Assignment 1 - CSC/DSC 265/465 - Spring 2017 - Due February 14

Q1. We are given a simple linear regression model $Y = \beta_0 + \beta_1 X$.

(a) Let $\hat{\beta}_i$ be the least squares estimates of β_i, $i = 0, 1$. Suppose a constant c is added to each response and the model refit. What will be the new least squares estimates of β_i, expressed in terms of the old estimates? Verify your answer analytically.

(b) The coefficient β_0 is referred to as the *intercept term* (where the Y-axis is intercepted by the regression line). It can be interpreted as a summary of the vertical location of the response Y, since the effect of changing the constant c of part (a) is directly observable in β_0. Of course, $\beta_0 = \mu_0$, where $\mu_x = \beta_0 + \beta_1 x$, so we may construct a new intercept μ_x at any vertical line $X = x$ for the same purpose. The least squares estimate will be

$$\hat{\mu}_x = \hat{\beta}_0 + \hat{\beta}_1 x.$$

What analytical criterion can be used to select x, and what would be the resulting optimal choice?

Q2. In this question, we will explore various aspects of the 'hat matrix' (see Chapter 5-6 of the lecture notes)

$$H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{1}$$

associated with multiple linear regression with q predictors (including the intercept term). We can think of H as a linear transformation of an n-dimensional response vector $\mathbf{y}$ to the n-dimensional fitted vector $\hat{\boldsymbol{y}} = H\mathbf{y}$. Furthermore, since $\hat{\boldsymbol{y}}$ is obtained by minimizing the SSE, the action taken by H on $\mathbf{y}$ is to *project* it onto the q-dimensional subspace $\mathcal{S}_q \subset \mathbb{R}^n$ spanned by the q predictors (equivalently, $\mathcal{S}_q$ is the set of all linear combinations of the predictors, see Appendix A.2). This means $\hat{\boldsymbol{y}}$ is the point in $\mathcal{S}_q$ closest to $\mathbf{y}$.

(a) Given Equation (1) prove that

$$trace(H) = q,$$

assuming matrices are invertible where indicated. [HINT: First prove the following identity. If A, B are $n \times m$ matrices, then $trace(AB^T) = trace(B^T A)$].

(b) A square matrix A is *idempotent* if and only if $A = AA$. Show that H is idempotent, using two arguments:

(i) Analytically, using matrix algebra following Equation (1). Assume matrices are invertible where indicated.

(ii) Logically, using the fact that, for any $\mathbf{y} \in \mathbb{R}^n$, $H\mathbf{y}$ is the point in $\mathcal{S}_q$ closest to $\mathbf{y}$.

(c) The 'hat matrix' structure appears in many modeling techniques, both parametric and nonparametric. Denote the least squares projection matrix defined in (1) H_{LS}. The following three modeling techniques yield various forms of fitted vectors $\hat{\boldsymbol{y}} = H'\mathbf{y}$ as linear transformations of response vector $\mathbf{y}$. In each case, describe precisely H', and give its trace.

(i) We have $\hat{\boldsymbol{y}} = (\bar{\mathbf{y}}, \ldots, \bar{\mathbf{y}}) \in \mathbb{R}^n$, where $\bar{\mathbf{y}}$ is the sample mean of the elements of $\mathbf{y}$.

(ii) Here, the elements of $\mathbf{y}$ are assumed sorted in some sequence, either by a time index or some other predictor. We take *moving average*

$$\begin{aligned}
\hat{\boldsymbol{y}}_1 &= \frac{\mathbf{y}_1 + \mathbf{y}_2}{2} \\
\hat{\boldsymbol{y}}_i &= \frac{\mathbf{y}_{i-1} + \mathbf{y}_i + \mathbf{y}_{i+1}}{3}, \quad i = 2, 3, \ldots, n-1, \\
\hat{\boldsymbol{y}}_n &= \frac{\mathbf{y}_{n-1} + \mathbf{y}_n}{2}
\end{aligned}$$

(iii) We can define a *saturated model* as one which fits the original reponse vector exactly, that is, $\hat{\boldsymbol{y}} = \mathbf{y}$.

(d) The quantity $trace(H)$ is sometimes referred to as the *effective degrees of freedom*. In linear regression it equals the model degrees of freedom (which is equal to the number of regression coefficients). But it has a similar interpretation for other forms of models, and can be taken as an index of model complexity. Order the four 'hat matrices' considered in part (c) by effective degrees of freedom, and describe briefly how this ordering relates to the complexity of the models considered. Note that we must assume $n \geq 2$, otherwise H_{LS} is not defined. Furthermore, for the ordering to hold, we may need to assume $n \geq n_{lower}$, where n_{lower} might be larger than 2.

Q3. For this question we will explore the `influence.measures()` function. The input is a fitted model object. In our example this will be a class `lm` object. It returns a `infl` class object, which has list elements `infmat`, `is.inf`, `call`. Here, `infmat` is a matrix with n rows, with columns containing a number of influence measures. They are as follows:

(i) Given q predictors (including the intercept), the first q columns give the quantities $DFBETAS_{ij}$ in row i column j.

(ii) The next column gives $DFFITS_i$.

(iii) The remaining columns give the covariance ratio, Cook's distance D_i and leverage H_{ii}.

Then `is.inf` is a logical matrix of the same dimensions as `infmat`. The i, jth entry of `is.inf` is `TRUE` if the i, jth entry of `inmat` indicates that observation i has been flagged as an influential point according to the diagnostic measure of column j. Finally `call` gives the unevaluated expression which produced the fit (this is a type of object of mode `call`).

We will make use of the `birthwt` data set from the `MASS` package.

(a) Do a simple linear regression fit of response birth weight (`bwt`) against predictor age (`age`).

(b) Apply the `influence.measures()` function to the fit.

(c) Create a plot with the following elements (use a comparable scheme if you would rather create a black and white graphic):

(i) The plot contains a scatter plot of `bwt` against `age`. Each point should be represented using a solid circle (`pch` = 20).

(ii) The fitted regression line should be superimposed on the scatter plot.

(iii) A point should be black, unless is it flagged as a high leverage point, in which case it should be red.

(iv) A point flagged by $DFFITS$ should have a triangle (pointing up) surrounding it (`pch` = 2). The triangle should be blue if $DFFITS < 0$ and green otherwise.

(v) A point flagged by the covariance ratio value should have a triangle (pointing down) surrounding it (`pch` = 6). The triangle should be blue if the covariance ratio is < 1 and green otherwise.

(d) What distinguishes high (> 0) from low (< 0) $DFFITS_i$ values?

(e) Is there any flag criterion which dominates, that is, one which is flagged when any of the others is flagged (if the answer is yes, this might, or might not, mean that we only need that criterion)?

(f) What distinguishes high (> 1) from low (< 1) covariance ratio values? Refer to the plot, and also to the original definition in Section 6.5.

(g) If we can select the values of the predictor variable, we refer to this as a design, which is ideally guided by some objective criterion. Suppose we wish to estimate the relationship $Y = \beta_0 + \beta_1 X + \epsilon$ and, under the standard assumptions for linear regression, we are able to observe $n = 10$ responses $Y_1, \ldots, Y_{10}$ for any 10 *design points* $X_1, \ldots, X_{10}$ we wish, provided $X_i \in [0, 1]$. We'll compare two designs:

$$\begin{aligned} X' &= (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), \\ X'' &= (0, 0, 0, 0, 0, 1, 1, 1, 1, 1). \end{aligned}$$

The comparison can be based on the variance of the coefficient estimates $\hat{\beta}_0$, $\hat{\beta}_1$ (see Section 5.1). Carry out this comparison. Does either design yield uniformly lower variances?

(h) Examine the form of the variances $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_0}$ in Equations (3.1), (3.3). What effect does the sample variance of the predictor variable have on these quantities?

(i) Would it be a good idea to rely exclusively on the covariance ratio to flag anomalies? Justify your answer.

Q4. For this question, use the data set `birthwt` data set from the `MASS` package used in **Q3**. A review of the examples in Chapter 4 of the lecture notes is recommended.

(a) Create a subset of this data by removing all observations flagged by the $DFFITTS$ diagnostic calculated in **Q3**. Use the `subset()` function.

(b) We will examine 6 regression models, using `bwt` as a response and `age` and `smoke` as predictors. Note that `smoke` is an binary variable. Formally, it a numeric vector, but it should be interpreted as an indicator variable, with `smoke` = 1 if the subject smokes. The models are

$$\begin{aligned}
(M1) \quad bwt &= \beta_0 + \beta_1 \times age \\
(M2) \quad bwt &= \beta_0 + \beta_1 \times smoke \\
(M3) \quad bwt &= \beta_0 + \beta_1 \times smoke + \beta_2 \times age \\
(M4) \quad bwt &= \beta_0 + \beta_1 \times smoke + \beta_2 \times smoke \times age \\
(M5) \quad bwt &= \beta_0 + \beta_1 \times smoke + \beta_2 \times (1 - smoke) \times age \\
(M6) \quad bwt &= \beta_0 + \beta_1 \times smoke + \beta_2 \times age + \beta_3 \times smoke \times age
\end{aligned}$$

Model (M1) does not distiguish by smoking group. Otherwise, the remaining models essentially create separate linear fits for the two groups, but with a pooled estimate for σ^2. What distinguishes the models is the inclusion or exclusion of an `age` term for each group. Create a linear regression fit for each model. Construct a table with a row for each model, and columns as follows:

(i) Columns 1-2 should contain the intercept and slope for the nonsmoking group (the slope may be 0).

(ii) Columns 3-4 should contain the intercept and slope for the smoking group (the slope may be 0).

(iii) Column 5 should contain R^2_{adj}.

(iv) Columns 6-9 should contain the F statistic, numerator d.f., denominator d.f., and p-value for a goodness-of-fit F-test comparing each model to the reduced model $Y = \beta_0 + \epsilon$.

It is recommended that the fitting of models is automated (see program file `CH3b.R`). Also, the quantities in columns 1-4 of the of the summary tables will be linear combinations of the coefficients estimated by the `lm()` function. For each model a single $4 \times q$ matrix can be constructed which will calculate all four quantities at once. A good strategy would therefore be to construct a list of 6 model formula, and a list of 6 coefficient transformation matrices. When creating formula, consider the following passage from the `R` documentation from `help(formula)`:

```
To avoid this confusion, the function I() can be used to bracket those portions
of a model formula where the operators are used in their arithmetic sense. For example,
in the formula y ~ a + I(b+c), the term b+c is to be interpreted as the sum of b and c.
```

(c) For each model construct a scatter-plot of `bwt` against `age`. Use something like `par(mfrow=c(3,2))` to display all plots on a single page. Superimpose the linear relationship separately for smokers and nonsmokers. Use appropriate coloring or symbols, with a legend, to distinguish between the two. Make sure the model represented by each plot is clearly indicated.

(d) Which model has the highest R^2_{adj}? Write explicitly the estimated linear relationship between `bwt` and `age` separately for smokers and nonsmokers.

(e) Of the remaining models, which is the submodel of the one identified in the previous part with the largest number of model degrees of freedom (which equals the number of coefficients to be estimated)? Include the null model $Y = \beta_0 + \epsilon$ if necessary. Do a goodness-of-fit test F-test to compare the two models (you can use the `anova()` function for this). Does the conclusion yield the same conclusion as the R^2_{adj} ranking?