HW6 Hints

COSC6323/Spring 2024

2024-03-01

Contents

Inspect Data —- 2
Linearity —- 2

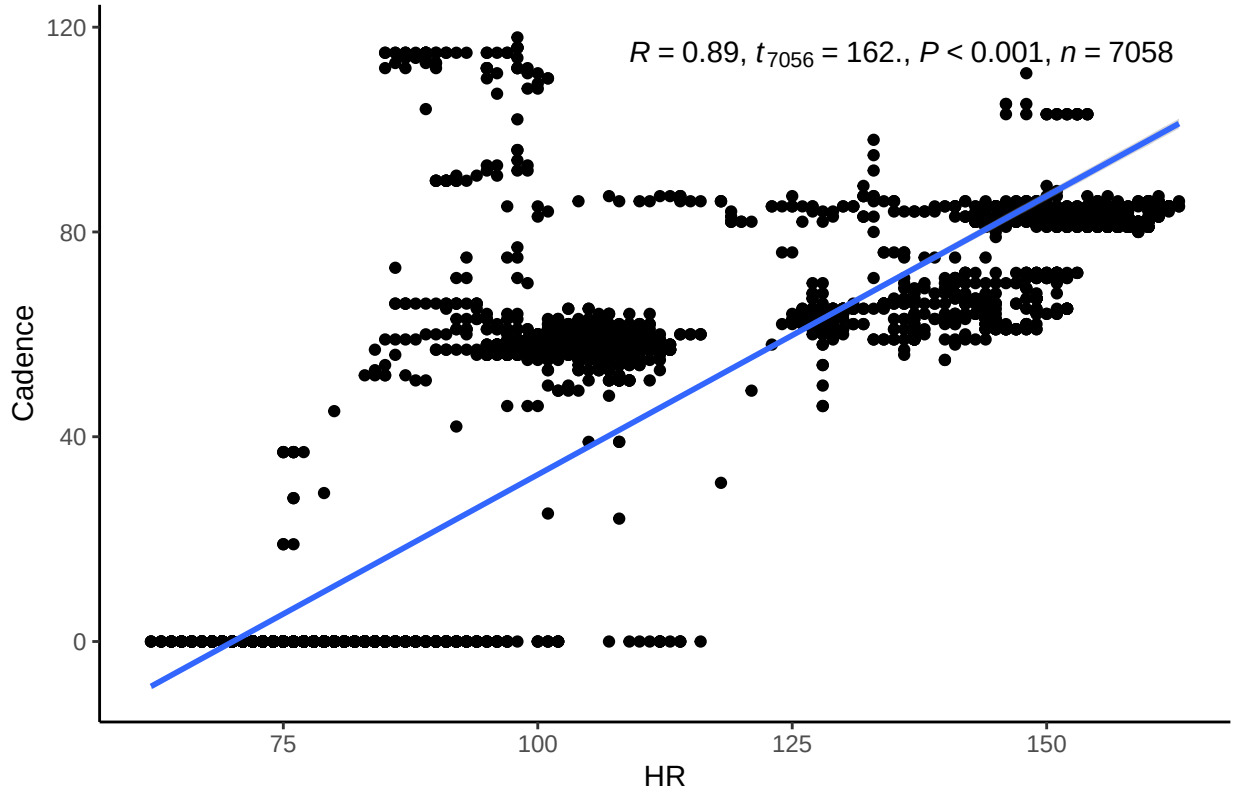
Normality Check (Row data): 2
Box Plots | QQplots —- 2
Shapiro Test —- 4
Anova Test 5

Regression diagnostics 5
Linear Regression —- 5
Fitted values and residuals —- 6
Regressions line scatter plot —- 7

Cooks distance, outliers and influence —- 8
Before and After Cook's Distance 9

Inspect Data —-

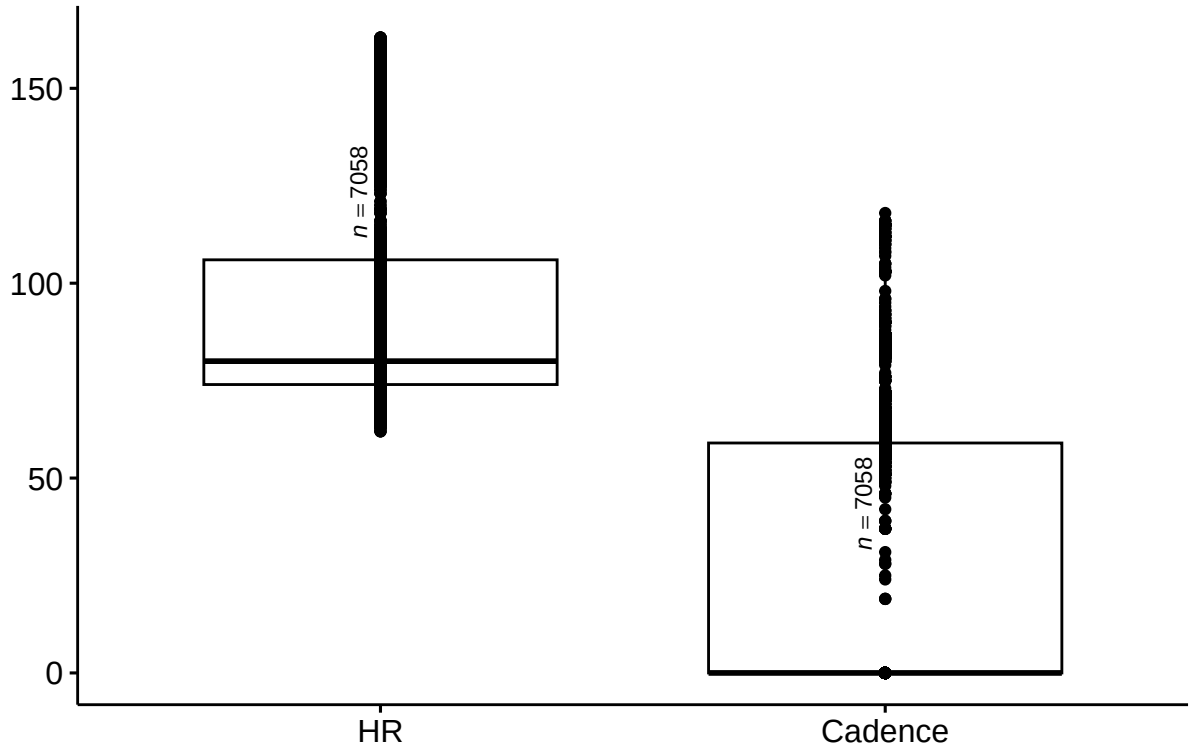
Linearity —-



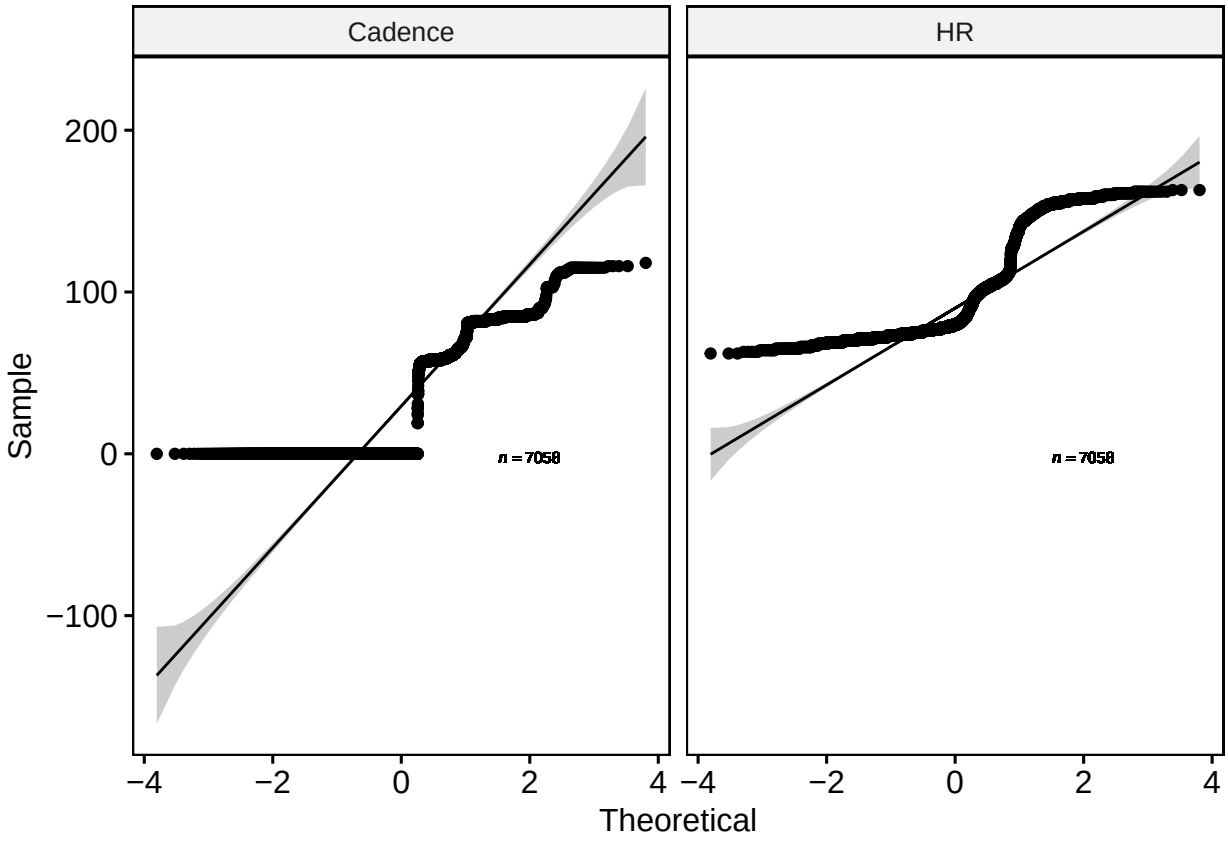
Normality Check (Row data):

Box Plots | QQplots —-

```
# Normality : Box Plots | QQplots ----

n_fun <- function(x) {
  return(data.frame(y = median(x) +30, label = paste0("~italic(n)", " == ", length(x))))
}

#Box Plots ----
all.df2 %>%
  select(HR, Cadence, ID) %>%
  gather(key = "variable", value = "value", -ID) %>%
  mutate(n= n()) %>%
  ggboxplot( x= "variable", y = "value", add = "point")+
  labs(title = "", x = "", y = "") +
  stat_summary( fun.data = n_fun, geom = "text", fun.y = median, angle =90,
                parse= T, size = 3, vjust = -0.75, hjust = 0,
                position = position_dodge(width = 0.75) )
```




```
## QQplots ----

all.df2 %>%
  select(HR, Cadence, ID) %>%
  #distinct() %>%
  mutate(n= n()) %>%
  gather(key = "variable", value = "value", -ID , -n) %>%
  ggqqplot("value", facet.by = "variable") +
  #ggqqplot("value", facet.by = "variable", color = "ID") +
  geom_text(aes(label = paste0("italic(n) ==  ", n)),
            x = 2, y = -2, parse = T, size = 2, bold = F)
```



Shapiro Test —-

Why do we do, bc we did not inspect our reference data in regards to normality.

```
## Shapiro Test ----

all.df2 %>%
  select(HR, Cadence, ID) %>%
  sample_n(5000) %>%
  shapiro_test(HR)
```

```
## # A tibble: 1 x 3
##   variable statistic        p
##   <chr>        <dbl>    <dbl>
## 1 HR           0.800 1.45e-61
```

```
all.df2 %>%
  select(HR, Cadence, ID) %>%
  sample_n(5000) %>%
  shapiro_test(Cadence)
```

```
## # A tibble: 1 x 3
##   variable statistic        p
##   <chr>        <dbl>    <dbl>
## 1 Cadence      0.718 2.74e-68
```

Anova Test

```
# Anova Test ----
result.aov<- all.df2 %>% anova_test(HR ~ Cadence)
result.aov
```

```
## ANOVA Table (type II tests)
##
##    Effect DFn  DFd        F p p<.05   ges
## 1 Cadence   1 7056 26115.48 0     * 0.787
```

Regression diagnostics

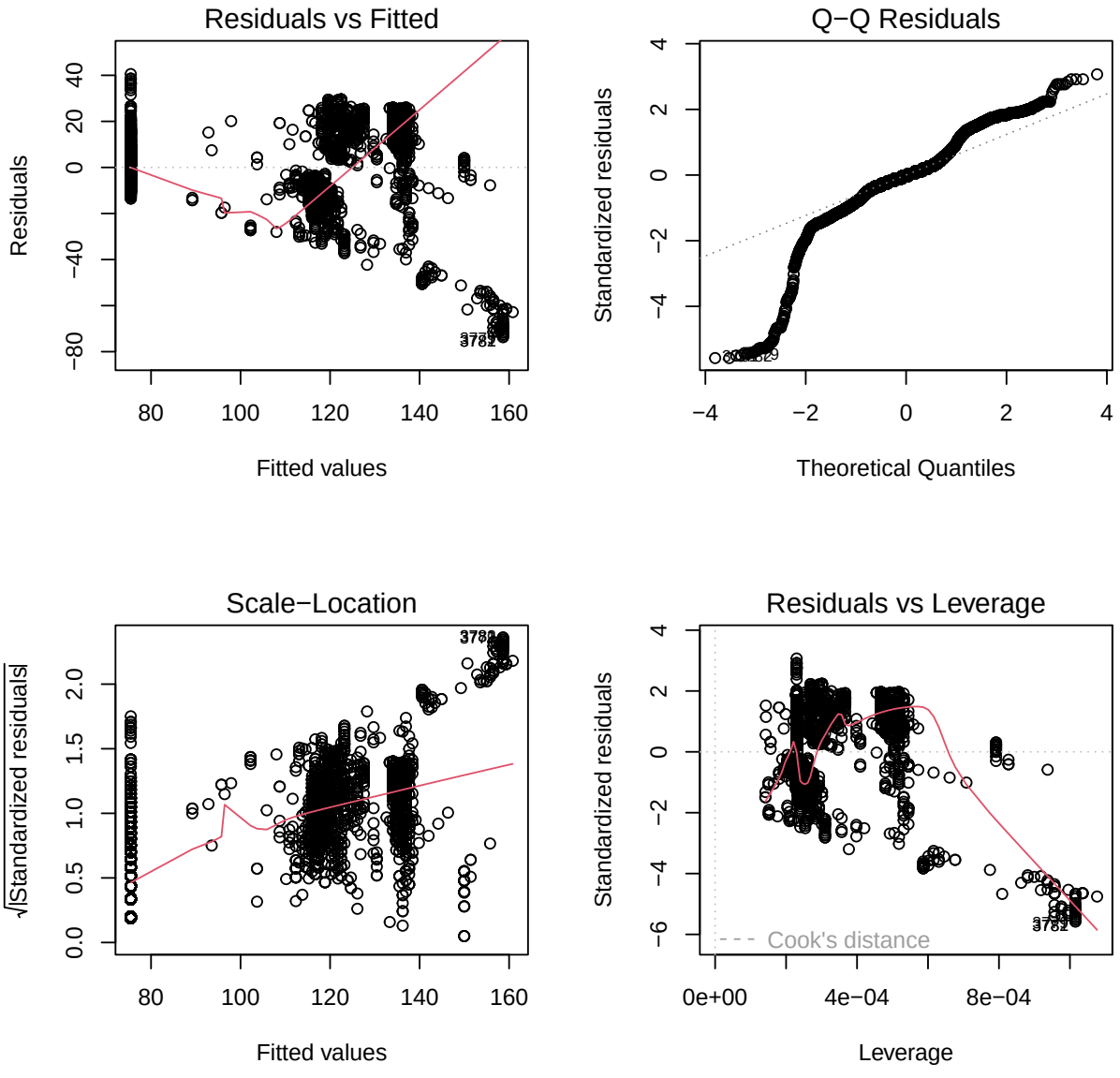
Linear Regression —-

```
# Linear Regression Analysis ----

# https://library.virginia.edu/data/articles/diagnostic-plots

hr_cad.lm <- lm(HR ~ Cadence, data = all.df2)
summary(hr_cad.lm)
```

```
##
## Call:
## lm(formula = HR ~ Cadence, data = all.df2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -73.65  -5.47  -0.47   5.53  40.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 75.470428   0.200321   376.7   <2e-16 ***
## Cadence      0.723271   0.004476   161.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.22 on 7056 degrees of freedom
## Multiple R-squared:  0.7873, Adjusted R-squared:  0.7873
## F-statistic: 2.612e+04 on 1 and 7056 DF,  p-value: < 2.2e-16
```

Fitted values and residuals —-

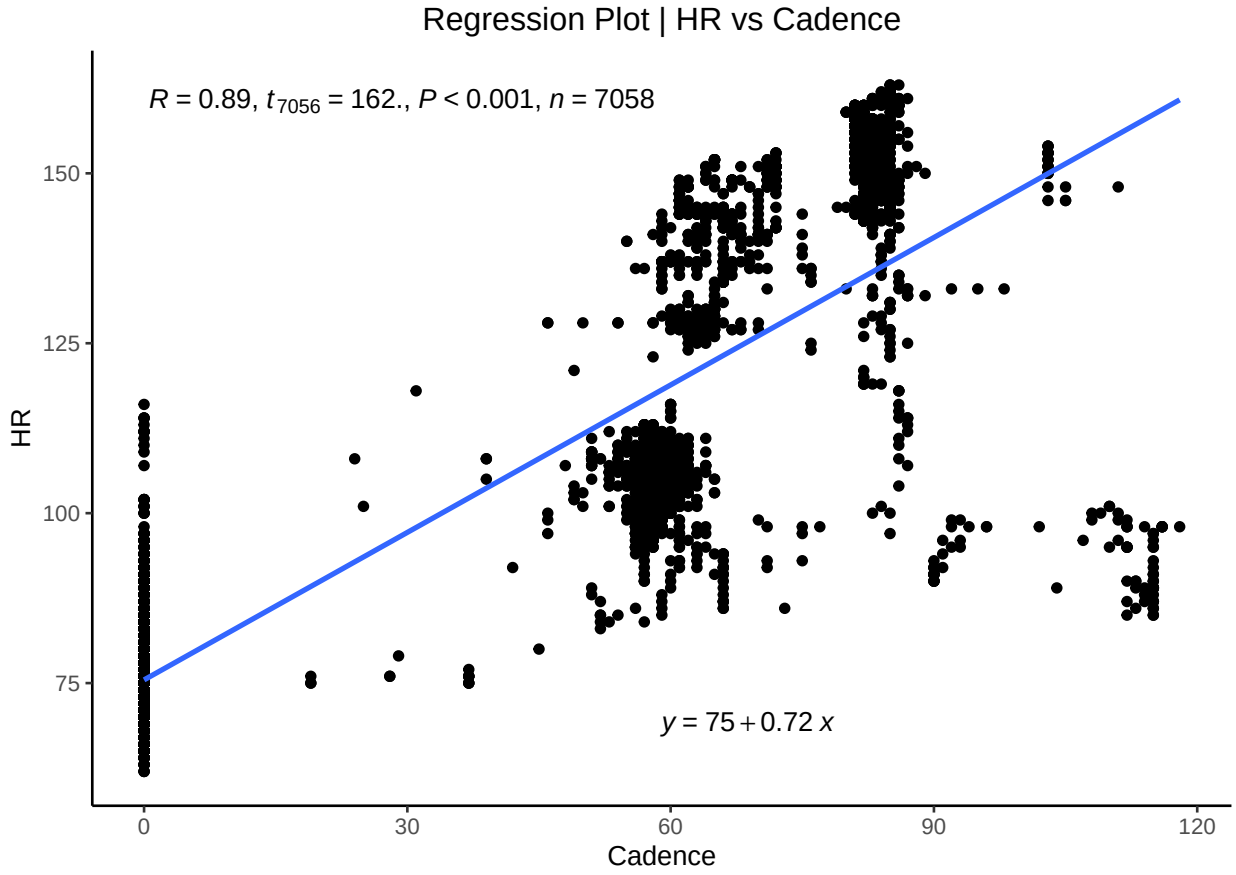
```
par(mfrow = c(2,2))
plot(hr_cad.lm)
```




```
par(mfrow = c(1,1))


# plot(hr_cad.lm, which = 1) # residuals_plot
# plot(hr_cad.lm, which = 2) # qqplot
# plot(hr_cad.lm, which = 3) # scale-location
# plot(hr_cad.lm, which = 4) # cook's distance
```

Regressions line scatter plot —-

```
all.df2 %>%
  ggplot(aes(x = Cadence, y = HR)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Regression Plot | HR vs Cadence",
       x = "Cadence",
       y = "HR") +
  stat_correlation(mapping = use_label(c("R", "t", "P", "n"))) +
  stat_regline_equation(label.y.npc = "bottom", label.x.npc = "center")
```



Cooks distance, outliers and influence —-

```
# Remove outlier with 95%
cooksD <- cooks.distance(hr_cad.lm)
cooksD.95 <- quantile(cooksD, prob = c(.95))

influential <- cooksD[(cooksD > cooksD.95)]
names_of_influential <- names(influential)
# influential
df_outlier <- all.df2[names_of_influential, ]

# Remove outliers
all.df3 <- all.df2 %>% anti_join(df_outlier)

# Update the index after filters
rownames(all.df2) <- 1:nrow(all.df2)

# Table of outliers
table(df_outlier$ID)
```

```
##
##      R001_biking     R001_driving R001_office_work     R001_running
##               20               20                0              231
##     R001_walking
##               69
```

```
## After Cooks Distance ----
signal.lm <- hr_cad.lm

aftrCD.plot <- all.df3 %>%
  ggplot(aes(x = HR, y = Cadence)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "",
       x = "HR",
       y = "Cadence") +
  stat_correlation(mapping = use_label(c("R", "t", "P", "n")),label.x = "right")
```

Before and After Cook's Distance