Quiz 4

02/25/2019

Instructions:

Write your name at the top right. You are to work on this quiz alone without any help from any other resource *except for a single* 8.5×11 *inch page of handwritten notes.*

Problems:

1 **In each part of problem 1:** suppose we try to fit a linear model by least squares for relationship between "age" in the first column and the explanatory variables in the remaining columns. Answer the following questions:

- If $n =$"the number of observations" and $p - 1 =$"the number of explanatory variables", what size is the matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$?
- In 1-2 sentences describe, what issues would you encounter with this model?

1.1 In this problem we consider "*cats_data_1*" as below:

$$
cats\_data\_1 \leftarrow \begin{array}{ccccc}
 & age & weight\_kg & weight\_lbs & height \\
1 & 1 & 2.1 & 4.62 & 7 \\
2 & 4 & 5.0 & 11.00 & 15 \\
3 & 6 & 3.2 & 7.04 & 9 \\
4 & 7 & 4.0 & 8.8 & 10 \\
5 & 7 & 3.5 & 7.7 & 9
\end{array}
$$

This problem utilizes three explanatory variables and one intercept, so the size of $\mathbf{X}$ will be 5×4. We notice that weight is listed in both *kg* and *lbs*, where one is a scalar multiple of the other — this is linear dependence among explanatory variables which will make the problem non-solvable as written.

1.2 In this problem we consider "*cats_data_2*" as below:

$$
cats\_data\_2 \leftarrow \begin{array}{ccccccc}
 & age & weight\_kg & blood\_pressure & height & t-cell & glucose \\
1 & 1 & 2.1 & 50 & 7 & 150 & 35 \\
2 & 4 & 5.0 & 45 & 15 & 167 & 37 \\
3 & 6 & 3.2 & 60 & 9 & 162 & 40
\end{array}
$$

This problem has five explanatory variables and one intercept, therefore the $\mathbf{X}$ matrix is size 3×6. This model is super-saturated and therefore we can never identify all parameters in the model.

1.3 In this problem we consider "*cats_data_3*" as below:

$$cats\_data\_3 \leftarrow \begin{array}{cccc} & age & weight\_kg & height \\ 1 & 1 & 2.1 & 7 \\ 2 & 4 & 5.0 & 15 \\ 3 & 6 & 3.2 & 9 \end{array}$$

This model has two explanatory variables and one intercept, so the $\mathbf{X}$ matrix is size 3×3. This model is saturated, so we can find a fit, but we cannot estimate the standard error without any degrees of freedom.

2 Suppose for an arbitrary set of data, we fit a linear relationship with least squares. The R^2 score is equal to .65. Does this represent a good fit of the model to the data? Why?

This may be a good fit, but it is unclear without context. In some applications like social sciences this may be considered a good score, but in engineering problems we will generally expect a much higher score to represent a "good fit".

3 Suppose that we have some data set with variables $x_1, \cdots, x_{p-1}$. After computing the anomalies of these variables, for x_i defined as

$$\mathbf{a}_i \triangleq \begin{pmatrix} x_{1,i} - \overline{x_1} & \cdots & x_{n,i} - \overline{x_n} \end{pmatrix}^{\mathrm{T}}, \tag{1}$$

we notice that for $i \neq j$ the **dot product**

$$\mathbf{a}_i^{\mathrm{T}} \mathbf{a}_j = 0. \tag{2}$$

How does this relate to statistical (in)dependence? How does this fact help our analysis?

We can see that the correlation (or covariance) of the variables x_i and x_j are all zero. This means that these variables are statistically independent. This is useful in our analysis because the explanatory variables are especially well conditioned giving a more clear response in terms of each predictor.