Chapter 12 Lab Answer Key

Tianwei Liu

December 11

Preparation

```
require(knitr)
require(haven)  ## install.packages("haven")
require(car)    ## install.packages("car")
require(AER)    ## install.packages("AER")
library(Hmisc)  ## use the describe command
library(mfx)

opts_chunk$set(echo = TRUE)
options(digits = 6)
```

(a) Use a LPM to estimate the effect of passenger class on survival.

```
# Create dummies
dta$pclass_2 <- (dta$pclass == 2)
dta$pclass_3 <- (dta$pclass == 3)

# Run LMP
reg.1a <- lm(survived ~ pclass_2 + pclass_3, data = dta)
summary(reg.1a)
```

```
##
## Call:
## lm(formula = survived ~ pclass_2 + pclass_3, data = dta)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.619 -0.255 -0.255  0.381  0.745
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.6192     0.0257   24.08  < 2e-16 ***
## pclass_2TRUE  -0.1896     0.0378   -5.01  6.2e-07 ***
## pclass_3TRUE  -0.3639     0.0310  -11.73  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.462 on 1306 degrees of freedom
## Multiple R-squared:  0.0977, Adjusted R-squared:  0.0963
## F-statistic: 70.7 on 2 and 1306 DF,  p-value: <2e-16
```

- The reference category in the model we selected is the first class. The result indicates that the probability of first class passengers surviving Titanic is 62%, as the p-value associated with the intercept is very small, it is statistically significant as 0.001 level of significant.

- Passengers in Class 2 are 19% less likely to survive than first class passengers, so 43% probability of surviving. The difference in means is statistically significant because the p-value is smaller than 0.001 level of significance.
- Passengers in Class 3 are 36% less likely to survive compared with first class passengers, so 25% probability of surviving. The difference in means is also statistically significant because the p-value is very small (smaller than 0.001 level of significance).

(b) Assess the following: did being a women or child affected survival? Did boarding location (a rough proxy for country of origin) affect survival? Ireland is Queenstown ("Q"), France is Cherbourg ("C") and the Englad is Southampton ("S"). Treat Southampton as the reference category. Control for age in your model.

```
## Create dummy variables
dta$female    = (dta$sex == "female")
dta$child     = (dta$age < 17)
dta$Queenstown= (dta$embarked == "Q")
dta$Cherbourg = (dta$embarked == "C")
```

```
reg.1b <- lm(survived ~ pclass_2 + pclass_3 + female + child + Queenstown + Cherbourg + age, data = dta)
summary(reg.1b)
```

```
##
## Call:
## lm(formula = survived ~ pclass_2 + pclass_3 + female + child +
##     Queenstown + Cherbourg + age, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0699 -0.2387 -0.0725  0.2427  1.0243
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.50200    0.05490    9.14  < 2e-16 ***
## pclass_2TRUE   -0.16197    0.03670   -4.41  1.1e-05 ***
## pclass_3TRUE   -0.31691    0.03451   -9.18  < 2e-16 ***
## femaleTRUE      0.48862    0.02549   19.17  < 2e-16 ***
## childTRUE       0.08680    0.04448    1.95  0.05128 .
## QueenstownTRUE -0.10059    0.05747   -1.75  0.08037 .
## CherbourgTRUE   0.11348    0.03248    3.49  0.00050 ***
## age            -0.00375    0.00112   -3.35  0.00083 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.388 on 1038 degrees of freedom
##   (263 observations deleted due to missingness)
## Multiple R-squared:  0.381,  Adjusted R-squared:  0.377
## F-statistic: 91.4 on 7 and 1038 DF,  p-value: <2e-16
```

- Being a woman and boarding at Cherbourg affect survival.
- The coefficient for being female and boarding at Cherbourg are very statistically significant because the p-values associated with the coefficients are very small (smaller than 0.001 level of significance).
- The coefficients for boarding at Queenstown and being a child are not statistically significant because the p-values are larger than 0.05 level of significance. Therefore, we fail to reject the null and conclude that there is no effect.

(c) For the model from above, what are the minimum and maximum predicted probabilities of survival?

```
max(predict(reg.1b))
```

```
## [1] 1.13086
```

```
min(predict(reg.1b))
```

```
## [1] -0.180043
```

- The maximum probability is 1.13; the minimum probability is -0.18. These probability values do not make sense because probability values range from 0 to 1.

(d)What is the name, age, gender and passenger class of the person with the lowest probability of surviving?

```
id <- as.numeric(names(predict(reg.1b)[predict(reg.1b) == min(predict(reg.1b))]))
c(dta$name[id], dta$age[id], dta$sex[id], dta$pclass[id])
```

```
## [1] "Connors, Mr. Patrick" "70.5"                 "male"
## [4] "3"
```

(e) What is the name of the person with the highest probability of surviving?

```
id2 <- as.numeric(names(predict(reg.1b)[predict(reg.1b) == max(predict(reg.1b))]))
dta$name[id2]
```

```
## [1] "Hippach, Miss. Jean Gertrude"
```

(f) Estimate a probit model where survival is a function of (only) passenger class. Treat passenger class as a nominal variable. Compare statistical significance to a similar LPM model. Is there an easy way to interpret the coefficients?

```
reg3 <- glm(survived ~ pclass_2 + pclass_3, data = dta, family = binomial(link = 'probit'))
summary(reg3)
```

```
##
## Call:
## glm(formula = survived ~ pclass_2 + pclass_3, family = binomial(link = "probit"),
##     data = dta)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.390  -0.768  -0.768   0.979   1.653
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.3034     0.0709    4.28  1.9e-05 ***
```

```
## pclass_2TRUE  -0.4808     0.1038   -4.63  3.6e-06 ***
## pclass_3TRUE  -0.9613     0.0873  -11.01  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1613.3  on 1306  degrees of freedom
## AIC: 1619
##
## Number of Fisher Scoring iterations: 4
```

In the LPM models, all three coefficients are statistically at 0.001 level of significance (p-values are smaller than 0.001). Similarly, in the probit model, all three coefficients are statistically significant at 0.001 level.

There is no easy of to intepret the coefficients. Because by nature, the prediction of a probit model depends on the particular values of Xi and also the value of other independent variables.

(g) Estimate a probit model where survival is a function of passenger class (treated as a nominal variable) age, gender, child and embarkation location. What is the minimum and maximum fitted value?

```
reg4 <- glm(survived ~ pclass_2 + pclass_3 + age + female + child + Queenstown + Cherbourg, data = dta,
summary(reg4)
```

```
##
## Call:
## glm(formula = survived ~ pclass_2 + pclass_3 + age + female +
##     child + Queenstown + Cherbourg, family = binomial(link = "probit"),
##     data = dta)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.669  -0.696  -0.424   0.681   2.522
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.20231    0.21119    0.96  0.33810
## pclass_2TRUE   -0.58732    0.13771   -4.27    2e-05 ***
## pclass_3TRUE   -1.11893    0.13324   -8.40  < 2e-16 ***
## age            -0.01496    0.00443   -3.38  0.00072 ***
## femaleTRUE      1.50184    0.09552   15.72  < 2e-16 ***
## childTRUE       0.23096    0.16872    1.37  0.17103
## QueenstownTRUE -0.38235    0.22862   -1.67  0.09444 .
## CherbourgTRUE   0.41236    0.12201    3.38  0.00073 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1414.6  on 1045  degrees of freedom
## Residual deviance:  967.1  on 1038  degrees of freedom
```

```
##   (263 observations deleted due to missingness)
## AIC: 983.1
##
## Number of Fisher Scoring iterations: 5
```

```
min(predict(reg4))
```

```
## [1] -2.35363
```

```
max(predict(reg4))
```

```
## [1] 2.10811
```

```
print (pnorm(min(predict(reg4))))
```

```
## [1] 0.00929558
```

```
print (pnorm(max(predict(reg4))))
```

```
## [1] 0.982489
```

- The minimum fitted value is -2.35, which corresponds to a 0.009 probability.
- The maximum fitted value is 2.11, which corresponds to a 0.982 probability.

(h) What is the name of the person with the lowest probability of surviving?

```
id3 <- as.numeric(names(predict(reg4)[predict(reg4) == max(predict(reg4))]))
dta$name[id3]
```

```
## [1] "Hippach, Miss. Jean Gertrude"
```

(i) For the above model, what is the effect of growing one year older (for an adult)? (Do this "manually", using the observed-value, discrete difference method described in the book/lecture.)

```
p1 = pnorm(reg4$coefficients[1] + reg4$coefficients[2]*dta$pclass_2 + reg4$coefficients[3]*dta$pclass_3
```

```
p2 = pnorm(reg4$coefficients[1] + reg4$coefficients[2]*dta$pclass_2 + reg4$coefficients[3]*dta$pclass_3
```

```
diffage = p2 - p1
describe(diffage)
```

```
## diffage  Format:%8.0g
##         n   missing  distinct      Info      Mean       Gmd       .05
##      1046       263       470         1 -0.003881  0.001693 -0.005929
##       .10       .25       .50       .75       .90       .95
## -0.005846 -0.005321 -0.003959 -0.002590 -0.001999 -0.001588
##
## lowest : -0.005967884 -0.005967859 -0.005967724 -0.005967699 -0.005966684
## highest: -0.000972610 -0.000958824 -0.000758606 -0.000444821 -0.000367526
```

Using the observed value, discrete difference method, the average effect of age is -0.00388, meaning that a one-year growth in age is associated with 0.00388 decrease in the rate of survival.

(j) Compare the probit effect of age to the LPM effect of age in part (b)

This result is similar to the coefficient given by the LPM model, which is -0.00375.

(k) What is the effect of the passenger class, female and child variables in the above probit model? Use the mfx package as described in the lecture. Compare the predicted effects of these variables in the probit model to the results in the LPM in part (b). You need only discuss one of these variables, but please note all of them as you do the work.

```
probitmfx(formula = survived ~ pclass_2 + pclass_3 + age + female + child + Queenstown + Cherbourg, data
```

```
## Call:
## probitmfx(formula = survived ~ pclass_2 + pclass_3 + age + female +
##     child + Queenstown + Cherbourg, data = dta, atmean = FALSE)
##
## Marginal Effects:
##                     dF/dx Std. Err.      z    P>|z|
## pclass_2TRUE   -0.144721  0.031401 -4.609 4.05e-06 ***
## pclass_3TRUE   -0.306976  0.034492 -8.900  < 2e-16 ***
## age            -0.003887  0.001135 -3.424 0.000618 ***
## femaleTRUE      0.488694  0.028049 17.423  < 2e-16 ***
## childTRUE       0.061386  0.045677  1.344 0.178973
## QueenstownTRUE -0.095398  0.054236 -1.759 0.078585 .
## CherbourgTRUE   0.112297  0.034272  3.277 0.001050 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "pclass_2TRUE"   "pclass_3TRUE"   "femaleTRUE"     "childTRUE"
## [5] "QueenstownTRUE" "CherbourgTRUE"
```

- Being in passenger class 2 or 3 has a negative effect on the probability of survival. Both of the two coefficients are statistically significant because two p-values are much lower than 0.001 level of significance.
- Being female has a positive effect on the probability of survival, and the coefficient is strongly significant as the z-score is very large and p-value very samll.
- Being a child does not have an effect on the probability of survival because the p-value is greater than conventional 0.05 level of significance. Therefore we fail to reject the null and conclude that being a child does not have an effect.
- The predicted effects of these variables in the probit model are pretty similar to ones in the LPM. For example, the marginal effect of being female is 0.489 in the probit model with a very large z-score. In the LPM, being female also has 0.489 effect on the probability of survival and also has a very large t-statistic.