Homework 7

Gianni Spiga

2023-03-10

Question 12

a.)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.0960976	0.6918325	-11.702395	0.0000000
PGC	0.0346091	0.0036276	9.540604	0.0000000
DBP	-0.0125217	0.0051709	-2.421558	0.0154542
Tricep	0.0027234	0.0068139	0.399689	0.6893856
Serum.Insulin	-0.0011215	0.0008975	-1.249524	0.2114733
BMI	0.0889731	0.0148382	5.996201	0.0000000
Age	0.0334305	0.0080611	4.147112	0.0000337

We can see from the model that Tricept skin fold thickeness and serum insulin variables are non signifcant in this model. We would only select predictors plasma glucose concentration, diastolic blood pressure,BMI and age for the model.

b.)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
s(PGC, 4)	1.0000	96.3309789	96.3309789	107.7905051	0.0000000
s(DBP, 4)	1.0000	0.1485510	0.1485510	0.1662227	0.6836083
s(Tricep, 4)	1.0000	0.3995085	0.3995085	0.4470340	0.5039545
s(Serum.Insulin, 4)	1.0000	1.7968990	1.7968990	2.0106580	0.1566168
s((BMI), 4)	1.0000	18.7036526	18.7036526	20.9286377	0.0000056
s(Age, 4)	1.0000	14.7870048	14.7870048	16.5460658	0.0000526
Residuals	742.9997	664.0091924	0.8936871	NA	NA

Comparing our GLM fit to our GAM fit, with additive terms of 4th degree for all predictors, we can see the GAM model also finds Serum Insulin as well as Tricep non signficant. However, in this model, Diastolic Blood Pressure is also found to not be signifcant in predicting diabetes.

c.)

Yes, I would replace Diastolic Blood Pressure in the GAM model with a linear function. Let's show how this model would come out.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
s(PGC, 4)	1.0000	96.3609151	96.3609151	108.1182144	0.0000000
DBP	1.0000	0.1571733	0.1571733	0.1763505	0.6746495
s(Tricep, 4)	1.0000	0.3901907	0.3901907	0.4377991	0.5083910
s(Serum.Insulin, 4)	1.0000	1.8400744	1.8400744	2.0645877	0.1511749
s((BMI), 4)	1.0000	18.4808426	18.4808426	20.7357485	0.0000062
s(Age, 4)	1.0000	14.7626951	14.7626951	16.5639382	0.0000521
Residuals	745.9999	664.8762281	0.8912551	NA	NA

Interestingly, our GPLAM model still finds that diastolic blood pressure is not signifcant in the GAM model, even as a linear effect. The remaining signficant predictors are Plasma Glucose Concentration, BMI, and age.

Question 13

Out of the four classifiers, the logistic classifier with the quadratic terms had the lowest error rate, at 0.156. A logistic classifier with linear predictors did slightly worse, while LDA and QDA performed 3rd and 4th respectively in error.

Question 14

a.)

We would perform a logistic regression on this data set, such that we have GLM and GAM

$$logit(E(Y|X)) = \beta_0 + \beta_{GLUC}X_{GLUC} + \beta_{AGE}X_{AGE}$$

$$logit(E(Y|X)) = \beta_0 + f_{GLUC}(x_{GLUC}) + f_{AGE}(X_{AGE})E(f_j(X_j)) = 0, j = 1, 2$$

b.)

GAM overcomes the curse of dimensionality by retaining the one-dimensional MSe convergence rate $n^{-4/5}$, thus not falling subject to slow convergence by other non-parametric techniques.

c.)

When presenting the results of GAM, we can show the parametric ANOVA table, including the slopes of predictors. With this, we are able to show which predict from the GAM smoothing, while we can fit non-significant terms from the non-parametric ANOVA as only linear effects.

d.)

The Generalized Additive Partial Linear model would be the best case, where we can have the GAM predictors for continuous variables but keep linearity in factor variables, in this case Gender.

e.)

In the GLM, we can test for interactions by checking if the coefficient for the three way interaction, call it β_4, would have slope 0, such that:

$$H_0 : \beta_4 = 0 H_a : \beta_4 \neq 0$$

We could use a test statistic $z_4 = \frac{\beta_4}{se(\beta_4)}$, which would be distributed as $N(0, 1)$ under the null hypothesis.