NTHU STAT 5410 Midterm Examination Nov 18, 2021

Instructions: Attempt all questions. Short and specific answers are preferred. Given explanation when required, but keep it as short and simple as possible. Give only one answer to each question – if you give alternative answers, the worst answer will be graded.

Dataset background: The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan. After a patient was admitted to the Stanford program, a donor heart, matched on blood type, was then sought. We chose to present here only patients who received a transplant and had been followed until their death, because that avoids some problems arising from dealing with censored data. (For patients still alive, we know only that their survival time will be longer than it is to date, so the observation is incomplete or "censored".) The data reported here cover the deaths following heart transplantation during the period January, 1968 through April, 1974. The variables in the data were:

`Survival:` number of days the patient survived after the operation,

`Reject:` transplant rejection (移植排斥反應) occurred before death or not (0=no rejection occurred; 1=rejection occurred),

`Mismatch:` a measure of the degree to which donor and recipient were mismatched for tissue type,

`Age:` age at time of operation,

`Waiting:` number of days from entry into the program until the operation was performed,

`Calendar:` number of days after January 1, 1968 that the operation was performed.

The results of some initial data analyses are given below, including scatter plots between these variables:

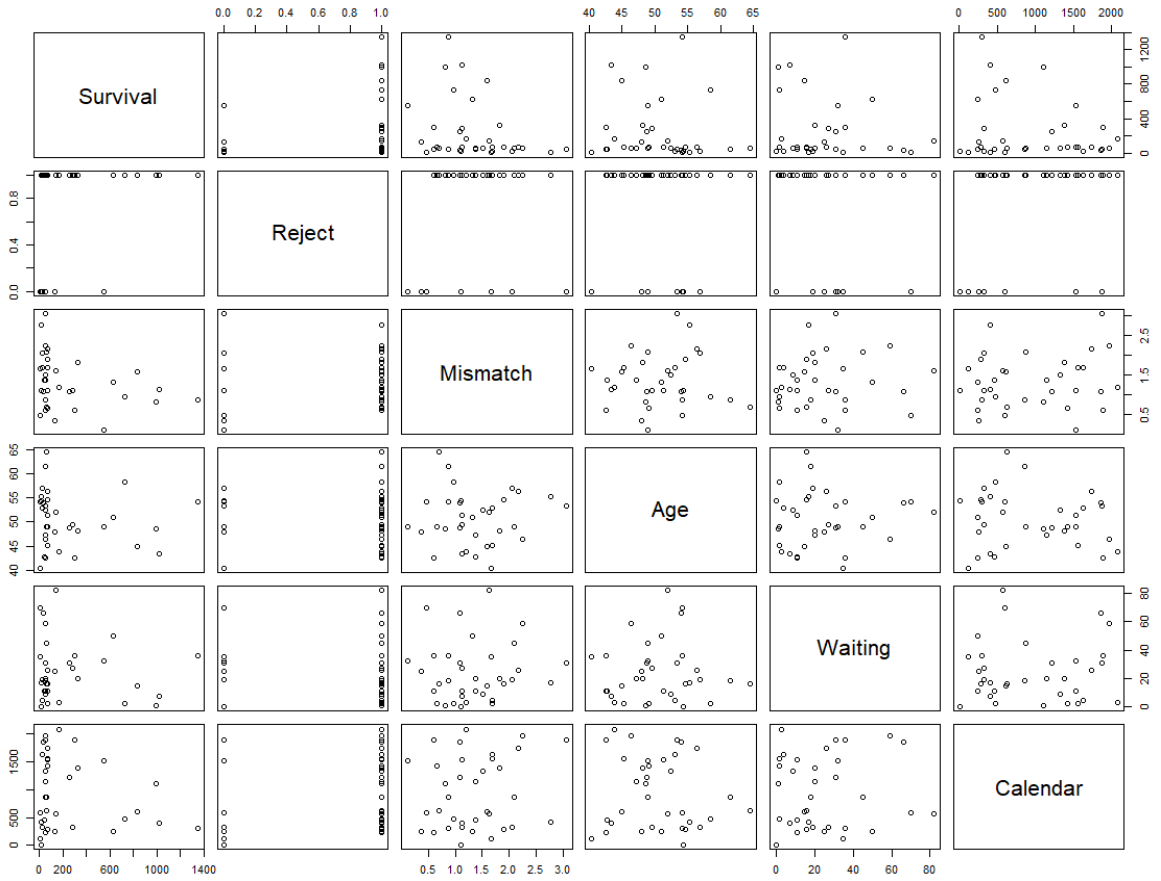

Figure 1

a brief descriptive summary of these variables:

```
    Survival                Reject              Mismatch
 Min.    :   1.0        Min.   :0.0         Min.   :0.120
 1st Qu. :  46.5        1st Qu.:1.0         1st Qu.:0.870
 Median  :  64.0        Median :1.0         Median :1.200
 Mean    : 244.6        Mean   :0.8         Mean   :1.335
 3rd Qu. : 288.5        3rd Qu.:1.0         3rd Qu.:1.680
 Max.    :1350.0        Max.   :1.0         Max.   :3.050
     Age                   Waiting             Calendar
 Min.    :40.40         Min.   : 0.00       Min.   :   6.0
 1st Qu. :46.80         1st Qu.:10.00       1st Qu.: 366.0
 Median  :49.50         Median :19.00       Median : 864.0
 Mean    :50.48         Mean   :24.43       Mean   : 949.8
 3rd Qu. :54.15         3rd Qu.:33.50       3rd Qu.:1531.5
 Max.    :64.50         Max.   :82.00       Max.   :2087.0
```

and the (unbiased) sample variances of these variables and a log transformation of `Survival`:

Survival	log(Survival)	Reject	Mismatch	Age	Waiting	Calendar
121493.3	2.646	0.1647	0.4391	31.3377	439.25	412823.7

(1) (1 pt) What is a better graphical representation for the relationship between `Reject` and the other variable(s) than their scatter plots shown in Figure 1?

(2) (2 pts) Comment on the following conclusion: "Figure 1 suggests that the occurrence of rejection is a good thing for the patients with heart transplant, since the patients with rejection tended to survive longer on average after their operations than the patients without rejection". If you feel this is a correct conclusion, find the plot in Figure 1 that supports it. If you feel this is an incorrect conclusion, explain your reasons and why it happened. [**Hint**. After an organ transplant, the patient must take a high dose of anti-rejection medicines for a period of time to avoid the immune system trying to destroy the new organ.]

Output from a multiple linear regression analysis of the model, called `Model a`:

```
log(Survival) ~ Mismatch + Age + Waiting + Calendar
```

is given below:

```
Coefficients:
             Estimate   Std. Error     t value    Pr(>|t|)
(Intercept)  7.008385   2.655283       2.64       0.013
Mismatch     -0.590004  0.429718       -1.37      0.180
Age          -0.032542  0.050609       -0.64      0.525
Waiting      -0.012850  0.013410       -0.96      0.346
Calendar     0.000213   0.000444       0.48       0.635
---
```

```
Residual standard error: ?? on 30 degrees of freedom
Multiple R-squared:  0.111, Adjusted R-squared:  -0.00796
F-statistic: ?? on ? and ?? DF,  p-value: 0.458
```

(3) (1 pt) In this data, how many patients had rejection before death? Explain how you get your answer. [**Hint**. Do not use Figure 1 to get the answer.]

(4) (1 pt) What is the RSS (residual sum of squares) of `Model a`? Explain how you get your answer.

(5) (1 pt) Suppose that a physician was interested in estimating the sum of the coefficients of the 4 predictors under `Model a`. What is the estimated value you would offer to them? What condition(s) must be satisfied to guarantee it is a good estimated value?

(6) (2 pts) Based on the results of `Model a`, how would you advise a physician about trade-offs between accepting a poor mismatch score or increasing waiting time? Explain.

(7) (1 pt) State and test the hypotheses related to the question of whether none of the *linear combinations* of the 4 predictors have an effect on the response.

(8) (2 pts) Under `Model a`, all the *t*-tests of the 4 predictors are insignificant. Suppose that the physicians believed that at least one of the predictors are important for survival. Based on the information available, give the physicians at least two reasons to explain why this (i.e., all predictors are insignificant) can happen even if the perception of the physicians on these predictors is correct.

(9) (1 pt) The predictor `Mismatch` has the largest absolute coefficient, **`0.590004`**. Can we use this finding to conclude that it has the most significant contribution in reducing RSS compared to the other predictors in `Model a`? Explain.

(10) (2 pts) Assume that the cases in this data were actually randomly drawn from a much larger data set. If `model a` were fit on the larger data set, would you expect $\hat{\sigma}$ (residual standard error) to increase, decrease, or stay about the same? Explain. How about RSS? Explain.

An alternative model adding `Reject` as a predictor, called `Model b`, is fit to this data:

```
log(Survival) ~ Reject + Mismatch + Age + Waiting + Calendar
```

and its output is given below:

```
Coefficients:
             Estimate   Std. Error     t value    Pr(>|t|)
(Intercept)  5.792184   2.498164       2.32       0.028
Reject       1.650093   ??             ??         0.018
Mismatch     -0.626368  0.396796       -1.58      0.125
Age          -0.031701  0.046701       -0.68      0.503
Waiting      -0.007990  0.012527       -0.64      0.529
Calendar     -0.000015  0.000420       -0.03      0.973
---
Residual standard error: ?? on 29 degrees of freedom
Multiple R-squared:  0.268, Adjusted R-squared:  0.142
F-statistic: ?? on ? and ?? DF,  p-value: 0.091
```

(11) (2 pts) It seems that there is a contradiction in the results of `Model b`. Under the significant level 0.05, the overall *F*-test is insignificant but the *t*-test of `Reject` is significant. For this data, what is the possible reason causing this contradiction under `Model b`? How could you possibly do to remove this contradiction?

(12) (1 pt) Let $\hat{Y}_a$ and $\hat{Y}_b$ be the vectors of the fitted values of `log(Survival)` under `Models` a and b respectively. Which vector, $\hat{Y}_a$ or $\hat{Y}_b$, has a larger length? Explain.

(13) (2 pts) Find the ratio of the (sample) variance of $\hat{Y}_b$ to the (sample) variance of $\hat{Y}_a$ and interpret it.

(14) (1 pt) Under this fit of `Model b`, for the patients with *no rejection*, what is the sum of their residuals? Explain how you get your answer.

(15) (1 pt) Suppose a physician would like to predict the survival days of a patient *before* a designated heart transplant operation. Which model, a or b, would you suggest the physician to adopt for this prediction purpose? Explain.

(16) (1 pt) Do we need to control for a time trend in this data? Explain.

Under `Model b`, a ***95%*** confidence region for the coefficients of `Reject` and `Calendar` is shown in Figure 2.

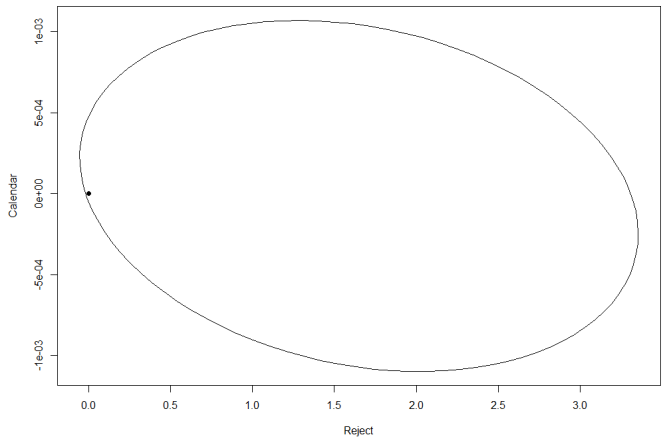

Figure 2

(17) (1 pt) Can we use Figure 2 to determine whether the null hypothesis of both the coefficients of `Reject` and `Calendar` being zero would be rejected or not *under the significant level **0.1***? If yes, do it. If no, explain why.

(18) (1 pt) Suppose we fit a simple regression model with `Calendar` as the response and `Reject` as the predictor. Use Figure 2 to identify the sign (i.e., positive or negative) of the estimated coefficient of `Reject` in this simple regression model? Explain how you get your answer.

A third model with an offset, called `Model c`, is fit to this data:

```
log(Survival) ~ offset(Reject) + Reject + Mismatch + Age + Waiting + Calendar
```

and its output is given below:

```
Coefficients:
             Estimate   Std. Error     t value    Pr(>|t|)
(Intercept)  5.792184   2.498164       2.32       0.028
Reject       ??         ??             0.98       0.333
```

Mismatch	**-0.626368**	**0.396796**	**-1.58**	**0.125**
Age	**-0.031701**	**0.046701**	**-0.68**	**0.503**
Waiting	**-0.007990**	**0.012527**	**-0.64**	**0.529**
Calendar	**-0.000015**	**0.000420**	**-0.03**	**0.973**

(19) (1 pt) In this fit of `Model c`, what is the percentage of the variation in the response explained by the 5 predictors together with the offset? Explain.

(20) (2 pts) Under the fit of `Model b` to all patients in this data, report the values of t-statistic and p-value for testing the null hypothesis of the coefficient of `Reject` being 1. Explain how you get your answer.

The `Model b` is fit to a *subset* of this data, which contains only the patients *with rejection*, and its output is given below:

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	**9.398887**	**2.257781**	**4.16**	**0.00038**
Reject	**NA**	**NA**	**NA**	**NA**
Mismatch	**-0.927335**	**0.436163**	**-2.13**	**0.04444**
Age	**-0.056395**	**0.041776**	**-1.35**	**0.19016**
Waiting	**0.004863**	**0.011368**	**0.43**	**0.67278**
Calendar	**-0.000603**	**0.000389**	**-1.55**	**0.13434**

(21) (1 pt) Why are the coefficient estimate, std. error, t value, and p-value corresponding to `Reject` not available in this result? Explain.

(22) (1 pt) Consider a model, called `Model d`, that contains the 5 predictors in `Model b` but *no intercept*. If we fit `Models b` and `d` to this subset of data (i.e., patients with rejection), which one or ones of the following quantities would stay invariant under the fits of the 2 models: (i) RSS, (ii) overall *F*-statistic, (iii) fitted values $\hat{Y}$, (iv) coefficient of determination R^2.

(23) (1 pt) Based on *all* the results given above, how would you tell the physicians about what factor(s) seem important for longer survival? Explain.