­Advanced Topics in Research Methods and Design

Yi Chen

Columbia University

Exercise Two

Advanced Topics in Research Methods and Design

# Problem 1: questions in chapter 11

## Question 1

Develop a research question for the preceding scenario.

*Answer*: Instead of a single research question, I proposed a series of research questions.

* Can individuals’ job satisfaction (i.e., very satisfied or not very satisfied) be correctly predicted from knowledge of their age, education, weekly working hour, income, number of brothers and sisters, and opinion about their life (i.e., dull or routine/exciting)?
* If job satisfaction status can be predicted correctly, which variables are central in the prediction of that status? Does the inclusion of a particular variable increase or decrease the probability of the specific outcome?
* How good is the model at classifying cases for which the outcome is unknown? In other words, how many very satisfied job satisfactions are classified correctly? How many not very satisfied job satisfactions are classified correctly?

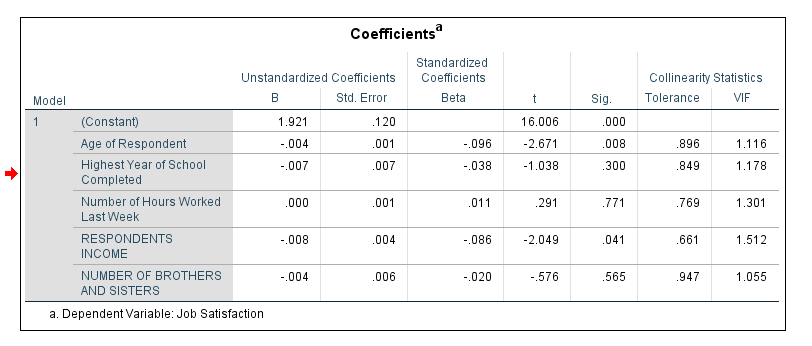
## Question 2

Conduct a preliminary Linear Regression to identify outliers and evaluate multicollinearity among the five continuous variables. Complete the following:

1. Using the Chi-Square table in Appendix B near the end of this book, identify the critical value at p < .001 for identifying outliers. Use Explore to determine if there are outliers. Which cases should be eliminated?

*Answer*: There are six independent variables in this study. Thus, the critical value to identify the outlier is . Eight cases include 50, 121, 192, 268, 406, 632, 689, and 1129 should be eliminated. Further, I use the select case function with “HAM 20.515”.

2. Is multicollinearity a problem among the five continuous variables?

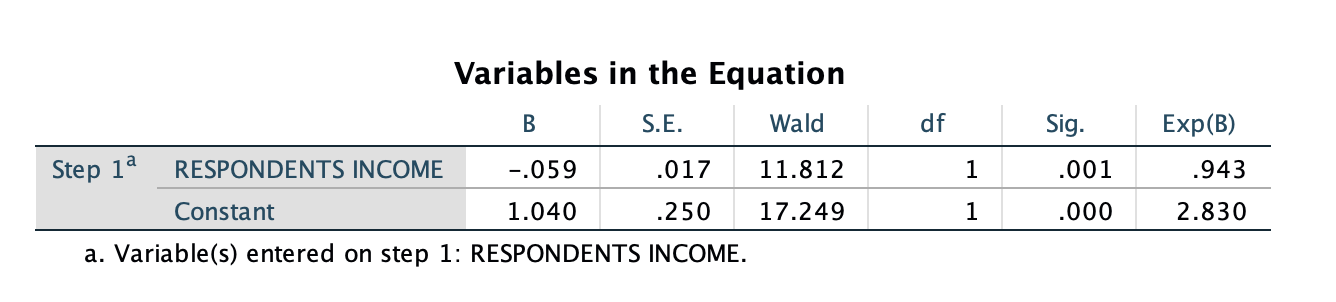
*Answer*: After eliminating the outlier cases, the evidence based on the regression analysis indicates there is no significant collinearity among the five continuous variables since all of them have tolerance bigger than 0.1.

*Plot 1* Multicollinearity analysis from SPSS

## Question 3

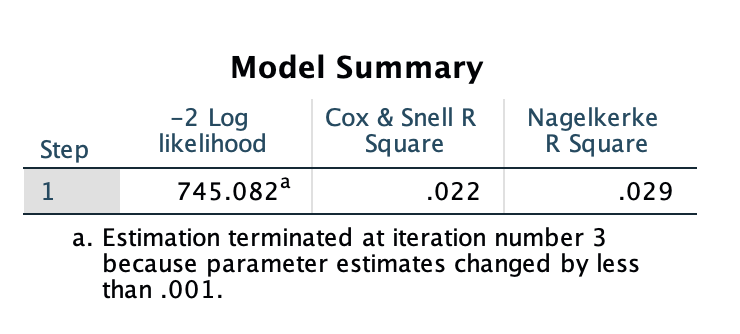
Conduct Binary Logistic Regression using the Forward: LR method.

1. Which variables were entered into the model?

*****Answer*: only respondents’ income is entered into the model with the constant.

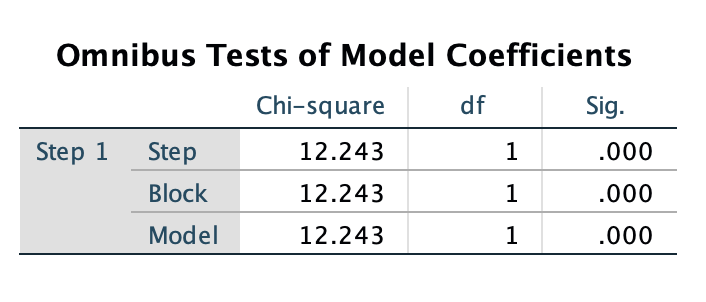
*Plot 2* Results from SPSS

2. To what degree does the model fit the data? Explain.

*****Answer*: The model does not fit the data well since the -2-loglikelihood (745.082) is extremely large. Meanwhile, Cox & Snell R square (0.22) and Nagelkerke R square (0.029) are all very low.

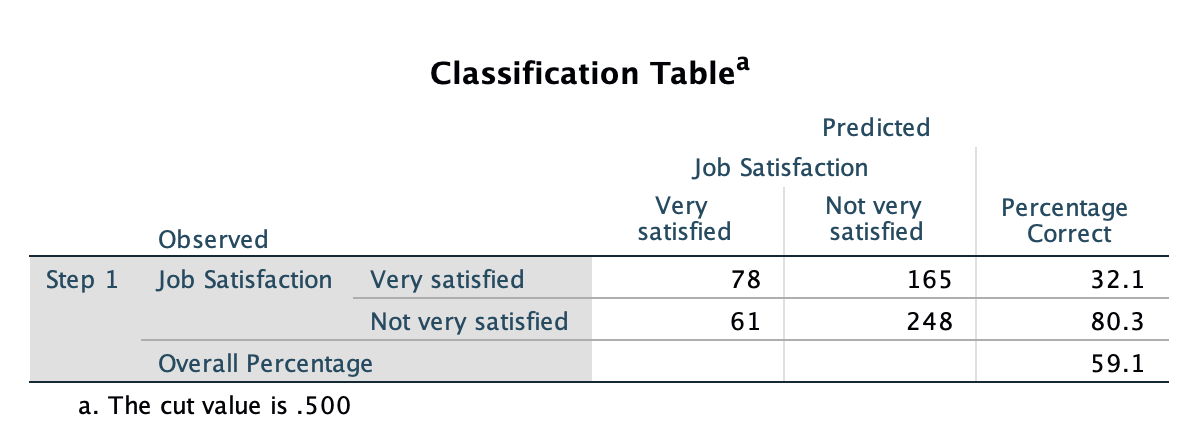
*Plot 3* Results from SPSS

3. Is the generated model significantly different from the constant-only model?

*****Answer*: Based on the likelihood ratio test, if we remove the only variable enter into the model (income), the change in the -2-log likelihood is 12.243, which is bigger than 1. The p-value of the test is 0, which reject the hypothesis that there is no difference between the models. In summary, there exists significantly difference from the constant-only model.

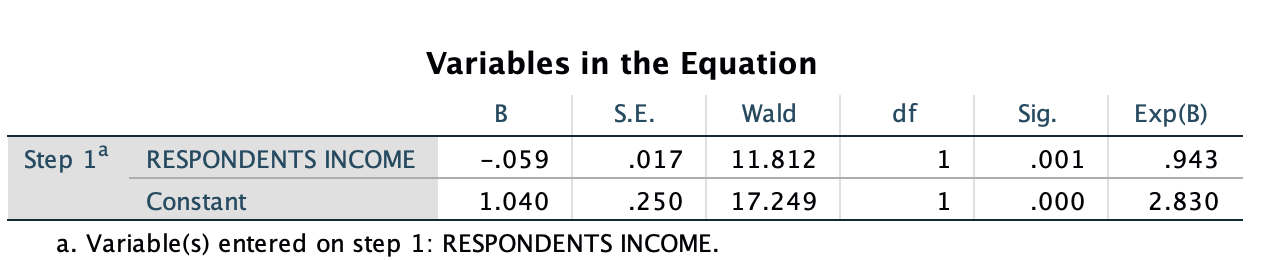
*Plot 4* Results from SPSS

4. How accurate is the model in predicting job satisfaction?

*Answer*: The overall classification accuracy is 59.1%, which is only slightly better than random guess.

*Plot 5* Results from SPSS

5. What are the odds ratios for the model variables? Explain.

*****Answer*: The odd ratio for variable income is 0.943, and for constant is 2.83. The odds ratios for the model variables are small.

*Plot 6* Results from SPSS

# Problem 2: questions in S&W (2003)

## Data set

The data set download is called “teachers.sav”.

## Life Table & Interpretations

* The SPSS code is:

get file '/Users/yichen/Desktop/teachers.sav'.

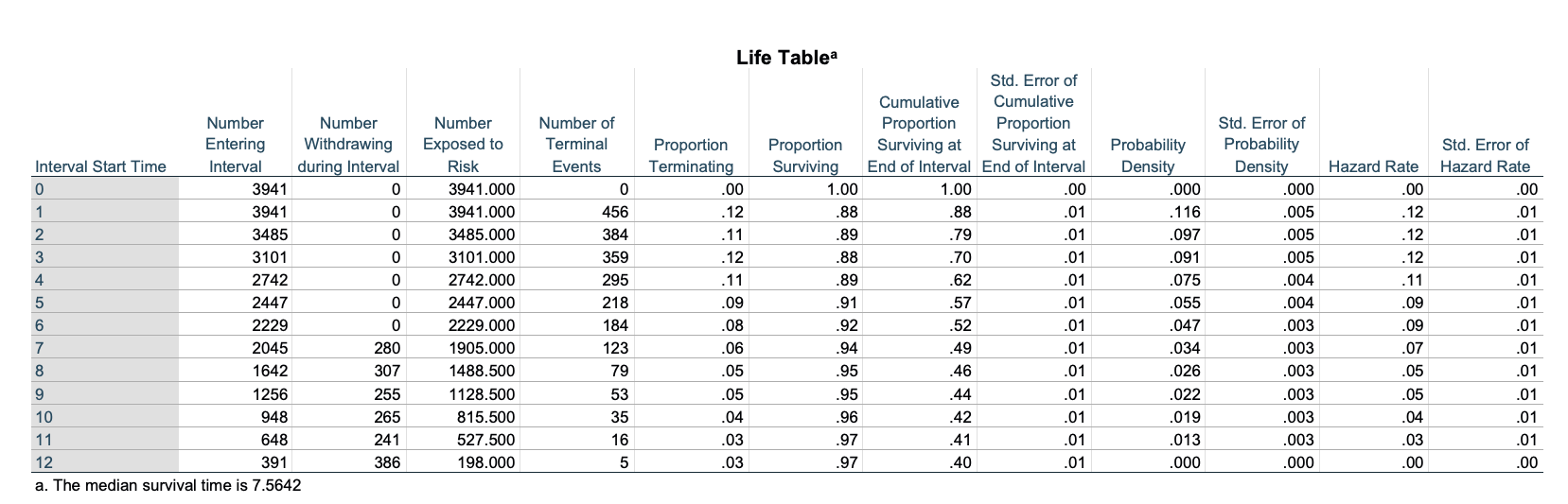
procedure output outfile=surv.

survival table = t

/interval = thru 12 by 1

/status = censor (0)

/write=tables.

* The SPSS result is:

*Plot 7* Life table from SPSS

* The narrative form of result:

*Table 1*: Life table (narrative form)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Year | Time Interval | Number | | | Proportion of | |
| Number entering interval | Number of Terminal Events | Number Withdrawing during Interval | Hazard Rate | Survival Function |
| 0 | [0,1) | 3941 | - | - | - | 1.00 |
| 1 | [1,2) | 3941 | 456 | 0 | 0.12 | 0.88 |
| 2 | [2,3) | 3485 | 384 | 0 | 0.11 | 0.79 |
| 3 | [3,4) | 3101 | 359 | 0 | 0.12 | 0.70 |
| 4 | [4,5) | 2742 | 295 | 0 | 0.11 | 0.62 |
| 5 | [5,6) | 2447 | 218 | 0 | 0.09 | 0.57 |
| 6\* | [6,7) | 2229 | 184 | 0 | 0.08 | 0.52 |
| 7 | [7,8) | 2045 | 123 | 280 | 0.06 | 0.49 |
| 8 | [8,9) | 1642 | 79 | 307 | 0.05 | 0.46 |
| 9 | [9,10) | 1256 | 53 | 255 | 0.04 | 0.44 |
| 10 | [10,11) | 948 | 35 | 265 | 0.04 | 0.43 |
| 11 | [11,12) | 648 | 16 | 241 | 0.02 | 0.42 |
| 12 | [12,13) | 391 | 5 | 386 | 0.01 | 0.41 |

Note \*: Median lifetime

* Interpretation

1. Time interval: time interval indicates the “beginning of the time” and the “end of the time” for discrete time window. In this case, time interval is year. It provides the base on which all the other indicators are calculated.

2. Number withdrawing during the interval / censoring: censoring is the number of observations withdraw during the time range. These are the individuals we cannot know when the events will happen for them. But censoring data provide partial information for estimation.

3. Risk set/ Number of exposed to risk: risk set is the number of subjects who are at the risk of experiencing an event, which is usually consist at each point in time of individuals who have been followed-up till that time and have not yet experience the event of interest just bore that time point. Hazard function is calculated based on risk sets.

4. Hazard rate/ function：hazard function is the conditional probability that individual will experience the event in the time period , given that he or she did not experience it in any earlier time.

5. Survival function: survival function is the cumulate the period-by-period risks of event together. The survival function for time can be calculated by the products of all .

6. median lifetime: median lifetime is the estimated time when the survival function equals to 0.5.

## Visualization & Explanations

## *Plot 8* Hazard function plot (left) and survival function plot (right)

## Interpretation

The hazard function plot on the left show the conditional probability of individual (teachers) experience the event (leaving the school) at each year, which do not need to be monotonous. Higher value of hazard function means higher probability that the individuals who survive in the previous years will experience the event at that year. Survival function on the right show a cumulative risk of event over time, which is always monotonous.

## SPSS Code

get file '/Users/yichen/Desktop/SURV.sav'.

compute p1 = left/begin.

do if $casenum = 1.

compute p2 = 1.

else.

compute temp = lag(p2).

compute p2 = (1-p1) \* temp.

format p1 p2 (f6.4) year begin left (f4.0) censored (f3.0).

end if.

if $casenum = 1 p1 =$sysmis.

list year begin left censored p1 p2.

graph

scatterplot(bivar) year with p1.

graph

scatterplot(bivar) year with p2.

\* Note 1: I mortified the results of life table manually and save the result into a sav (SURV.sav) file directly, instead of read the TXT output as what UCLA web suggest.

\* Note 2: the underline two codes are the main codes for drawing the plot, which is scatter plot in default.

## Report

Table 2. Discrete-time Hazard Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model A | Model B | Model C | Model D |
| Parameter estimates and asymptotic standard errors | | | | |
|  | -2.398\*\*\*  (0.270) | -2.994\*\*\*  (0.318) | -2.465\*\*\*  (0.274) | -2.893\*\*\*  (0.321) |
|  | -3.117\*\*\*  (0.387) | -3.700\*\*\*  (0.421) | -3.159\*\*\*  (0.389) | -3.585\*\*\*  (0.423) |
|  | -1.720\*\*\*  (0.222) | -2.281\*\*\*  (0.272) | -1.730\*\*\*  (0.225) | -2.150\*\*\*  (0.276) |
|  | -1.287\*\*\*  (0.210) | -1.823\*\*\*  (0.259) | -1.285\*\*\*  (0.213) | -1.693\*\*\*  (0.265) |
|  | -1.163\*\*\*  (0.230) | -1.654\*\*\*  (0.270) | -1.136\*\*\*  (0.232) | -1.518\*\*\*  (0.276) |
|  | -.731\*\*\*  (0.239) | -1.179\*\*\*  (0.272) | -.642\*\*\*  (0.243) | -1.010\*\*\*  (0.281) |
|  |  | .874\*\*\*  (0.217) |  | .661\*\*\*  (0.237) |
|  |  |  | .443\*\*\*  (0.114) | .296\*\*\*  (0.125) |
| Goodness-of-fit | | | | |
|  | -325.98 | -317.33 | -318.59 | -314.57 |
| Deviance | 651.96 | 634.66 | 637.17 | 629.15 |
| N parameter | 6 | 7 | 7 | 8 |
| AIC | 663.96 | 648.66 | 651.17 | 645.15 |
| BIC | 681.00 | 668.54 | 671.05 | 667.87 |

Note: all the model results are based comes from SPSS. Hypothesis tests are ignored in this table since it is not directly related to the model results.

Based on the t tests, the parameter estimation in all of four models are good. Meanwhile, the result from AIC and BIC indicates that model D may be the best model.

* **SPSS Code**

get file '/Users/ yichen /Desktop/ firstsex\_pp.sav'.

sort cases by pt.

temporary.

split file by pt.

logistic regression var = event

/method = enter d7 to d12

/origin

/save pred (pred).

logistic regression var = event

/method = enter d7 d8 d9 d10 d11 d12

/origin

logistic regression var = event

/method = enter d7 d8 d9 d10 d11 d12 pt

/origin.

logistic regression var = event

/method = enter d7 d8 d9 d10 d11 d12 pas

/origin.

logistic regression var = event

/method = enter d7 d8 d9 d10 d11 d12 pt pas

/origin.

\* Note 1: firstsex\_pp.sav already dummy code all the time interval.

# Problem 3: Research Proposal

## Purpose

In this study, I will introduce how the survival model could be used for measuring the online discussion behavior. I take a video-based learning platform “[Vialogues](https://www.vialogues.com/)” as an example.

Vialogues (Agarwala, Hsiao, Chae, & Natriello, 2012) is a video-driven discussion tool developed by EdLab at Teachers College, which is used for many courses (e.g., HUDK 4052) for discussions and team projects in real class environment. A vialogue consists of a video and all the discussions associated with the video.

## Literature Review and Theory

Video-driven discussion (VDD) has been widely used for decades in education to promote reflection, critical thinking, and constructive learning (Copeland & Decker, 1996; Koc, Peker, & Osmanoglu, 2009; Close, Scherr, Close, & McKagan, 2012). While traditional video learning materials alone support learning passively, video-driven discussion platforms provide an active learning environment through asynchronous discussion and content sharing (Sherin, 2003b).

The use of video as an instructional tool has become widely employed in digital learning environments. Technologies enabling video based education have been applied in a broad range of educational contexts (Giannakos et al., 2015), e.g., broadcasting lectures in distance education (Maag, 2006), delivering recordings of in-class lectures for the purpose of review- ing (Brotherton & Abowd, 2004), and providing supplementary video learning materials for self-study (Dhonau & McAlpine, 2002). Traditional video platforms primarily support passive learning, but with innovative technologies available today, video tools have evolved to provide active learning environments for learners to discuss video and share content col- laboratively. Examples include YouTube (http://youtube.com), Vimeo (http://vimeo.com), TED Ed (http://ed.ted.com). A number of benefits to using video in education have been reported over several decades of research (Passey, 2006; Poquet et al., 2018; Schwan & Riempp, 2004; Vieira et al., 2014; Lee & Sharma, 2008; Traphagan et al., 2010; Ljuboje- vic et al., 2014). These benefits include a) the opportunity to manipulate and interpret principles and processes situated in the video, b) linking content and concepts to every- day experience, c) self evaluation, modifying, testing and revising one’s own knowledge, d) inspiring and engaging students.

With the growing number of online learning tools and increasing volume of user data, understanding students’ engagement with such technologies has become an important sub- ject. Such understanding allows researchers / developers to support their learning strategies and approaches. In the past, many researchers studied learners’ engagement and behavioral patterns in various e-learning environments. Del Valle & Duffy (2009) examined profiles of student approaches to online courses developed for distance learning. In Lust et al. (2013) and Hung & Zhang (2008), researchers studied learner patterns in the context of learning management systems (LMS). Brooks et al. (2011) explored students’ engagement patterns with video lectures and Mirriahi et al. (2016) analyzed user profiles for video annotation tools. In Hou (2012), user profiles for game-based learning were studied.

With the development of innovative technologies, a growing number of online video-discussion learning tools have been created and widely applied in education (Giannakos, Chorianopoulos, & Chrisochoides, 2015). Examples including YouTube (http://youtube.come), Vimeo (http://vimeo.com), TED-Ed (http://ed.ted.com), and Vialogues (https://vialogues.com). However, the majority of these online resources have not been consistently applied in a real class environment. Furthermore, the increasing volume of user data has not been explored fully to understand the students’ learning behavior.

**Research Question**

Can students’ discussion behavior on the video-based learning platform be correctly predicted from knowledge of students’ background and video characteristics? If it can be predicted correctly, which variables are central in the prediction of comment posting behavior intensity?

## Method

Survival model have strength to modeling the longitudinal user behavior with the possible censor, like visitor leaves the page before the video playing finished (right censoring). Covariates can also be incorporated into the model (even in a hierarchical modeling design) to improve the interpretation ability and reduce the estimation bias.

## Data

I will take HUDK 4052 as example. Forty-eight Students are required to watch 28 Vialogues and discuss the video during the whole semester as part of their homework. Web analytics applications (e.g., Google Analytics) have been applied in Vialogues to track website activity such as session duration, pages per session, bounce rate etc. of individuals using the site, along with the information on the source of the traffic. These data set provide the detailed behavior logs of the participants in Vialogues. Since the visitors have to login to common on the video, it is straightforward to identify the students across different visiting times through their visitor ID. Meanwhile, professors who use the Vialogues during the class have many background information (e.g., gender, department, and weekly homework score) of the students. The characteristic of each video (e.g., video topic) may also make an effect on students’ discussion behaviors. These data could be used as the predictor or additional in the model.

The target event is the common posting. This event can be defined binary (make the comment or not make the comment), but better to be polytomous (number of words for each comment). Meanwhile, event can be spell (appear only once like death), but better to be repeatable. Since the length of video is not always the same, it makes it hard to compare the event directly across different Vialogues. I can normalize the time interval as the proportional value. We can split the continuous time range into multiple discrete time components. For example, the time between 0% to 10% of the video can be the first interval, and so on.

In this study, I propose the Cox proportional hazards model which assumes the time between events in a Poisson process. The hazard function for this model has the form:

, where **X** represents the covariates matrix for the participant , **Y** is the covariate matrix for the video , and is the baseline change over time. This model can also be called Cox Regression without a time-dependent covariate.

## Implication

This model will help to identify the association between student, class, professor, or video variables and students’ discussion in VDD with the real student behavior data.