A4_Q1

Joohyeok

2023-11-08

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.2
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

```
df=read.table("concrete.txt", header=TRUE)

fit=lm(CompressiveStrength~Cement+Slag+FlyAsh+Water+SP+CoarseAggr+FineAggr,df)

summary(fit)
```

```
##
## Call:
## lm(formula = CompressiveStrength ~ Cement + Slag + FlyAsh + Water +
##     SP + CoarseAggr + FineAggr, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8411 -1.7063 -0.2831  1.2986  7.9424
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.78150   71.10128   1.966  0.05222 .
## Cement        0.06141    0.02282   2.691  0.00842 **
## Slag         -0.02971    0.03176  -0.935  0.35200
## FlyAsh        0.05053    0.02316   2.182  0.03159 *
## Water        -0.23270    0.07166  -3.247  0.00161 **
## SP            0.10315    0.13459   0.766  0.44532
## CoarseAggr   -0.05562    0.02744  -2.027  0.04546 *
## FineAggr     -0.03908    0.02882  -1.356  0.17833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.609 on 95 degrees of freedom
## Multiple R-squared:  0.8968, Adjusted R-squared:  0.8892
## F-statistic:   118 on 7 and 95 DF,  p-value: < 2.2e-16
```

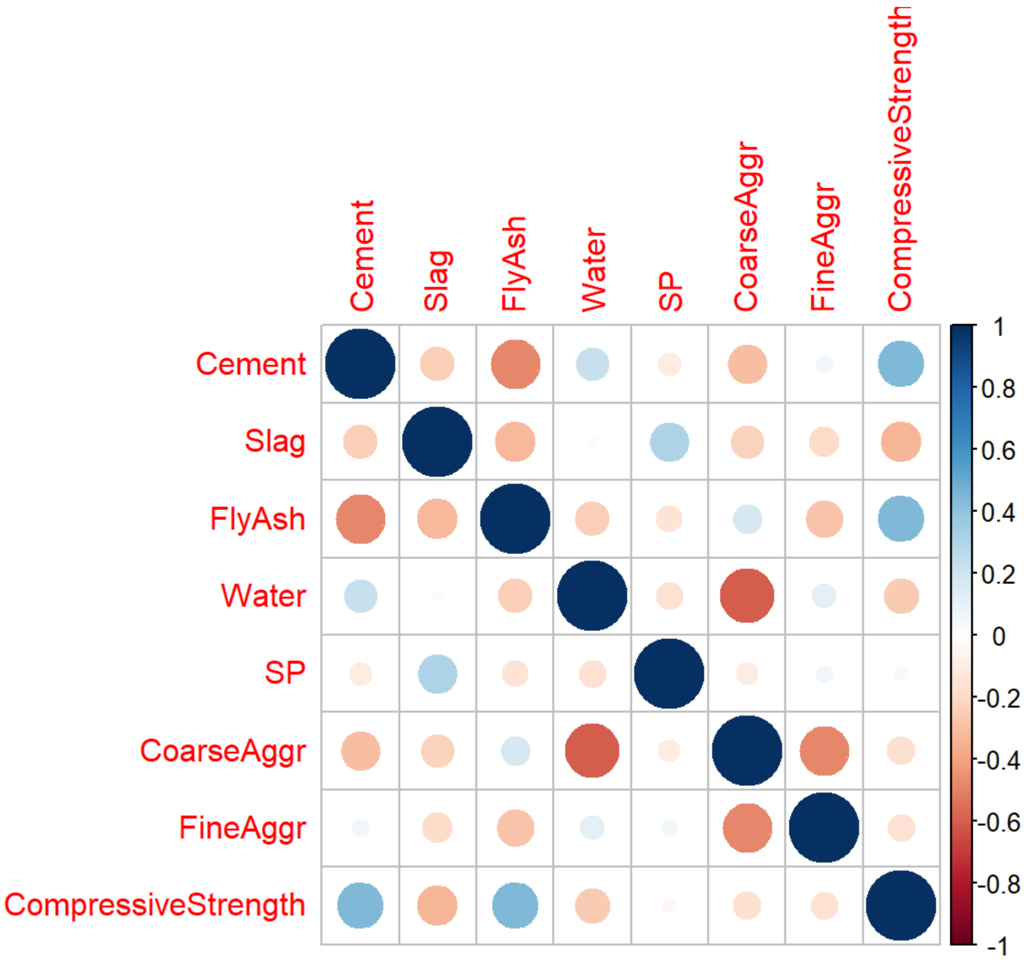
```
vif(fit)
```

```
##     Cement       Slag     FlyAsh      Water         SP CoarseAggr   FineAggr
##  48.570807  55.276977  58.649500  31.431899   2.139998  88.171895  49.961057
```

a. This shows severe multicollinearity because most VIFs are greater than 10.

b. In a data analysis, severe multicollinearity can cause instability in the parameter estimates, their standard errors, and p-values.

```
cc<-cor(df)
corrplot(cc)
```




```
fit1=lm(CompressiveStrength~Cement+Slag+FlyAsh+Water+SP+FineAggr,df)

vif(fit1)
```

```
##   Cement     Slag   FlyAsh    Water       SP FineAggr
## 1.886804 1.832936 2.318386 1.117704 1.158483 1.323080
```

c. The correlation between water and coarse aggregates is the highest. So, I suggest excluding coarse aggregates variable then we can avoid problems with multicollinearity because The VIFs are all less than 5.

Date/ Page/

Q2. a) Let Y_i and x_i be the depression score and BMI, respectively, of the i^{th} child.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

where $\epsilon_i \sim iid\, N(0, \sigma^2)$ and the ϵ_i's are independent.

b) No, In order to use this sort of interpretation, the other predictors in the model must be held fixed. But we can't simultaneously fix the value of BMI and increase the value of standardized BMI.

A4_Q2

Joohyeok

2023-11-08

```r
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```r
df=read.table("depression.txt",header=TRUE)
fit=lm(Depression~BMI+I(BMI^2),df)
result <- summary(fit)
```

c.

```r
result$coefficients
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept) 0.655284204 0.003088142 212.193651 0.000000e+00
## BMI         0.004676904 0.002566182   1.822514 6.856614e-02
## I(BMI^2)    0.013646925 0.001852625   7.366266 2.815475e-13
```

d. From the summary output, R^2 is 0.03649. The deterministic part of the model (the quadratic polynomial of BMI) explains 3.65% of the observed variability in depression score.

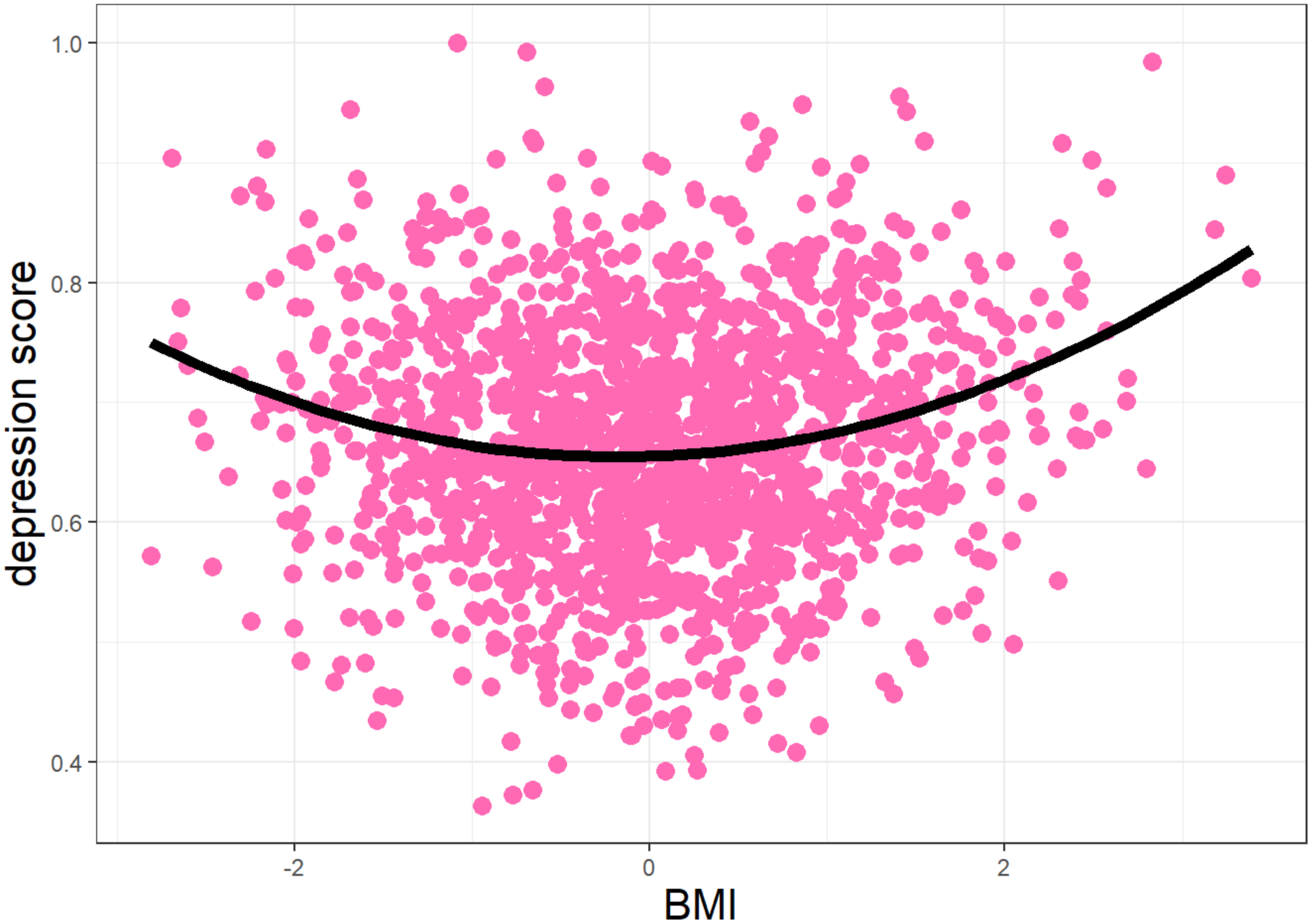
```r
result
```

```
##
## Call:
## lm(formula = Depression ~ BMI + I(BMI^2), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29909 -0.06602 -0.00357  0.06856  0.33370
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.655284   0.003088 212.194  < 2e-16 ***
## BMI         0.004677   0.002566   1.823   0.0686 .
## I(BMI^2)    0.013647   0.001853   7.366 2.82e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1001 on 1578 degrees of freedom
## Multiple R-squared:  0.03649,    Adjusted R-squared:  0.03527
## F-statistic: 29.88 on 2 and 1578 DF,  p-value: 1.832e-13
```

e.

```
ggplot(df,aes(x=BMI,y=Depression))+
  geom_point(size=3,colour="hotpink")+
  stat_smooth(method="lm",se=FALSE, formula=y~poly(x,2,raw=TRUE),
              colour="black",size=2)+
  labs(y="depression score", x="BMI")+
  theme_bw()+
  theme(axis.title=element_text(size=16))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



f. i. H0: beta2=0 vs. Ha: beta2 is not equal to 0

ii. The value of the test statistic(from the summary output) is t = 7.366

iii. p-value = 2.82e-13 (from the output)

iv. Since p-value<0.01, we reject H0. We have evidence at the 1% level that depression is more serious in children who are underweight or overweight.

Date/ Page/

Q3 a) Let Y_i be the selling price of a diamond of the i^{th} diamond

Let x_{1i} be the its weight of the i^{th} diamond

Let $x_{2i} = 1$ if the i^{th} diamond colour is I and 0 otherwise.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and the ϵ_i's are independent.

b) The effect of weight depends on whether the colour is H.

The main effect of weight in the model represents the estimated change in the selling price of a diamond for a one-carat increase in weight, holding the color constant.

CamScanner로 스캔하기

A4_Q3

Joohyeok

2023-11-15

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
df=read.table("diamond-red.txt",header=TRUE)

fit1=with(df,lm(Price~Weight*Colour))

result<-summary(fit1)
```

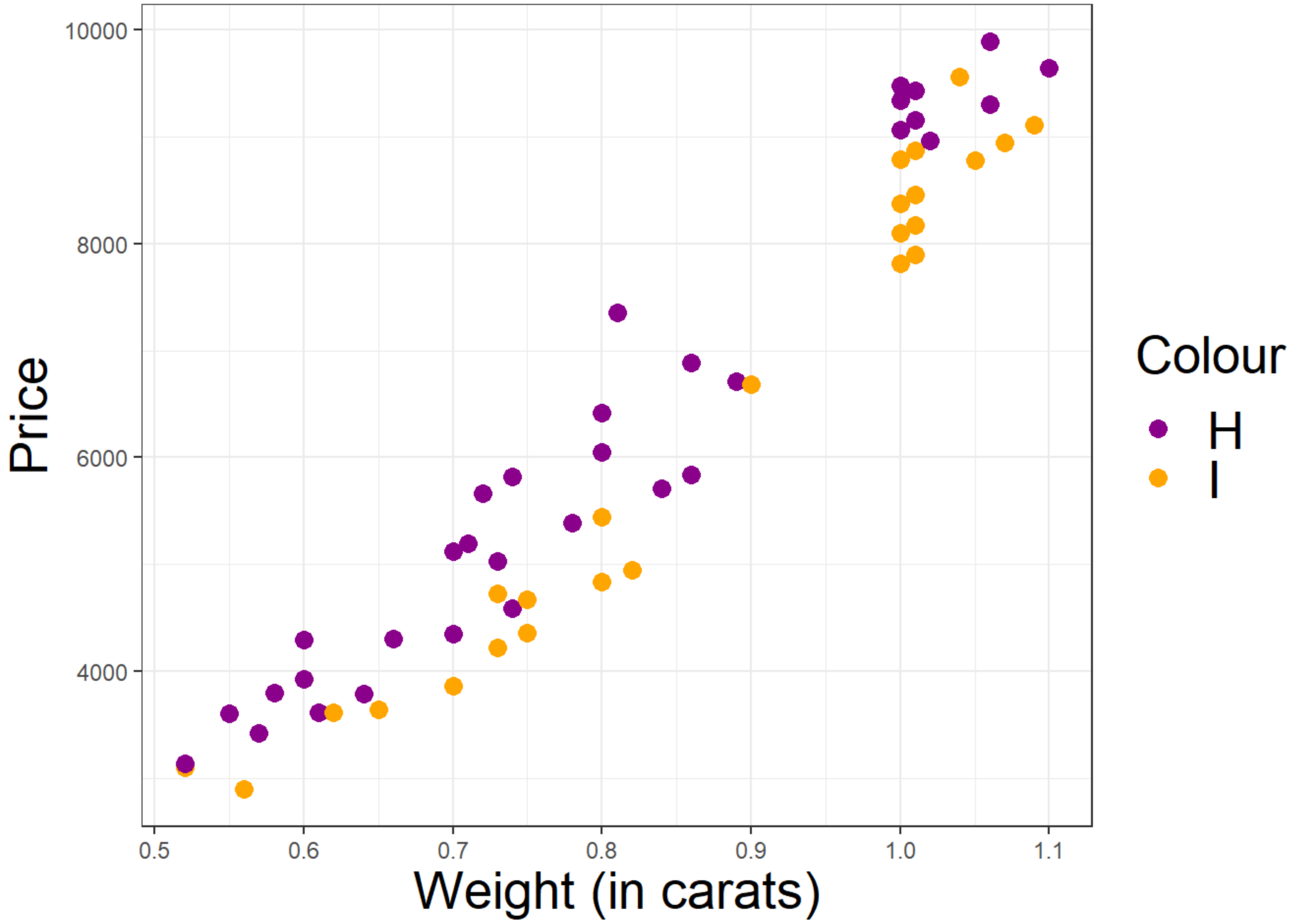
c.

```
result$coefficients
```

```
##                  Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)    -3824.4007   340.7803 -11.2224812 2.656630e-17
## Weight         12595.9012   415.7672  30.2955645 5.728360e-42
## ColourI         -831.6573   581.1933  -1.4309477 1.568945e-01
## Weight:ColourI   190.0849   668.3727   0.2843995 7.769438e-01
```

d.

```
ggplot(df,aes(y=Price,x=Weight,colour=Colour))+
  geom_point(size=3)+
  labs(x="Weight (in carats)", y="Price")+
  scale_colour_manual(values=c("darkmagenta","orange"))+
  theme_bw()+
  theme(axis.title=element_text(size=20),
        legend.text=element_text(size=20),
        legend.title=element_text(size=20))
```



e.

```
result
```

```
##
## Call:
## lm(formula = Price ~ Weight * Colour)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1173.07  -260.60    -0.03   306.46  1102.35
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3824.4      340.8 -11.222   <2e-16 ***
## Weight          12595.9      415.8  30.296   <2e-16 ***
## ColourI          -831.7      581.2  -1.431    0.157
## Weight:ColourI    190.1      668.4   0.284    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 468.8 on 70 degrees of freedom
## Multiple R-squared:  0.9561, Adjusted R-squared:  0.9543
## F-statistic: 508.6 on 3 and 70 DF,  p-value: < 2.2e-16
```

i. H0: beta2=0 vs Ha: beta2 is not equal to 0

ii. From the summary output, test statistic is -1.431

iii. t 1,70

iv. p-value = 0.157

v. Since p-value>0.05, we do not reject Ho. We have no evidence that the overall effect of colour is significant at the 5% level.

f.

```
predict(fit1,data.frame(Weight=1,Colour="I"))
```

```
##        1
## 8129.928
```

The estimated mean price of a 1 carat diamond of colour I is 8129.93

g. The discrepancy between the overall effect of color being significant and the main effect of color and the interaction term not being significant might be explained by the presence of an interaction effect in the model. This means that the effect of color on price depends on the weight of the diamond. When considering the main effect of color alone, it may not be significant because it does not account for this interaction. However, the overall effect test (which includes the interaction term) captures the combined effect of color and weight, making it significant.

Date/ Page/

Q4

a) Let Y_i be the spiciness of the i^{th} pepper

Let $x_{1i} = 1$ if the i^{th} pepper is an chipotle
and $x_{1i} = 0$ otherwise

Let $x_{2i} = 1$ if the i^{th} pepper is a Jalapeno
and $x_{2i} = 0$ otherwise

Let $x_{3i} = 1$ if the i^{th} pepper is a Serrano
and $x_{3i} = 0$ otherwise

$$\therefore Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and the ϵ_i's are independent

A4_Q4

Joohyeok

2023-11-15

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.3     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

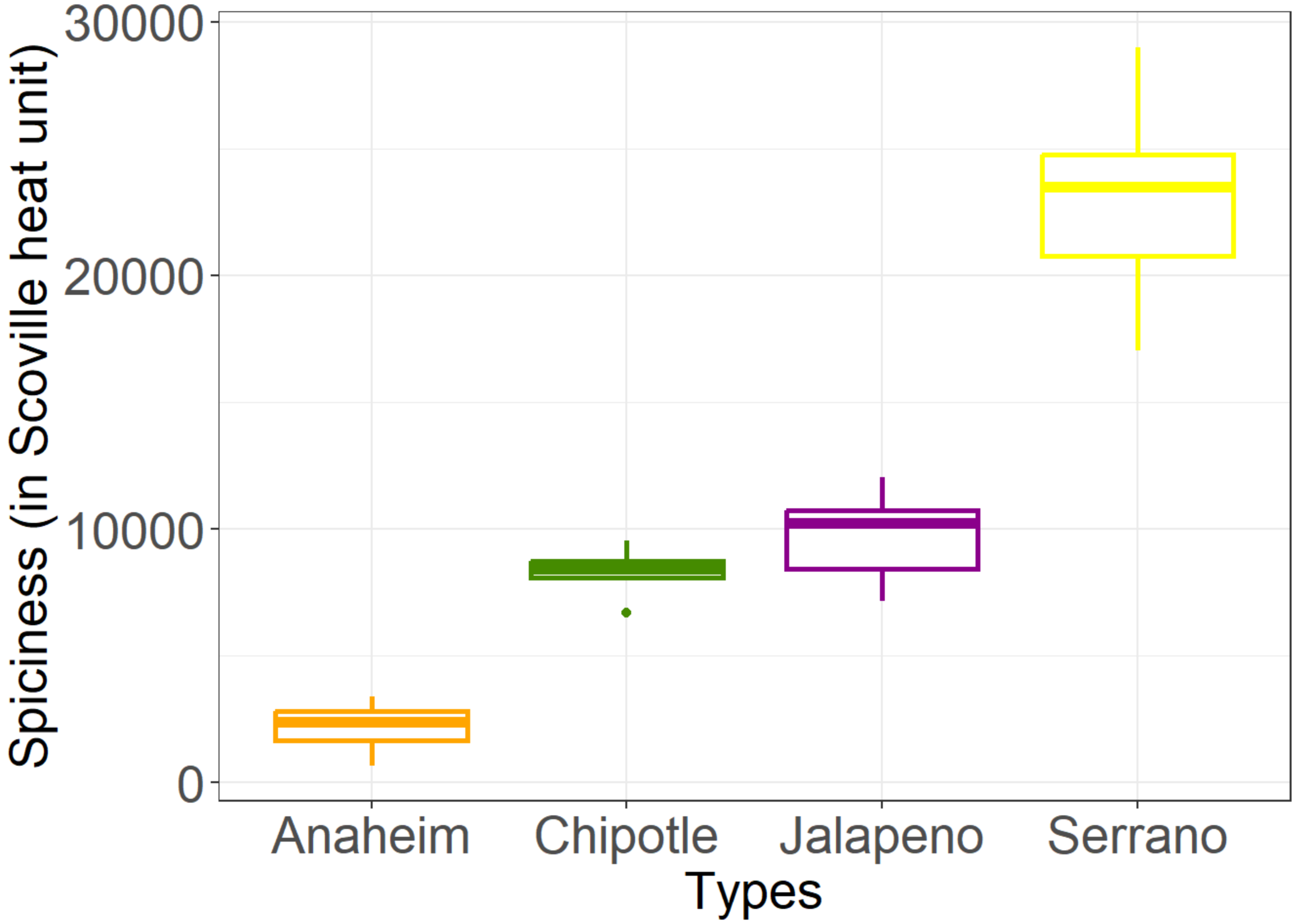
```
df=read.table("pepper.txt",header=TRUE)

df = df %>% mutate(Type=factor(Type))

contrasts(df$Type)
```

```
##          Chipotle Jalapeno Serrano
## Anaheim         0        0       0
## Chipotle        1        0       0
## Jalapeno        0        1       0
## Serrano         0        0       1
```

```
ggplot(df,aes(y=Spiciness,x=Type))+
  geom_boxplot(colour=c("orange","chartreuse4","darkmagenta","yellow"),size=1)+
  labs(y="Spiciness (in Scoville heat unit)", x="Types")+
  theme_bw()+
  theme(axis.title=element_text(size=20),axis.text=element_text(size=20))
```




```
fit=lm(Spiciness~Type,df)

predict(fit,data.frame(Type = 'Serrano'))
```

```
##        1
## 23120.69
```

b. The estimated mean spiciness of Serrano peppers is 23120.7

```
summary(fit)
```

```
##
## Call:
## lm(formula = Spiciness ~ Type, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6102  -1014    317    978   5868
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2246.5      697.0   3.223  0.00253 **
## TypeChipotle   6161.9     1012.7   6.084 3.60e-07 ***
## TypeJalapeno   7464.0      943.8   7.909 1.05e-09 ***
## TypeSerrano   20874.2      927.1  22.515  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2204 on 40 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.9287
## F-statistic: 187.6 on 3 and 40 DF,  p-value: < 2.2e-16
```

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Spiciness
##           Df     Sum Sq   Mean Sq F value    Pr(>F)
## Type       3 2734307433 911435811   187.6 < 2.2e-16 ***
## Residuals 40  194333652   4858341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. P-value is very low (p-value < 0.05). Since p-value < 0.05, we reject the null hypothesis. We have evidence at the 5% level that spiciness varies by type of pepper. Therefore, the types of peppers have a significant impact on spiciness.