Math 583 HW 7 Fall 2017. Due Wednesday, Oct. 25.
Quiz 7 on Friday, Oct. 27 is similar to HW 7. Use 3 sheets of notes.

Problem numbers are from the Olive text.

A) 4.8. In a generalized additive model (GAM), $Y \perp\!\!\!\perp \boldsymbol{x}|AP$ where $AP = \alpha + \sum_{i=1}^{k} S_i(x_i)$. In a generalized linear model (GLM), $Y \perp\!\!\!\perp \boldsymbol{x}|SP$ where $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Note that a GLM is a special case of a GAM where $S_i(x_i) = \beta_i x_i$. A GAM is useful for showing that the predictors $x_1, ..., x_k$ in a GLM have the correct form, or if predictor transformations or additional terms such as x_i^2 are needed. If the plot of $\hat{S}_i(x_i)$ is linear, do not change x_i in the GLM, but if the plot is nonlinear, use the shape of $\hat{S}_i$ to suggest functions of x_i to add to the GLM, such as $\log(x_i), x_i^2$, and x_i^3. Refit the GAM to check the linearity of the terms in the updated GLM. Wood (2017, pp. 125-130) describes heart attack data where the response Y is the *number of heart attacks* for m_i patients suspected of suffering a heart attack. The enzyme ck (creatine kinase) was measured for the patients. A binomial logistic regression (GLM) was fit with predictors $x_1 = ck$, $x_2 = [ck]^2$, and $x_3 = [ck]^3$. Call this the Wood model I_2. The predictor ck is skewed suggesting $\log(ck)$ should be added to the model. Then output suggested that ck is not needed in the model. Let the binomial logistic regression model that uses $x = \log(ck)$ as the only predictor be model I_1. a) The *R* code for this problem from the URL above Problem 4.7 makes 4 plots. Plot a) shows $\hat{S}$ for the binomial GAM using ck as a predictor is nonlinear. Plot b) shows that $\hat{S}$ for the binomial GAM using $\log(ck)$ as a predictor is linear. Plot c) shows the EE plot for the binomial GAM using ck as the predictor and model I_1. Plot d) shows the response plot of ESP versus $Z_i = Y_i/m_i$, the proportion of patients suffering a heart attack for each value of $x_i = ck$. The logistic curve $= \hat{E}(Z_i|x_i)$ is added as a visual aid. Include these plots in *Word*.

Do the plotted proportions fall about the logistic curve closely?

b) The command for b) gives AIC(outw) for model I_2 and AIC(out) for model I_1. Include the two AIC values below the plots in a).

A model I_1 with j fewer predictors than model I_2 is "better" than model I_2 if $AIC(I_1) \le AIC(I_2) + 2j$. Is model I_1 "better" than model I_2?

Table 1: PIs for modt = 1, Problem 4.9

error type	n	95% slen	PI olen	95% dlen	PI scov	95% ocov	PI dcov	adf
1	100	4.7095	4.6949	5.0585	0.9660	0.9604	0.9736	6.27

B) 4.9. The smoothing spline simulation compares the PI lengths and coverages of 3 PIs for $Y = m(x) + e$ and a single measurement x. Values for the first PI were denoted by scov and slen, values for 2nd PI were denoted by ocov and olen, and values for third PI (2.15) by dcov and dlen. The second PI replaces d by 1 in PI (2.15). Three model types were used 1) $m(x) = x + x^2$, 2) $m(x) = \sin(x) + \cos(x) + \log(|x|)$, and 3) $m(x) = 3\sqrt{|x|}$. The smoothing spline is flexible so the df $> p$. The estimated df is given by adf. Copy and paste the R commands for this problem and make a table like the one below. The pimenlen gives slen, olen, and dlen.

a) In Table 1 above, which PI worked best?
b) For the table you make from the R output, which PI worked best?

C) 4.10. This problem does lasso for binary regression for artificial data with $n = 100$, $p = 101$ and 5 active population nontrivial predictors. If $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$, then the 100 nontrivial predictors are in $\boldsymbol{x}$ and $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, ..., 0)^T$.

a) Copy and paste the source and library commands into R. Then copy and paste the commands for this part into R. Relaxed lasso gets the binary logistic regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. Include the plot in *Word*.
Does the step function track the logistic curve?
b) Copy and paste the commands for this part into R. These commands to MLR lasso, then the relaxed lasso gets the binary logistic regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. For this data set, one more predictor was used than that in a). Include the plot in *Word*.
Does the step function track the logistic curve?
c) Copy and paste the commands for this part into R. The commands for this part use MLR forward selection with EBIC, and only nontrivial predictor x_4 was selected. Then the binary logistic regression if fit using this variable and the response plot is made. Include the plot in *Word*.
Is the plot in c) worse than the plots in a) and b)?

D) 4.11. This problem does lasso for Poisson regression for artificial data with $n = 100$, $p = 101$ and 5 active population nontrivial predictors. If $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$, then the 100 nontrivial predictors are in $\boldsymbol{x}$ and $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, ..., 0)^T$.

a) Copy and paste the source and library commands into R. Then copy and paste the commands for this part into R. Relaxed lasso gets the Poisson regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. Include the plot in *Word*. The horizontal line is $\overline{Y}$ and the jagged curve is lowess which tracked the exponential curve well until ESP > 3. Lasso overfit using 26 variables instead of 5.
b) Copy and paste the commands for this part into R. These commands to MLR lasso, then the relaxed lasso gets the Poisson regression model to the predictors corresponding to the nonzero lasso coefficients. Then the response plot is made. For this data set, 20 variables were used. Include the plot in *Word*.
c) Copy and paste the commands for this part into R. The commands for this part use MLR forward selection with EBIC, and only nontrivial predictor x_5 was selected. Then the Poisson regression if fit using this variable and the response plot is made. Include the plot in *Word*.
If the Poisson regression model is good, we would like the vertical scale to be not more than 10 times the horizontal scale in the OD plot. (This happened in a) and b).) Is the vertical scale more than 10 times the horizontal scale in the OD plot for this model?