Homework 5

Shun-Lung Chang, Dilip Hiremath

```r
# import packages
library(magrittr)
library(dplyr)
library(ggplot2)
library(car)
```

1. First of all, load the data frame Wage from the library ISLR. You start out with a close look at wage differences between the two health levels.

```r
data(Wage, package = "ISLR")
```

(a) (1.5 points) Compute mean and standard deviation of wage for each health level separately. Summarize the result in an English sentence.

```r
wage_stats <- Wage %>%
    group_by(health) %>%
    summarise(mean_wage = mean(wage),
              sd_wage = sd(wage),
              counts = n())
wage_stats
```

```
# A tibble: 2 x 4
          health mean_wage  sd_wage counts
          <fctr>     <dbl>    <dbl>  <int>
1      1. <=Good  101.6613 35.18500    858
2 2. >=Very Good  115.7262 43.43896   2142
```

There are 858 observation for good health or below workers and their average wage is 101.6613 with the standard deviation of 35.18500. There are 2142 observation for very good health or above workers and their average wage is 115.7262 with the standard deviation of 43.43896.

(b) (1 point) Compute the standard errors for the mean wages in the two groups.

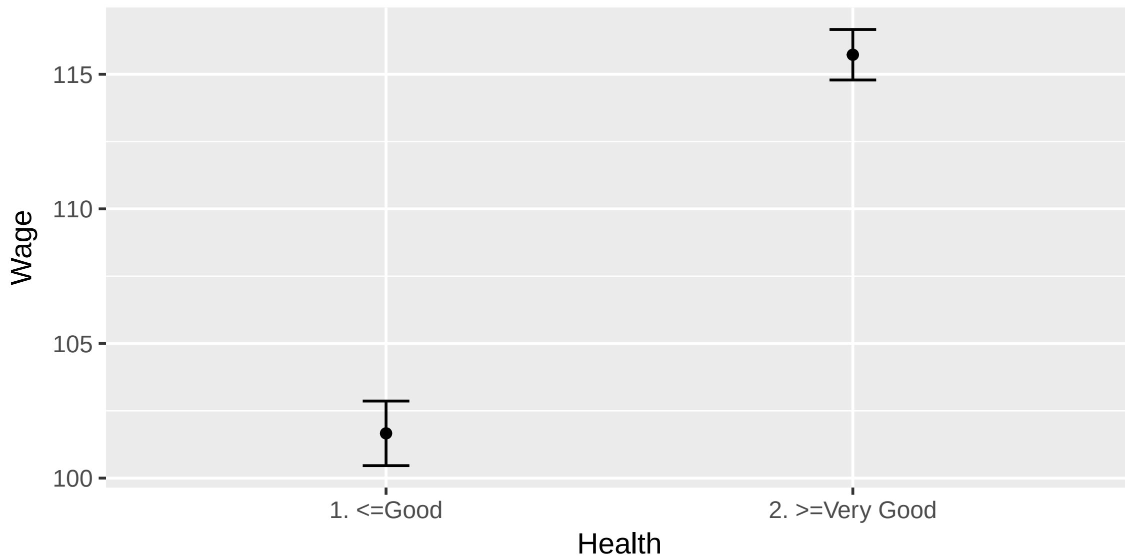
```r
wage_stats$se_wage <- wage_stats$sd_wage / sqrt(wage_stats$counts)
wage_stats
```

```
# A tibble: 2 x 5
          health mean_wage  sd_wage counts   se_wage
          <fctr>     <dbl>    <dbl>  <int>     <dbl>
1      1. <=Good  101.6613 35.18500    858 1.2011960
2 2. >=Very Good  115.7262 43.43896   2142 0.9385766
```

The standard errors for the mean wage for workers with good is 1.2011960 and workers with very good health is 0.9385766.

2. (2.5 points) Create a plot showing the mean wages for the two groups and corresponding error bars, i.e. add lines of length one standard error of the mean to both sides of the mean.

```r
ggplot(wage_stats, aes(x = health, y = mean_wage, group = 1)) +
    geom_errorbar(aes(x = health, ymin = mean_wage - se_wage, ymax = mean_wage + se_wage),
                  width = 0.1) +
    geom_point() +
    labs(x = 'Health' , y = 'Wage')
```



The plot is shown above.

3. (2.5 points) Using an appropriate statistical procedure, test whether average wage is the same for workers with health level "1. at most Good" and workers with health level "2. at least Very Good". Formulate the null and alternative hypothesis and report the results in an English sentence refering to the relevant numbers.

$H_0 : Average\ wage_{level_1} = Average\ wage_{level_2}$

$H_1 : Average\ wage_{level_1} \neq Average\ wage_{level_2}$

```r
# Assume that population variances of the two classes are not equal
t.test(Wage$wage ~ Wage$health, var.equal = FALSE)
```

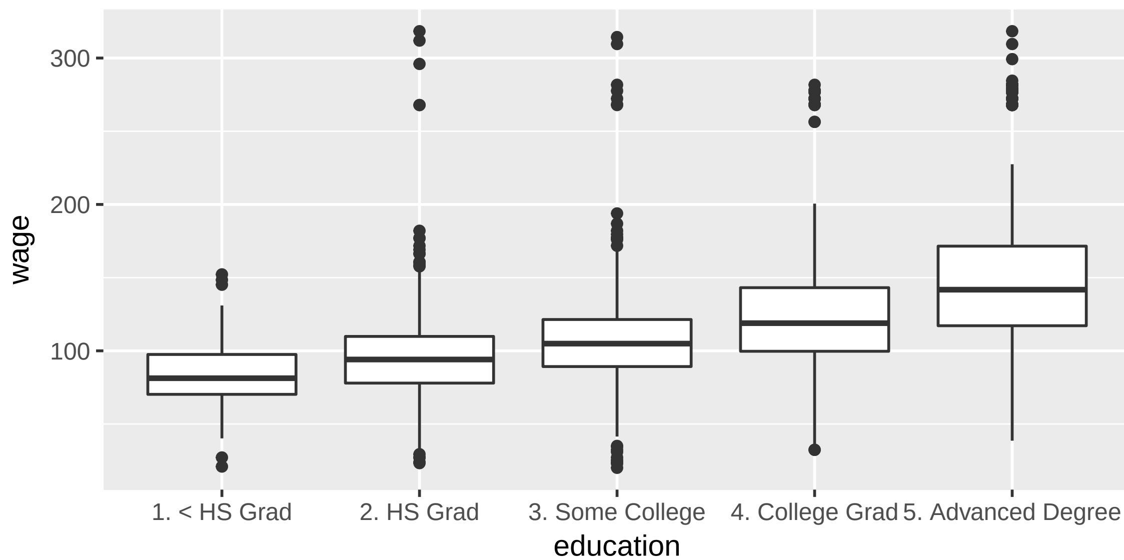
```

    Welch Two Sample t-test

data:  Wage$wage by Wage$health
t = -9.2265, df = 1934.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.05452 -11.07524
sample estimates:
     mean in group 1. <=Good mean in group 2. >=Very Good
```

```
                    101.6613                      115.7262
```

The null hypothesis is that the average wage for level 1 at most Good is equal to average income of workers with health level 2 at least Very Good. The alternative hypothesis states that the average wage for level 1 at most Good is not the same as average income of workers with health level 2 at least Very Good. Based on the test results, we can reject the null hypothesis in favour of the alternative hypothesis. The difference in mean wages between level 1 at most Good (101.6613) and average income of workers with health level 2 (115.7262) is significant in a two-sided, Welch t-test (t = -9.2265, df = 1934.3, p-value < 2.2e-16).

4. Plot a box plot of the workers raw wage (variable wage using the education level (variable education) as grouping.

```r
ggplot(Wage) +
    geom_boxplot(aes(x = education, y = wage))
```



(a) (half a point) Are half of the wages for workers who have less than a high school degree below the first quartile of the wage for workers with some college degree?

No.

(b) (half a point) Do half of the workers with a HS degree have higher wages than three quarters of the high school dropouts?

No.

(c) (half a point) The minimum wage of workers with advanced degree is larger than the median wage of high school dropouts?

No. The minimum wage of workers with advanced degree is 38.6059145, while the median wage of high school dropouts is 81.2832533.

(d) (half a point) The interquartile range differs substantially between all groups.

No, at least the interquartiles in levels 1, 2 and 3 are quite the same. And the interquartiles in level 4 and 5 are slightly wider.

(e) (half a point) Spread as measured by the length of the whiskers differs substantially between all groups.

No, the length of the whiskers are similar.

5. You want to assess the wage difference between educational groups. Before you run the appropriate statistical test, you check some of the assumptions for ANOVA. In particular, you assess homoscedasticity.

(a) (1.5 points) Looking at the boxplot in Question 4. Does homoscedasticity hold for the five groups? Give reasons for your answer!

No, by looking at group 1 and 2, we can see that the data points in group 1 lie closer to the box, but there exist a number of outliers in group 2. Thus we cannot safely conclude that the homoscedasticity holds.

(b) (1 point) Select a suitable variance test to check on this. Does the test confirm homoscedasticity?

```
leveneTest(Wage$wage, Wage$education)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    4  50.021 < 2.2e-16
      2995
```

No, this result does not confirm homoscedasticity, which is in line with the observation from the box plot. We observe that the p-value is very small and hence we can reject the null hypothesis in favour of the alternative.

6. Now, you assess the wage difference between educational groups using a statistical test.

(a) (1 point) Using an appropriate statistical test check whether wages are equal across education groups. Report the result in a complete English sentence including the relevant numbers!

```
wage_education <- aov(wage ~ education, data = Wage)
anova(wage_education)
```

```
Analysis of Variance Table

Response: wage
            Df  Sum Sq Mean Sq F value    Pr(>F)
education    4 1226364  306591  229.81 < 2.2e-16
Residuals 2995 3995721    1334
```

There is a highly significant difference in wages across educational groups as given by the ANOVA test with a test-statistic of F = 229.81 with 4 numerator and 2995 denominator degrees of freedom yielding a p-value of $p < 2.2e\text{-}16$.

(b) (1 point) From the ANOVA table derive the total sum of squares for wages and compare this result with the variance of wage when multiplied by 2999.

```r
sum(anova(wage_education)[, 2])
```

```
[1] 5222086
```

```r
var(Wage$wage) * 2999
```

```
[1] 5222086
```

They are equal as shown in the above calculations.

(c) (half a point) Which proportion of total variation in wages is due to the group differences in education?

```r
anova(wage_education)[1, 2] / sum(anova(wage_education)[, 2])
```

```
[1] 0.2348419
```

0.2348419 is the proportion of total variation in wages due to the group differences in education.

7. Having found an overall difference, you now want to use a post-hoc test with Holm correction, to asses which marital status groups do actually differ significantly in wages?

```r
pairwise.t.test(Wage$wage, Wage$education, p.adjust.method = "holm")
```

```

    Pairwise comparisons using t tests with pooled SD

data:  Wage$wage and Wage$education

                   1. < HS Grad 2. HS Grad 3. Some College 4. College Grad
2. HS Grad         3.7e-06      -          -               -
3. Some College    < 2e-16      2.3e-10    -               -
4. College Grad    < 2e-16      < 2e-16    3.5e-16         -
5. Advanced Degree < 2e-16      < 2e-16    < 2e-16         < 2e-16

P value adjustment method: holm
```

```r
pairwise.t.test(Wage$wage, Wage$maritl, p.adjust.method = "holm")
```

```

    Pairwise comparisons using t tests with pooled SD

data:  Wage$wage and Wage$maritl

             1. Never Married 2. Married 3. Widowed 4. Divorced
```

```
2. Married   < 2e-16          -           -           -
3. Widowed   1.00             0.22        -           -
4. Divorced  0.01             0.000001    1.00        -
5. Separated 0.67             0.01        1.00        1.00

P value adjustment method: holm
```

(a) (2 points) According to the post hoc test which groups differ significantly?

For education groups, since all p-values are far less than 0.001, we can say that the wages differ significantly across all groups.

For Martial groups:

- never married - married
- never married - divorced
- married - divorced
- married - separated

(b) (half a point) According to the post hoc test which groups do not differ significantly?

As mentioned in (a), there is no group differing significantly.

For Martial groups:

- never married - widowed
- never married - separated
- married - widowed
- widowed - divorced
- widowed - separated
- divorced - separated

8. You now investigate the relationship between wage and the two predictors education and health status.

(a) (1 point) First, calculate a main effects model only. Give a verbal summary of the model result!

```r
wage_edu_heal <- aov(wage ~ education + health, data = Wage)
anova(wage_edu_heal)
```

```
Analysis of Variance Table

Response: wage
            Df  Sum Sq Mean Sq F value      Pr(>F)
education    4 1226364  306591 231.248   < 2.2e-16
health       1   26239   26239  19.791 0.000008956
Residuals 2994 3969483    1326
```

There is a highly significant difference in wages across education groups as given by the ANOVA test with a test-statistic of F = 231.248 with 4 numerator and 2995 denominator degrees of freedom yielding a p-value of < 2.2e-16.

There is also a significant difference in wages across health groups as given by the ANOVA test with a test-statistic of F = 19.791 with 1 numerator and 2995 denominator degrees of freedom yielding a p-value of 8.956e-06.

(b) (1 point) Second, calculate a model with interaction. Give a verbal summary of the model result!

```r
wage_edu_heal_inter <- aov(wage ~ education + health + education:health, data = Wage)
anova(wage_edu_heal_inter)
```

```
Analysis of Variance Table

Response: wage
                   Df  Sum Sq Mean Sq  F value      Pr(>F)
education           4 1226364  306591 231.3193   < 2.2e-16
health              1   26239   26239  19.7967 0.000008928
education:health    4    6530    1632   1.2316      0.2952
Residuals        2990 3962953    1325
```

The difference in wages caused by the interaction effect of health and educational groups is not significant as given by the ANOVA test with a test-statistic of F = 1.2316 with 4 numerator and 2995 denominator degrees of freedom yielding a p-value of 0.2952.

(c) (half a point) Using the TukeyHSD post-hoc tests, which education levels do actually differ significantly in wages?

```r
TukeyHSD(wage_edu_heal)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = wage ~ education + health, data = Wage)

$education
                                         diff       lwr      upr     p adj
2. HS Grad-1. < HS Grad              11.67894  4.821323 18.53655 0.0000343
3. Some College-1. < HS Grad         23.65115 16.436562 30.86574 0.0000000
4. College Grad-1. < HS Grad         40.32349 33.162914 47.48407 0.0000000
5. Advanced Degree-1. < HS Grad      66.81336 59.064785 74.56194 0.0000000
3. Some College-2. HS Grad           11.97221  6.935590 17.00884 0.0000000
4. College Grad-2. HS Grad           28.64456 23.685608 33.60351 0.0000000
5. Advanced Degree-2. HS Grad        55.13443 49.358811 60.91004 0.0000000
4. College Grad-3. Some College      16.67234 11.230411 22.11427 0.0000000
5. Advanced Degree-3. Some College 43.16221 36.966958 49.35746 0.0000000
5. Advanced Degree-4. College Grad 26.48987 20.357598 32.62214 0.0000000

$health
                               diff      lwr      upr     p adj
2. >=Very Good-1. <=Good 6.443038 3.558529 9.327548 0.0000123
```

As the results indicate, all pair of educations differ significantly in wages given the small p-values.