Standard Errors Bootstrapping

1.

Choose one of the multiple regression fits you performed in HW3 or HW4 where you observed violations of the standard assumptions, such as nonlinearity, multicollinearity, heteroscedasticity, etc. Report the standard errors of the model coefficients as automatically provided by the linear model fit under the standard assumptions.

For this question, we choose the linear model using combined continuous and discrete predictors from HW4, which violated the standard assumption of homoscedasticity. First, we load in the data:

```
hours_clean = read.csv('../data/hour_clean.csv')
hours = hours_clean[, c("yr", "mnth", "workingday", "weathersit", "atemp", "hum", "windspeed", "cnt")]
```

Then, we preprocess the data and fit the linear model:

```
# one-hot-encode and standardize
library(fastDummies)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# these are outliers that affect the scaling
hours$weathersit[hours$weathersit == 4] = 3

hours = dummy_cols(hours,
                   select_columns = c("mnth", "weathersit"),
                   remove_first_dummy = TRUE,
                   remove_selected_columns = TRUE)
hours = dplyr::select(hours, -cnt, cnt)
head(hours)
```

```
##   yr workingday  atemp  hum windspeed mnth_2 mnth_3 mnth_4 mnth_5 mnth_6 mnth_7
## 1  0          0 0.2879 0.81    0.0000      0      0      0      0      0      0
## 2  0          0 0.2727 0.80    0.0000      0      0      0      0      0      0
## 3  0          0 0.2727 0.80    0.0000      0      0      0      0      0      0
## 4  0          0 0.2879 0.75    0.0000      0      0      0      0      0      0
## 5  0          0 0.2879 0.75    0.0000      0      0      0      0      0      0
## 6  0          0 0.2576 0.75    0.0896      0      0      0      0      0      0
##   mnth_8 mnth_9 mnth_10 mnth_11 mnth_12 weathersit_2 weathersit_3 cnt
## 1      0      0       0       0       0            0            0  16
## 2      0      0       0       0       0            0            0  40
## 3      0      0       0       0       0            0            0  32
```

```
## 4      0      0       0       0       0            0            0  13
## 5      0      0       0       0       0            0            0   1
## 6      0      0       0       0       0            1            0   1
```

```
model = lm(cnt ~ ., data=hours)
summary(model)
```

```
##
## Call:
## lm(formula = cnt ~ ., data = hours)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -397.40  -99.38  -25.49   69.43  651.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    68.351      7.579   9.019  < 2e-16 ***
## yr             75.015      2.277  32.951  < 2e-16 ***
## workingday      4.610      2.444   1.887  0.05924 .
## atemp         543.034     12.878  42.168  < 2e-16 ***
## hum          -281.919      7.177 -39.281  < 2e-16 ***
## windspeed      57.875      9.955   5.813 6.23e-09 ***
## mnth_2        -17.052      5.759  -2.961  0.00307 **
## mnth_3        -12.044      5.852  -2.058  0.03961 *
## mnth_4        -20.593      6.258  -3.291  0.00100 **
## mnth_5        -12.788      7.069  -1.809  0.07049 .
## mnth_6        -65.889      7.589  -8.682  < 2e-16 ***
## mnth_7       -102.514      8.193 -12.512  < 2e-16 ***
## mnth_8        -57.434      7.773  -7.389 1.55e-13 ***
## mnth_9          5.655      7.253   0.780  0.43564
## mnth_10        38.356      6.461   5.937 2.96e-09 ***
## mnth_11        31.435      5.851   5.373 7.87e-08 ***
## mnth_12        29.561      5.703   5.184 2.20e-07 ***
## weathersit_2   12.525      2.767   4.526 6.04e-06 ***
## weathersit_3   -2.816      4.621  -0.609  0.54228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 148.6 on 17259 degrees of freedom
## Multiple R-squared:  0.3314, Adjusted R-squared:  0.3307
## F-statistic: 475.3 on 18 and 17259 DF,  p-value: < 2.2e-16
```

The summary output above provides the model coefficients and their standard errors (in the `Std. Error` column) as provided by the linear model fit under the standard assumptions. For example, the estimated coefficient for `atemp` is 543.034 with a standard error of 12.878. In HW4, we saw that this model observes heteroscedasticity, meaning the variance of residuals is not constant across fitted values. Due to this heteroscedasticity, the standard errors reported above may be biased and unreliable, which affects the accuracy of t-values and p-values for inference.

2.

Estimate the standard errors of the coefficients by bootstrapping. Decide and explain whether it is appropriate to resample cases or residuals. Compare to the classical estimates from Part (a).

Since we observed heteroscedasticity in the model, it is appropriate to resample cases, since resampling

residuals assumes homoscedasticity. Now, we use the `boot` R package to perform the bootstrapping 1000 times:

```
library(boot)

boot_fn = function(data, indices) {
  resample = data[indices, ]
  model = lm(cnt ~ ., data = resample)
  return(coef(model))
}

set.seed(123)
boot_results = boot(data = hours, statistic = boot_fn, R = 1000)
boot_results
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = hours, statistic = boot_fn, R = 1000)
##
##
## Bootstrap Statistics :
##         original       bias    std. error
## t1*    68.351289 -0.252154328    6.732752
## t2*    75.015380 -0.034468142    2.112617
## t3*     4.610080 -0.009779601    2.252051
## t4*   543.034471  0.500277566   13.197820
## t5*  -281.918884  0.139133608    7.438311
## t6*    57.874898  0.301391717    9.772045
## t7*   -17.052333 -0.141498046    4.066606
## t8*   -12.043622 -0.031272230    4.666873
## t9*   -20.593342 -0.193358873    5.456335
## t10*  -12.787709 -0.207795966    6.506753
## t11*  -65.888766 -0.265436376    7.706425
## t12* -102.513558 -0.360481963    8.013148
## t13*  -57.433997 -0.132859838    7.800117
## t14*    5.654579  0.002029682    7.261635
## t15*   38.356460  0.006347488    6.187349
## t16*   31.434722  0.068181926    4.946683
## t17*   29.561016 -0.056802253    4.416626
## t18*   12.524680  0.095484598    2.636512
## t19*   -2.816099 -0.060517138    4.324994
```

Here is a table and side-by-side barplot comparing the results of the bootstrap with the classical:

Coefficient	Classical SE	Bootstrap SE
(Intercept)	7.579	6.733
yr	2.277	2.113
workingday	2.444	2.252
atemp	12.878	13.198
hum	7.177	7.438
windspeed	9.955	9.772
mnth_2	5.759	4.067

Coefficient	Classical SE	Bootstrap SE
mnth_3	5.852	4.667
mnth_4	6.258	5.456
mnth_5	7.069	6.507
mnth_6	7.589	7.706
mnth_7	8.193	8.013
mnth_8	7.773	7.800
mnth_9	7.253	7.262
mnth_10	6.461	6.187
mnth_11	5.851	4.947
mnth_12	5.703	4.417
weathersit_2	2.767	2.637
weathersit_3	4.621	4.325

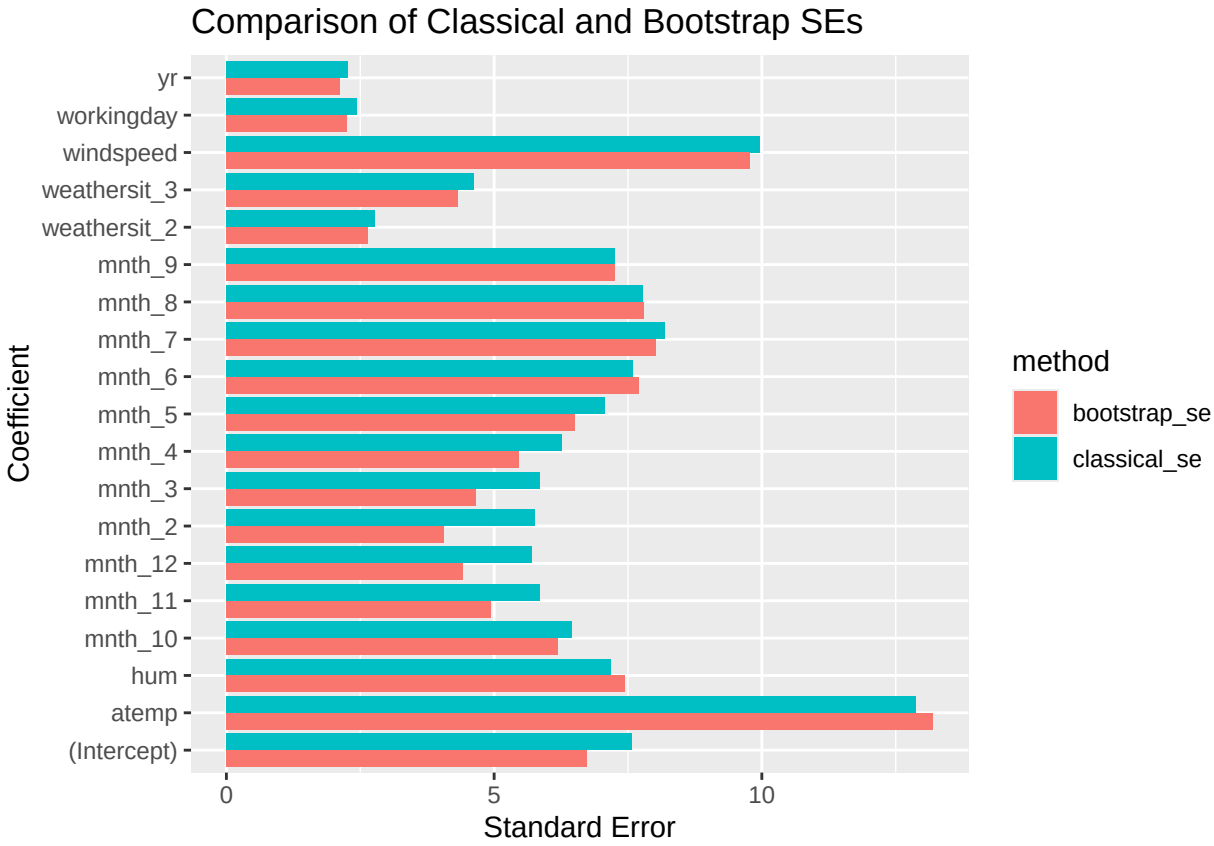
```r
library(ggplot2)
library(tidyr)
library(dplyr)

classic_se = summary(model)$coefficients[, 'Std. Error']

data_se = data.frame(
  coefficient = names(classic_se),
  classical_se = as.numeric(classic_se),
  bootstrap_se = apply(boot_results$t, 2, sd)
)

data_long <- data_se %>%
  gather(key = "method", value = "se", classical_se, bootstrap_se)

ggplot(data_long, aes(x = coefficient, y = se, fill = method)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +  # horizontal
  labs(x = "Coefficient", y = "Standard Error", title = "Comparison of Classical and Bootstrap SEs")
```



For each coefficient, the bootstrap standard error (computed as the standard deviation of 1000 bootstrap estimates) was broadly similar to the classical standard error reported by the linear model. This overall agreement suggests that the standard assumptions for a linear model may hold reasonably well across most predictors However, we do observe moderate differences in the standard errors for the following coefficients: `mnth_5`, `mnth_4`, `mnth_3`, `mnth_2`, `mnth_12`, `mnth_11`, and the intercept, where the bootstrap standard errors are consistently smaller than those obtained from the classical approach. These smaller bootstrap standard errors may suggest that the classical method is overestimating the uncertainty in these coefficient estimates.