Stat 230 HW 4

Name: Victor Huang

### worked with: No one

Homework 4 is due **by 3pm Thursday, Oct. 14**. Please complete the assignment in this Markdown document, filling in your answers and R code below. I didn’t create answer and R chunk fields like I did with homework 1, but please fill in your answers and R code in the same manner as hw 1. Submit a hard copy of the **compiled pdf or word doc** either

* in class
* in drop-in office hours
* in the paper holder outside my CMC 222 office door

Tips for using Markdown with homework sets:

* Work through a problem by putting your R code into R chunks in this .Rmd. Run the R code to make sure it works, then knit the .Rmd to verify they work in that environment.
  + Make sure you load your data in the .Rmd and include any needed library commands.
* Feel free to edit or delete questions, instructions, or code provided in this file when producing your homework solution.
* For your final document, you can change the output type from html\_document to word\_document or pdf\_document. These two to output types are better formatted for printing.
  + on maize: you may need to allow for pop-ups from this site
* If you want to knit to pdf while running Rstudio from your computer (*not* from maize), you will need a LaTeX compiler installed on your computer. This could be [MiKTeX](https://miktex.org/), [MacTeX](http://www.tug.org/mactex/) (mac), or TinyTex. The latter is installed in R: first install the R package tinytex, then run the command tinytex::install\_tinytex() to install this software.
  + If you are using maize, you don’t need to install anything to knit to pdf!

## Problem 1: Depression and Education: ch. 9 exercise 19 parts (a), (b) and (c)

* Part (c) is missing its label in my version of your textbook, if this is the case in your textbook, then part (c) is the question that follows the part (b) question.
* There **is no data for this problem**. Write down a *specific model mean function* **with indicator variables for factor levels**.
* **Hint:** You must carefully write a mean function for both parts (a) and (b). You will be using dummy variables for your categorical variable, so you must define what the dummy variable(s) are indicating (i.e. which level are they identifying?). When trying to answer the questions about divergence, make sure to review the info provided for the problem - they are wondering if the difference in mean depression score for college (i) and high school only (iii) widens as people age.

1. As we can see from the equation, the would be the parameter that measures the diverging gap between categories (i) and (iii) with age.
2. As we can see from the equation, a possible parameter would be would be a parameter that measures the diverging gap.
3. Yes, by measuring the parameter we can see how age and education correlate with depression, thus, characteizing the divergence hypothesis.

## Problem 2: Bat Echolocation

Use the data from case study 10.1.2 (case1002) to answer the questions below.

> library(Sleuth3)  
> library(dplyr)  
> library(ggplot2)  
> library(ggResidpanel)  
> library(broom)  
> library(knitr)  
> case1002 <- case1002

### (2a)

Use R to fit the parallel regression lines model for the regression of log energy on log mass and vertebrate type. (The same model fit in the case study). Note that your model summary will be different from display 10.6 because the baseline Type group set by R is different than the baseline used by the book. **Do not** change the baseline group in R to make your results look like display 10.6.

> case1002 <- case1002  
> case1002\_lm <- lm(log(Energy) ~ log(Mass) + Type, data=case1002)  
> summary(case1002\_lm)  
  
Call:  
lm(formula = log(Energy) ~ log(Mass) + Type, data = case1002)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.23224 -0.12199 -0.03637 0.12574 0.34457   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1.49770 0.14987 -9.993 2.77e-08 \*\*\*  
log(Mass) 0.81496 0.04454 18.297 3.76e-12 \*\*\*  
Typenon-echolocating bats -0.07866 0.20268 -0.388 0.703   
Typenon-echolocating birds 0.02360 0.15760 0.150 0.883   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.186 on 16 degrees of freedom  
Multiple R-squared: 0.9815, Adjusted R-squared: 0.9781   
F-statistic: 283.6 on 3 and 16 DF, p-value: 4.464e-14

Here we have the general estimated mean model

Here we have the estimated mean model for echo locating bats

Here we have the estimated mean model for non-echo locating bats

Here we have the estimated mean model for non-echo locating birds

Use the summary results, write down the estimated mean models for each vertebrate type: (i) echolocating bats, (ii) non-echo locating bats and (iii) non-echo locating birds. Simplify your models/equations as much as possible (e.g. give a formula with estimated slope and intercept for *each* of the three mean log energy lines, but no need to “unlog” your variables.)

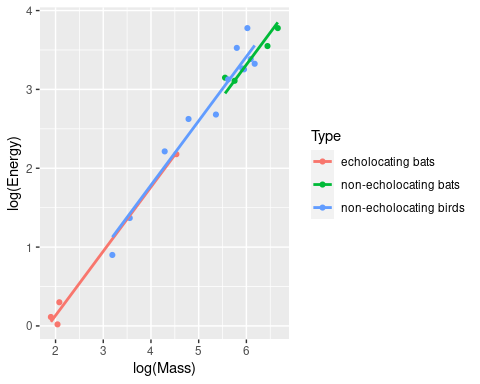
### (2b)

With results from (a), test whether the lines for the echo locating bats and non-echo locating birds coincide (e.g. are the same).

> case1002\_lm <- lm(log(Energy) ~ log(Mass) + Type, data=case1002)  
> kable(tidy(case1002\_lm))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -1.4976965 | 0.1498690 | -9.9933702 | 0.0000000 |
| log(Mass) | 0.8149575 | 0.0445414 | 18.2966182 | 0.0000000 |
| Typenon-echolocating bats | -0.0786637 | 0.2026793 | -0.3881190 | 0.7030432 |
| Typenon-echolocating birds | 0.0235982 | 0.1576005 | 0.1497345 | 0.8828453 |

> case1002\_augment <- augment(case1002\_lm) # add fitted values  
> ggplot(case1002\_augment, aes(`log(Mass)`, `log(Energy)`, color = Type)) +   
+ geom\_point() +   
+ geom\_line(aes(y = .fitted), size = 1)

 Looking at the p-value, since they are significantly larger than 0.05, we fail to reject the null hypothesis and conclude that the lines for echo locating bats and non-echo locating birds coincide.

### (2c)

Do the following test: vs.  where is the coefficient for the non-echo locating bats and is the coefficient for the non-echo locating birds. Report the test stat and p-value for this test, then state your conclusion in context - carefully explaining what these hypotheses are stating about your model for log energy.

> case1002\_lm <- lm(log(Energy) ~ log(Mass) + Type, data=case1002)  
> vcov(case1002\_lm)  
 (Intercept) log(Mass) Typenon-echolocating bats  
(Intercept) 0.022460721 -0.005235301 0.009482580  
log(Mass) -0.005235301 0.001983939 -0.006869742  
Typenon-echolocating bats 0.009482580 -0.006869742 0.041078883  
Typenon-echolocating birds 0.004914868 -0.005138789 0.026439562  
 Typenon-echolocating birds  
(Intercept) 0.004914868  
log(Mass) -0.005138789  
Typenon-echolocating bats 0.026439562  
Typenon-echolocating birds 0.024837918  
> est <- -0.07866 - 0.0235982  
> se.est <- sqrt(0.009482580 + 0.009482580 \*(-1)^2 + 2\*(-1)\*0.004914868)  
> tstat <- est/se.est  
> 2\*(1- pt(-tstat, df = 18))  
[1] 0.2988123

From the calculations shown above we get a t-stat value of 0.3 and a corresponding p-value of 1.701. Since our p-value is much larger than 0.05, we fail to reject the alternative hypothesis and accept the null hypothesis. As such, we cannot conclude that the effect of body weight on mean total energy expenditure differs between non-echo locating bats and non-echo locating birds.

### (2d)

Draw a scatterplot of the data for this problem (colored by animal type) and add the parallel regression lines to the plot for each type of animal. **Hint:** See the Stat 230 Notes [Section 3.4.1.1 example](https://kstclair.github.io/stat-230-notes/mlr.html#mlr-penguins1) for the correct way to draw a parallel line model in ggplot2.

> library(skimr)  
> skim(case1002)

Data summary

|  |  |
| --- | --- |
| Name | case1002 |
| Number of rows | 20 |
| Number of columns | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 1 |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| Type | 0 | 1 | FALSE | 3 | non: 12, ech: 4, non: 4 |

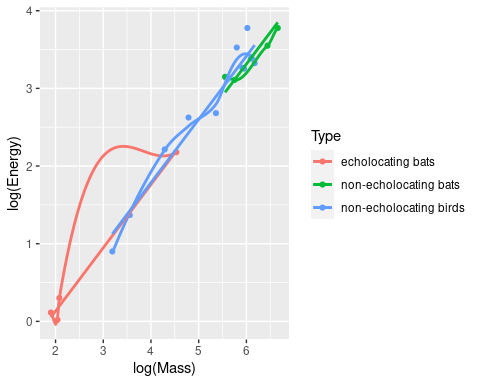
**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| Mass | 0 | 1 | 262.68 | 220.9 | 6.70 | 63.35 | 266.5 | 391.00 | 779.0 | ▇▃▅▁▂ |
| Energy | 0 | 1 | 19.52 | 14.0 | 1.02 | 7.61 | 22.6 | 28.23 | 43.7 | ▇▂▆▅▂ |

> case1002\_lm <- lm(log(Energy) ~ log(Mass) + Type, data=case1002)  
> kable(tidy(case1002\_lm))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -1.4976965 | 0.1498690 | -9.9933702 | 0.0000000 |
| log(Mass) | 0.8149575 | 0.0445414 | 18.2966182 | 0.0000000 |
| Typenon-echolocating bats | -0.0786637 | 0.2026793 | -0.3881190 | 0.7030432 |
| Typenon-echolocating birds | 0.0235982 | 0.1576005 | 0.1497345 | 0.8828453 |

> case1002\_augment <- augment(case1002\_lm) # add fitted values  
> ggplot(case1002\_augment, aes(`log(Mass)`, `log(Energy)`, color = Type)) +   
+ geom\_point() +   
+ geom\_smooth(se = FALSE)+   
+ geom\_line(aes(y = .fitted), size = 1)  
`geom\_smooth()` using method = 'loess' and formula 'y ~ x'  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : span too small. fewer data values than degrees of freedom.  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : pseudoinverse used at 1.889  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : neighborhood radius 0.19049  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : reciprocal condition number 0  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : There are other near singularities as well. 6.2727  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : span too small. fewer data values than degrees of freedom.  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : pseudoinverse used at 5.5474  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : neighborhood radius 0.89511  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : reciprocal condition number 0  
Warning in simpleLoess(y, x, w, span, degree = degree, parametric =  
parametric, : There are other near singularities as well. 0.82985



## Problem 3: Agstrat data revisited again

Recall the agstrat.csv data used for homework 1 and 2. This was a stratified random sample of US counties. We will consider the regression of farm acreage in 1992 (y) on farm acreage in 1982 (x).

> agstrat <- read.csv("http://people.carleton.edu/~kstclair/data/agstrat.csv")

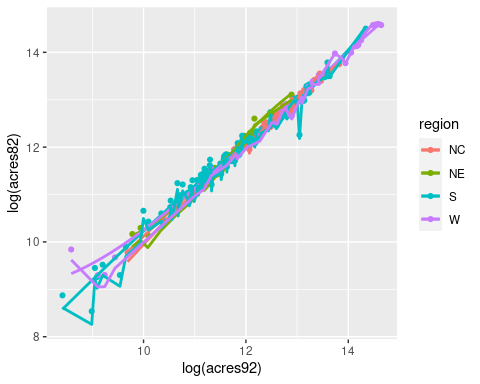
### (3a)

Fit the regression of the log of acres92 on the log of acres82 and region with rows 119 (New York County, NY) and 168 (Parish County, LA) removed from the data. (These counties have very low or 0 acres). Write down the fitted mean functions (models) for each of the four regions. Simplify your models/equations as much as possible (e.g. give a formula with estimated slope and intercept for each of the regions, but no need to “unlog” your variables.)

> agstrat\_new <- agstrat %>% slice(-119,-168)  
> agstrat <- read.csv("http://people.carleton.edu/~kstclair/data/agstrat.csv")  
> agstrat\_new <- agstrat %>% slice(-119,-168)  
> agstrat\_lm <- lm(log(acres92) ~ log(acres82) + region, data = agstrat\_new)  
> kable(tidy(agstrat\_lm))

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -0.7251775 | 0.1061376 | -6.8324272 | 0.0000000 |
| log(acres82) | 1.0539321 | 0.0084236 | 125.1165159 | 0.0000000 |
| regionNE | -0.0920789 | 0.0352412 | -2.6128212 | 0.0094432 |
| regionS | -0.0126336 | 0.0190963 | -0.6615755 | 0.5087632 |
| regionW | -0.0248363 | 0.0258738 | -0.9599040 | 0.3378950 |

> agstrat\_augment <- augment(agstrat\_lm) # add fitted values  
> ggplot(agstrat\_augment, aes(`log(acres92)`, `log(acres82)`, color = region)) +   
+ geom\_point() +   
+ geom\_smooth(se = FALSE) +   
+ geom\_line(aes(y = .fitted), size = 1)  
`geom\_smooth()` using method = 'loess' and formula 'y ~ x'



> glance(agstrat\_lm) %>% select(df.residual, nobs) %>% kable()

|  |  |
| --- | --- |
| df.residual | nobs |
| 293 | 298 |

Note: you can omit one or more cases from a lm by adding a subset argument to the command. For example, adding subset=-c(119,168) to the lm command will omit rows 119 and 168 from your regression fit!

### (3b)

Consider the t-test for the parameter (the parameter for the NE region). Using the p-value given for the t-test, carefully interpret the result of this test.

From the calculations shown above we get a t-stat value of -2.61282 and a corresponding p-value of 1.990557 Since our p-value is much larger than 0.05, we fail to reject the alternative hypothesis and accept the null hypothesis. As such, we cannot conclude that there is an effect of acreage in 82 in the NE region on mean acreage in 92 in the NE region, after controlling all other variables.

## Problem 4: Wages and Race

Consider ch.10 exercise 29 to answer the following questions. Review the background info for this exercise and data coding provided by the exercise description. We will still consider whether the CPS provides evidence of racial discrimination with respect to wages after accounting for factors like education, experience and location. Here is the model we are considering for this problem:

### (4a)

Write down the model mean function for Black males in the midwest region. Then repeat this for non-Black males in the midwest. Which **one** parameter () measures the difference in mean earnings between Black and non-Black populations in the midwest region?

By subtracting the two, we see that it is the that measures the difference in mean earnings between Black and non-Black populations in the midwest region.

### (4b)

Write down the model mean function for Black males in the northeast region. Then repeat this for non-Black males in the northeast. Which **combination** of parameters (’s) measures the difference in mean earnings between Black and non-Black populations in the northeast region?

### (4c)

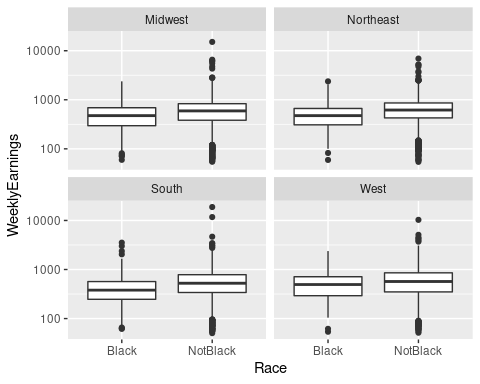
Researchers are particularly interested in determining if the effect of race (Black/non-Black) on earnings of males differs by region, after controlling for region, education, experience, and metropolitan status. Explain why the model above can be used to answer this question.

The model can be used to explain this effect because, with a baseline set for black employees in the Midwest region, every factor and combination of factors exists a parameter that shows its effects on the mean (parameters specific to the region as well as parameters specific to non-Black in a region). All the variables can be considered categorical and to determine effect of race in each region all that would need to be done is to keep the controlled region, education, experience, and metropolitan status constant while changing the categorical not related to the region to “0” while keeping the variables that are related to the region to “1” (as shown in the example above for NE and MW). This way, we can determin the effect of race on earnings of male differing from region to region.

### (4d)

Using the data for exercise 29 (ex1029), create the ggplot2 package boxplot of log earnings vs. race *faceted* by region (make sure to turn eval=TRUE in your homework):

> ex1029 <- ex1029  
> ggplot(ex1029, aes(x=Race, y=WeeklyEarnings)) +   
+ geom\_boxplot() +   
+ facet\_wrap(~Region) +   
+ scale\_y\_log10()

 Explain what this plot shows about the relationship between earnings, race and region. Does it support the idea that the effect of race on earnings differs by region?

(note: you will actually fit this model and analyze it in HW 5!)

## Looking at the boxgraph, we see that the NotBlack mean weekly earnings is consistenly higher than their black counterparts. However, as we move from region to region while keeping all other variables controlled, we do not see a significant shift or change casued by the chagne of region. As scuh, we can conclude that while ones race and ethnicity plays a role in their weekly earnings, the region in which they work does not have a significant effect on their earnings (the difference of race persists throughout all regions despite changing regions).

## Problem 5: Woodpecker nesting

**No R work is needed (expect for qt command) - all the info for this problem is given below.** Recall the woodpecker example from the day 3 notes. Below is the output for the regression of nest depth (cm) on air temp (Celsius), along with covariance matrix for the coefficients. Compute a 95% confidence interval for the mean response when temp is 8 degrees Celsius using the facts about inference for linear combinations of coefficients from ch. 10 (section 10.4.3). Your CI should match (except for rounding error) the CI obtained in the day 3 notes using the predict command: 15.83 to 18.94 cm.

> wpdata<- read.csv("http://people.carleton.edu/~kstclair/data/woodpeckers.csv")  
> wood\_lm<- lm(depth ~ temp, data = wpdata)  
> summary(wood\_lm)  
  
Call:  
lm(formula = depth ~ temp, data = wpdata)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.8066 -1.3321 -0.6529 0.6811 4.8512   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 20.12228 0.94024 21.401 1.11e-09 \*\*\*  
temp -0.34218 0.05961 -5.741 0.000188 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2.335 on 10 degrees of freedom  
Multiple R-squared: 0.7672, Adjusted R-squared: 0.7439   
F-statistic: 32.96 on 1 and 10 DF, p-value: 0.0001875  
> vcov(wood\_lm)  
 (Intercept) temp  
(Intercept) 0.88406062 -0.039081046  
temp -0.03908105 0.003552822  
> est <- 20.12228 - 8\*0.34218  
> se\_est <- sqrt(0.88406062 + (-1)^2\*0.03908105)  
> glance(wood\_lm)$df.residual   
[1] 10  
> tstar <- qt(.975, df = 10)  
> est - tstar\*se\_est   
[1] 15.24404  
> est + tstar\*se\_est   
[1] 19.52564

## Problem 6: Gender stereotypes

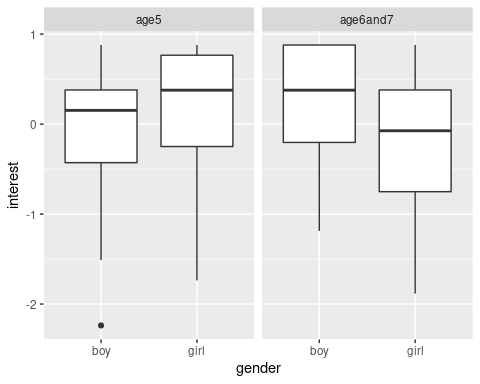
A 2017 *Science* article studied gender stereotypes around intellect in 5-7 year old boys and girls ([link](http://science.sciencemag.org/content/355/6323/389)). For this problem we will look at some of the data collected for studies 3 and 4, contained in the data file SteroTypeStudy34.csv. For both these studies, researchers measured subject interest in either a game for “smart” people or a game for people who “try-hard.” The response interest is a standardized measure of subject interest in the “smart” games, so higher interest means more interest expressed in that type of game than other study subjects. The variable gender records gender (boy/girl) and age2 groups subjects into 5-year-old (age5) participants and 6- to 7-year-old (age6and7) participants. This mimics the use of age in some of the paper’s results.

> stereo <- read.csv("http://math.carleton.edu/kstclair/data/StereoTypeStudy34.csv")

### (6a)

Create an EDA graph that explores whether the effect of gender on interest level is affected by the age of the subject (measured by age2 *not* by age). Explain what your graph shows you.

> stereo <- read.csv("http://math.carleton.edu/kstclair/data/StereoTypeStudy34.csv")  
> ggplot(stereo, aes(x=gender, y=interest)) +   
+ geom\_boxplot() +   
+ facet\_wrap(~age2)

 The graph shows that while at age 5 girls show higher interest than boys at age 5, boys show more interest thatn girls in games from age 5 to 6

### (6b)

Determine the appropriate regression model to model interest as a function of gender and age2 that allows the effect of gender on interest to differ by age2 group. Write down your model mean function in terms of parameters first (like was done at the start of problem 4), then fit the model in R. Then, holding age constant, write down the estimated mean interest level for girls and for boys.

> stereo <- read.csv("http://math.carleton.edu/kstclair/data/StereoTypeStudy34.csv")  
> stereo\_lm<- lm(interest ~ gender \* age2, data = stereo)  
> summary(stereo\_lm)  
  
Call:  
lm(formula = interest ~ gender \* age2, data = stereo)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.1334 -0.4280 0.1794 0.6485 1.0997   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.1044 0.1725 -0.605 0.5465   
gendergirl 0.1334 0.2331 0.572 0.5683   
age2age6and7 0.3270 0.2087 1.567 0.1199   
gendergirl:age2age6and7 -0.5769 0.2888 -1.997 0.0481 \*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.752 on 116 degrees of freedom  
Multiple R-squared: 0.05831, Adjusted R-squared: 0.03396   
F-statistic: 2.394 on 3 and 116 DF, p-value: 0.07192

### (6c) Use your model in (b) to determine if interest level in “smart” games differs by gender for the 5-year old age group. Give a p-value to support your conclusion and, if statistically significant, give and interpret a confidence interval for the effect of gender in this age group.

According to the lm() above, we get that the term is of p-value 0.5683. Since this is significantly larger than the 0.05 threshold, we reject the alternative hypothesis and keep the null hypothesis. As such, we an conclude that there exists no difference in interest levels in “smart” games between genders for the 5-year old age group.

### (6d)

Repeat (c) but for the 6 to 7-year old age group.

$$
\mu(interest|age = 6and7, gender = girl) - \mu(interest|age = 6and7, gender = girl) = \beta\_0 + \beta\_1 gendergirl + \beta\_2 age2age6and7 + \beta\_3 gendergirl \* age2age6and7 - (\beta\_0 + \beta age2age6and7) = \beta\_1 gendergirl - \bet\_3 gendergirl \* age2age6and7
Var1 <- (-1)^2 \* 0.0543549)
Var2 <- (-1)^2 \* 0.08343398
Cov <- 2\*(-1)\*(-1)\* -0.05435493
sqrt(Var1 + Var2 + Cov)
$$

> summary(stereo\_lm)  
  
Call:  
lm(formula = interest ~ gender \* age2, data = stereo)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-2.1334 -0.4280 0.1794 0.6485 1.0997   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -0.1044 0.1725 -0.605 0.5465   
gendergirl 0.1334 0.2331 0.572 0.5683   
age2age6and7 0.3270 0.2087 1.567 0.1199   
gendergirl:age2age6and7 -0.5769 0.2888 -1.997 0.0481 \*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.752 on 116 degrees of freedom  
Multiple R-squared: 0.05831, Adjusted R-squared: 0.03396   
F-statistic: 2.394 on 3 and 116 DF, p-value: 0.07192  
> vcov(stereo\_lm)  
 (Intercept) gendergirl age2age6and7  
(Intercept) 0.0297658 -0.02976580 -0.0297658  
gendergirl -0.0297658 0.05435493 0.0297658  
age2age6and7 -0.0297658 0.02976580 0.0435597  
gendergirl:age2age6and7 0.0297658 -0.05435493 -0.0435597  
 gendergirl:age2age6and7  
(Intercept) 0.02976580  
gendergirl -0.05435493  
age2age6and7 -0.04355970  
gendergirl:age2age6and7 0.08343398  
> Var1 <- (-1)^2 \* 0.0543549  
> Var2 <- (-1)^2 \* 0.08343398  
> Cov <- 2\*(-1)\*(-1)\* -0.05435493  
> sqrt(Var1 + Var2 + Cov)  
[1] 0.1705257  
> tstar <- (-0.1334 - (-0.5769))/(sqrt(Var1 + Var2 + Cov))  
> tstar  
[1] 2.600781  
> 2\*(1-pt(tstar, df = 116))  
[1] 0.0105117

According to the calculations above, we get a p-value of 0.0105. Since this is significantly less than 0.05, we can reject the null hypothesis and accept the alternative. As such, we conclude that there is a difference in interest between genders from the ages 6 to 7.