The Analytics Edge

Test Your Knowledge of SVD and Censored Data

Note to all. I have compiled the answers in the following format – for each question, the qualitative or "written" solutions will be provided together with their sub-questions. The R scripts (as well as the console outputs) will be provided *after* each whole question, followed by all the relevant plots. If I have missed anything in the solutions, or if you have any questions, you may email me at benjamin_tanwj@mymail.sutd.edu.sg. Thank you!

1. In this exercise, you will study a technique called latent semantic indexing, which applies singular value decomposition to create a low dimensional representation of data that is designed to capture semantic similarity of words. A list of all 460 unique words/terms that occurs in a set of 9 documents is provided in lsiWords.txt. A document by term matrix is in lsiMatrix.txt.

   (a) Use the read.table() function to read in the data from the lsiMatrix.txt file. Create a matrix X that is the transpose of the lsiMatrix, so that each column represents a document. Compute the singular value decomposition of X and make an approximation to it using the first 2 singular values and vectors. Plot the low dimensional representation of the 9 documents in two dimensions by using the right singular vectors. Which two documents appear to be the closest to each other in this low dimensional representation?

   Solution. Plot found here. Documents 7 and 8 are the closest in this low-dimensional representation.

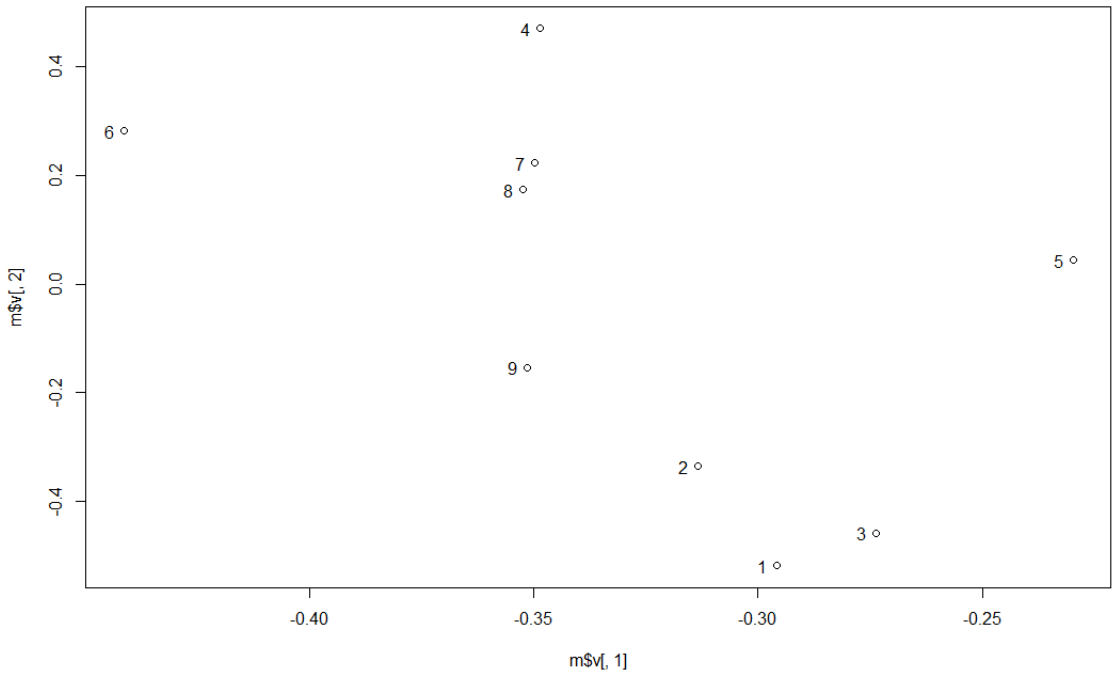
   (b) Consider finding documents that are about alien abductions. If you look at the words in lsiWords.txt, there are three versions of this word, term 23 ("abducted"), terms 24 ("abduction"), term 25 ("abductions"). Suppose you want to find documents containing the word "abducted", documents 2 and 3 contain it but document 1 does not. In the original dataset however document 1 is also related to the topic. Thus LSI should also find document 1. Create a test document q containing the one word "abducted" and project it into the 2D subspace to make $\hat{q}$. Now compute the cosine similarity between $\hat{q}$ and the low dimensional representation of all the documents. What are the top 3 closest matches?

   Solution. The decreasing order of cosine similarities is: 1, 3, 2, 9, 5, 8, 7, 6, 4. The top 3 matches are documents 1, 3, and 2.

R Scripts.

```
> #1)
> #a)
> lsiMatrix <- read.table("lsiMatrix.txt")
> X <- as.matrix(t(lsiMatrix))
> str(X)
 int [1:460, 1:9] 2 0 0 0 0 0 0 2 0 4 ...
 - attr(*, "dimnames")=List of 2
  ..$ : chr [1:460] "V1" "V2" "V3" "V4" ...
  ..$ : NULL
> dim(X)
[1] 460   9
> m <- svd(X)
> mreduced <- m$u[,1:2]%*%diag(m$d[1:2])%*%t(m$v[,1:2])
> plot(m$v[,1],m$v[,2])
> text(m$v[,1],m$v[,2],c(1:9),adj=2)

> #b)
> q <- matrix(0,nrow=460,ncol=1)
> q[23] <- 1
> hatq <- solve(diag(m$d[1:2]))%*%t(m$u[,1:2])%*%q
> str(hatq)
 num [1:2, 1] -3.45e-05 -6.40e-04
> cosine <- matrix(0,9,1)
> for(i in 1:9){
+ cosine[i] <- sum(hatq*m$v[i,1:2])/sqrt(sum(hatq^2)*sum(m$v[i,1:2]^2))
+ }
> order(cosine,decreasing=T)
[1] 1 3 2 9 5 8 7 6 4
```



Plot for Q1a. Click here to go back to the question.

2. The data in the file **mroz.csv** includes data on hours worked for 753 married women. This dataset is taken from the paper by Mroz, T.A. (1987) titled "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions", Econometrica, 55, 765-799. The variables in the dataset include:

   - **hours**: Hours worked
   - **kidslt6**: Number of children less than six years old
   - **kidsge6** Number of kids between 6 and 18 years of age
   - **age**: Age
   - **educ**: Years of education
   - **exper**: Past years of labor market experience
   - **nwifeinc**: Husband's earnings, measured in thousands of dollars

   (a) Quadratic functions are often used in labor economics to capture increasing or decreasing marginal rates. Start by defining a new variables exper2 which is defined as the square of the exper variable. Run a linear regression on the hours worked using all the variables including exper2 and report on the R-squared and adjusted R-squared. Write down the fitted equation. Does the result indicate that the exper2 variable is significant at the 5% level? Does the results indicate that experience has an increasing or diminishing marginal effect on wage?

   Solution. The fitted equation is

   $$hours = 1330.4 - 442 kidslt6 - 32.77 kidsge6 - 30.51 age + 28.76 educ + 65.67 exper - 3.44 nwifeinc - 0.70 exper2.$$

   The $R^2 = 0.2656$ and the adjusted $R^2 = 0.2587$. The coefficient of the *exper2* variable has a p-value of 0.031, indicating that it is significant at the 0.05 level. Since the coefficient is negative, this shows a decreasing marginal effect on wage.

   (b) What is the range of fitted values for hours from the result in part (a)? How many of the fitted values are below 0? How many observations in the dataset have hours = 0? How do these two numbers compare?

   Solution. The fitted values range from -719.8 to 1614.7. 39 out of 753 observations have negative fitted values in the linear regression. On the other hand, 325 out of the 753 observations have actual zero values, which is significantly larger than the number of negative fitted values.

   (c) Estimate a Tobit model to predict hours using all the variables including exper2. Compare the signs of the coefficients with the results in part (a) and check if the results match?

Solution. The given linear equation is

$$hours = 965 - 894 kidslt6 - 16 kidsge6 - 54.4 age + 80.65 educ + 131.5 exper - 8.81 nwifeinc - 1.86 exper2.$$

The signs of this linear equation is consistent with the results for the linear regression.

(d) We will now compare the R-squared values from the results in part (a) and (c). Remember that for the linear regression model, the R-squared value is the squared correlation between the fitted values of hours and the actual value. We now define the R-squared for a Tobit model in a similar manner. Given the estimated values of β and the scale parameter σ, the predicted mean value of the dependent variable is given as:

$$E(y_i|x_i) = \beta' x_i \Phi(\beta' x_i/\sigma) + \sigma\phi(\beta' x_i/\sigma).$$

Compute this value for all observations (remember that the predict function in survreg only returns the $\beta' x_i$ values). Now define the R-squared value by computing the correlation of these predicted value with the actual value of the hours. Compare the R-squared values of the two models and comment on which is preferred.

Solution. The R^2 value for the Tobit model is 0.274 which is slightly better than the linear regression. However, it is to be noted that the Tobit model was not designed to maximise the R^2 value (its objective is to maximise log-likelihood).

(e) We now compare the estimates from the two models as a function of education. Assume that all the variables other than educ are set at their mean values. Write down the linear equation that describes the average hours worked as a function of the education level from the linear regression model. What is the estimated value at 8 and 12 years of education from the linear regression model? Compare this with the estimated values from the Tobit model. Is there increasing or decreasing marginal effect of education on the hours worked in the Tobit model?

Solution. For the linear regression model, the equation is given as $E(hours) = 387.19 + 28.76 educ$. At 8 and 12 years of education, this gives an estimated value of 617 and 732 respectively.

For the Tobit model, the equation is nonlinear:

$$E(hours) = (80.54 educ - 694.1)\Phi\Big(\frac{80.54 educ - 694.1}{1122}\Big) + 1122\,\phi\Big(\frac{80.54 educ - 694.1}{1122}\Big)$$

At 8 and 12 years of education, this gives estimated values of 423.57 and 597.63 respectively.

The values from Tobit gives lower estimates of expected hours worked for these levels of education. The marginal effects are also increasing from education, on the hours worked.

Note that the difference bewteen the Tobit estimates is larger than that of the linear estimate.

R Scripts.

```
> #2)
> #a)
> mroz <- read.csv("mroz.csv")
> mroz$exper2 <- mroz$exper^2
> m1 <- lm(hours~.,data=mroz)
> summary(m1)

Call:
lm(formula = hours ~ ., data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-1511.3  -537.8  -146.9   538.1  3555.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1330.4824   270.7846   4.913 1.10e-06 ***
kidslt6     -442.0899    58.8466  -7.513 1.66e-13 ***
kidsge6      -32.7792    23.1762  -1.414   0.1577
age          -30.5116     4.3639  -6.992 6.04e-12 ***
educ          28.7611    12.9546   2.220   0.0267 *
exper         65.6725     9.9630   6.592 8.23e-11 ***
nwifeinc      -3.4466     2.5440  -1.355   0.1759
exper2        -0.7005     0.3246  -2.158   0.0312 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 750.2 on 745 degrees of freedom
Multiple R-squared:  0.2656,Adjusted R-squared:  0.2587
F-statistic:  38.5 on 7 and 745 DF,  p-value: < 2.2e-16


> #b)
> summary(m1$fitted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -719.8   417.5   737.7   740.6  1093.1  1614.7
```

```
> table(m1$fitted>0)

FALSE  TRUE
   39   714
> table(mroz$hours>0)

FALSE  TRUE
  325   428

> #c)
> library(survival)
> m2 <- survreg(Surv(hours,hours>0,type="left")~.,data=mroz,dist="gaussian")
> summary(m2)

Call:
survreg(formula = Surv(hours, hours > 0, type = "left") ~ .,
    data = mroz, dist = "gaussian")
               Value Std. Error      z       p
(Intercept)  965.3053   446.4361   2.16 0.03060
kidslt6     -894.0217   111.8780  -7.99 1.3e-15
kidsge6      -16.2180    38.6414  -0.42 0.67470
age          -54.4050     7.4185  -7.33 2.2e-13
educ          80.6456    21.5832   3.74 0.00019
exper        131.5643    17.2794   7.61 2.7e-14
nwifeinc      -8.8142     4.4591  -1.98 0.04808
exper2        -1.8642     0.5377  -3.47 0.00053
Log(scale)     7.0229     0.0371 189.51 < 2e-16

Scale= 1122

Gaussian distribution
Loglik(model)= -3819.1   Loglik(intercept only)= -3954.9
Chisq= 271.59 on 7 degrees of freedom, p= 7e-55
Number of Newton-Raphson Iterations: 4
n= 753


> #d)
> summary(m1)

Call:
```

```
lm(formula = hours ~ ., data = mroz)

Residuals:
    Min      1Q  Median      3Q     Max
-1511.3  -537.8  -146.9   538.1  3555.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1330.4824   270.7846   4.913 1.10e-06 ***
kidslt6     -442.0899    58.8466  -7.513 1.66e-13 ***
kidsge6      -32.7792    23.1762  -1.414   0.1577
age          -30.5116     4.3639  -6.992 6.04e-12 ***
educ          28.7611    12.9546   2.220   0.0267 *
exper         65.6725     9.9630   6.592 8.23e-11 ***
nwifeinc      -3.4466     2.5440  -1.355   0.1759
exper2        -0.7005     0.3246  -2.158   0.0312 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 750.2 on 745 degrees of freedom
Multiple R-squared:  0.2656,Adjusted R-squared:  0.2587
F-statistic:  38.5 on 7 and 745 DF,  p-value: < 2.2e-16

> cor(m1$fitted,mroz$hours)^2
[1] 0.2656245
> p2 <- predict(m2,newdata=mroz)
> predict2 <- (p2*pnorm(p2/m2$scale)) + (m2$scale*dnorm(p2/m2$scale))
> cor(predict2,mroz$hours)^2
[1] 0.2742441

> #e)
> const1 <- m1$coef[1] + m1$coef[2]*mean(mroz$kidslt6) + m1$coef[3]*mean(mroz$kidsge6)
           + m1$coef[4]*mean(mroz$age) + m1$coef[6]*mean(mroz$exper)
           + m1$coef[7]*mean(mroz$nwifeinc) + m1$coef[8]*mean(mroz$exper2)
> const1 + m1$coef[5]*8
(Intercept)
   617.2817
> const1 + m1$coef[5]*12
(Intercept)
   732.3262
> const2 <- m2$coef[1] + m2$coef[2]*mean(mroz$kidslt6) + m2$coef[3]*mean(mroz$kidsge6)
```

```
           + m2$coef[4]*mean(mroz$age) + m2$coef[6]*mean(mroz$exper)
           + m2$coef[7]*mean(mroz$nwifeinc) + m2$coef[8]*mean(mroz$exper2)

> ((m2$coef[5]*8+const2)*pnorm(((m2$coef[5]*8+const2))/m2$scale))
 + (m2$scale*dnorm(((m2$coef[5]*8+const2))/m2$scale))
    educ
423.5725

> ((m2$coef[5]*12+const2)*pnorm(((m2$coef[5]*12+const2))/m2$scale))
 + (m2$scale*dnorm(((m2$coef[5]*12+const2))/m2$scale))
    educ
597.6833
```