STA 210: Lab 4

Jackson Hubbard

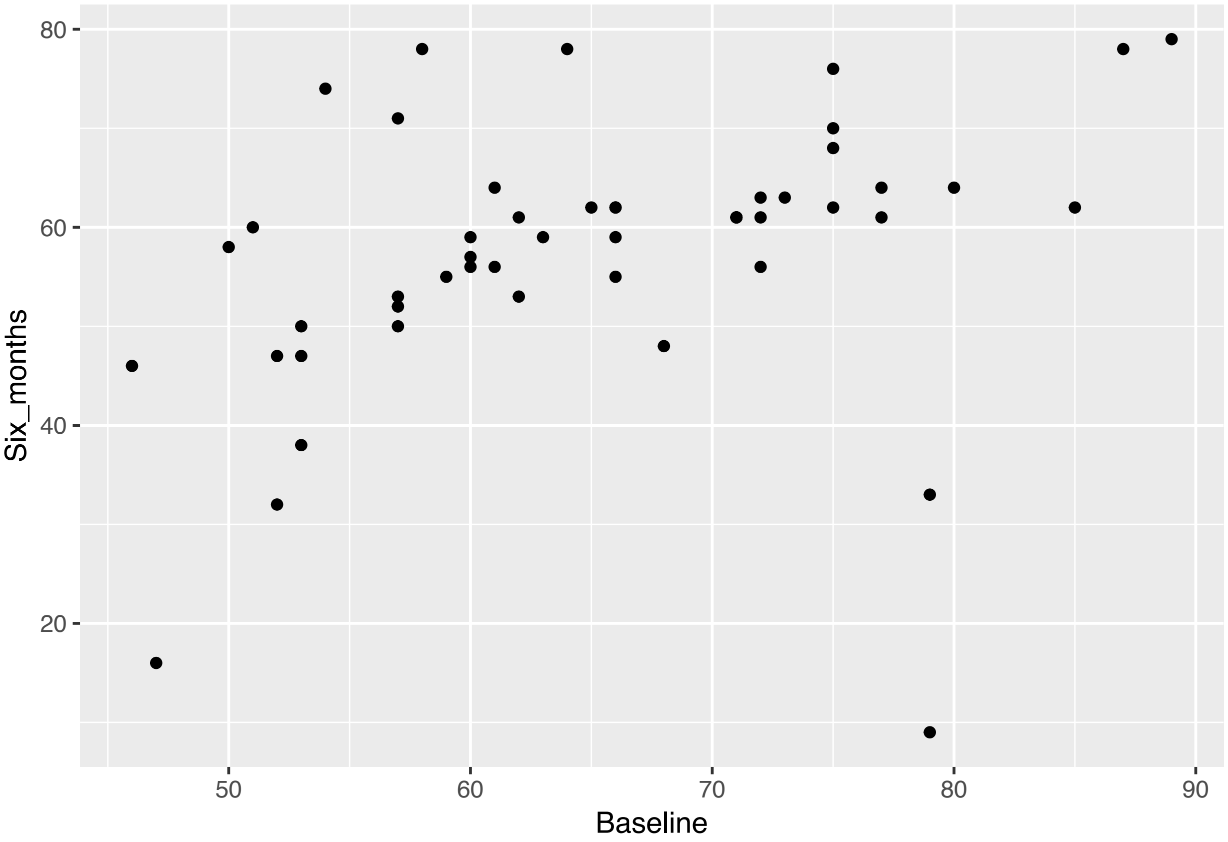
September 24, 2018

```
setwd("~/")
prostate <- read.csv("lab04_prostate.csv")
```

```
setwd("~/")
textmessages <- read.csv("lab04_textmessages.csv")
```

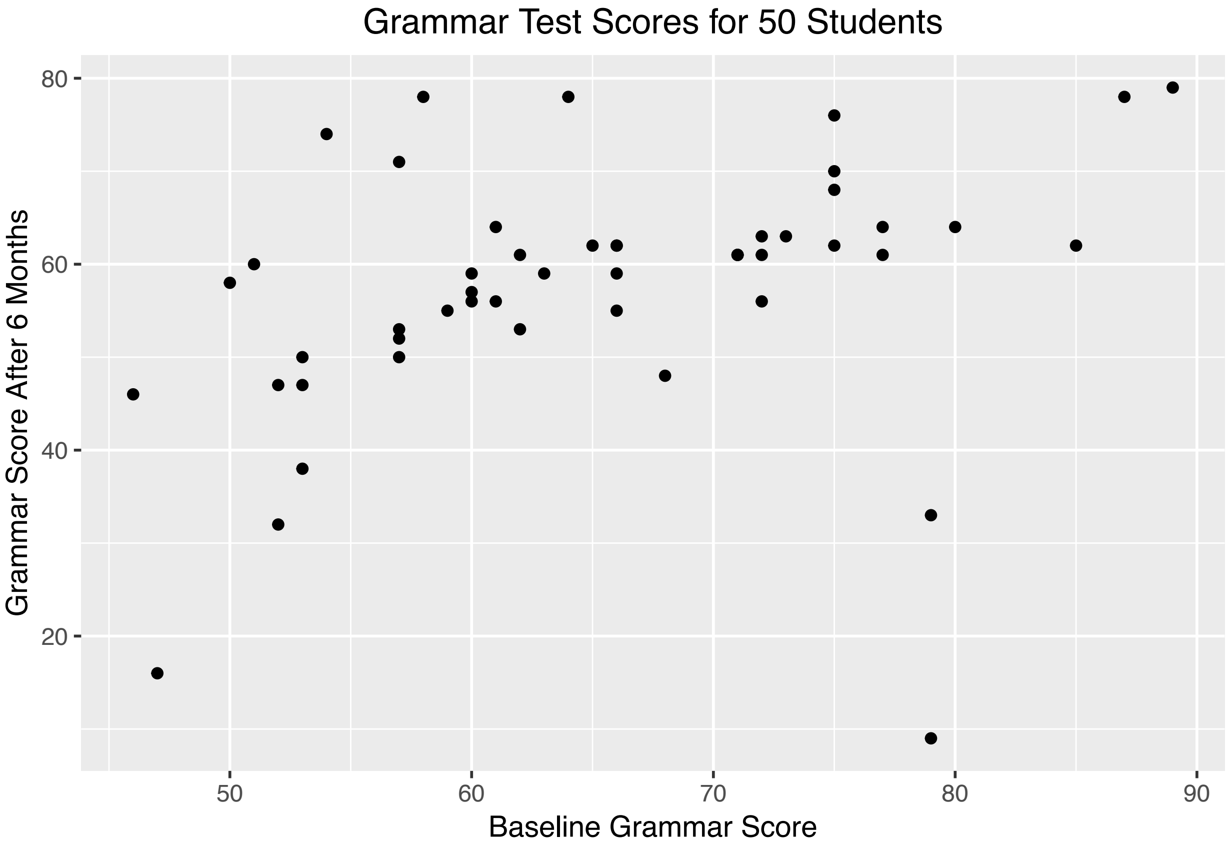
Question 1

```
ggplot(data = textmessages, aes(x = Baseline, y = Six_months)) + geom_point()
```



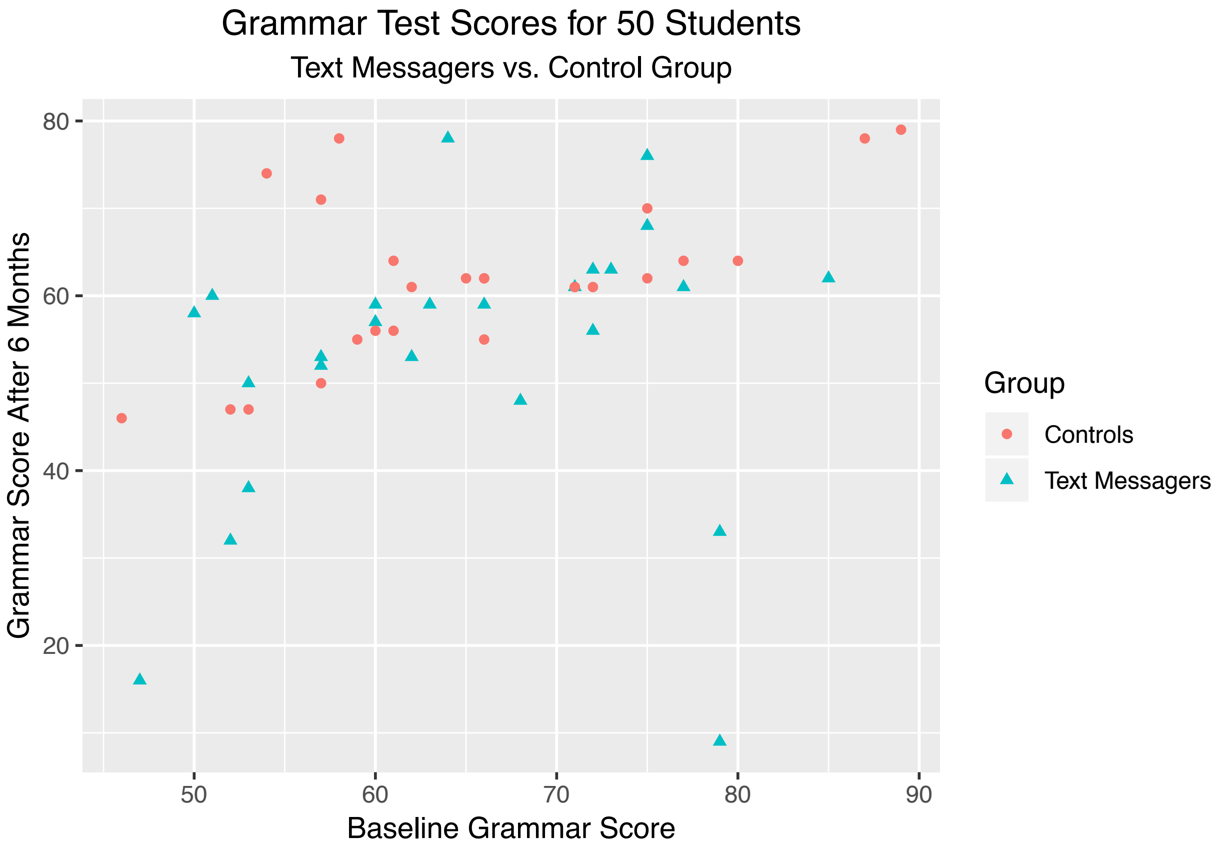
Question 2

```
ggplot(data= textmessages, aes(x = Baseline, y = Six_months)) + geom_point() + labs( x=
```



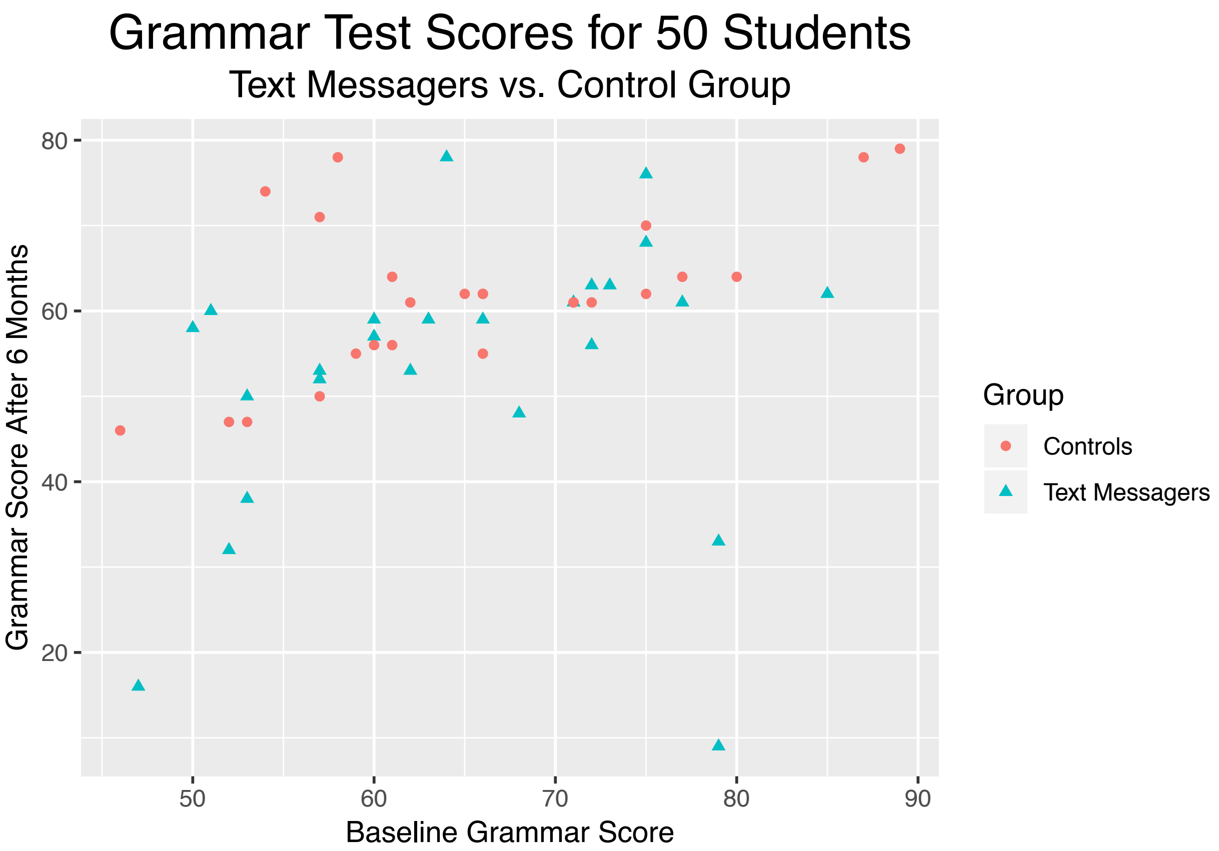
Question 3

```
ggplot(data= textmessages, aes(x = Baseline, y = Six_months, color = Group, shape = Grou
```



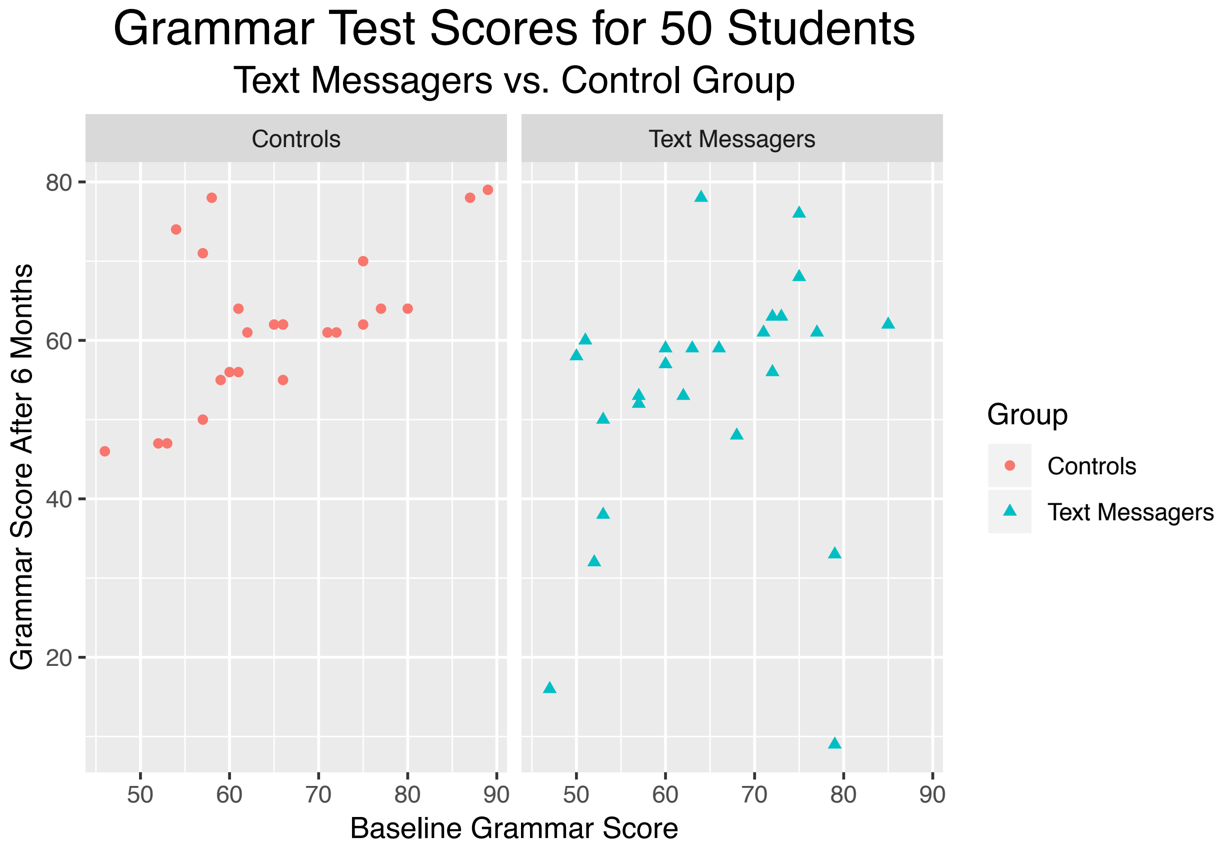
Question 4

```
ggplot(data= textmessages, aes(x = Baseline, y = Six_months, color = Group, shape = Grou
```



Question 5

```
ggplot(data= textmessages, aes(x = Baseline, y = Six_months, color = Group, shape = Grou
```



Question 6

```
sum_stat <- textmessages %>% group_by(textmessages$Group) %>% summarise(n=n(), mean = me
kable(sum_stat,format="latex",digits=2)
```

textmessages$Group	n	mean	sd
Controls	25	61.84	9.41
Text Messagers	25	52.96	16.33

###Question 7

The variables from graph D are most appropriate for linar regression. On this graph, we have log(cavol) as the explanatory variable and log(psa) as the response variable. Graph A is decent, however there are a couple of outliers. Graphs B and C have a curvature to them so the linearity condition is not met.

Question 8

```
prostate <- prostate %>% mutate(logcavol= log(cavol), logpsa = log(psa))
model1 <- lm(logpsa ~ logcavol, data = prostate)
tidy(model1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.52     0.148      10.3  3.12e-15
## 2 logcavol       0.713    0.0820      8.69 1.73e-12
```

```
kable(tidy(model1), format= "latex", digits= 2)
```

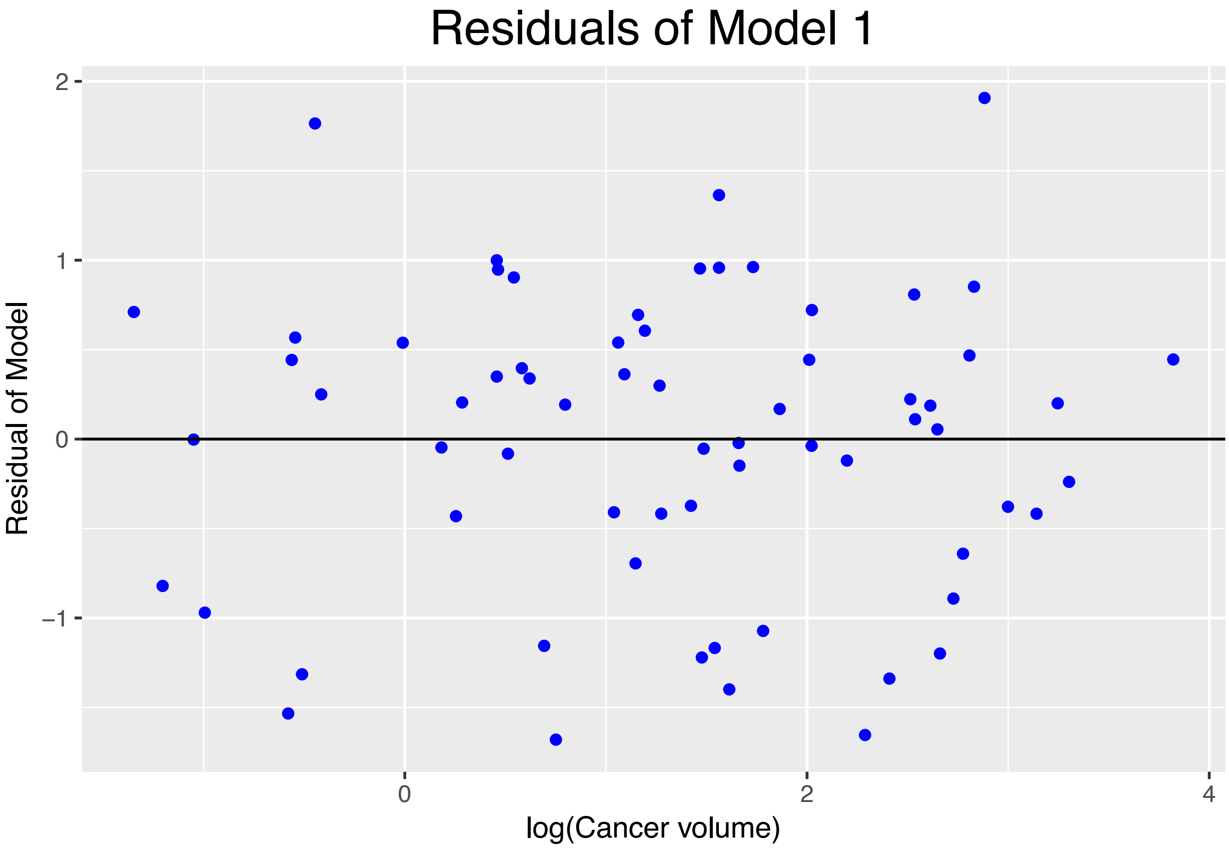
term	estimate	std.error	statistic	p.value
(Intercept)	1.52	0.15	10.26	0
logcavol	0.71	0.08	8.69	0

Question 9

#a

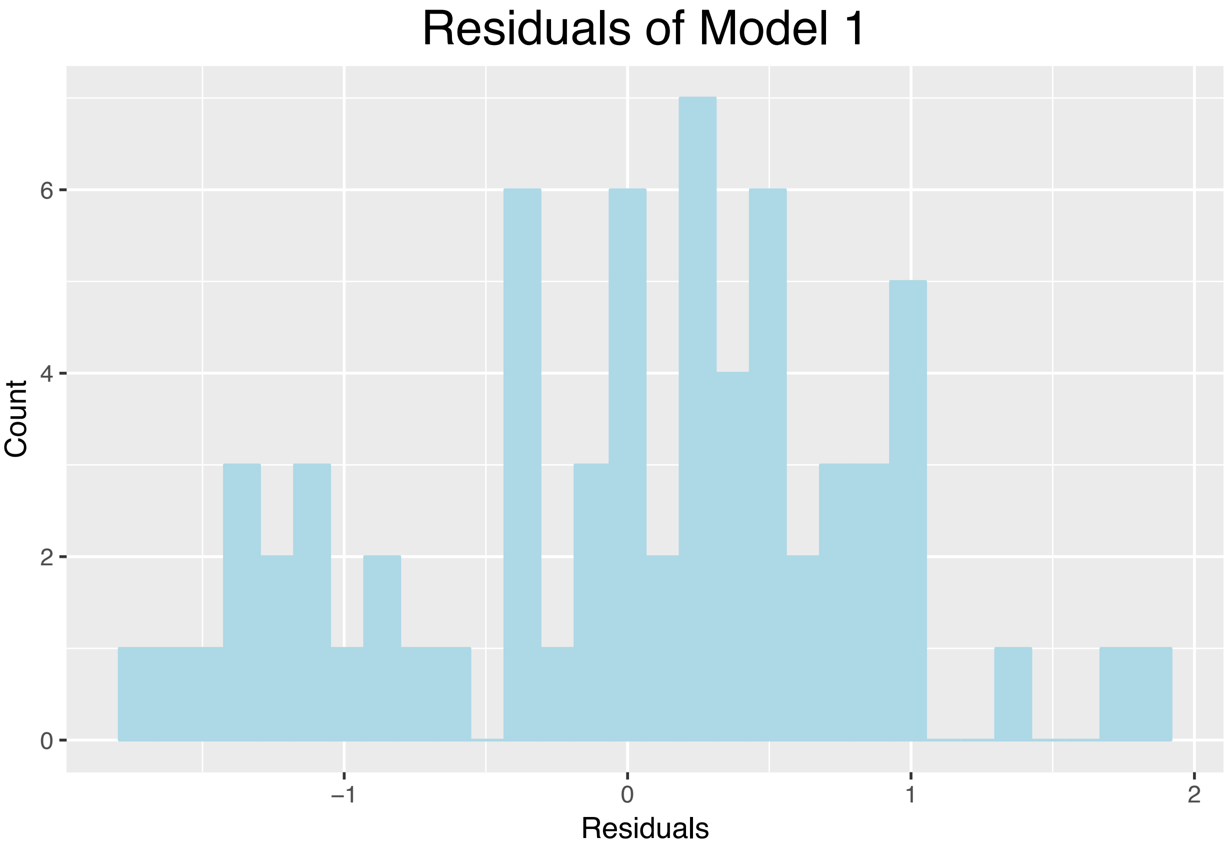
```
resid1 <- resid(model1)
prostate <- prostate %>% mutate(residual = resid1)


ggplot(prostate, aes(x = logcavol, y = residual)) + geom_point(color = "blue") +
labs(title= "Residuals of Model 1", x = "log(Cancer volume)", y= "Residual of Model", le
```



#b

```
ggplot(prostate, aes(x = residual)) + geom_histogram(stat= "bin", color= "light blue", 
labs(title= "Residuals of Model 1", x = "Residuals", y= "Count") + theme(plot.title = el
```



#c

Yes, the conditions for linear regressionn are met. Looking at the plot of the log(cavol) vs log(psa), we see that there is a linear relationship between the two variables. We also know the the condition of independence is met because each man's cancer characteristics is independent of another man's. Next, looking at the plot of log(cavol) vs. the residuals we see that there is constant variance as it does not follow any pattern.Further, the histogram of the residuals show that the normality assumption is also met.

Question 10

#a Use your regression model to predict the mean prostate-specific antigen (psa) for men with cancer volume (cavol) of 10. Calculate a 95% confidence interval for your prediction.

```
cavol <- 10
newdata=data.frame(cavol=cavol,logcavol=log(cavol))
predict.lm(model1,newdata, level = 0.95, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 3.157208 2.898339 3.416076
```

Looking at the output we see that the median (median used since we used a log transformation) prostate specific antigen for a cancer volume of 10 is e^3.157208 (23.5046). We are 95%

confident that the true median prostate specific antigen for a cancer volume of 10 is between (18.1439, 30.4496), which is e^2.898 and e^3.416.

#b

```
predict.lm(model1,newdata, level = 0.95, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.157208 1.484068 4.830348
```

Using the model to predict the PSA for an indivual man with a cancer volume of 10 results in the same predicted value of e^3.157208, which is 23.5046. However, perfroming a 95% confidence interval on this prediction results in a wider interval (4.41085258189, 125.254541652) compared to (18.1439, 30.4496) before. This is because the standard error is larger now.