) + \beta\_{36}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3011}}) +\\ \beta\_{37}(\operatorname{gender}\_{\operatorname{Male}} \times \operatorname{examiner\\_id\\_factor}\_{\operatorname{3012}}))
$$

# Brief diagnostics

library(tidyverse)  
  
outcomes <- bind\_cols(model.frame(health\_model\_examiner\_intgender),   
 predicted=ifelse(predict(health\_model\_examiner\_intgender,type="response")>.5,1,0))  
  
outcome\_table <- outcomes %>% count(health\_status2,predicted) %>% rename(Actual = health\_status2, Predicted = predicted)  
  
# Accuracy  
outcome\_table %>% group\_by(Actual == Predicted) %>% summarize(n=sum(n)) %>% mutate(n = n/sum(n))

## # A tibble: 2 x 2  
## `Actual == Predicted` n  
## <lgl> <dbl>  
## 1 FALSE 0.152  
## 2 TRUE 0.848

# Sensitivity  
outcome\_table %>% group\_by(Actual) %>% summarize(Predicted= Predicted,n = n/sum(n)) %>%   
 bind\_cols(desc = c("True negative rate", "False positive rate","False negative rate","True positive rate"))

## # A tibble: 4 x 4  
## # Groups: Actual [2]  
## Actual Predicted n desc   
## <dbl> <dbl> <dbl> <chr>   
## 1 0 0 0.873 True negative rate   
## 2 0 1 0.127 False positive rate  
## 3 1 0 0.190 False negative rate  
## 4 1 1 0.810 True positive rate

# Pseudo R2  
# Per http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/  
1-health\_model\_examiner\_intgender$deviance/health\_model\_examiner$null.deviance

## [1] 0.4881164

# Alternative formulation from Statistical Modeling class last year  
# Very similar results, though not identical  
1-exp((health\_model\_examiner\_intgender$deviance-health\_model\_examiner\_intgender$null.deviance)/dim(model.frame(health\_model\_examiner\_intgender))[1])

## [1] 0.4801316

# Consider R2 of no examiner  
1-health\_model\_lim$deviance/health\_model\_lim$null.deviance

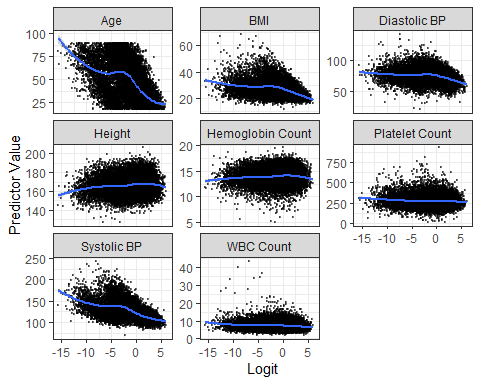
## [1] 0.2455187

# Further diagnostics

library(tidyverse)  
library(broom)  
  
# Assess relationship between continuous covariates and logit of outcome  
probabilities <- predict(health\_model\_examiner\_intgender, type = "response")  
predicted.classes <- ifelse(probabilities > 0.5, "pos", "neg")  
head(predicted.classes)

## 1 2 3 4 5 6   
## "neg" "neg" "pos" "pos" "neg" "neg"

mydata <- full\_data %>%  
 drop\_na(black, gender, age, west,systolic\_bp, diastolic\_bp, locomotion, paralysis, impairment, infection,  
 bmi, height,wbc\_count, plt\_count, hemoglobin,examiner\_id\_factor, health\_status2) %>%  
 dplyr::select(age, systolic\_bp, diastolic\_bp, bmi, height, wbc\_count, plt\_count, hemoglobin)   
predictors <- colnames(mydata)  
  
mydata <- mydata %>%  
 mutate(logit = log(probabilities/(1-probabilities))) %>%  
 gather(key = "predictors", value = "predictor.value", -logit)  
  
pred <- c("Age", "BMI", "Systolic BP", "Diastolic BP", "Height", "WBC Count", "Platelet Count", "Hemoglobin Count")  
names(pred) <- c("age", "bmi", "systolic\_bp", "diastolic\_bp", "height", "wbc\_count", "plt\_count", "hemoglobin")  
pred\_val <- c("Predictor Value")  
names(pred\_val) <- c("predictor.value")  
  
ggplot(mydata, aes(logit, predictor.value))+  
 geom\_point(size = 0.5, alpha = 0.5) +  
 geom\_smooth(method = "loess") +   
 theme\_bw() +   
 facet\_wrap(~predictors, scales = "free\_y", labeller = labeller(predictors = pred)) +   
 labs(y = "Predictor Value", x = "Logit")



# Assess multicollinearity  
car::vif(health\_model\_examiner\_intgender)

## GVIF Df GVIF^(1/(2\*Df))  
## black 1.302813 1 1.141408  
## gender 4.140482 1 2.034817  
## age 1.588497 1 1.260356  
## west 1.139894 1 1.067658  
## scale(systolic\_bp) 1.850557 1 1.360352  
## scale(diastolic\_bp) 1.619194 1 1.272476  
## locomotion 1.070420 1 1.034611  
## paralysis 1.024794 1 1.012321  
## impairment 1.011713 1 1.005839  
## infection 1.042225 1 1.020894  
## scale(bmi) 1.125440 1 1.060867  
## scale(height) 1.995945 1 1.412779  
## scale(wbc\_count) 1.193521 1 1.092484  
## scale(plt\_count) 1.192807 1 1.092157  
## scale(hemoglobin) 1.899631 1 1.378271  
## examiner\_id\_factor 5355.476912 11 1.477375  
## gender:examiner\_id\_factor 7533.634575 11 1.500471

#Note: We don't seem to have a problematic amount of multicollinearity.

# Final model table (*health\_model\_examiner\_intgender*)

## Warning: package 'kableExtra' was built under R version 4.0.3

Variable

Estimate

SE

sig

sigstar

(Intercept)

1.224

0.105

0.000

\*\*\*

black

-0.392

0.062

0.000

\*\*\*

genderMale

0.025

0.099

0.801

age

-0.050

0.002

0.000

\*\*\*

west

0.339

0.062

0.000

\*\*\*

scale(systolic\_bp)

-0.795

0.044

0.000

\*\*\*

scale(diastolic\_bp)

-0.258

0.034

0.000

\*\*\*

locomotion

-1.706

0.170

0.000

\*\*\*

paralysis

-2.622

0.664

0.000

\*\*\*

impairment

-1.453

0.621

0.019

* infection
* -1.173
* 0.187
* 0.000
* \*\*\*
* scale(bmi)
* -0.824
* 0.030
* 0.000
* \*\*\*
* scale(height)
* 0.135
* 0.035
* 0.000
* \*\*\*
* scale(wbc\_count)
* -0.187
* 0.029
* 0.000
* \*\*\*
* scale(plt\_count)
* -0.070
* 0.027
* 0.011
  + scale(hemoglobin)
  + 0.084
  + 0.034
  + 0.013
    - examiner\_id\_factor3001
    - 2.606
    - 0.154
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3002
    - -6.090
    - 0.509
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3003
    - 0.712
    - 0.386
    - 0.065
    - .
    - examiner\_id\_factor3004
    - 0.805
    - 0.110
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3005
    - -3.616
    - 0.330
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3006
    - 3.240
    - 0.310
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3008
    - 1.355
    - 0.136
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3009
    - 1.313
    - 0.141
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3010
    - 1.027
    - 0.381
    - 0.007
    - \*\*
    - examiner\_id\_factor3011
    - 2.683
    - 0.139
    - 0.000
    - \*\*\*
    - examiner\_id\_factor3012
    - 1.372
    - 0.117
    - 0.000
    - \*\*\*
    - genderMale:examiner\_id\_factor3001
    - 0.073
    - 0.215
    - 0.735
    - genderMale:examiner\_id\_factor3002
    - 0.640
    - 0.639
    - 0.316
    - genderMale:examiner\_id\_factor3003
    - 0.326
    - 0.543
    - 0.548
    - genderMale:examiner\_id\_factor3004
    - 0.647
    - 0.153
    - 0.000
    - \*\*\*
    - genderMale:examiner\_id\_factor3005
    - 1.936
    - 0.395
    - 0.000
    - \*\*\*
    - genderMale:examiner\_id\_factor3006
    - -0.130
    - 0.452
    - 0.774
    - genderMale:examiner\_id\_factor3008
    - -0.037
    - 0.181
    - 0.840
    - genderMale:examiner\_id\_factor3009
    - 0.507
    - 0.190
    - 0.008
    - \*\*
    - genderMale:examiner\_id\_factor3010
    - 0.301
    - 0.542
    - 0.579
    - genderMale:examiner\_id\_factor3011
    - -0.235
    - 0.193
    - 0.223
    - genderMale:examiner\_id\_factor3012
    - -0.424
    - 0.178
    - 0.017