Exercise 2

# Answer 1: Linear regression

Let there be n linear equations with n unknowns i.e. yi = f(xi) , where βi are coefficients of the independent variable xi and εi is the noise. It is assumed that noise is distributed as iid and follows a normal Gaussian distribution i.e. ε = N(0,σ2)

Y1 = β0 + β1 X1 + ε1

Y2 = β0 + β1 X2 + ε2

…………………………………………

Yn = β0 + β1 xn + εn

These linear equations can be written in matrix form as

= +

= . +

This can be written as **Y = X + ε ;** note bold characters refer to matrixes

Or **ε = Y –**

**ε 2 = εT ε**

We want to find elements of such that **ε2** is minimized.

**ε 2 = εT ε = (Y – X)T(Y – X) ….. (1)**

To minimize **ε 2 ,** we will take its partial derivative w.r.t. and set it equal to zero.

**ε2 = (Y – X)T(Y – X) ….. (2)**

Let **S** = **(Y – X)T(Y – X) ….. (3)**

**= (YT - TXT) (Y –X)**

**= YT Y - YTX - TXT Y + TXT X ……(4)**

**(YTX) = YT1xnXnx22x1** = 1x1 matrix

Also**, TXT Y = T1x2XT 2xnYnx1**= 1 x 1 matrix

For a 1x1 matrix, A = AT

Therefore, **(YTX) = (YTX)T = TXTY**

Therefore, (4) reduces to

**S** = **YT Y -**2 **TXT Y + TXT X …….(5)**

Now, according to theorem of matrix calculus where x is n × 1, A is n × n, and A does not depend on x, then

1. If **α = yTAx,** then **α = xTy**

**y**

1. **If α = xTAx,** then **α = xT(AT + A)**

**x**

**S = -2 YT X + TXT X)” ;** using “ to denote partial derivative

**TXT X)” = 2TXT X ;** using second point of above theorem

Putting **S = 0,**

We get,

**= (XTX)-1(XTY)**

# Answer 2

The 50 items in the data set for training is split in 75%:25% ratio so that 75% (38 entries) of entries are used for training and remaining 12 entries are used for validation.

Regression model of upto 5 orders are fitted on the 38 entries of the training set and mean of squared errors is computed.

The regression model for upto 5 orders is as follows:

|  |  |
| --- | --- |
|  | Model |
| 1st Order | 0.9726429 + 0.3200254 X |
| 2nd Order | 0.8012937 + 1.0156365 X - -0.4644960 X^2 |
| 3rd Order | 0.9632519 + -0.3195257 X + 1.8010990 X ^2 - -1.0146796 X^3 |
| 4th Order | 1.05416420 - -1.51462702 X + 5.45151544 X^2 - 4.90249923 X^3 + 1.32960573 X^4 |
| 5th Order | 0.98867379 - 0.31487545 X - 0.09292972 X^2 + 5.05491793 X^3 - 6.27506523 X^4 + 2.07199640 X^5 |

The mean of standard error per regression model is also computed as follows on the training and validation set

|  |  |  |
| --- | --- | --- |
|  | Mean Standard Error – Training Set | Mean Standard Error – Validation Set |
| 1st order | 0.1232344 | 0.08690818 |
| 2nd order | 0.0939629 | 0.05096113 |
| 3rd order | 0.0733392 | 0.07940147 |
| 4th order | 0.06825019 | 0.06375078 |
| 5th order | 0.06596547 | 0.06640959 |

The regression model of 2nd order is chosen as that model gives the minimum Mean of Standard Error and tested on the test set. The Mean Standard Error on the test set is computed as 0.1153557.

The R code for the program is attached.

# Answer 3

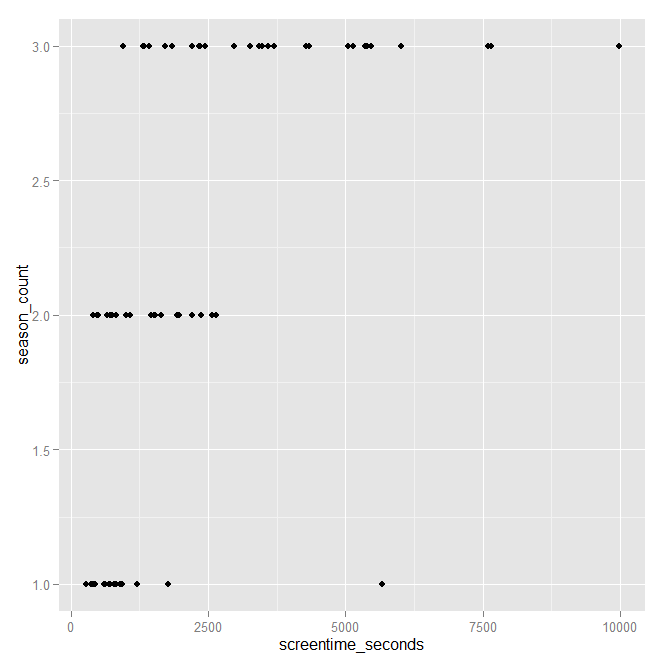
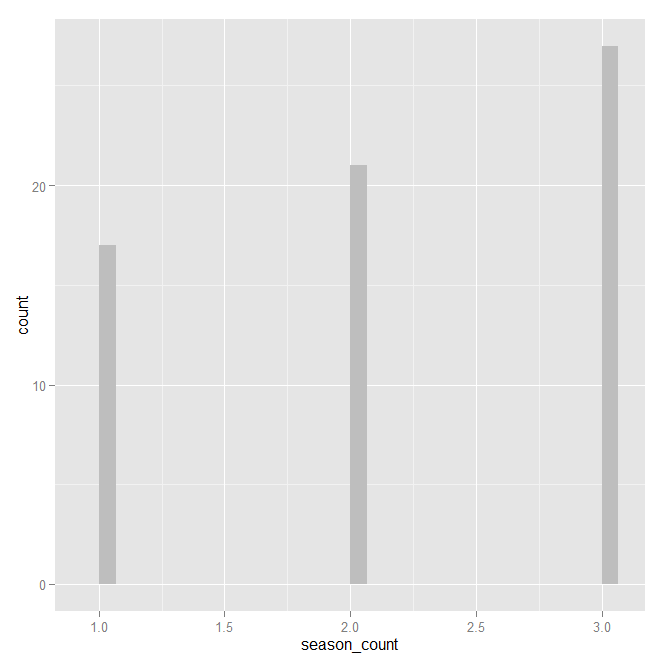
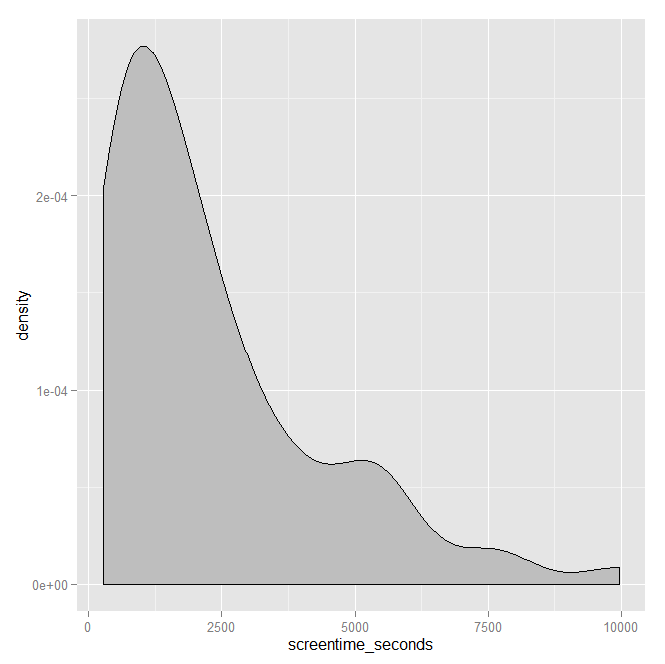


Figure 1 Figure 2 Figure 3

The summary statistics of screen\_time\_seconds shows

Min. 1st Qu. Median Mean 3rd Qu. Max.

277 788 1640 2321 3262 9975

Since the mean > median , therefore the distribution of screen\_time\_seconds is right skewed.

Figure 1 is the density distribution of screen\_time\_seconds

Figure 2 shows that most of the participants have done 3 shows.

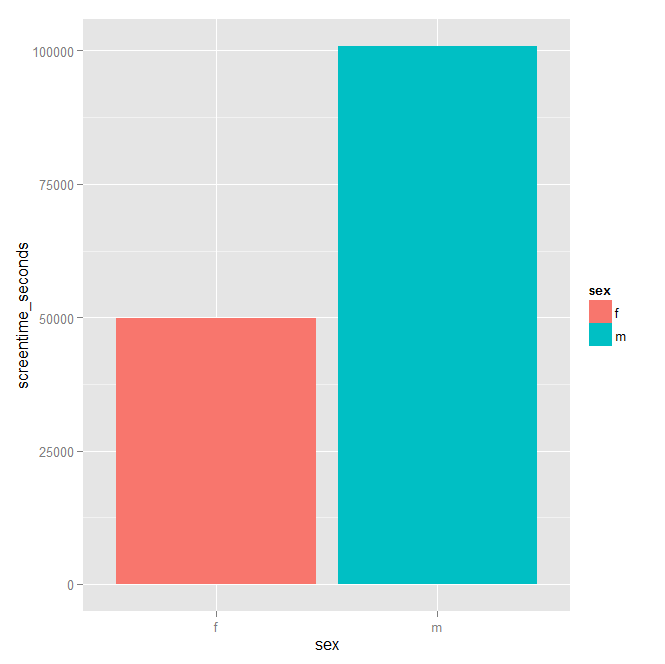
Figure 3 shows the relationship between screen\_time\_seconds and screen\_count. The following table shows that generally people who have higher screen\_count have higher screen\_time.

season\_count screentime\_seconds.mean screentime\_seconds.median

1 1 1004.647 699.000

2 2 1384.524 1462.000

3 3 3877.556 3476.000



This figure shows that males have had more screentime than females.