WRITTEN ASSIGNMENT 1 SOLUTIONS

Spring 2017

1.

a)

Table 1. Summary statistics for blood calcium levels (mmol/l) for all subjects and subjects separated by sex.

Subset	*n*	Min	Q1	Median	Q3	Max	Mean	SD
All	178	1.9	2.25	2.35	2.45	3.75	2.361	0.169
Females	87	2	2.3	2.4	2.495	3.75	2.405	0.199
Males	91	1.9	2.25	2.33	2.4	2.58	2.318	0.122

b)

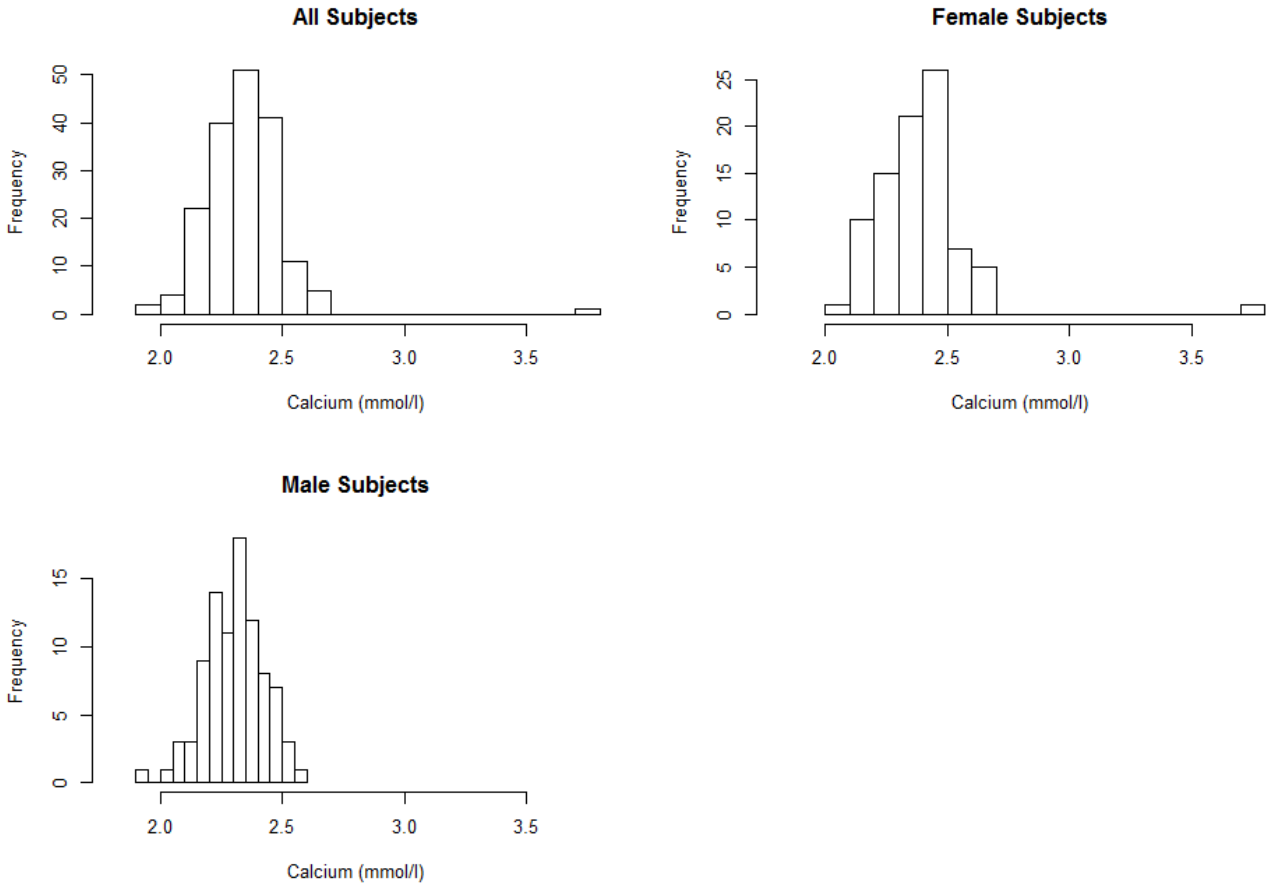


Figure 1. Histograms of calcium concentration (mmol/l) in all subjects, and separated by sex.

c)

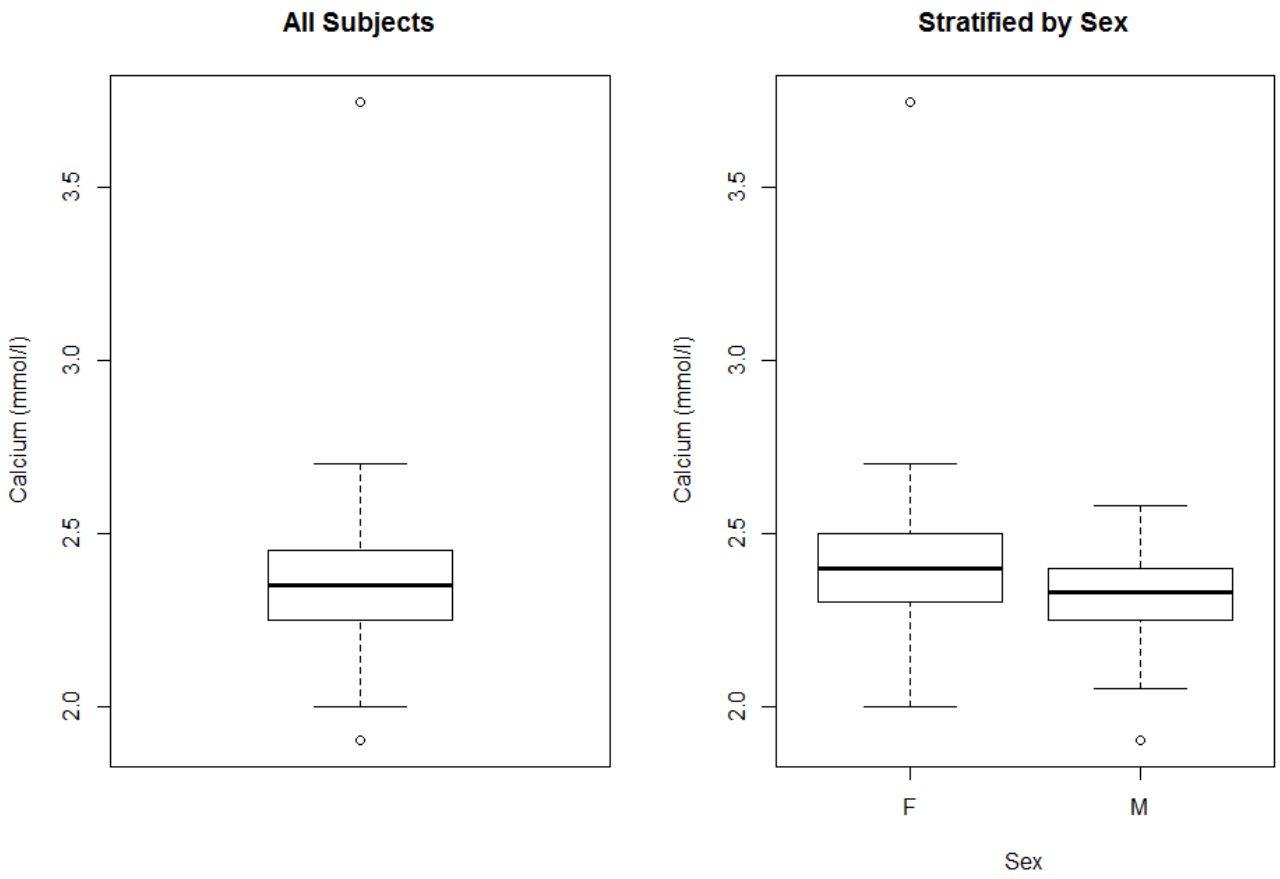


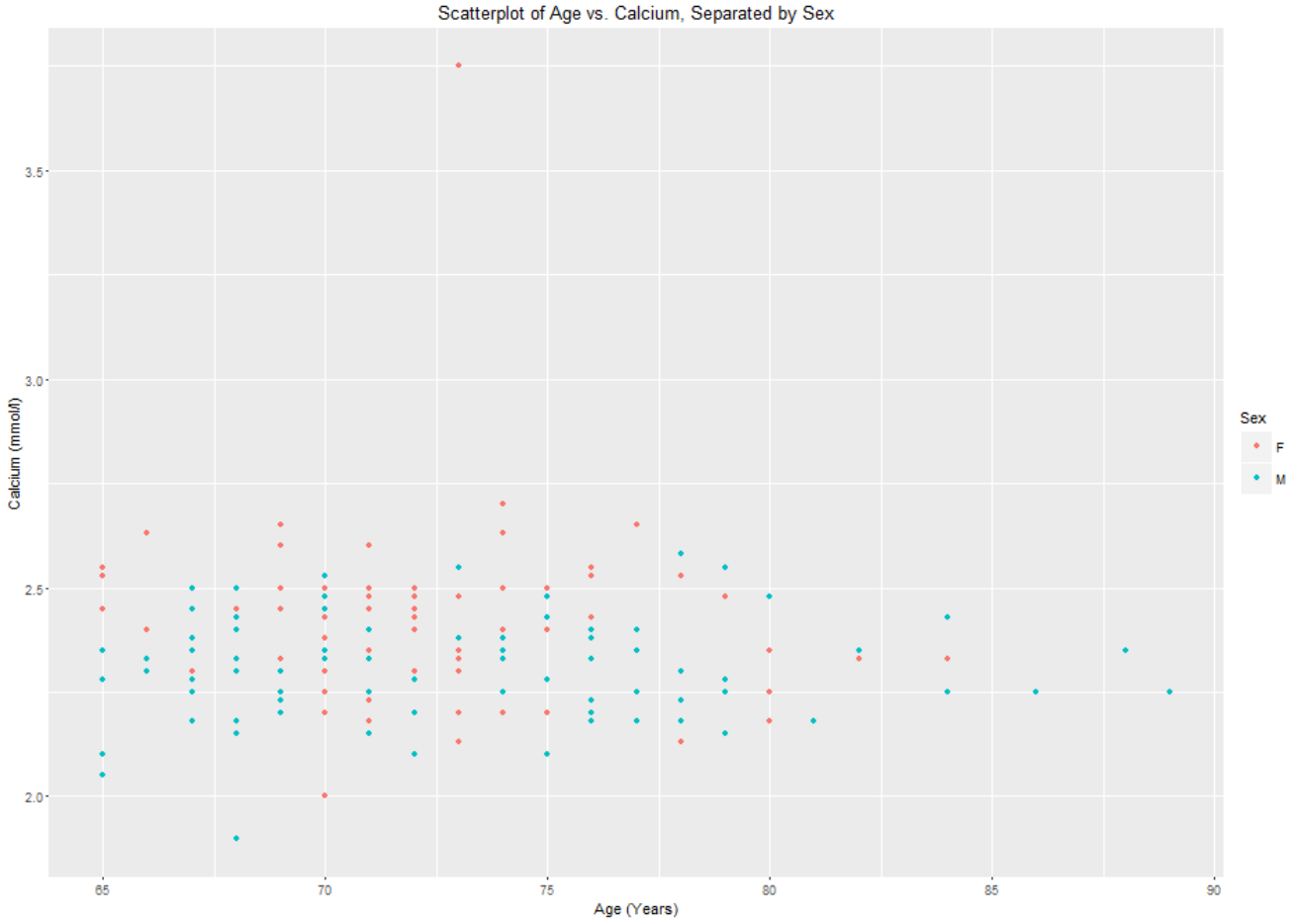
Figure 2. Boxplots of calcium levels (mmol/l) for all subjects (left) and stratified by sex (right).

d)
All three calcium level distributions (for all subjects and separately for female and male subjects) are unimodal. This can be seen from the histograms presented in Figure 1. These also seem fairly symmetric, although there is a slight amount of skew. The distributions for all subjects and female subjects are skewed to the right, and the distribution for male subjects is skewed to the left. However, Table 1 shows that the means and medians are very similar. This is especially the case for female subjects, where the mean is greater than the median, but the values are 2.405 mmol/l and 2.4 mmol/l, respectively. From Figure 2, we see that there are two outliers when considering all subjects (one high and one low). When considering calcium levels for each sex separately, there is one high outlier for female subjects and one low outlier for male subjects. Figure 2 also helps us to visualize the shape and location of the distributions for female and male subjects. The median blood calcium level is greater for female subjects; in fact, the first quartile is almost as high as the median value for male subjects. The bulk of the distribution (the middle 50%) is shifted higher for female subjects. The distribution for female subjects also shows more spread. This is in agreement with Table 1, which shows that the SD and the IQR are higher for female subjects.

e)
Any calcium level that is 1.5 IQRs beyond the first and third quartiles is considered an outlier. The IQR for female subjects is 2.495 – 2.3 = 0.195 mmol/l. Anything less than 2.3 – 1.5 * 0.195 = 2.0075 mmol/l or greater than 2.495 + 1.5 * 0.195 = 2.7875 mmol/l is considered an outlier. Therefore, a value

of 1.9 mmol/l would be considered an outlier.

2.
a)



Calcium concentrations seem pretty constant with age. If there is a linear association, it is very slightly negative for female subjects and very slightly positive for male subjects. Consistent with Figures 1 and 2, calcium concentrations seem to be higher for female subjects.

b)
The correlation coefficient between calcium and age is -0.0295. When separated by sex, the correlation is -0.0660 for female subjects and 0.0290 for male subjects. The fact that the coefficient is negative for female subjects means that older individuals are associated with a lower blood calcium level. The opposite trend is true for male subjects. Since the number of male and female subjects is similar and the trend is stronger for female subjects, the overall correlation between age and calcium is negative. However, the magnitudes of these correlation coefficients are very small, so we can say there is basically no linear relationship.

c)

```
Call:
lm(formula = Calcium[indexM] ~ Age[indexM])

Residuals:
     Min       1Q   Median       3Q      Max
-0.41527 -0.07612  0.01281  0.08154  0.25835

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.2718930  0.1697104  13.387   <2e-16 ***
Age[indexM] 0.0006379  0.0023347   0.273    0.785
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1224 on 89 degrees of freedom
Multiple R-squared:  0.0008382, Adjusted R-squared:  -0.01039
F-statistic: 0.07466 on 1 and 89 DF,  p-value: 0.7853
```

The regression equation is Calcium_hat = 2.272 + 0.000638 * Age

d)

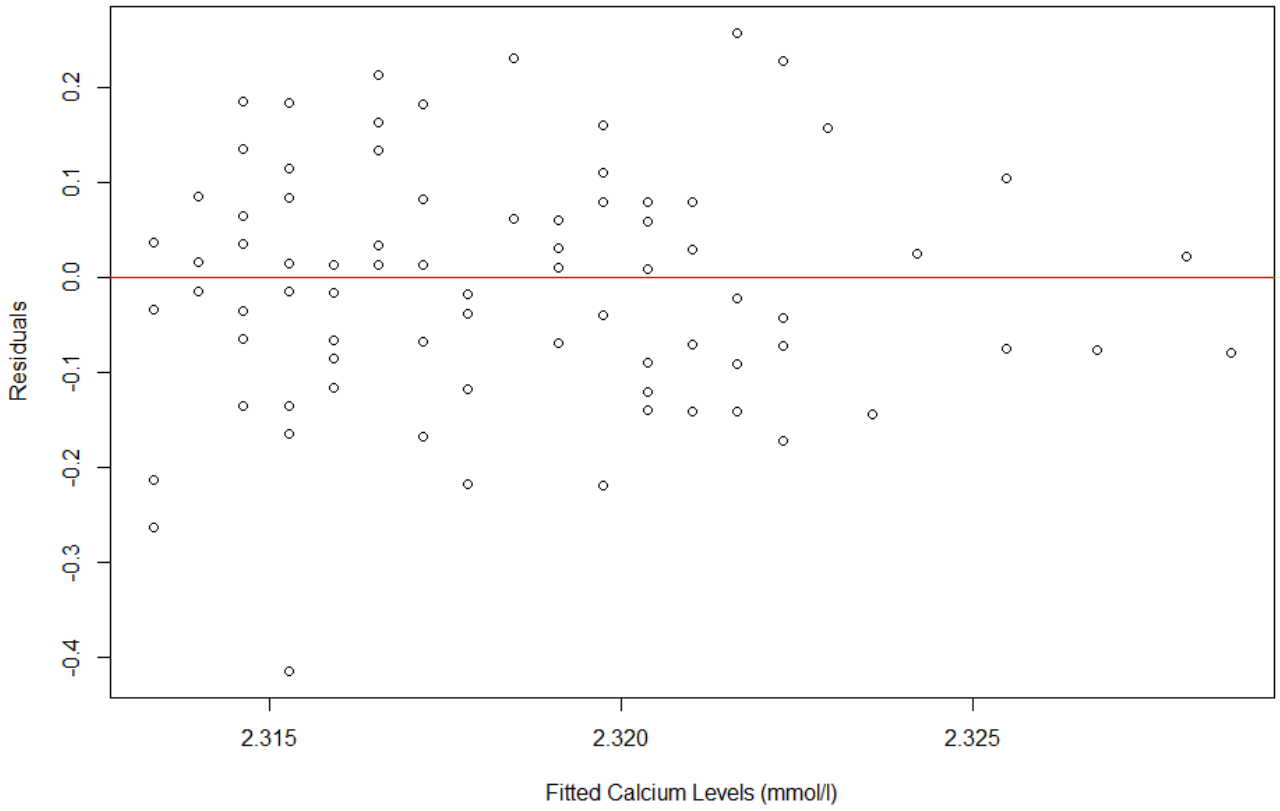


Figure 3. Residuals plot for regression of calcium on age for male subjects.

The residuals plot from the regression looks sufficiently random. The residuals seem to be scattered around y = 0 fairly randomly and symmetrically. The variance seems to be lower for the largest fitted values, but the number of points is small so this is not necessarily cause for concern.

e)
The coefficient of determination is 0.000838. This means that 0.0838% of the variation in calcium levels for male subjects can be explained through a relationship with age.

f)
The mean age is 72.484. The standard deviation is 5.524. The slope can be calculated using the formula b1 = r*sy/sx. Since r = 0.0290 from part b and sy = 0.122 from 1a), we can see that b1 = 0.000638.

We can calculate the intercept by using the formula b0 = y_bar – x_bar * b1. From 1a), y_bar is 2.318, so b0 = 2.272. This gives Calcium_hat = 2.272 + 0.000638 * Age, which is the same regression equation that we got by using the lm() function in part c.

g)
Calcium_hat = 2.272 + 0.000638 * 83 = 2.325 mmol/l

h)
No, we do not recommend that this regression equation be used to predict calcium levels for male subjects between 65-89 years old. The slope is very small, and the spread in the data overwhelms the association of calcium and age. This can be seen visually in the plot in part a. This is supported by the value of the coefficient of determination, which as we showed in part e is very small.

3.
We first create the following contingency table, which will help in solving parts a through e.

Table 2. Contingency Table for Sex vs. Age Group

Sex	**Age Group**					
	65-69	70-74	75-79	80-84	85-89	
F	21	46	15	5	0	87
M	35	24	23	6	3	91
	56	70	38	11	3	178

a)
The marginal distribution for sex is:
F: 87/178 = 48.9%
M: 91/178 = 51.1%

b)
The marginal distribution for age group is:
65-69: 56/178 = 31.5%
70-74: 70/178 = 39.3%
75-79: 38/178 = 21.3%
80-84: 11/178 = 6.2%
85-89: 3/178 = 1.7%

c) There are 35 subjects who are male and in the 65-69 year old age group, so the joint percentage is 35/178 = 19.7%

d)
The conditional distribution for age group given that a subject is female is:
65-69: 21/87 = 24.1%
70-74: 46/87 = 52.9%
75-79: 15/87 = 17.2%
80-84: 5/87 = 5.7%
85-89: 0/87 = 0.0%

e)
The conditional distribution for sex given that a subject is 70-74 years old is:
F: 46/70 = 65.7%
M: 24/70 = 34.3%