Homework 3

IS 6489

Himanshi Sharma

September 21, 2017

Getting started

Question 1

```
n <- 1000
set.seed(197)
x <- rnorm(n, mean = 100, sd = 5)
y <- rnorm(n, mean = 100.5, sd =10)

#calculating mean
mean(x);mean(y)
```

```
## [1] 99.94635
```

```
## [1] 100.1929
```

```
mean(x) - mean(y)
```

```
## [1] -0.2465723
```

```
x[1]
```

```
## [1] 103.0459
```

```
#performing t-test
t <- data.frame(n = as.numeric(), p.value = as.numeric(), stringsAsFactors=FALSE)

for(i in 2:1000){
  t <- rbind(t,list(n = i , p.value =t.test(x[1:i], y[1:i], alternative = "two.sided", paired = F, var.
}

subset(t, p.value < 0.05)[1,]
```

```
##       n    p.value
## 128 129 0.04379248
```

Question 2

```
# value of t-statistics in t-test for n =132
round(t.test(x[1:129], y[1:129], alternative = "two.sided", paired = F, var.equal = F)$statistic, 2)
```

```
##     t
## -2.03
```

Question 3

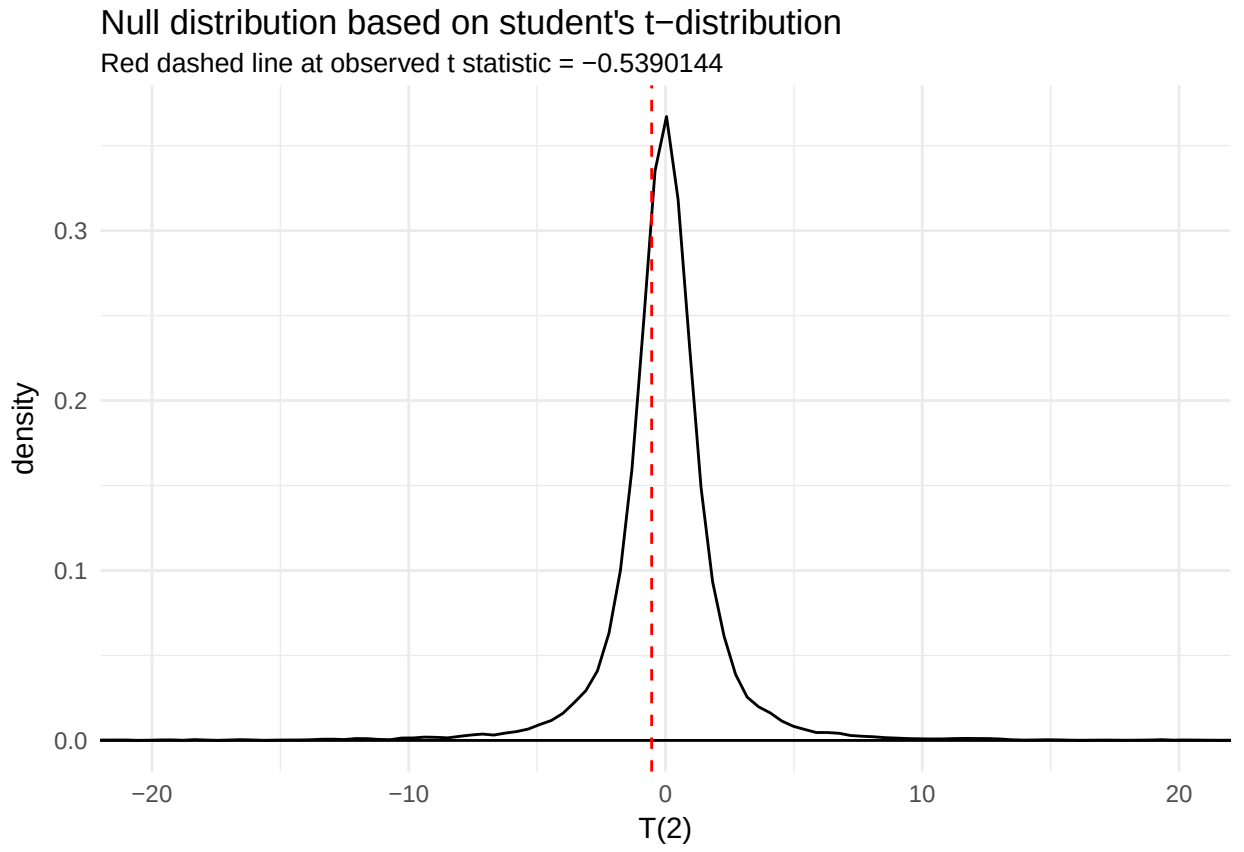
Looking at the t-test for sample x and y at n= 2 we get,

t-statistics = -0.53901 p-value = 0.6432 95% CI -43.54196 to 34.42516 The difference is not statistically significant as per the t-test. Because:

1) Looking at the t-statistics for this t-test if we plot a t-distribution of df = nx+ny - 2.

```
tt <- t.test(x[1:2], y[1:2], alternative = "two.sided", paired = F, var.equal = F)
n <- 2
t_null <- data.frame(t_dist = rt(10000, df = n+n-2))

ggplot(t_null, aes(t_dist)) +
  geom_density() +
  coord_cartesian(xlim = c(-20, 20))+
  geom_vline(xintercept = tt$statistic, col = "red", lty = 2) +
  labs(title = "Null distribution based on student's t-distribution",
       subtitle = "Red dashed line at observed t statistic = -0.5390144",
       x = "T(2)") +
  theme_minimal()
```



This graph shows the probability distribution function, meaning it shows the likelihood of t-value in t-distribution.

T-distribution assumes that the null hypothesis i.e. the difference between the means is 0, and any difference visible is by chance and is not because of the population difference. Our t-value is -0.53901 for this t-test falls close to 0, implying the difference like this happens mostly under the null hypothesis, i.e. the probability is quite high. Hence, the null hypothesis is not rejected and that is why the difference is statistically not

significant.

2) The p-value here represents the probability that the observed difference between the sample falls under the null hypothesis being true. Here the probability is 64.32% which is pretty high for us to reject null hypothesis and that is why the difference in mean is not statistically significant.

Question 4

```
max(x);max(y)
```

```
## [1] 116.4009
```

```
## [1] 129.0692
```

```
mean(x);mean(y)
```

```
## [1] 99.94635
```

```
## [1] 100.1929
```

```
sd(x);sd(y)
```

```
## [1] 5.11136
```

```
## [1] 9.773452
```

```
#95% CI for maximum x
xdf <-data.frame(x)
xboot_max <- NULL
set.seed(197)
for(i in 1:1000){
  boot_sample <- sample(xdf$x, replace=T)
  xboot_max[i] <- max(boot_sample)
}
round(quantile(xboot_max,probs = c(.025,.975)))
```

```
##  2.5% 97.5%
##   112   116
```

```
#95% CI for maximum y
ydf <-data.frame(y)
yboot_max <- NULL
set.seed(197)
for(i in 1:1000){
  boot_sample <- sample(ydf$y, replace=T)
  yboot_max[i] <- max(boot_sample)
}
round(quantile(yboot_max,probs = c(.025,.975)))
```

```
##  2.5% 97.5%
##   125   129
```

Question 5

Loading the Bikeshare date.

```
day <- read_csv("~/day.csv")
View(day)
```

Linear Regression

```
#modeling using lm function
model1 <- lm(cnt~factor(season)+yr+factor(mnth)+holiday+workingday+
               factor(weathersit)+temp+hum+windspeed, day)

#Function to calculate rmse
rmse <- function(fitted, actual){
  sqrt(mean((fitted - actual)^2))
}

#calculating rmse
round(rmse(fitted(model1), day$cnt),2)
```

```
## [1] 766.09
```

Question 6

```
summary(model1)
```

```
##
## Call:
## lm(formula = cnt ~ factor(season) + yr + factor(mnth) + holiday +
##     workingday + factor(weathersit) + temp + hum + windspeed,
##     data = day)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3976.3  -375.2    68.3   471.5  3133.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1805.96     229.97   7.853 1.51e-14 ***
## factor(season)2       892.93     181.54   4.919 1.08e-06 ***
## factor(season)3       841.89     215.56   3.906 0.000103 ***
## factor(season)4      1562.19     183.04   8.535  < 2e-16 ***
## yr                   2013.76      58.88  34.199  < 2e-16 ***
## factor(mnth)2         140.39     145.40   0.966 0.334593
## factor(mnth)3         545.14     167.35   3.257 0.001178 **
## factor(mnth)4         446.47     250.30   1.784 0.074899 .
## factor(mnth)5         712.46     270.49   2.634 0.008623 **
## factor(mnth)6         473.14     284.69   1.662 0.096963 .
## factor(mnth)7         -22.66     316.44  -0.072 0.942930
## factor(mnth)8         381.11     304.52   1.252 0.211160
## factor(mnth)9         979.90     267.42   3.664 0.000267 ***
## factor(mnth)10        533.29     244.27   2.183 0.029352 *
## factor(mnth)11        -85.12     233.30  -0.365 0.715331
## factor(mnth)12        -63.00     184.21  -0.342 0.732455
## holiday              -561.73     180.38  -3.114 0.001919 **
## workingday            124.76      64.55   1.933 0.053652 .
```

```
## factor(weathersit)2  -440.18      77.68  -5.667 2.12e-08 ***
## factor(weathersit)3 -1919.57     197.84  -9.702  < 2e-16 ***
## temp                 4544.49     415.71  10.932  < 2e-16 ***
## hum                 -1630.39     293.60  -5.553 3.97e-08 ***
## windspeed           -2946.38     410.57  -7.176 1.81e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 778.4 on 708 degrees of freedom
## Multiple R-squared:  0.8434, Adjusted R-squared:  0.8385
## F-statistic: 173.3 on 22 and 708 DF,  p-value: < 2.2e-16
```

```
model2 <- lm(cnt~factor(season)+yr+factor(mnth)+holiday+
               factor(weathersit)+temp+hum+windspeed, day)
summary(model2)
```

```
##
## Call:
## lm(formula = cnt ~ factor(season) + yr + factor(mnth) + holiday +
##     factor(weathersit) + temp + hum + windspeed, data = day)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3925.2  -374.6    85.4   487.7  3031.6
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1888.37     226.42   8.340 3.86e-16 ***
## factor(season)2       899.00     181.86   4.943 9.59e-07 ***
## factor(season)3       847.39     215.95   3.924 9.56e-05 ***
## factor(season)4      1566.51     183.38   8.542  < 2e-16 ***
## yr                   2011.99      58.99  34.108  < 2e-16 ***
## factor(mnth)2         139.38     145.68   0.957 0.339000
## factor(mnth)3         536.62     167.61   3.202 0.001428 **
## factor(mnth)4         423.88     250.51   1.692 0.091079 .
## factor(mnth)5         684.73     270.63   2.530 0.011617 *
## factor(mnth)6         437.43     284.63   1.537 0.124788
## factor(mnth)7         -67.18     316.21  -0.212 0.831813
## factor(mnth)8         347.52     304.61   1.141 0.254302
## factor(mnth)9         947.38     267.41   3.543 0.000422 ***
## factor(mnth)10        512.40     244.51   2.096 0.036468 *
## factor(mnth)11        -94.43     233.70  -0.404 0.686272
## factor(mnth)12        -71.92     184.50  -0.390 0.696817
## holiday              -649.39     174.92  -3.712 0.000221 ***
## factor(weathersit)2  -429.32      77.62  -5.531 4.48e-08 ***
## factor(weathersit)3 -1899.66     197.96  -9.596  < 2e-16 ***
## temp                 4624.37     414.45  11.158  < 2e-16 ***
## hum                 -1656.81     293.85  -5.638 2.48e-08 ***
## windspeed           -2969.32     411.19  -7.221 1.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 779.9 on 709 degrees of freedom
## Multiple R-squared:  0.8426, Adjusted R-squared:  0.8379
## F-statistic: 180.7 on 21 and 709 DF,  p-value: < 2.2e-16
```

```
rmse <- function(fitted, actual){
  sqrt(mean((fitted - actual)^2))
}
#rmse for model 1 (with working day)
rmse1 <- rmse(fitted(model1), day$cnt)
rmse1
```

```
## [1] 766.0938
```

```
#rmse for model 2 (without working day)
rmse2 <- rmse(fitted(model2), day$cnt)
rmse2
```

```
## [1] 768.1124
```

When we fit a linear model without working day, the model does not improve. If we look at the Root Mean Square error (rmse) for both the model we get :

rmse for model 1 (with working day): 766.0938

rmse for model 2 (without working day): 768.1124

It means that with the working day the model1 was off by 766.0938 riders and without working day the model2 was off by 768.1124 riders. There is no overall improvement in the model but a very slight decrease in the efficiency. Moreover, if we look at the R-Squared value there is a change of -0.0008, which is very small and further facilitates that removing of the predictor does not necessarily affect the model in a positive way.

Question 7

```
round(model2$coefficients[1], 2)
```

```
## (Intercept)
##     1888.37
```

```
round(model2$coefficients[5], 2)
```

```
##      yr
## 2011.99
```

Question 8

The intercept in the linear model is 1888.37. It represents the average number of bike ridership (cnt) in year 0: 2011 of season 1: Springer, month 1: January on a non-holiday with weather condition 1: Clear/ Few clouds/Partly cloudy/Partly cloudy at 0-degree Celsius temperature, 0 humidity, and 0 wind speed.

Question 9

The regression coefficient for yr (year) is 2011.99. It represents the predictive change with the increase of year from 0 to 1 while keeping other predictors constant.

Question 10

```
ridership <- round(model2$coefficients[1] + model2$coefficients[3] +
                     model2$coefficients[5] + model2$coefficients[11] +
                     model2$coefficients[20]*mean(subset(day, mnth == 7)$temp) +
                     model2$coefficients[21]*mean(subset(day, mnth == 7)$hum) +
                     model2$coefficients[22]*mean(subset(day, mnth == 7)$windspeed))
ridership
```

```
## (Intercept)
##        6691
```

Question 11

```
boot_predict <- NULL
n <- nrow(day)
set.seed(512)
for( i in 1:1000){
  boot <- sample(n, replace = T)
  model <- lm(lm(cnt~factor(season)+yr+factor(mnth)+holiday+
                   factor(weathersit)+temp+hum+windspeed, day[boot,]))
  prediction <- model$coefficients[1] + model$coefficients[3] + model$coefficients[5] +
    model$coefficients[11] + model$coefficients[20]*mean(subset(day, mnth == 7)$temp) +
    model$coefficients[21]*mean(subset(day, mnth == 7)$hum) +
    model$coefficients[22]*mean(subset(day, mnth == 7)$windspeed)
boot_predict[i] <- prediction
}
round(quantile(boot_predict, probs =c(0.025,0.975)))
```

```
##  2.5% 97.5%
##  6414  6943
```