Data 602 - Assignment Five

Michael Ellsworth, ID 30101253

October 14, 2019

Question 1

Refer to Question 6 from Assignment Four:

Consider your estimation of the model

$$R_{Suncor,i} = \beta_0 + \beta_1 R_{TSE,i} + e_i$$

a.

From these data, can you infer that the monthly rate of return of Suncor stock can be expressed as a positive linear function of the monthly rate of return of the TSE Index? State your statistical hypotheses, compute (and report) both the test statistic and the P-value and provide your decision.

$$\begin{aligned} \mathrm{H}_0 &: \beta_1 = 0 \;\; \text{(Suncor stock CANNOT be expressed as a positive linear function of TSE Index)} \\ \mathrm{H}_A &: \beta_1 > 0 \;\; \text{(Suncor stock CAN be expressed as a positive linear function of TSE Index)} \end{aligned}$$

The F_{obs} statistic:

$$\begin{aligned} F_{obs} &= \frac{MSR}{MSE} \sim F_{1,n-2} \\ df_R &= 2 - 1 = 1 \\ df_E &= n - 2 = 59 - 2 = 57 \\ MSR &= \frac{SSR}{df_R} = \frac{0.0638}{1} \\ MSE &= \frac{SSE}{df_E} = \frac{0.4612}{57} \\ F_{obs} &= 7.89 \end{aligned}$$

The P-value:

$$P - \text{value} = P(F_{1,57} > 7.89) = 0.0068$$

```
capmdata <- read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/capm.csv")
predict_suncor <- lm(Suncor ~ TSE.Index, data = capmdata)
summary(aov(predict_suncor))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## TSE.Index    1 0.0638 0.06384    7.89 0.0068 **
## Residuals   57 0.4612 0.00809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P value is less than 0.05, Suncor stock CAN be expressed as a positive linear function of TSE Index. We can reject H0.

b.

Compute a 95% confidence interval for β_1, then interpret its meaning in the context of these data.

Using the ‘confint’ function, the results from the TSE.Index row represent the 95% confidence interval for β_1

```
confint(predict_suncor, level = 0.95)
```

```
##                    2.5 %     97.5 %
## (Intercept) -0.006928949 0.04022482
## TSE.Index    0.154658904 0.92272309
```

Therefore, the 95% confidence interval for β_1 is:

$$0.155 \leq \beta_1 \leq 0.923$$

In the context of these data, as the monthly rate of return for the TSE Index increases by 1%, the average monthly rate of return for the Suncor share will increase between 0.155 and 0.923% with 95% confidence.

c.

Compute a 95% confidence interval for the mean monthly rate of return of Suncor stock when the TSE has a monthly rate of return of 3%.

```
predict(predict_suncor, newdata=data.frame(TSE.Index = 0.03), interval = "conf", conf.level = 0.95)
```

```
##          fit         lwr        upr
## 1 0.03280867 0.007660256 0.05795708
```

The 95% confidence interval for the mean monthly rate of return of Suncor stock when the TSE index has a monthly rate of return of 3% is $0.00766 \leq \widehat{R}_{Stock} \leq 0.0580$

d.

In a month of September, the TSE Index had a rate of return of 1.16%. With 95% confidence, compute the September rate of return for Suncor stock.

```
predict(predict_suncor, newdata=data.frame(TSE.Index = 0.0116), interval = "predict", conf.level = 0.95)
```

```
##          fit        lwr       upr
## 1 0.02289675 -0.1587618 0.2045553
```

The 95% confidence interval for $Y|_{X=x_p}$ is $-0.159 \leq Y|x = 0.016 \leq 0.205$

e.

Consider the bootstrap statistic r_{boot}. Using 1000 bootstraps, provide a 95% bootstrap confidence interval for the value of the ρ, the ***population correlation*** *that measures the degree of linear association between Suncor's monthly rate of return and the TSE Index monthly rate of return.*

```
Nbootstraps_1e = 1000 #resample n =  XX, 3000 times
cor.boot_1e = numeric(Nbootstraps_1e) #define a vector to be filled by the cor boot stat
a.boot_1e = numeric(Nbootstraps_1e) #define a vector to be filled by the a boot stat
b.boot_1e = numeric(Nbootstraps_1e) #define a vector to be filled by the b boot stat
ymean.boot_1e = numeric(Nbootstraps_1e) #define a vector to be filled by the predicted y boot stat
```

```
nsize_1e = dim(capmdata)[1]  #set the n to be equal to the number of bivariate cases, number of rows
xvalue_1e = 60000 #set x = 60000
#start of the for loop
for(i in 1:Nbootstraps_1e)
{   #start of the loop
    index = sample(nsize_1e, replace=TRUE)  #randomly picks n- number between 1 and n, assigns as index
    CAPM.boot = capmdata[index, ] #accesses the i-th row of the CAPM data frame
    #
    cor.boot_1e[i] = cor(~TSE.Index, ~Suncor, data=CAPM.boot) #computes correlation for each bootstrap sample
    CAPM.lm = lm(Suncor ~ TSE.Index, data = CAPM.boot)  #set up the linear model
    a.boot_1e[i] = coef(CAPM.lm)[1] #access the computed value of a, in position 1
    b.boot_1e[i] = coef(CAPM.lm)[2] #access the computed valeu of b, in position 2
    ymean.boot_1e[i] = a.boot_1e[i] + (b.boot_1e[i]*xvalue_1e)
}
#end the loop
#create a data frame that holds the results of teach of he Nbootstraps
bootstrapresultsdf_1e = data.frame(cor.boot_1e, a.boot_1e, b.boot_1e, ymean.boot_1e)
```

```
favstats(~cor.boot_1e, data = bootstrapresultsdf_1e)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
-0.2036122	0.1798418	0.3472199	0.4601895	0.7462704	0.3219679	0.1788879	1000	0

1 row

```
qdata(~cor.boot_1e, c(0.025, 0.975), data = bootstrapresultsdf_1e)
```

	quantile	p
	<dbl>	<dbl>
2.5%	-0.02538979	0.025
97.5%	0.60972309	0.975

2 rows

The 95% bootstrap confidence interval for ρ is $-0.0247 \leq \rho \leq 0.635$

Question 2

Refer to Question 7 from Assignment Four, where you wished to estimate the model

$$Balance_{Student,i} = A + (B * Income_{Student,i}) + e_i$$

a.

Compute the value of S_e, then interpret its meaning on the context of these data.

```
predict_balance <- lm(balance ~ income, filter(Default, student == "Yes"))
aov(predict_balance)
```

```
## Call:
##    aov(formula = predict_balance)
##
## Terms:
##                    income Residuals
## Sum of Squares     232619 686080187
## Deg. of Freedom         1      2942
##
## Residual standard error: 482.9099
## Estimated effects may be unbalanced
```

$$SSE = 686080187$$
$$n = 2944$$
$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{686080187}{2944-2}} = 482.91$$

b.

Compute the coefficient of determination, followed by its interpretation in the context of these data.

$$r^2 = \frac{SSR}{SST} = \frac{232619}{232619 + 686080187} = 0.0003389402$$

```
rsquared(predict_balance)
```

```
## [1] 0.00033894
```

c.

From what you have done with these data - Assignment Four and now - can you infer that a student's credit card balance can be expressed as a linear function of their income? Ensure you state your statistical hypotheses, provide both the value of your test statistic and P-value, and a decision and conclusion in the context of these data.

Based on the coefficient of determination, the statistical model does not mimic the actual relationship between student's credit card balance and income very well.

In order to further our understanding of the significance of the model, we can perform an F-test.

```
summary(aov(predict_balance))
```

```
##                Df    Sum Sq Mean Sq F value Pr(>F)
## income          1    232619  232619   0.997  0.318
## Residuals    2942 686080187  233202
```

Based on the results above:

$$F_{obs} = \frac{\frac{232619}{1}}{\frac{686080187}{2942}} = 0.997$$

$$P - \text{value} = P(F_{1,2942} > 0.997) = 0.318$$

Since the P-Value is greater than 0.05, we cannot assume that Balance can be expressed as a linear function of Income.

d.

(Perhaps read both this question and part (e) before you attempt to complete both.) Consider the coefficient of determination as a bootstrap statistic. Use 1000 resamples to generate the bootstrap distribution this statistic. Then, compute a 95% bootstrap confidence interval.

```r
Nbootstraps_2d = 1000 #resample n =  XX, 3000 times
rsquared.boot_2d = numeric(Nbootstraps_2d) #define a vector to be filled by the cor boot stat
a.boot_2d = numeric(Nbootstraps_2d) #define a vector to be filled by the a boot stat
b.boot_2d = numeric(Nbootstraps_2d) #define a vector to be filled by the b boot stat
ymean.boot_2d = numeric(Nbootstraps_2d) #define a vector to be filled by the predicted y boot stat
```

```r
nsize_2d = dim(filter(Default, student == "Yes"))[1]  #set the n to be equal to the number of bivariate cases, nu
mber of rows
xvalue_2d = 60000 #set x = 60000
#start of the for loop
for(i in 1:Nbootstraps_2d)
{   #start of the loop
    index = sample(nsize_2d, replace=TRUE)  #randomly picks n- number between 1 and n, assigns as index
    studentbalance.boot = filter(Default, student == "Yes")[index, ] #accesses the i-th row of the Default data f
rame
    studentbalance.lm = lm(balance ~ income, data = studentbalance.boot)  #set up the linear model
    rsquared.boot_2d[i] = rsquared(studentbalance.lm) #computes coefficient of determination for each bootstrap s
ample
    a.boot_2d[i] = coef(studentbalance.lm)[1] #access the computed value of a, in position 1
    b.boot_2d[i] = coef(studentbalance.lm)[2] #access the computed valeu of b, in position 2
}
#end the loop
#create a data frame that holds the results of teach of he Nbootstraps
bootstrapresultsdf_2d = data.frame(rsquared.boot_2d, a.boot_2d, b.boot_2d)
```

```r
favstats(~rsquared.boot_2d, data = bootstrapresultsdf_2d)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>
3.907224e-10	8.33583e-05	0.0003539672	0.0009464335	0.005479053	0.0006635297	0.0008310649	1000

1 row | 1-9 of 10 columns

```r
qdata(~rsquared.boot_2d, c(0.025, 0.975), data = bootstrapresultsdf_2d)
```

	quantile	p
	<dbl>	<dbl>
2.5%	9.377889e-07	0.025
97.5%	3.112043e-03	0.975

2 rows

The 95% bootstrap confidence interval for the coefficient of determination is $0.000000781 \leq r^2 \leq 0.002971762$

e.

Using 1000 different resamples, estimate the model above with a_{boot} and b_{boot}.

```
favstats(~a.boot_2d, data = bootstrapresultsdf_2d)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
898.7828	994.7439	1020.9	1045.703	1123.815	1020.996	36.57809	1000	0

1 row

```
favstats(~b.boot_2d, data = bootstrapresultsdf_2d)
```

min	Q1	median	Q3	max	mean	sd	n
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
-0.007846439	-0.003265622	-0.00187083	-0.000521237	0.004288318	-0.001880637	0.00200431	1000

1 row | 1-9 of 10 columns

Using the means for a_{boot} and b_{boot}, we can estimate the model as

$$\widehat{Balance}_{Student,i} = 1024.977 + (-0.0021 * Income_{Student,i})$$

Question 3

Refer to Question 9 of Assignment 1, where you were asked to refer to certain variables of the General Society Survey of 2002. For your convenience, the data file is linked below.

```
gss = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/GSS2002.csv")
```

a.

*Is there a relationship between one's support for gun laws (variable name is **GunLaw**) and their opinion about current government spending on Science (variable name is **SpendSci**)? State the appropriate statistical hypotheses.*

$$H_0 : \text{Support of gun laws and opinion about government spending on science are independent}$$
$$H_A : \text{Support of gun laws and opinion about government spending on science are NOT independent}$$

b.

Use R Studio to create the contingency table.

```
gun_science <- gss %>%
  filter(!is.na(GunLaw), !is.na(SpendSci))
gun_science_tally <- tally(~GunLaw + SpendSci, data = gun_science)
gun_science_tally
```

```
##         SpendSci
## GunLaw   About right Too little Too much
##   Favor          166        117       42
##   Oppose          35         37       12
```

	Science Spending		
Gun Law Support	About right	Too little	Too much
Favor	166	117	42
Oppose	35	37	12

c.

Carry out the appropriate statistical test, providing both the test statistic and the P-value.

```
xchisq.test(gun_science_tally, correct=FALSE)
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  x
## X-squared = 2.4447, df = 2, p-value = 0.2945
## 
##    166      117       42
## (159.72) (122.37) ( 42.91)
## [0.247]  [0.236]  [0.019]
## < 0.50>  <-0.49>  <-0.14>
## 
##     35       37       12
## ( 41.28) ( 31.63) ( 11.09)
## [0.956]  [0.912]  [0.075]
## <-0.98>  < 0.96>  < 0.27>
## 
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

From this output, we observe the value of the test statistic $\chi^2_{obs} = 2.4447$ and the P-value is 0.2945.

d.

What can you conclude? Do these data support your null hypothesis in part (a)? State your decision and conclusion.

Since the P-value is greater than 0.05, we cannot reject the null hypothesis and we must assume that support on gun laws and opinion on government spending on science are independent.

e.

*Re-trace a result, in the form of a bar-graph, that was provided in Assignment 2, Question 9. Can you infer from these data that one's level of **Education** is independent of their **Race**? Present your findings in the form of a paragraph, outlining the decision you have made, why you made the decision you made, and the P-value.*

$$H_0 : \text{Education and Race are independent}$$
$$H_A : \text{Education and Race are NOT independent}$$

```r
race_education <- gss %>%
  filter(!is.na(Race), !is.na(Education))
race_education_tally <- tally(~Race + Education, data = race_education)
race_education_tally
```

```
##        Education
## Race    Bachelors Graduate   HS Jr Col Left HS
##   Black        27       15  231     34     101
##   Other        27       17   81     17      24
##   White       389      198 1173    151     275
```

```r
xchisq.test(race_education_tally, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 79.05, df = 8, p-value = 7.59e-14
##
##      27        15       231        34       101
## (   65.49) (   34.00) ( 219.52) (   29.86) (   59.13)
## [2.3e+01] [1.1e+01] [6.0e-01] [5.7e-01] [3.0e+01]
## <-4.756> <-3.258> < 0.775> < 0.757> < 5.445>
##
##      27        17        81        17        24
## (   26.64) (   13.83) (   89.32) (   12.15) (   24.06)
## [4.8e-03] [7.2e-01] [7.7e-01] [1.9e+00] [1.4e-04]
## < 0.069> < 0.851> <-0.880> < 1.392> <-0.012>
##
##     389       198      1173       151       275
## ( 350.87) ( 182.17) (1176.16) ( 159.99) ( 316.81)
## [4.1e+00] [1.4e+00] [8.5e-03] [5.1e-01] [5.5e+00]
## < 2.036> < 1.173> <-0.092> <-0.711> <-2.349>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

From the results of the xchisq test above, we can infer that one's level of Education is NOT independent of their Race. This decision was made based on the results of the test statistic and the P-value presented below:

$$\chi^2_{obs} = 79.05 \text{ with a } P - \text{value} = P(\chi^2_{df=8} > 79.05) \approx 0$$

Since the P-value is almost 0, we can infer that one's level of Education is NOT independent of their Race.

Question 4

A group of patients with a binge-eating disorder were randomly assigned to take either the experimental drug fluvoxamine or the placebo in a nine-week-long, double-blinded clinical trial. At the end of the trial the condition of each patient was classified into one of four categories: no response, moderate response, marked response, or remission. The table below shows a cross-classification, or contingency table, of these data.

	No Response	Moderate Response	Marked Response	Remission
Fluvoxamine	15	7	3	15
Placebo	22	7	3	11

Do these data provide statistically significant evidence to conclude that there is an association between the type of treatment received and a patient's response?

Ensure you provide your statistical hypotheses, test statistic and P-value in your finding(s).

$$H_0 : \text{The type of treatment received and the patient's response are independent}$$
$$H_A : \text{The type of treatment received and the patient's response are NOT independent}$$

Contingency table:

```
treatment_response <- rbind(c(15, 7, 3, 15), c(22, 7, 3, 11))
rownames(treatment_response) = c("Fluvoxamine", "Placebo")
colnames(treatment_response) = c("No Response", "Moderate Response", "Marked Response", "Remission")
treatment_response
```

```
##             No Response Moderate Response Marked Response Remission
## Fluvoxamine          15                 7               3        15
## Placebo              22                 7               3        11
```

```
xchisq.test(treatment_response, simulate.p.value=TRUE)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  x
## X-squared = 1.8337, df = NA, p-value = 0.6262
##
##  15.00     7.00     3.00    15.00
## (17.83)  ( 6.75)  ( 2.89)  (12.53)
## [0.4496] [0.0095] [0.0041] [0.4869]
## <-0.670> < 0.097> < 0.064> < 0.698>
##
##  22.00     7.00     3.00    11.00
## (19.17)  ( 7.25)  ( 3.11)  (13.47)
## [0.4182] [0.0088] [0.0038] [0.4529]
## < 0.647> <-0.094> <-0.062> <-0.673>
##
## key:
##  observed
##  (expected)
##  [contribution to X-squared]
##  <Pearson residual>
```

$$\chi^2_{obs} = 1.8337 \text{ with a } P - \text{value} = 0.6257$$

These data DO NOT provide statistically significant evidence to conclude that there is an association between the type of treatment received and a patient's response. We cannot reject the null hypothesis.

Question 5

Was Barry Bonds using Steroids? The following bivariate data set gives the year and the number of home runs divided by the number of at bats - attempts to hit the ball - for each season. The number of homeruns is not used as later in his career he was given intential walks, which do not count as an at bat.

In this exercise, you will build on your learning of model building and attempt to predict the number of home runs Barry Bonds would have hit in the 2001 season. These data are stored in the data file (http://people.ucalgary.ca/~jbstall/DataFiles/bondsdata.csv).

*Read these data into a data frame called **Ass5ques5data**, then look at the first three and the last three rows as a “check”.*

```
Ass5ques5data = read.csv("http://people.ucalgary.ca/~jbstall/DataFiles/bondsdata.csv")
head(Ass5ques5data, 3)
```

	season <int>	hrat <dbl>
1	1987	0.045372
2	1988	0.044610
3	1989	0.032759

3 rows

```
tail(Ass5ques5data, 3)
```

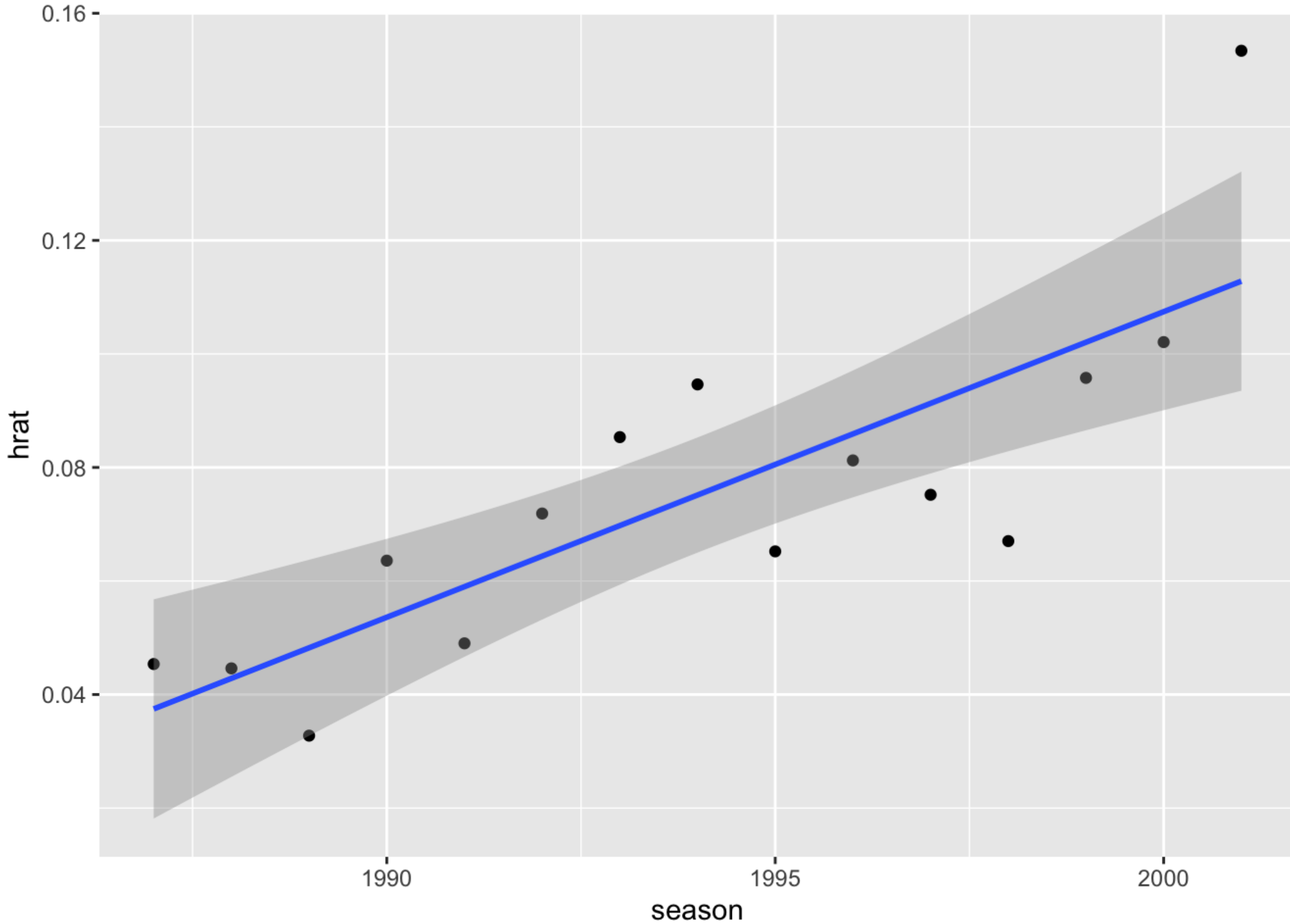
	season <int>	hrat <dbl>
13	1999	0.095775
14	2000	0.102083
15	2001	0.153400

3 rows

a.

*Create a scatter plot of these data, with **season** acting as your x-variable and **hrat** (home runs to at bat ratio) acting as your y-variable.*

```
Ass5ques5data %>%
  ggplot(aes(x = season, y = hrat)) +
  geom_point() +
  geom_smooth(method = "lm")
```



b.

Remove the data point that corresponds to the ***season == 2001****. After, you are attempting to build a statistical model of the following form:*

$$HRAT_i = A + B * Year_i + e_i \qquad i = 1993, 1994, \cdots, 2000.$$

Estimate this model and compute the S_e *as well as* r^2.

```
predict_hrat <- lm(hrat~season, data = filter(Ass5ques5data, season != 2001))
predict_hrat
```

```
##
## Call:
## lm(formula = hrat ~ season, data = filter(Ass5ques5data, season !=
##     2001))
##
## Coefficients:
## (Intercept)       season
##   -7.992499     0.004044
```

```
aov(predict_hrat)
```

```
## Call:
##    aov(formula = predict_hrat)
##
## Terms:
##                      season   Residuals
## Sum of Squares  0.003720832 0.002119886
## Deg. of Freedom           1          12
##
## Residual standard error: 0.01329124
## Estimated effects may be unbalanced
```

```
summary(aov(predict_hrat))
```

```
##             Df   Sum Sq  Mean Sq F value   Pr(>F)
## season       1 0.003721 0.003721   21.06 0.000622 ***
## Residuals   12 0.002120 0.000177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
rsquared(predict_hrat)
```

```
## [1] 0.6370504
```

$$\widehat{AverageHRAT}_i = -7.99 + (0.00404 * Season_i)$$

$$S_e = 0.0133$$

$$r^2 = 0.637$$

c.

From these data, can you conclude that Bonds' home-run-to-at-bat ratio ***hrat*** *can be expressed as a positive linear function of the number of seasons he has played? A comment based on your statistical hypotheses and subsequent P-value is sufficient here.*

$$\mathrm{H}_0 : B = 0 \text{ (Y cannot be expressed as a linear function of X)}$$
$$\mathrm{H}_A : B > 0 \text{ (Y can be expressed as a linear function of X)}$$

Since the P-value above is 0.0006, we can reject H_0 and can conclude that Bonds' home-run-to-at-bat ratio can be expressed as a positive linear function of the number of seasons he has played.

d.

Compute the 95% confidence interval for B, and interpret its meaning on the context of these data.

```
confint(predict_hrat, level=0.95)
```

```
##                     2.5 %       97.5 %
## (Intercept) -11.819970817 -4.165027763
## season        0.002124197  0.005964141
```

$$0.00212 \leq B \leq 0.00596$$

As Barry Bonds' seasons played increases by 1, then his home-run-to-at-bat ratio will increase by an average of anywhere between:

$$0.00212 \leq h_{rat} \leq 0.00596$$

As the number of seasons played increases by one, the average home-run-to-at-bat ratio will increase by an average from 0.00212 and 0.00596.

e.

Find a 95% prediction level for Bonds' homerun to at bat ratio in 2001. What does your interval represent?

```
predict(predict_hrat, newdata=data.frame(season = 2001), interval="predict", conf.level=0.95)
```

```
##           fit        lwr       upr
## 1 0.09988334 0.06662845 0.1331382
```

$$0.0666 \leq Y|x = 2001 \leq 0.133$$

f.

During the 2001 Season, the number of at bat Bonds had was 476. Since the $HRAT$ ratio is defined as

$$HRAT = \frac{\text{no. homeruns}}{\text{no. At Bats}}$$

Use the result you obtained in part (e) to predict the number of homeruns that Bonds would have hit in the 2001 season.

```
lb_5f <- 0.06662845 * 476
lb_5f
```

```
## [1] 31.71514
```

```
ub_5f <- 0.1331382 * 476
ub_5f
```

```
## [1] 63.37378
```

$$31.7 \leq homerun|at\ bats = 476 \leq 63.4$$

```
0.09988334 * 476
```

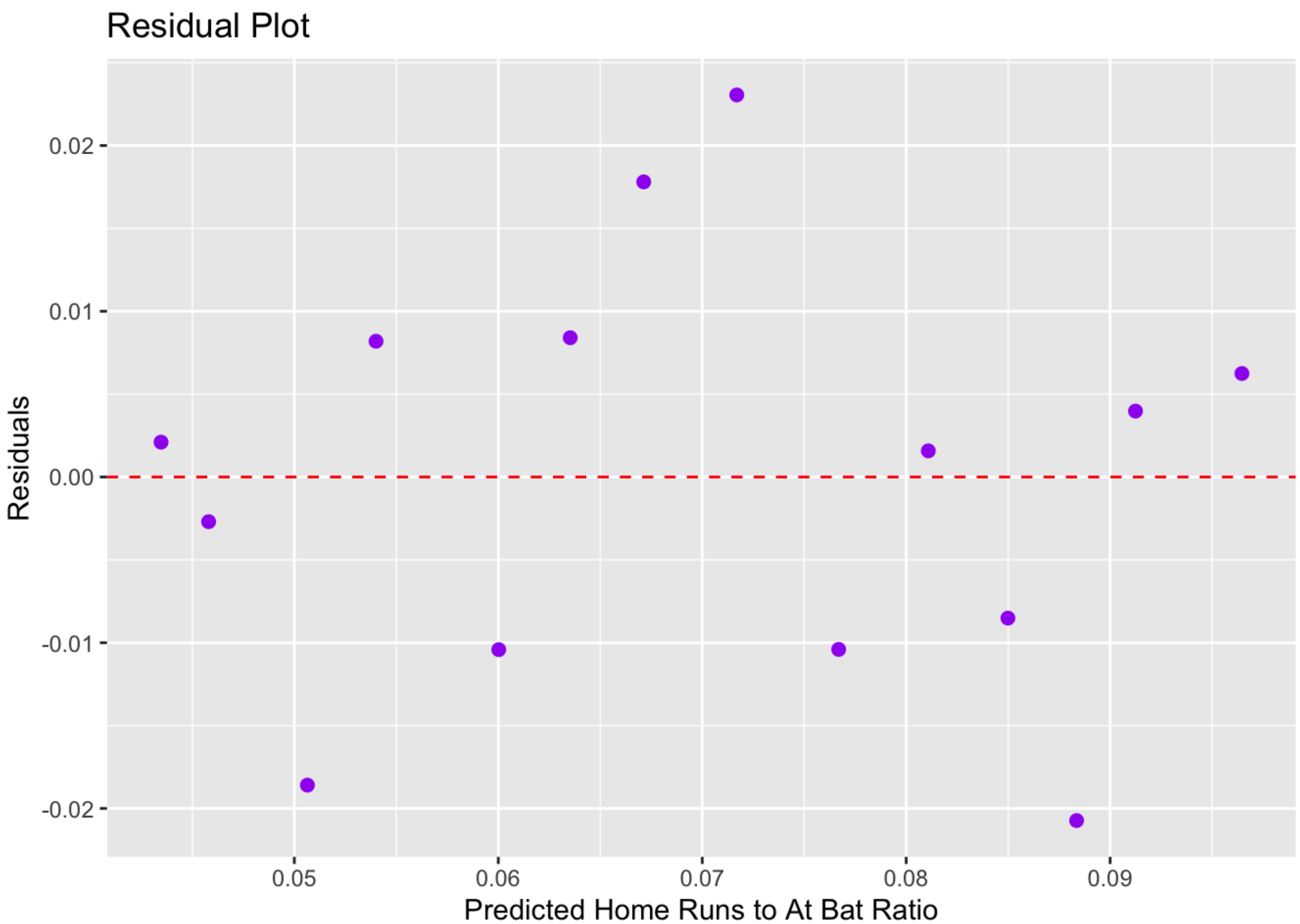
```
## [1] 47.54447
```

The number of homeruns would be ≈ 47.5

g.

Create a residual plot. What condition does this residual plot inspect? Does this condition appear to hold?

```
predictshrat <- predict_hrat$fitted.values
eishrat <- predict_hrat$residuals
diagnosticdf_5g <- data.frame(predictshrat, eishrat)
diagnosticdf_5g %>%
  ggplot(aes(x = predict_hrat$fitted.values, y = predict_hrat$residuals)) +
  geom_point(col = 'purple', size = 2, position = "jitter") +
  xlab("Predicted Home Runs to At Bat Ratio") +
  ylab("Residuals") +
  ggtitle("Residual Plot") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed")
```



This plot checks the homoscedasticity condition, or the error term is the same across the range of predicted HRAT. This condition appears to hold.

6.

Reconsider the data presented in Question 5. Use the bootstrap method (1000 resamples) as a means to estimate the model presented in Question 5.

```
Nbootstraps_6 = 1000
a.boot_6 = numeric(Nbootstraps_6)
b.boot_6 = numeric(Nbootstraps_6)
```

```
nsize_6 = dim(filter(Ass5ques5data, season != 2001))[1]
xvalue_6 = 60000

for(i in 1:Nbootstraps_6)
  {
    index = sample(nsize_6, replace=TRUE)
    HRAT.boot = filter(Ass5ques5data, season != 2001)[index, ]
    HRAT.lm = lm(hrat ~ season, data = HRAT.boot)
    a.boot_6[i] = coef(HRAT.lm)[1]
    b.boot_6[i] = coef(HRAT.lm)[2]
}
#end the loop
#create a data frame that holds the results of teach of he Nbootstraps
bootstrapresultsdf_6 = data.frame(a.boot_6, b.boot_6)
```

```
favstats(~a.boot_6, data = bootstrapresultsdf_6)
```

min <dbl>	Q1 <dbl>	median <dbl>	Q3 <dbl>	max <dbl>	mean <dbl>	sd <dbl>	n <int>	missing <int>
-14.26587	-8.91433	-7.933095	-6.992919	-1.923981	-7.945696	1.533114	1000	0

1 row

```
qdata(~a.boot_6, c(0.025, 0.975), data = bootstrapresultsdf_6)
```

	quantile	p
	<dbl>	<dbl>
2.5%	-10.743263	0.025
97.5%	-4.883155	0.975

2 rows

```
favstats(~b.boot_6, data = bootstrapresultsdf_6)
```

min	Q1	median	Q3	max	mean	sd	n	missing
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<int>	<int>
0.001001639	0.00354234	0.004013775	0.004509048	0.007195036	0.004020839	0.0007692043	1000	0

1 row

```
qdata(~b.boot_6, c(0.025, 0.975), data = bootstrapresultsdf_6)
```

	quantile	p
	<dbl>	<dbl>
2.5%	0.002484997	0.025
97.5%	0.005423301	0.975

2 rows

From the above bootstrap distribution of a_{boot} and b_{boot}, we can take the means to estimate the model.

$$\widehat{AverageHRAT}_i = -7.96 + (0.00403 * Season_i)$$

Ensure you have justified the computations and your findings.