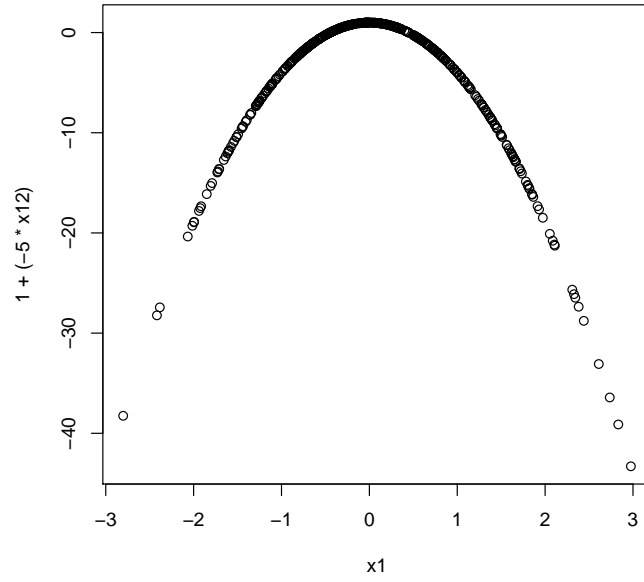
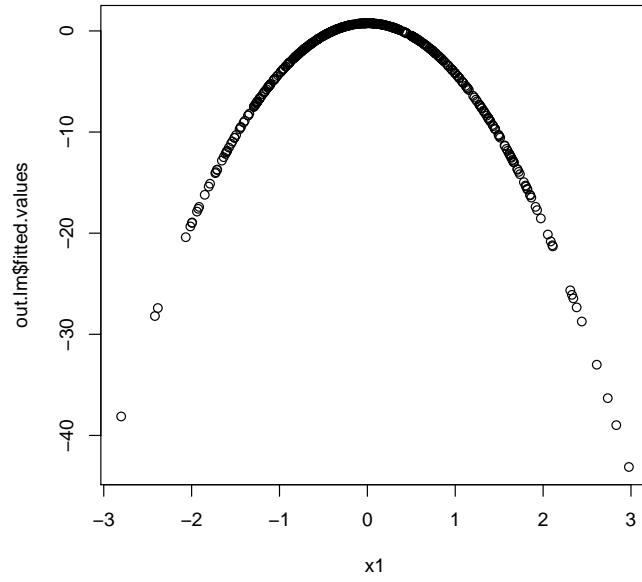


### Problem Set 1

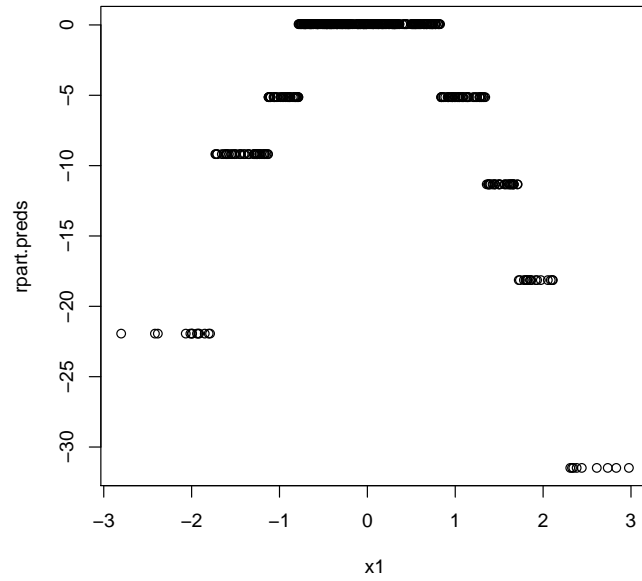
1. Below is the true  $f(X)$  we hope to recover from the data:



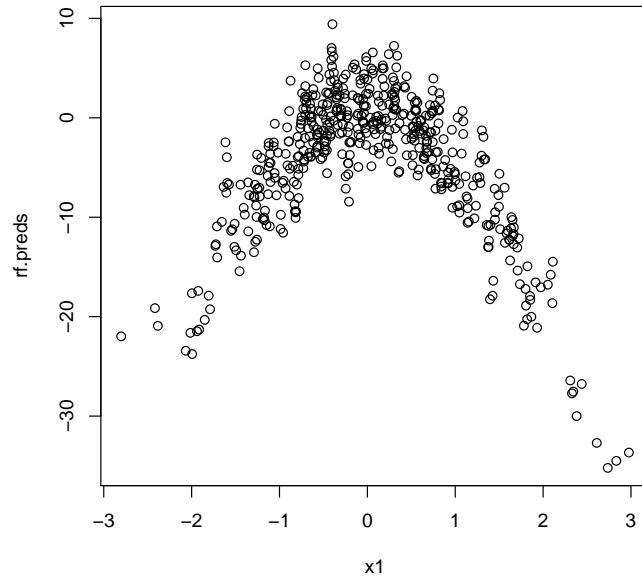
2. The plot below is nearly identical to the one shown in #1 above. When the true  $f(X)$  is known *a priori*, linear regression is a useful function estimation tool.



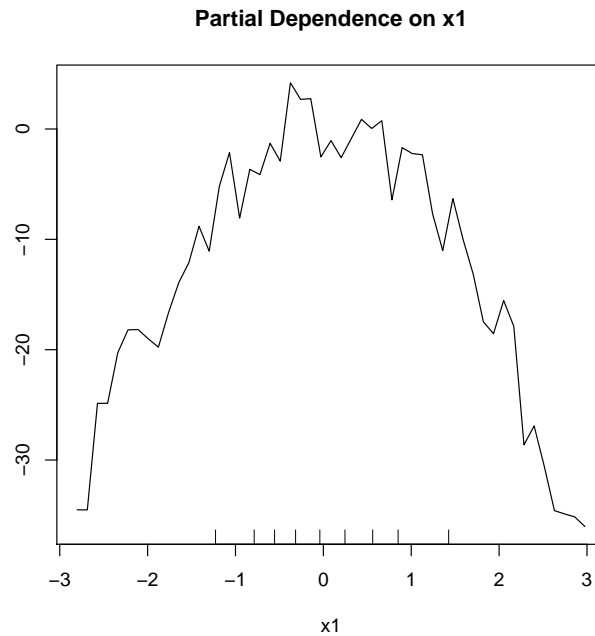
3. CART captures the parabolic relationship between  $y$  and  $x_1$  without knowing anything about the conditional distribution *a priori*. However, because the CART model includes just eight terminal nodes even though  $f(X)$  is a continuous function, the CART model is not smooth, which can lead to substantial amount of bias in the predictions.



4. Random forests does not in this fit the data quite as well as linear regression when the functional form is known (as evidenced by the scatter in the predictions). However, unlike CART, random forests captures the smoothness and the parabolic behavior in  $f(X)$  without specifying anything about the relationship between  $y$  and  $x_1$ .



5. The fitted values in CART consist of only a handful of distinct fitted values, whereas random forests has lots of different fitted values because of the use of a large number of samples that are averaged. This is akin to the impact of bagging on CART. As a result,  $f(X)$  is considerably smoother using random forests.
6. The partial dependence plot with  $x_1$  as the predictor does a pretty good job of capturing the true  $f(X)$ . Although there is some jaggedness, the parabolic behavior is very apparent. The jaggedness could be removed if a smoother with a wider span or a greater penalty for roughness were applied.



7. The random forests partial plot shows the conditional relationship of the response on the predictor, *holding other predictors constant*. Since we have defined  $y$  to be solely a function of  $x_1$  (and noise), there are no other predictors to hold constant. Thus, the partial plot of  $x_1$  and the plot of the fitted values on  $x_1$  will be similar.

## Problem Set 2

1. Using the default settings, random forests exhibits a mean squared error of 41.91 and a percent of response variance explained of 32.3.
2. By setting  $mtry = 4$ , the number of variables tried at each split is increased from 1 to 4. With 4 predictors total in the SLID data, random selection of predictors is not employed. In doing so, the MSE is 46.38 and the percent of the response variance explained is 25.07. This random forests model performs worse than the model constructed in #1 above. This is because weaker predictors are forced to compete more directly with strong predictors.
3. Using the default settings except specifying  $ntree = 100$ , we obtain a MSE of 42.32 and a percent response variance explained of 31.64. When we increase the number of trees to 1000, the MSE is 41.96 and the percent variance explained is 32.2. Thus, the fit quality is better with more trees grown but by a small amount. Incidentally, the MSE and fit quality are actually better using 500 trees than 1000 trees, again by insignificant amounts. Of course, growing 1000 versus 100 trees will require more computational time, approximately 10-fold.

Using  $mtry = 4$  and  $ntree = 100$ , the MSE is 46.68, and the percent variance explained is 24.59. Using  $mtry = 4$  and  $ntree = 1000$ , the MSE is 46.56, and the percent variance explained is 24.77. Just as when using  $mtry = 1$ , the biggest difference here is computational time. And just as in #2, these models using  $mtry = 4$  do not perform as well as those using  $mtry = 1$ , because we are no longer exploiting the ability to randomly sample predictors. The general lesson may be that number of predictors sampled can make an important difference in the fit. Several hundred trees is usually sufficient and beyond that more trees does not help much.

4. The unscaled and scaled variable importance plots are shown below. The first two plots show unscaled variable importance measured two ways.

The first is the average percent increase in MSE, as a result of permuting each predictor variable. That is, the out-of-bag MSE is computed for each tree in the usual random forests model, and the same measure is computed after permuting the predictor of interest. Here, the MSE increases an average of 15 percent if the values of *age* are shuffled randomly, and it is the most important predictor in this regard. Education is the second most important predictor; its inclusion in the random forests model corresponds to a nearly 10 percent reduction. *Sex* and *language* are the third and fourth most important variables.

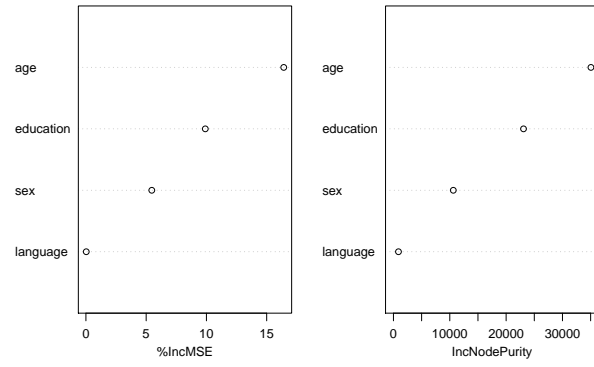
The second plot shows the total decrease in node impurities from splitting on each variable, averaged over all trees. Since the response is quantitative here, node impurity is the residual sum of squares. In this situation, both unscaled importance plots rank the predictors in the same order.

The third and fourth plots show the same measures as the first two plots, except these importance measures are normalized by the standard deviation over trees. The difference in prediction accuracy is computed for each tree, and so too is the standard deviation of this difference across all trees. The same is true for the total decrease in node impurities. In both plots, age is still the most important predictor and language is the least important predictor. However, age and education flip-flop in rank.

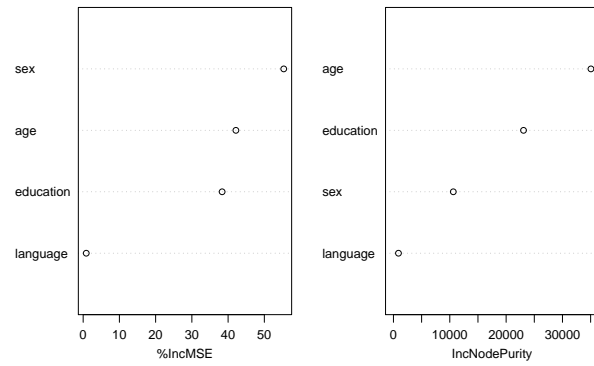
Although the variable ranks in the unscaled plot do not differ, it is possible for this to happen (just as it did in the scaled plot) because one importance measure involves permuting the predictor of interest, while the other method does not. The two measures can be quite different; one is measuring forecasting skill and the other is measure goodness-of-fit.

If it is important to assess statistical significance, then standardizing the predictor importance measures may make sense if there is reason to believe that the empirical distribution of the measure in question is close to normal. If substantive significance is the chief concern, standardization can lead to important interpretative complications.

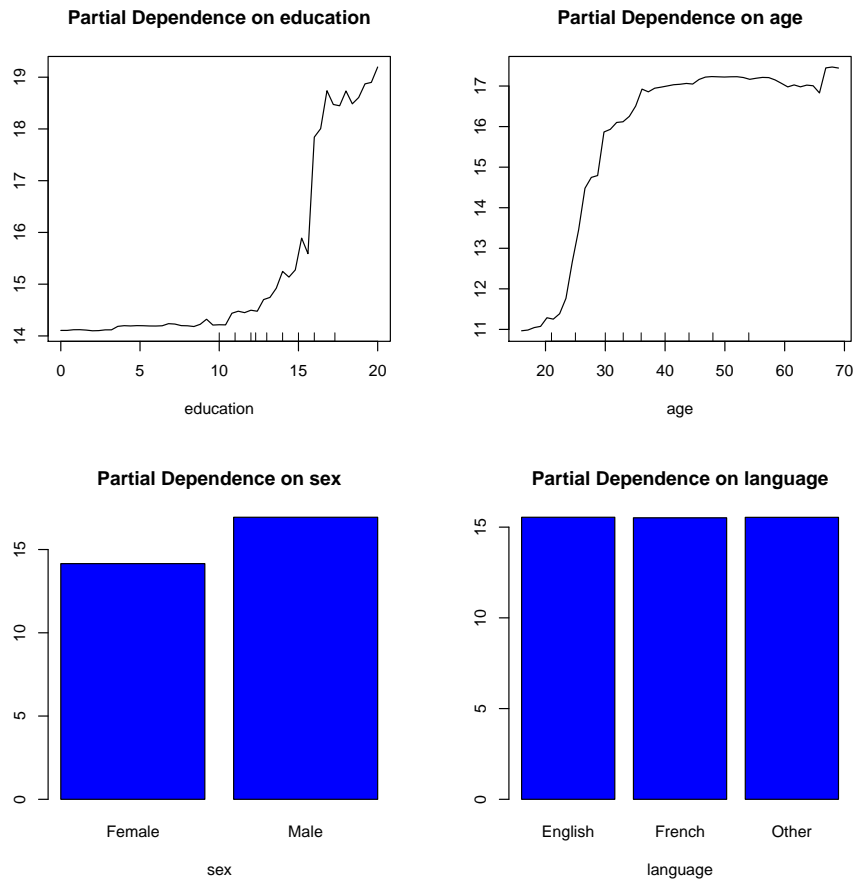
Variable Importance (Unscaled), SLID Data



Variable Importance (Scaled), SLID Data



5. Below are the partial plots for each of the four predictors:



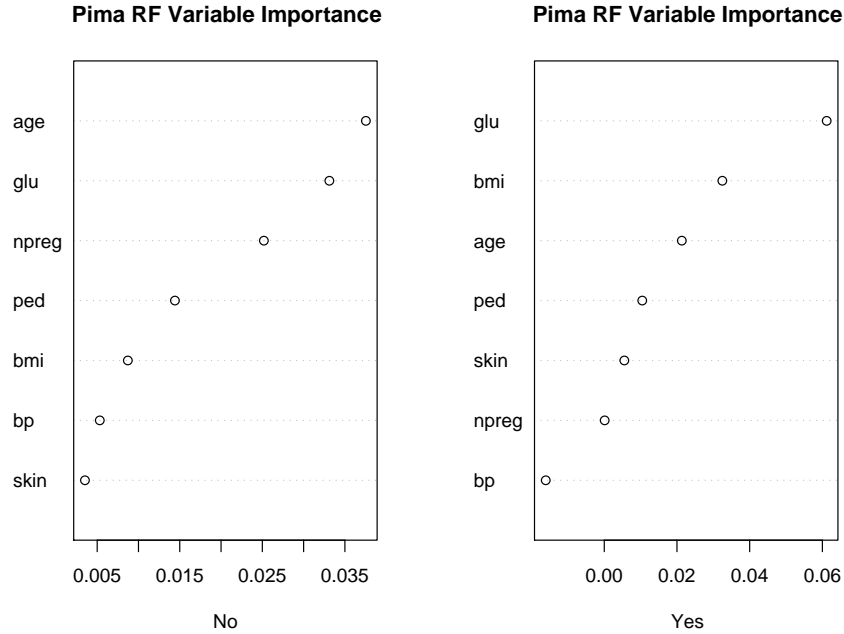
Education and age are positively related to wages; however, each predictor has a distinct partial dependence relationship with the response. Holding all other variable constant, wage does not change for education values of 0 to 10. In the SLID data set, there is just one value of education less than 10 years; thus, this portion of the partial plot should not be substantively interpreted. Wages increase at a considerably steep rate for education values between 10 and approximately 17 years of education, at which point they level off slightly and then increase once again. With respect to age, wages increase steeply for people between approximately 20 and 35 years of age. After roughly 35 years of age, the relationship is fairly flat. With respect to gender, wages are slightly higher (by a couple of Canadian dollars) for men. Finally, the partial relationship between wage and language is flat.

### Problem Set 3

1. Below is the confusion table for the Pima.tr data using the default settings in *randomForest*:

	Predicted No	Predicted Yes	Classification Error
No	110	22	.17
Yes	35	33	.51
Use Error	.24	.40	Overall Error = .29

- (a) The overall error rate is 28.5 percent.
- (b) Among those not diabetic, forecasts are correct 83.3 percent of the time. Among those who are diabetic, forecasts are correct 48.5 percent of the time. Computation is done within rows.
- (c) rates for the forecasted “No” and “Yes,” a forecast of “No” would be incorrect about 24 percent of the time and a forecast of “Yes” would be incorrect about 40 percent of the time. Computation is done within columns.
2. Below are the variable importance plots for each of the two classes. These plots show importance in terms of the unscaled mean decrease in accuracy.
- (a) Among the absence of diabetes class, the three most important predictors are *age*, *glu*, and *npreg*. Among the presence of diabetes class, the three most important predictors are *glu*, *bmi*, and *age*. These importance plots suggest that somewhat different predictors are useful for forecasting each class, which often happens. The reason is that the computations are based on the reductions in forecasting accuracy, which depend importantly the marginal distribution of the response. The order of importance can change depending in the class being forecasted because both the number of forecasting errors and the number of cases in a class can change depending on the class being forecasted.
- (b) Because there are fewer observations in the presence of diabetes class, the forecasting contributions are generally larger than those of the absence of diabetes class. All importance calculations are less than 0.04 for the absence class, while importance calculations are as high as 0.062 for the presence class.
3. Below are the partial dependence plots for each of the seven predictors in the Pima.tr data set. Because the response has just two classes, it suffices to show one partial plot for each predictor.



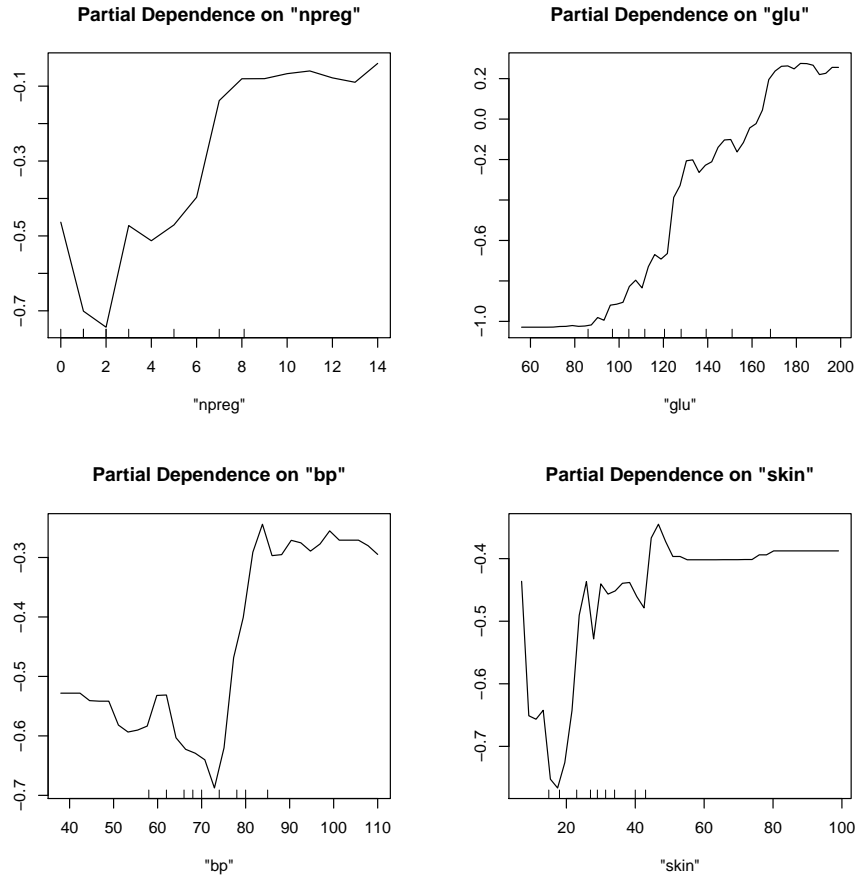
There is a negative association between being diabetic and the number of pregnancies, for women that have had 0 to 2 pregnancies. For women with more than 2 pregnancies, the relationship is positive. This partial plot should not be taken at face value for counts of *npreg* greater than 9, given the small number of women with more than 9 pregnancies.

There is a strong, positive relationship between being diabetic and the plasma glucose concentration. The relationship is roughly linear between values of 80 and 160.

For diastolic blood pressure less than 75 mm Hg, being diabetic and blood pressure have a negative relationship. For *bp* between than 75 - 85 mm Hg, the relationship is positive. For *bp* greater than approximately 85 mm Hg, the relationship is essentially flat.

For triceps skin thickness less than 10 mm and greater than 60 mm, the partial plot should not be taken at face value. There are very few observations. For *skin* between roughly 10 and 20 mm, there is a negative partial relationship with the response. After 20 mm, this relationship is positive for the most part, though the partial plot suggests some jaggedness.

With the exception of a downward trend for *bmi* less than 25, being diabetic and *bmi* are positively related, *ceteris paribus*. For *bmi* less than 25, the partial relationship with the response tends to be negative. There is

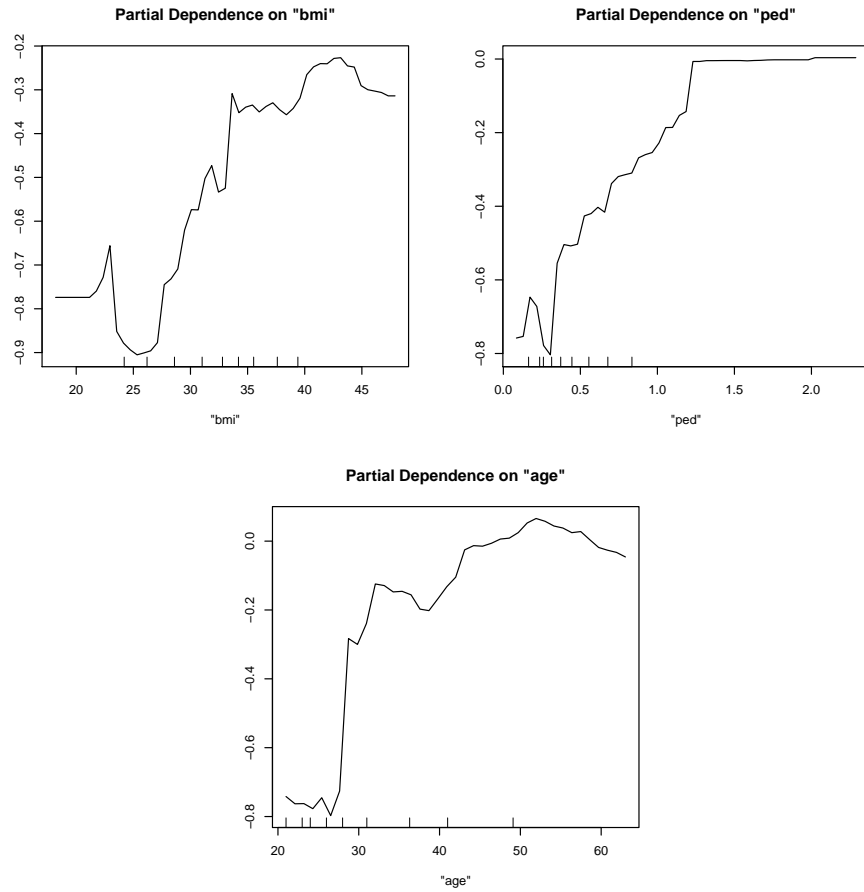


one region of *bmi* – roughly 35 to 40  $\text{kg}/\text{m}^2$  – for which the relationship is flat.

The relationship between the response and *ped* (diabetes pedigree function) is positive and strong. Much like the other covariates' respective partial plots, results in both tails of the distribution should not be taken at face value.

The partial relationship between age and the response is positive from about 25 years of age up to roughly 33. For younger ages, the relationship is flat. Between roughly 33 and 40 years old, the probability of being diabetic declines slightly. After approximately 40 years of age, the probability increases once again.

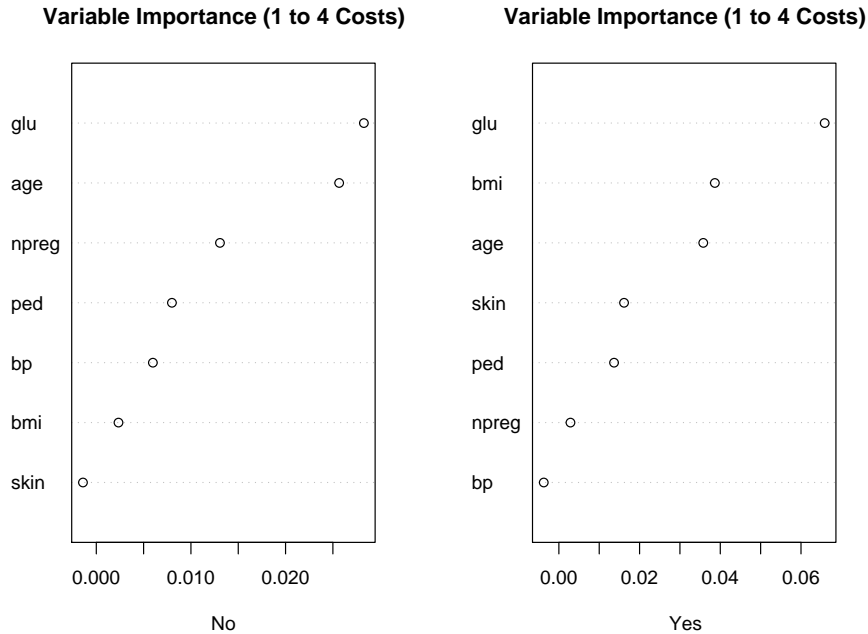
It is likely that many of the locations where the curves change direction quickly are chance or measurement artifacts of the data. But one would



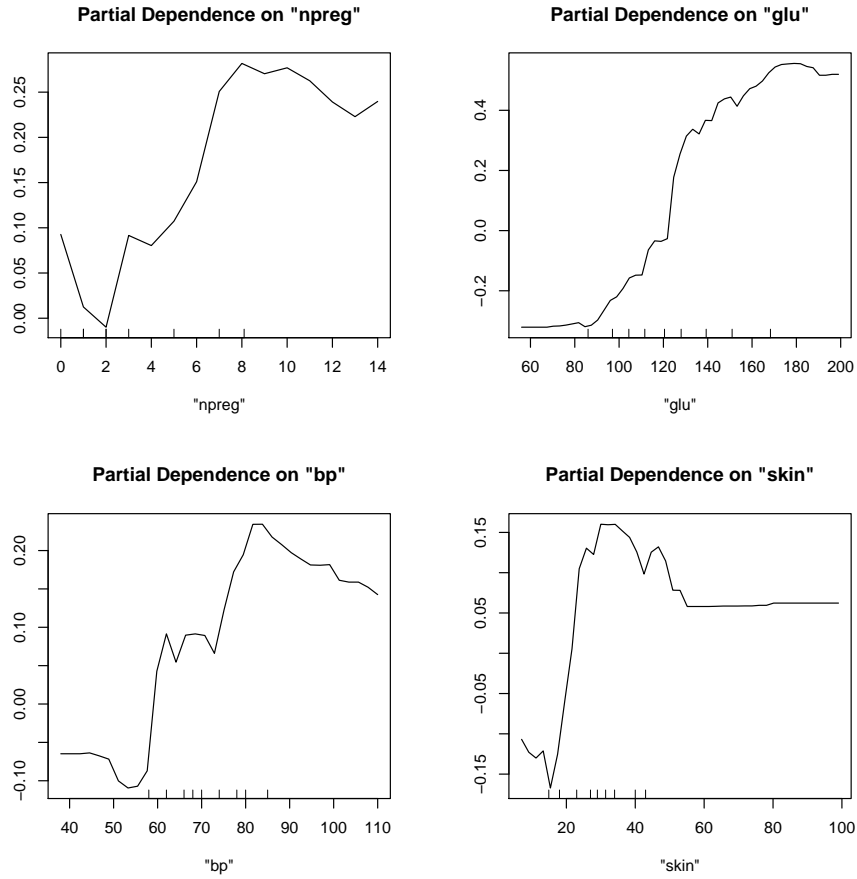
have to have real expertise in the subject matter to say with much credibility.

- Below is an analysis that is analogous to that in #s 1-3 above, except now we have taken a weighted stratified bootstrap sample. Specifically, we found that using `sampsize = c(22, 42)` produced a ratio of four false positives for every one false negative.

	No	Yes	Classification Error
No	90	42	.32
Yes	11	57	.16
Use Error	.11	.42	Overall Error = .27



- (a) The overall error actually decreases slightly from .29 to .27 when 1 to 4 costs are employed. Of course, overall error does not take the costs of those errors directly into account. The classification error for "no" class increases from .17 to .32, but the classification error for the "yes" class decreases from .51 to .19. Clearly, important tradeoffs are occurring. The ratio of false negatives to false positives is 22 to 35 when using default costs, while the ratio is 42 to 11 (i.e., close to 4 to 1) when oversampling the "yes" class. In short, although the overall error did not change much the distribution of those errors between the two classes changed a lot.
- (b) In this example, the magnitudes of the importance calculations change, but the ranking of predictors in each class remains the same in comparison to #2.
- (c) In this example, the general behavior of each of the partial dependence plots is the same across the default and 1:4 cost ratios. When



the cost ratio is altered to 1:4, the magnitudes on the y-axis shift upwards. For classification problems, the interpretation of partial dependence plots for is not trivial; however, upward shifts in the partial plot magnitudes are sensible, given that the cost ratio and prior distribution are altered in the direction of the "yes" class.

