DSA1101 Final Exam Solution

Nicholas Russell Saerang

Semester 2, 2016/2017

1. (a) Since we are only interested in itemsets with larger than 0.05 support, the itemsets count must be larger than $200,000 \times 0.05 = 10,000$.
   Therefore, the subsets with significant support are

   $$\{A\}, \{B\}, \{C\}, \{D\}, \{BC\}, \{AD\}, \{BD\}, \{CD\}, \{BCD\}$$

   (b) Recall that

   $$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \wedge Y)}{\text{Support}(X)}$$

   Hence,

   $$\text{Support}(X \wedge Y) = \text{Confidence}(X \to Y) \times \text{Support}(X)$$

   $$\begin{aligned}\text{Support}(\{BCDE\}) &= \text{Confidence}(\{BCD\} \to \{E\}) \times \text{Support}(\{BCD\}) \\ &= 0.667 \times \frac{15}{200} \\ &= 0.050025 < 0.075\end{aligned}$$

   The exact same thing happens to Support({ADE}).

   $$\begin{aligned}\text{Support}(\{BCE\}) &= \text{Confidence}(\{BC\} \to \{E\}) \times \text{Support}(\{BC\}) \\ &= 0.429 \times \frac{35}{200} \\ &= 0.075075 > 0.075\end{aligned}$$

   The exact same thing happens to Support({BDE}) and Support({CDE}).

   $$\begin{aligned}\text{Support}(\{AE\}) &= \text{Confidence}(\{A\} \to \{E\}) \times \text{Support}(\{AE\}) \\ &= 0.3 \times \frac{50}{200} \\ &= 0.075\end{aligned}$$

   Since it's not larger than 0.075, we exclude this from our result.

   $$\begin{aligned}\text{Support}(\{BE\}) &= \text{Confidence}(\{B\} \to \{E\}) \times \text{Support}(\{B\}) \\ &= 0.292 \times \frac{120}{200} \\ &= 0.1752 > 0.075\end{aligned}$$

   The exact same thing happens to Support({CE}).

   $$\begin{aligned}\text{Support}(\{DE\}) &= \text{Confidence}(\{D\} \to \{E\}) \times \text{Support}(\{D\}) \\ &= 0.333 \times \frac{150}{200} \\ &= 0.24975 > 0.075\end{aligned}$$

Hence, all the itemsets having support larger than 0.075 are

$$\{BCE\}, \{BDE\}, \{CDE\}, \{BE\}, \{CE\}, \{DE\}$$

with the respective supports computed above.

2. We can use linear regression by expressing $\ln(P)$ as a function of $\frac{1}{T}$.
Let
$$T = c(1030, 1048, 1067, 1082, 1084, 1112, 1132, 1133, 1134, 1135, 1135, 1150)$$
and
$$P = c(0.104, 0.123, 0.178, 0.236, 0.290, 0.398, 0.555, 0.523, 0.557, 0.581, 0.622, 0.724)$$
Running the following R code has given you the final answer.

```
simple_LS <- function(x,y){
beta_1 <- (sum(x*y)-mean(y)*sum(x))/(sum(x^2)-mean(x)*sum(x));
beta_0 <- mean(y)-beta_1*mean(x);
return(c(beta_0,beta_1));
}

simple_LS(1/T,log(P))
```

Which gives you $\alpha = 16.97211$ and $\beta = -19886.13535$.
PS : I know it shouldn't be done with R since it is worth 15 points. However, since R is allowed to be used, the question is a giveaway.

3. Name the respective test data values A, B, C, D and E. Since B is the only data with NSAL $\geq$ 5940, we can predict B as 0. The rest has NSAL $<$ 5940.

Next, since E is the only data remaining with BED $<$ 60, we can predict E as 1. The rest has BED $\geq$ 60.

They all also have MCDAYS $<$ 244 and PCREV $<$ 19e+3. Next, since C is the only data remaining with FEXP $\geq$ 4320, we can predict C as 1. The rest has FEXP $<$ 4320.

Both of them have NSAL $<$ 4402 and MCDAYS $<$ 164. Next, we find that D is the only data remaining with MCDAYS $\geq$ 145, hence we predict D as 0. This leaves A with MCDAYS $<$ 145.

Finally, we predict A as 1 as the value BED of A is $\geq$ 61.

4. (a) The residuals of each data point (x_i, y_i) is

$$\epsilon_i = y_i - \alpha x_i$$

because the mean of this error is zero and they have equal variances. Hence, the residual sum of squares is

$$h(\alpha) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha x_i)^2$$

Using the least squares method, we must have

$$\begin{aligned}\frac{dh(\alpha)}{d\alpha} &= \sum_{i=1}^{n} 2(y_i - \alpha x_i)(-x_i) \\ &= 2\left(\alpha \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i\right) = 0\end{aligned}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

(b) We have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i-\alpha x_i)^2 &= \frac{1}{n}\left(\sum_{i=1}^{n}y_i^2 - 2\hat{\alpha}\sum_{i=1}^{n}x_iy_i + \hat{\alpha}^2\sum_{i=1}^{n}x_i^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}y_i^2 - 2\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}\sum_{i=1}^{n}x_iy_i + \left(\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}\right)^2\sum_{i=1}^{n}x_i^2\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}y_i^2 - 2\frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2} + \frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2}\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}y_i^2 - \frac{\left(\sum_{i=1}^{n}x_iy_i\right)^2}{\sum_{i=1}^{n}x_i^2}\right)\\
&= \frac{\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}x_iy_i\right)^2}{n\sum_{i=1}^{n}x_i^2}
\end{aligned}$$

(c) Since the LHS is equivalent to $\frac{\sum_{i=1}^{n}\epsilon_i}{n}$, we do expect this value to become 0. This is due to the nature of linear regression that tries to fit a line with zero sum of residuals and minimum residual sum of squares.

5. (a) **Found a quicker solution so just read the red box instead of the whole question (update 25 November 2020).**

We use the fact that

$$\mathrm{var}(a+bx) = b^2\mathrm{var}(x)$$

$$\mathrm{var}(a+bx+cy) = b^2\mathrm{var}(x) + c^2\mathrm{var}(y) + 2bc\ \mathrm{cov}(x,y)$$

and

$$\mathrm{cov}(ax, by+cz) = ab\ \mathrm{cov}(x,y) + ac\ \mathrm{cov}(x,z)$$

also,

$$\mathrm{cov}(x,x) = \mathrm{var}(x)$$

Therefore,

$$\begin{aligned}
r_{zx} &= \mathrm{cor}(z,x)\\
&= \frac{\mathrm{cov}(z,x)}{\mathrm{sd}(z)} \qquad (\mathrm{sd}(x)=1)\\
&= \frac{\mathrm{cov}(x+\beta y, x)}{\mathrm{sd}(x+\beta y)}\\
&= \frac{\mathrm{cov}(x,x)+\mathrm{cov}(\beta y,x)}{\sqrt{\mathrm{var}(x+\beta y)}}\\
&= \frac{\mathrm{var}(x)+\beta\mathrm{cov}(y,x)}{\sqrt{\mathrm{var}(x)+\beta^2\mathrm{var}(y)+2\beta\mathrm{cov}(x,y)}}\\
&= \frac{1+\beta\cdot 0}{\sqrt{1+\beta^2\cdot 1+2\beta\cdot 0}} = \frac{1}{\sqrt{\beta^2+1}}
\end{aligned}$$

Recall

$$\mathrm{sd}(x) = \sqrt{\mathrm{var}(x)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

and

$$\mathrm{cov}(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

also

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\text{sd}(x)\text{sd}(y)}$$

Note that

$$z_i - \bar{z} = x_i + \beta y_i - \bar{x} - \beta \bar{y} = (x_i - \bar{x}) + \beta(y_i - \bar{y})$$

Therefore,

$$\begin{aligned}
r_{zx} &= \text{cor}(z, x) \\
&= \frac{\text{cov}(z, x)}{\text{sd}(z)} \qquad (\text{sd}(x) = 1) \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2}} \\
&= \frac{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x}) + \beta(y_i - \bar{y})](x_i - \bar{x})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[(x_i - \bar{x}) + \beta(y_i - \bar{y})]^2}} \\
&= \frac{\frac{1}{n-1}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \beta\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right]}{\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\beta\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) + \beta^2\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}} \\
&= \frac{\text{var}(x) + \beta\text{cov}(x, y)}{\sqrt{\text{var}(x) + 2\beta\text{cov}(x, y) + \beta^2\text{var}(y)}} \\
&= \frac{1 + \beta \cdot 0}{\sqrt{1 + 2\beta \cdot 0 + \beta^2 \cdot 1}} = \frac{1}{\sqrt{\beta^2 + 1}}
\end{aligned}$$

(b) **Found a quicker solution so just read the red box instead of the whole question (update 25 November 2020).**

Suppose that the fitted linear regression model is denoted with

$$y = \beta_0 + \beta_1 x$$

Then,

$$\beta_1 = r_{xy}\frac{\text{sd}(y)}{\text{sd}(x)} = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

and the residuals are computed as follows

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

Therefore, combining these facts with the ones we have used in Question 5(a), we have

$$\begin{aligned}
\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} &= \frac{\text{var}(y) - \text{var}(e)}{\text{var}(y)} \\
&= \frac{\text{var}(y) - \text{var}(y - \beta_0 - \beta_1 x)}{\text{var}(y)} \\
&= \frac{\text{var}(y) - (\text{var}(y) + \beta_1^2\text{var}(x) - 2\beta_1\text{cov}(x, y))}{\text{var}(y)} \\
&= \frac{2\beta_1\text{cov}(x, y) - \beta_1^2\text{var}(x)}{\text{var}(y)} \\
&= \frac{2\frac{\text{cov}(x,y)}{\text{var}(x)}\text{cov}(x, y) - (\frac{\text{cov}(x,y)}{\text{var}(x)})^2\text{var}(x)}{\text{var}(y)} \\
&= \frac{\text{cov}(x, y)^2}{\text{var}(x)\text{var}(y)} = r_{xy}^2
\end{aligned}$$

PS : Claimed the answer to be r_{xy}^2 so that the whole calculation is simply backtracking. Suppose that the fitted linear regression model is denoted with

$$y = \beta_0 + \beta_1 x$$

where

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \bar{y}\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})}$$

and

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Then, the residuals are computed as follows

$$e_i = y_i - (\beta_0 + \beta_1 x_i)$$

and thus

$$\bar{e} = \bar{y} - (\beta_0 + \beta_1 \bar{x}) = 0 \to e_i - \bar{e} = y_i - (\beta_0 + \beta_1 x_i)$$

Thus,

$$\frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2 - \frac{1}{n}\sum_{i=1}^n (e_i - \bar{e})^2}{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i-\bar{y})^2 - \sum_{i=1}^n (y_i - (\beta_0+\beta_1 x_i))^2}{\sum_{i=1}^n (y_i-\bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - ((\bar{y} - \beta_1\bar{x}) + \beta_1 x_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - ((\bar{y} - \beta_1\bar{x}) + \beta_1 x_i))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n ((y_i - \bar{y}) - \beta_1(x_i - \bar{x}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{2\beta_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Now we want to prove that

$$\operatorname{cov}(x, y) = \beta_1 \operatorname{var}(x)$$
$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} \sum (x_i - \bar{x})^2$$
$$\sum (x_i - \bar{x})(y_i - \bar{y}) \sum x_i(x_i - \bar{x}) = \sum x_i(y_i - \bar{y}) \sum (x_i - \bar{x})^2$$
$$\left[\sum y_i(x_i - \bar{x}) - \bar{y}\sum (x_i - \bar{x})\right] \sum x_i(x_i - \bar{x}) = \sum x_i(y_i - \bar{y}) \left[\sum x_i(x_i - \bar{x}) - \bar{x}\sum (x_i - \bar{x})\right]$$
$$\sum x_i(x_i - \bar{x}) \sum (y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) \sum (x_i - \bar{x})$$

Which is true since

$$\sum (y_i - \bar{y}) = \sum (x_i - \bar{x}) = 0$$

Thus, we can continue with our counting.

$$\frac{2\beta_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
$$= \frac{2\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}\operatorname{cov}(x, y) - (\frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)})^2 \operatorname{var}(x)}{\operatorname{var}(y)}$$
$$= \frac{2(\operatorname{cov}(x, y))^2}{\operatorname{var}(x)\operatorname{var}(y)}$$
$$= r_{xy}^2.$$