NTHU STAT 5410 Midterm Solution Nov 18, 2021

(1, 1pt) A (side-by-side) box-plot would be better than a scatter plot because Reject is a categorical variables and the other variables are quantitative variables.

(2, 2pts) This is an incorrect conclusion. If the interpretation (i.e., having rejection is a good thing) is correct, taking a high dose of anti-rejection medicines after operations would be unnecessary. Notice that both Survival and Reject are observational data, not experimental. The increasing pattern appeared in the scatter plot of Survival and Reject could be caused by some *unobserved* important variable(s), which bias the coefficient estimate of Reject. Such variable(s) must be positively (or negatively, respectively) correlated with Reject and has a positive (or negative, respectively) influence on the reaponse to generate this increasing pattern. Can you give a possible example of such variable(s)?

Some argued that the possible reason might be the collinearity between Reject and the other predictors in this data. This is not an acceptable answer for this case because in the output of Model b (on all 35 patients), the estimated coefficient of Reject is still positive and significant.

(3, 1pt) The number of patients in this data is $n = (n-p) + p = 30 + 5 = 35$, where p is the number of coefficients in Model a. There are $35 \times 0.8 = 28$ patients having rejection before death.

(4, 1pt) RSS $= (1 - R^2) \times$ [(sample variance of the response)$\times(n-1)$] $= (1 - 0.111) \times 2.646 \times (35 - 1) = 79.978$.

(5, 1pt) (a) The best estimate (i.e., BLUE) is

$$(-0.590004) + (-0.032542) + (-0.012850) + 0.000213 = -0.635183.$$

(b) Gauss-Markov condition.

(6, 2pts) Let a and b be two constants. From

$$\begin{aligned}&(-0.590004)(\text{Mismatch}+a)+(-0.012850)(\text{Waiting}+b)\\ >\ &(-0.590004)(\text{Mismatch})+(-0.012850)(\text{Waiting}),\end{aligned}$$

we can derive that

$$\frac{a}{b} < -\frac{(-0.012850)}{(-0.590004)} = -0.02177951 \quad \Rightarrow \quad \frac{a}{b/30} < -0.6533854.$$

That is, if an increase of one month (30 days) on the waiting time can guarantee to find a new heart with at least a reduction of 0.6534 unit on the mismatch score (compared to the current one), it would be worthy.

(7, 1pt) This is the overall F-test. Let $\beta_1, \ldots, \beta_4$ be the coefficients of the 4 predictors under Model a. The null and alternative hypotheses are

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs. } H_A : \text{at least one of these } \beta_i\text{'s not zero}$$

The p-value of the test is 0.458 so the null is not rejected.

(8, 2pts) The following are some acceptable reasons:

(a) Sample size is not large enough.

(b) There might exist some outliers.

(c) The mean structure $\boldsymbol{X\beta}$ is incorrect, e.g., some important predictor(s) are not included in this model, or some quadratic or interaction terms should be added, ...

Notice that in this case, the collinearity between the 4 predictors is not a reason causing the insignificance in these t-tests (i.e., it is impossible for a predictor to become significant after removing at least one of the other predicots from the model), because the overall F-test is far from significant. Also, Figure 1 does not show any severe collinearity between the 4 predictors.

(9, 1pt) No, it also depends on the scale. Adjusting unit or scale of a predictor can change the size of estimated coefficient but the predictor still explain the same amount of variation in the response. For example, if we change the unit of Waiting from days to minutes, its absolute estimated coefficient becomes $0.012850 \times 24 \times 60 = 18.504 > 0.590004$, but it still reduces the same amount in RSS because the two vectors, unscaled and scaled Waiting, are proportional.

(10, 2pts) We can regard the collection of all the patients in the larger data set as a population and the 35 patients in this data as a sample. From this viewpoint, the variance estimate under the fit on the larger data set can be regarded as a "population parameter", denoted by σ^2, and the $\hat{\sigma}^2$ obtained from the sample of 35 patients as an estimate of σ^2. Because it is a simple random sample, we know that $E(\hat{\sigma}^2) = \sigma^2$. We therefore expect the residual stand error in the fit on the larger data set to stay about the same.

The RSS would increase (dramatically) because (a) $E(RSS) = (n-p)E(\hat{\sigma}^2) = (n-p)\sigma^2$; (b) the n of the sample is 35 while the n of the population is a number much larger than 35. Furthermore, the residual vector of the larger data set would have a dimension much larger than that (35) of the smaller data set, which would result in a larger RSS in the former.

(11, 2pts) The denominators of the t-statistic and the overall F-statistic are identical. Their numerators are the "RSS reduced" divided by the degrees of freedom (d.f.) used, which might be realized as the "average RSS" reduced by each d.f.. The overall F use 5 d.f. to reduce about the same amount of RSS as the t-statistic does by using only 1 d.f.. The average RSS reduced in the overall F is smaller than that in the t-statistic, which causes this contradiction.

If we remove the preditor(s) with large p-value from the model (e.g., Calendar) so that the average RSS reduced in the overall F can increase, we should be able to get consistent results on t and F.

(12, 1pt) $\hat{Y}_b$, because Model a is a submodel of Model b, i.e., the model space Ω of Model b contains the model space ω of Model a. You can also infer this answer from $0.111 = R_a^2 < R_b^2 = 0.268$, where R_a^2 and R_b^2 are the coefficients of determination under Models a and b respectively.

(13, 2pts) According to the definition of R^2,

$$\frac{\text{variance of } \hat{Y}_b}{\text{variance of } \hat{Y}_a} = \frac{R^2 \text{ under Model b}}{R^2 \text{ under Model a}} = \frac{0.268}{0.111} = 2.414414.$$

It implies the variation in the response that can be explained by the systematic part of Model b is about 2.4 times larger than that of Model a.

(14, 1pt) The sum is zero, because the residual must be orthogonal to the intercept and the column of Reject in the model matrix $\boldsymbol{X}$, and therefore their difference. The inner product of this difference vector and the residual vector is the sum of residuals of patients with no rejection.

(15, 1pt) Model a. Although Model b seems to be a better fit, it cannot be used to predict this case because the value of Reject cannot be observed before an operation.

(16, 1pt) The predictor Calendar is actually a time index. Because Calendar is the most insignificant predictor in Models a and b, there is no evidence showing that we must adjust for a time trend for this data.

(17, 1pt) Because (0, 0) falls in the confidence region and is very close to its boundary, we know that the p-value should be only slightly larger than 0.05. Under a significant level 0.1, we are confident that the null would be rejected.

(18, 1pt) Figure 2 shows that the estimated coefficients (i.e., $\hat{\beta}_i$'s) of Reject and Calendar are negatively correlated. It implies Reject and Calendar (i.e., x_i's) should be positively correlated (Figure 1 also supports it), which indicates a positive slope in the simple regression line of Calendar against Reject.

(19, 1pt) 0.268, because the model spaces of Models b and c are identical and they all contain the intercept term.

(20, 2pts) t-statistic = 0.98 and p-value = 0.333. Because Model b can be written as

$$\begin{aligned}
\text{Survival} &= \beta_0 + \beta_1\text{Reject} + \beta_2\text{Mismatch} + \beta_3\text{Age} + \beta_4\text{Waiting} + \beta_5\text{Calendar} + \epsilon \\
&= \beta_0 + \text{Reject} + (\beta_1 - 1)\text{Reject} + \beta_2\text{Mismatch} + \beta_3\text{Age} + \beta_4\text{Waiting} \\
&\quad + \beta_5\text{Calendar} + \epsilon \\
&= \beta_0 + \text{Reject} + \beta_1^*\text{Reject} + \beta_2\text{Mismatch} + \beta_3\text{Age} + \beta_4\text{Waiting} \\
&\quad + \beta_5\text{Calendar} + \epsilon \quad (\leftarrow \text{this is Model c}),
\end{aligned}$$

testing $\beta_1 = 1$ under Model b is equivalent to testing $\beta_1^* = 0$ under Model c.

(21, 1pt) The appearance of NA's in this output indicates that Model b is unidentifiable under this subset of data. It occurs because in the model matrix $\boldsymbol{X}$ from this subset of data, the column corresopnding to Reject is identical to the intercept.

(22, 1pt) (i) RSS and (iii) $\hat{Y}$ stay invariant.

(23, 1pt) Based on all the results, for these predictors we can conclude that:

(a) Reject: Although it is the only significant predictor (under $\alpha = 0.05$) in Model b (all patients data), the significance is possibly due to some unobserved important variables (check the answer to problem (2)). On the other hand, Reject cannot be observed before an operation (check the answer to prolem (15)). Saying that Reject is important has no value to physicians before operations. Actually, Reject plays a role like a "blocking" variable in Model b, i.e., we are not interested in its coefficient estimate (because it is biased) and use Reject mainly to reduce RSS so that the predictors of interest can be more significant.

(b) Mismatch: It is the most significant predictor except Reject in every models (although most of its p-values are larger than 0.05). Furthermore, its coefficient estimates are quite consistent across these models. But, the variation in the response explained by Mismatch is relatively low (less than 11.1%). However, it is not uncommon to see important predictor(s) of this kind in medical data due to large error variance presented in many biological phenomena.

(c) Age: The range (40-65) of Age is relatively small, i.e., this data contained only middle-aged patients. If more cases of young and older patients were included, Age might become more significant. [Recall that in LNp.3-18, the R^2 increases (i.e., interpreting more variation in the response) when the range of a predictor becomes larger.]

(d) Waiting: It seems that waiting is not important from these results. But, notice that this data only contained patients receiving heart transplant operations. This data does not contain information about how many patients were on the waiting list but died before receiving a heart transplant operation during the period of this study. It is a risk of waiting which we cannot evaluate by using this data. From this viewpoint, Waiting is still a variable that should be explored in future study.

(e) Calendar: check the answer to Problem (16).

In summary, I might conclude that only Mismatch (and possibly Age) seems important for longer survival based on these results of data analysis, but more investigations would be required as well.