STAT408_HW3

2022-10-13

Question 1

(10 points) In this question, we will use the prostate dataset. Import this dataset and answer following questions.

```
prostate <- read.csv("prostate.csv")
```

part (a)

Compute a 95% CI for the parameter associated with age. Use the manual method.

```
fit5 <- lm(lpsa ~ ., data = prostate)
summary(fit5)$coef[4,1] + c(-1,1) * qnorm(0.975) * summary(fit5)$coef[4,2]
```

```
## [1] -0.041535314  0.002260963
```

part (b)

Compute a 90% CI for the parameter associated with age. Use the manual method.

```
summary(fit5)$coef[4,1] + c(-1,1) * qnorm(0.95) * summary(fit5)$coef[4,2]
```

```
## [1] -0.038014673 -0.001259678
```

part (c)

Based on these two CIs, what can we expect the p-value of this parameter in t-test? Compare your conclusion with the p-value output by summary function.

Based on these two confidence intervals, we can expect that the p-value for the age parameter given by the t test will likely not be significant at a level of alpha = 0.05 because the first confidence interval includes zero. However, the second confidence interval does not include zero and includes only negative values, suggesting that the parameter associated with age may be slightly significant.

```
summary(fit5)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

These expectations were correct as the parameter for age is not significant at the level alpha = 0.05. However, it does appear to be significant at the level alpha = 0.1, because the p-value of 0.08229 is less than 0.1.

part (d)

Conduct a permutation t-test for predictor age in this model.

```
# orginal t test
lm.model <- lm(lpsa~., data = prostate)
summary(lm.model)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
```

```
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

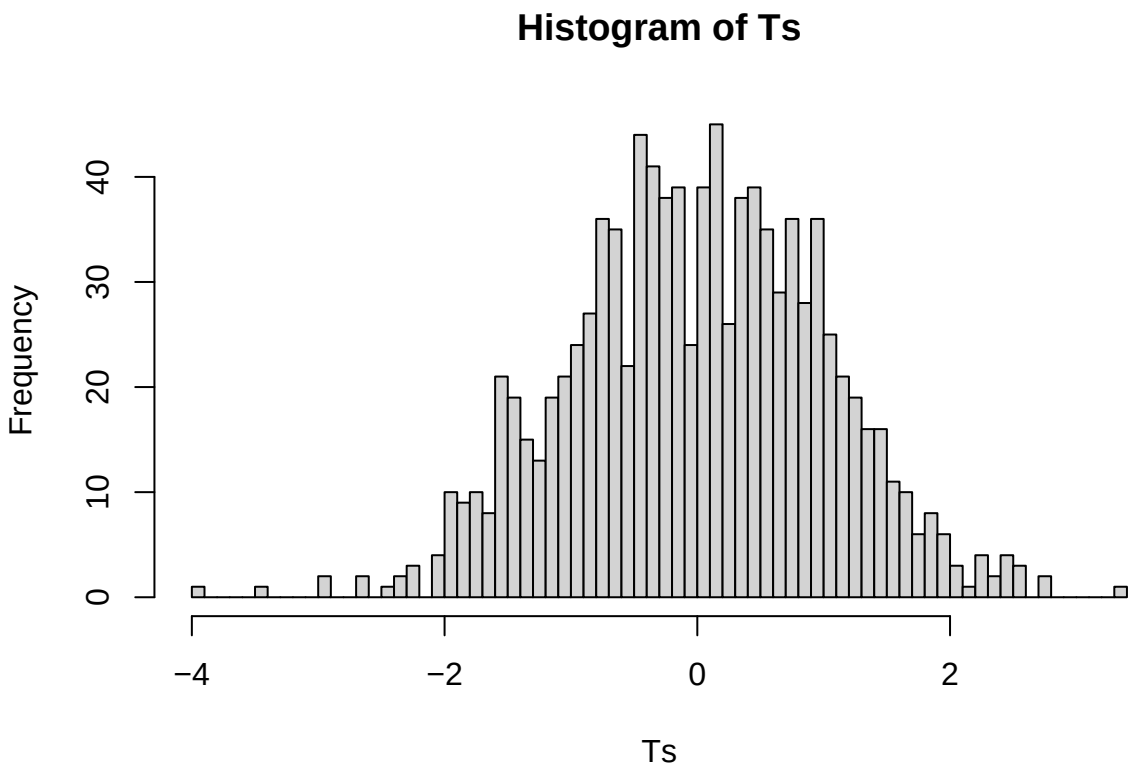
```
summary(lm.model)$coef[4,]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.01963718  0.01117272 -1.75759949  0.08229321
```

```
T.original <- summary(lm.model)$coef[4,3]

# set random seed to keep result same
set.seed(123)

# empty vector to save each permutation T statistic
Ts <- c()

for(i in 1:1000){
  # linear regression on shuffled pregnant
  lm.model <- lm(lpsa~lcavol+lweight+sample(age)+lbph+svi+lcp+gleason+pgg45, data = prostate)
  # save permutation T statistic
  Ts[i] <-  summary(lm.model)$coef[4,3]
}

# Calculate the proportion of less than the original T statistics
mean(abs(Ts) > abs(T.original))
```

```
## [1] 0.075
```

```
hist(Ts, breaks = 100)
```



The p-value for the permutation t-test for the predictor age in this model is 0.075, which means that age is not a significant predictor of lpsa at the level alpha = 0.05

part (e)

Remove all the predictors not significant at the 5% level. Use anova function to conduct an F test to test this model against the original full model. Which model is preferred? Give your reason.

```
fit6 <- lm(lpsa ~ lcavol + lweight + svi, data = prostate)
anova(fit6, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 47.785
## 2     88 44.163  5    3.6218 1.4434 0.2167
```

Based on the results of this F test, model 1 is preferred because the p-value of 0.2167 is greater than a significance level of alpha = 0.05, suggesting that the additional parameters in the full model are not significant or necessary.

Question 2

(10 points) In this question, we will use the cheddar dataset. Import this dataset and answer following questions.

```
cheddar <- read.csv('cheddar.csv')
```

part (a)

Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level.

```
fit <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(fit)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

H2S and Lactic are both significant predictors of taste at the level alpha = 0.05

part (b)

Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model. Hint: exponential function is exp().

```
fit1 <- lm(taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
summary(fit1)
```

```
##
## Call:
## lm(formula = taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.209  -7.266  -1.651   7.385  26.335
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.897e+01  1.127e+01  -1.684   0.1042
## exp(Acetic)  1.891e-02  1.562e-02   1.210   0.2371
## exp(H2S)     7.668e-04  4.188e-04   1.831   0.0786 .
## Lactic       2.501e+01  9.062e+00   2.760   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 26 degrees of freedom
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.5264
## F-statistic: 11.75 on 3 and 26 DF,  p-value: 4.746e-05
```

Only Lactic acid is a significant predictor of taste at the level alpha = 0.05 in this model.

part (c)

Can we use an F-test to compare these two models? Which model provides a better fit to the data? Explain your reasoning for these two questions.

No, we cannot use an F-test to compare these two models because they are not a smaller model and a full model. In other words, there is not a model that contains all of the predictors in the other model plus some additional predictors. Therefore, we cannot use an F-test to compare two models where one model is not "nested" within the other model.

However, we can conclude that the first model appears to be a better fit for the data because the adjusted R^2 value of 0.6116 is greater than 0.5264. Additionally, it has two predictors, H2S and Lactic, that are significant at the level alpha = 0.05.

part (d)

If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected?

```
3.9118*0.01
```

```
## [1] 0.039118
```

If H2S is increased by 0.01 for the model used in part a, taste is expected to increase by 0.039118 units.

Question 3

(10 points) In this question, we will use the teengamb dataset. Import this dataset and answer following questions.

```
teengamb <- read.csv("teengamb.csv")
```

part (a)

Fit a model with gamble as the response and the other variables as predictors. Which variables are statistically significant at the 5% level?

```
fit2 <- lm(gamble ~ ., data = teengamb)
summary(fit2)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

Sex and income are significant predictors of gambling at the level alpha = 0.05.

part (b)

Check the meaning of each variable. Does the variable significance in (a) make sense? Give your reasoning.

Yes, it makes sense that men may be more likely to spend more on gambling than women and that those with a higher income would spend more on gambling because they have more money to spend.

part (c)

Fit a model with just income as a predictor and use an F-test to compare it to the full model.

```
fit3 <- lm(gamble ~ income, data = teengamb)
anova(fit3, fit2)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the results of this F test, the full model appears to be preferred because the p-value of 0.01177 is less than a significance level of alpha = 0.05, suggesting that the additional parameters in the full model are significant and thus that the full model is a better fit for the data.

Question 4

(10 points) In this question, we will use the sat dataset. It was collected to study the relationship between expenditures on public education and test results. It contains the following variables

Expend: Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars) Ratio: Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994 Salary: Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars) Takers: Percentage of all eligible students taking the SAT, 1994-95 Verbal: Average verbal SAT score, 1994-95 Math: Average math SAT score, 1994-95 Total: Average total score on the SAT, 1994-95

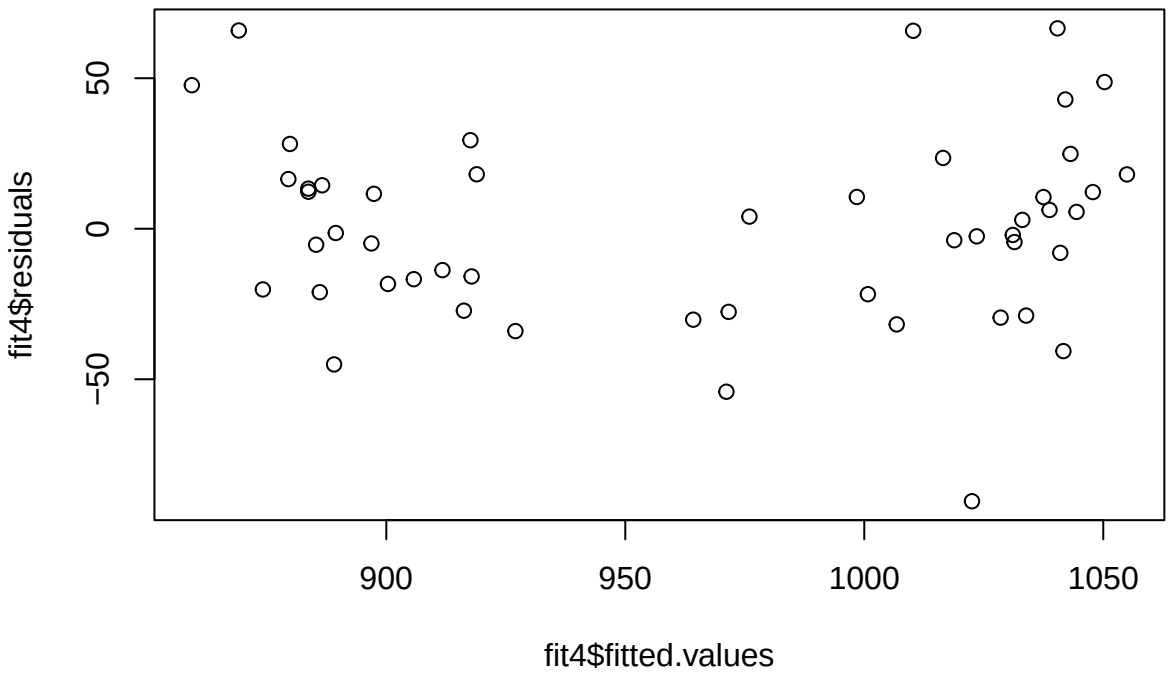
Using the sat dataset, fit a linear model with the total SAT score as the response and expend, salary, ratio, and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Some questions may be subjective. Show the most valid judgment and give your seasons.

```
sat <- read.csv("sat.csv")
fit4 <- lm(total ~ expend + salary + ratio + takers, data = sat)
```

part (a)

Plot residual vs. fitted response to check the constant variance assumption for the errors.

```
plot(fit4$fitted.values, fit4$residuals)
```

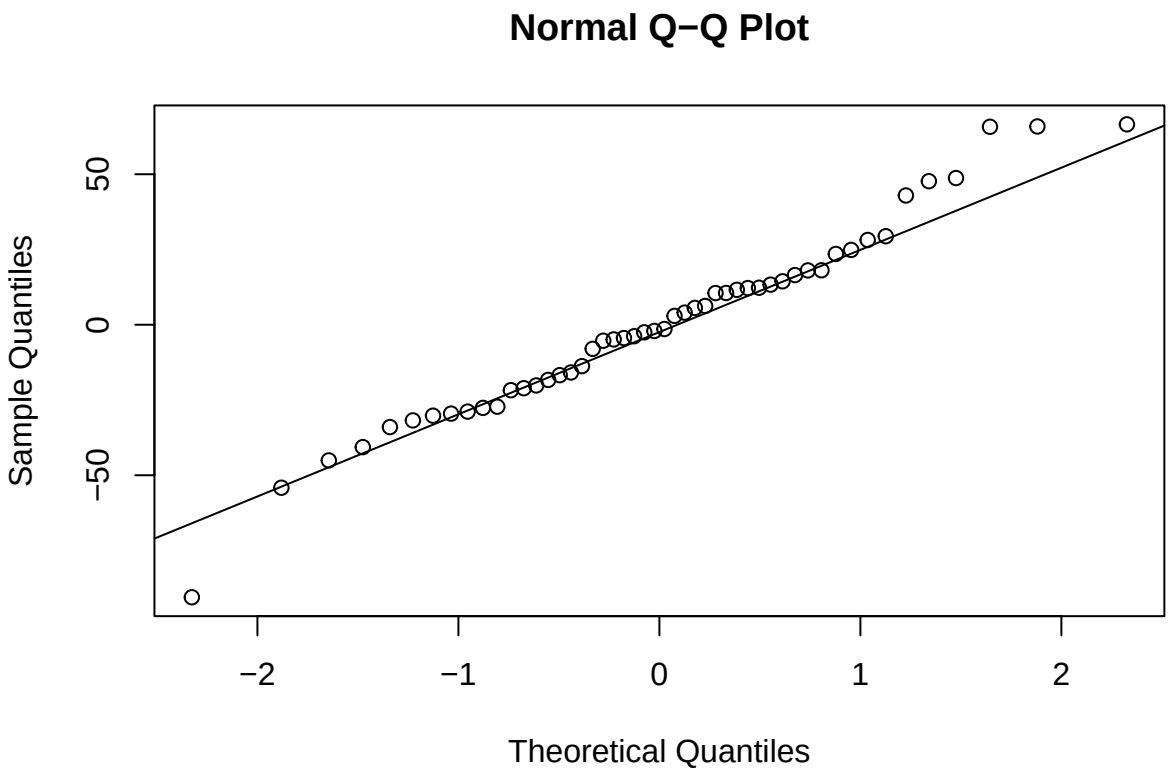


For the most part, the assumption of constant variance appears to be met based on this plot of the residuals versus the fitted values. There does appear to be an extremely slight megaphone shape in the graph but this seems to be due to a couple outliters (such as the point at the bottom of the graph between 1000 and 1050 on the x-axis) rather than the overall trend.

part (b)

Use Q-Q plot to check the normality assumption. What is the shape of the error distribution?

```
qqnorm(fit4$residuals)
qqline(fit4$residuals)
```



Based on the qq plot, the normality assumption does not appear to be met, as the qq plot appears to have short tails on either end that are not following the line.

part (c)

Use studentized residuals to check the outliers. Set the cutoff of being "large" as the 5% critical value in t distribution. Note that we need to consider the two sides.

```
# studentized residues
sr <- rstudent(fit4)
n <- dim(sat)[1]
df <- n - 4 - 1
# use 5% critical value as cutoff
which(abs(sr) > qt(0.975, df))
```

```
## 29 34 44 48
## 29 34 44 48
```

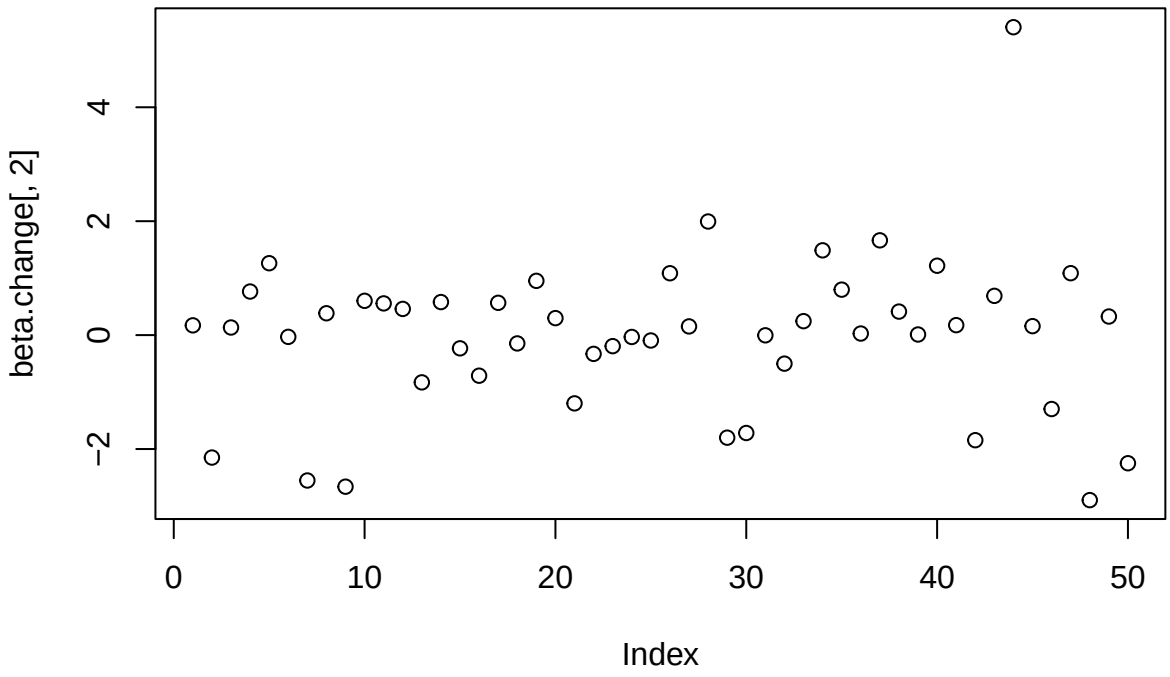
```
sum(abs(sr) > qt(0.975, df))
```

```
## [1] 4
```

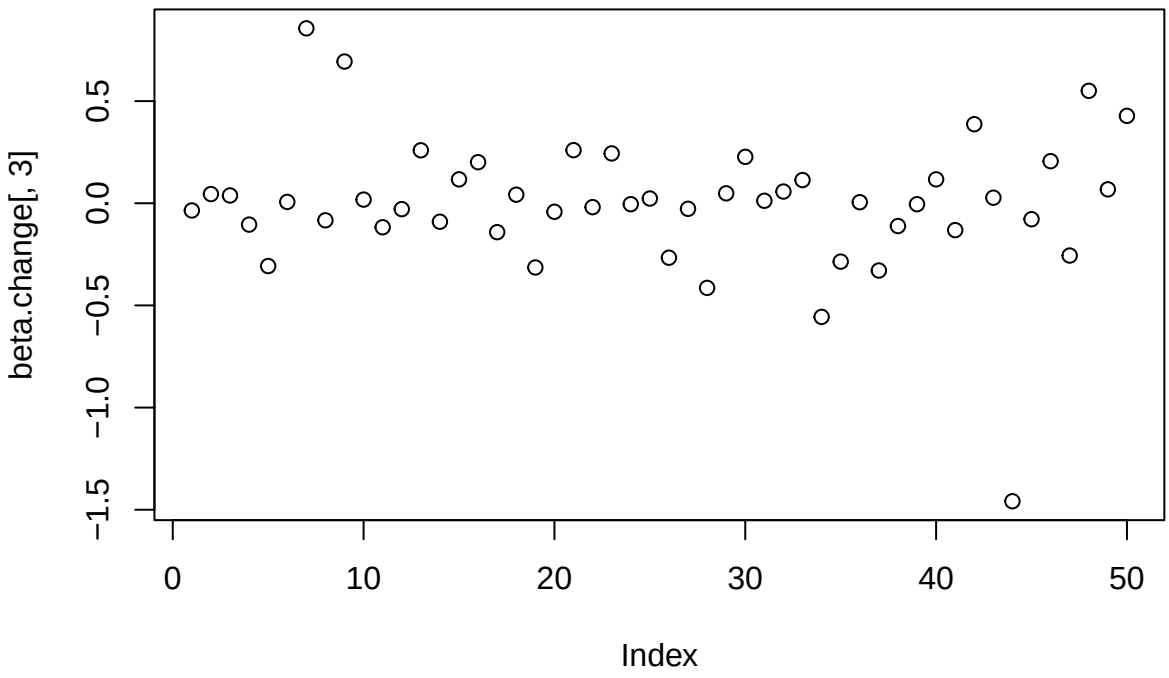
part (d)

Using defbeta function to plot the change of parameter estimation and check influential points. Check influential points in terms of each parameter except the intercept.

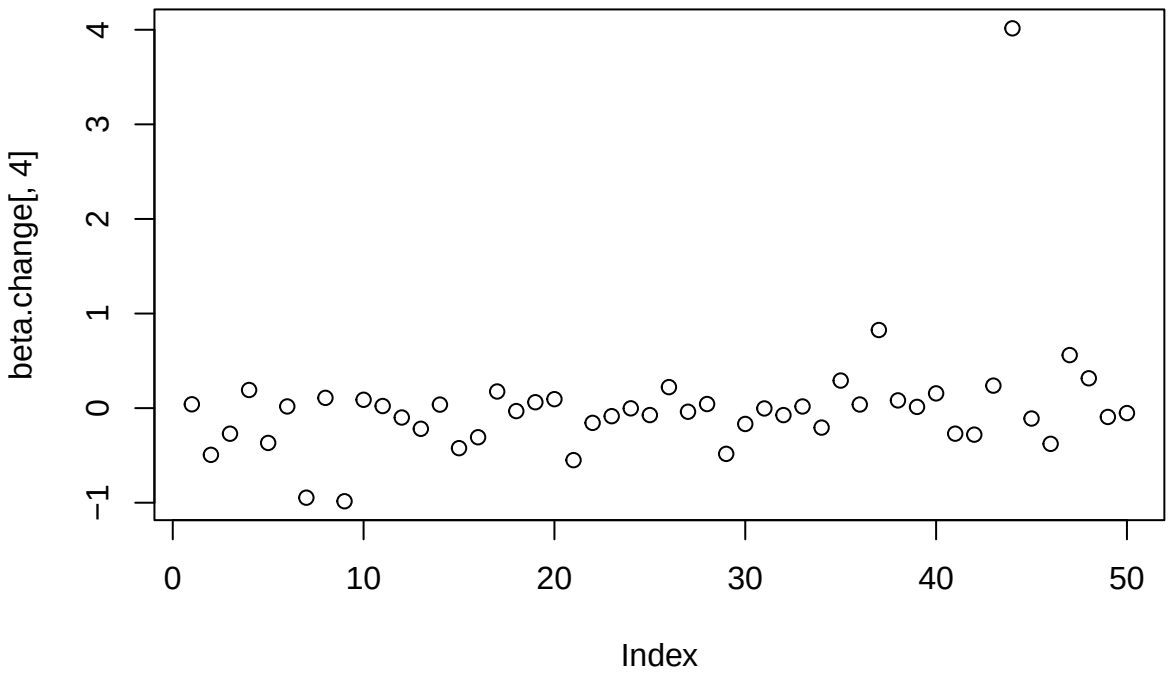
```
# influential observation
beta.change <- dfbeta(fit4)
# plot the change of expend parameter after removing each data point
plot(beta.change[,2])
```



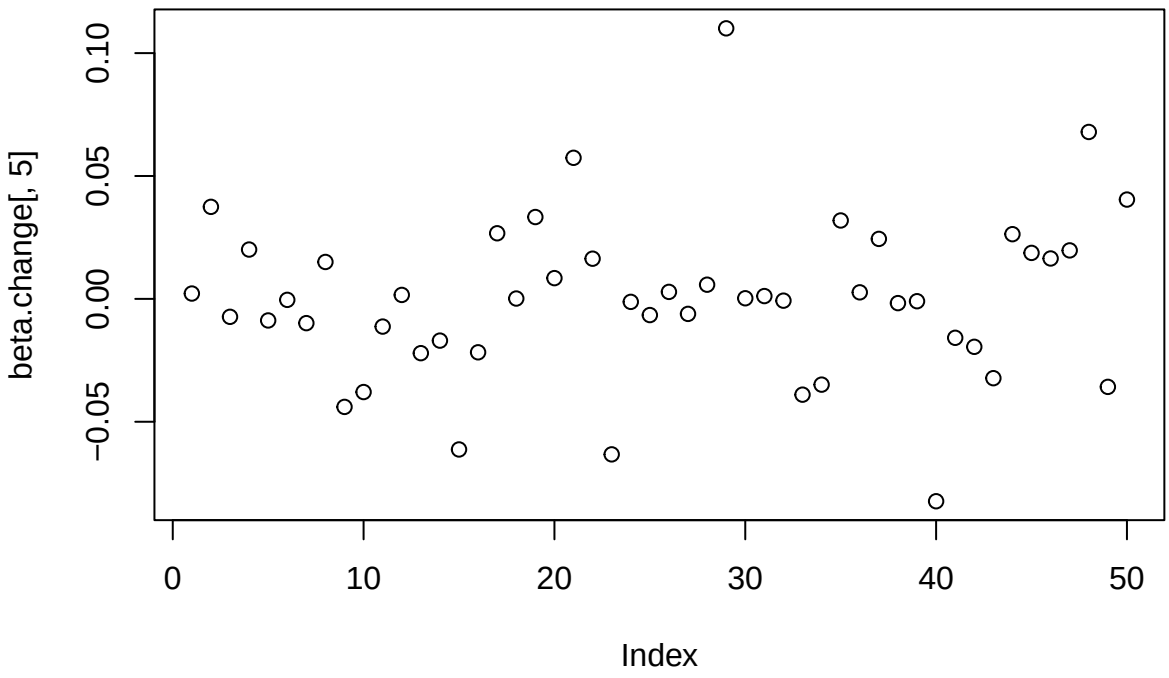

```
# plot the change of salary parameter after removing each data point
plot(beta.change[,3])
```




```
# plot the change of ratio parameter after removing each data point
plot(beta.change[,4])
```




```
# plot the change of takers parameter after removing each data point
plot(beta.change[,5])
```



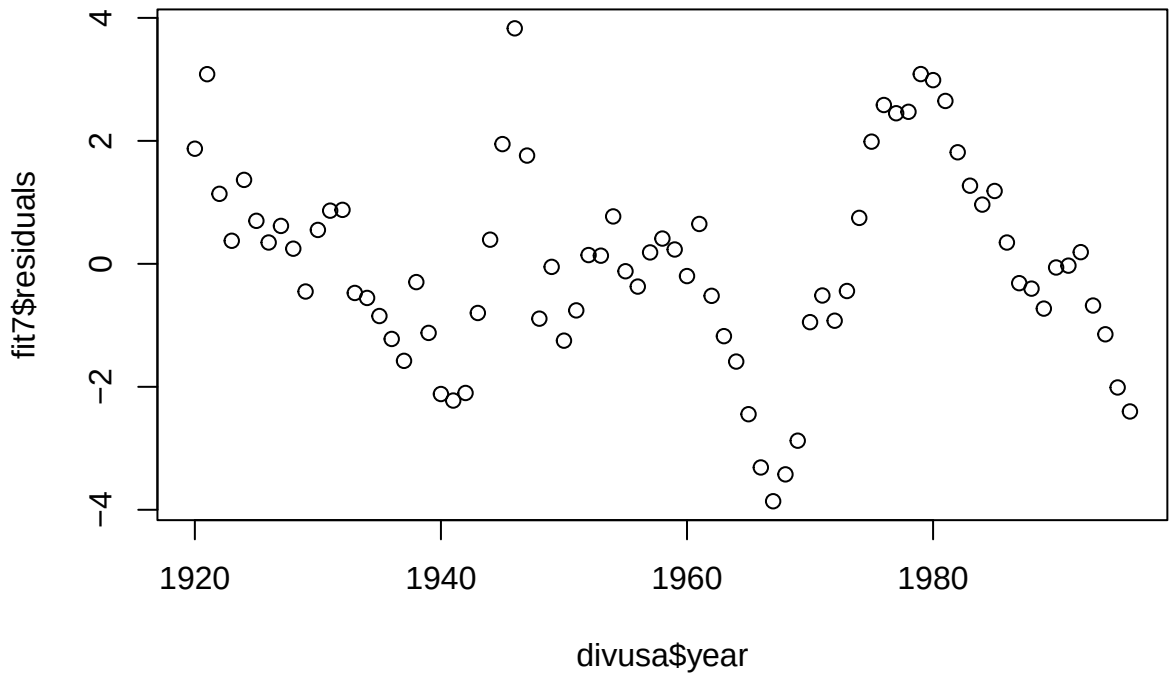
Question 5

(10 points) The divusa dataset records the US divorce and social-economic variables in 77 years.

Year: the year from 1920-1996 Divorce: divorce per 1000 women aged 15 or more Unemployed: unemployment rate Femlab: percent female participation in labor force aged 16+ Marriage: marriages per 1000 unmarried women aged 16+ Birth: births per 1000 women aged 15-44 Military: military personnel per 1000 population

Fit a model with divorce as the response and the other variables, except year, as predictors. Check for error correlation using three methods: plot residuals vs. years; plot residual (t+1) vs. residual (t); fit a linear model of residual (t+1) ~ residual (t). Give your insight and conclusion.

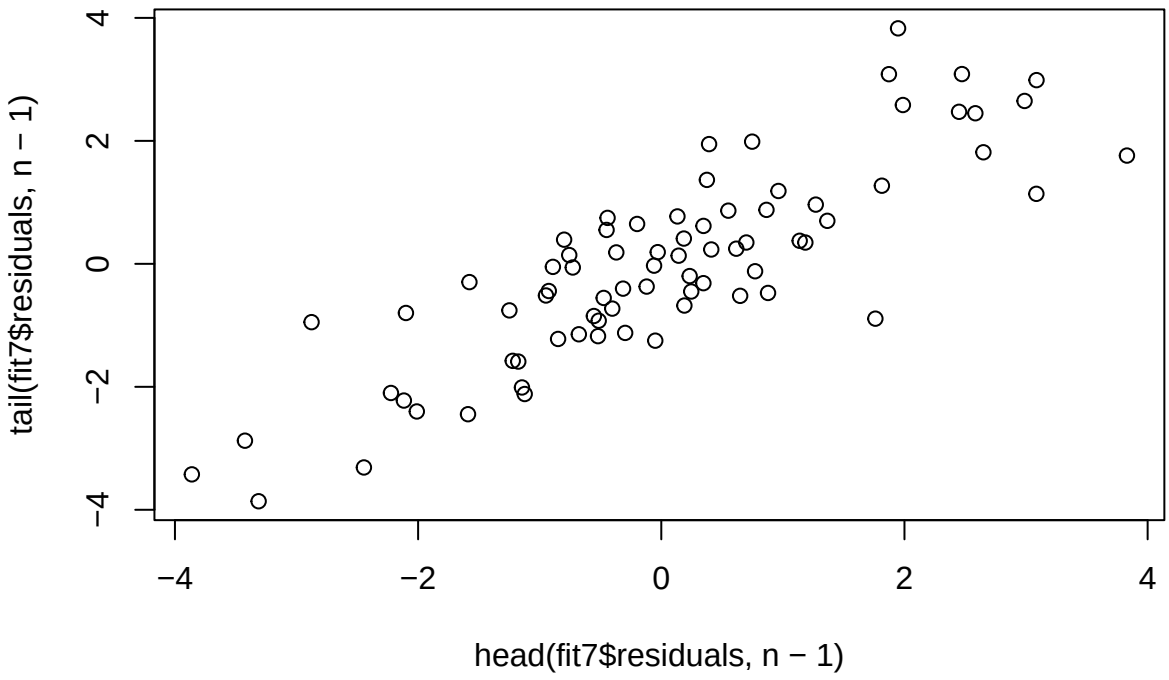
```
divusa <- read.csv("divusa.csv")
fit7 <- lm(divorce ~ unemployed + femlab + marriage + birth + military, data = divusa)
# plot residuals vs years
plot(divusa$year, fit7$residuals)
```




```
# plot residual (t+1) vs. residual (t)
n <- dim(divusa)[1]
cor(tail(fit7$residuals, n-1), head(fit7$residuals, n-1))
```

```
## [1] 0.8469792
```

```
plot(tail(fit7$residuals, n-1)~head(fit7$residuals, n-1))
```




```
# fit a linear model of residual (t+1) ~ residual (t)
fit8 <- lm(tail(fit7$residuals, n-1)~head(fit7$residuals, n-1))
summary(fit8)
```

```
##
## Call:
## lm(formula = tail(fit7$residuals, n - 1) ~ head(fit7$residuals,
##     n - 1))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34046 -0.54703 -0.08307  0.47315  2.22203
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -0.05155    0.09769  -0.528    0.599
## head(fit7$residuals, n - 1)  0.85213    0.06218  13.705   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8515 on 74 degrees of freedom
## Multiple R-squared:  0.7174, Adjusted R-squared:  0.7136
## F-statistic: 187.8 on 1 and 74 DF,  p-value: < 2.2e-16
```

The plot of residuals versus years does show a curved pattern, especially around 1980 where these is a distinct peak. The plot comparing the residuals of t and t+1 observations shows a somewhat linear pattern but the

points are definitely more spaced out toward the right side of the graph. Additionally, the linear model of residual (t+1) ~ residual (t) has a p-value of <2e-16 for the predictor, which is significant at the level alpha = 0.05. Therefore, these results do show indications of error correlation for this model.