MET CS555: Homework 6

Patrick Ryan



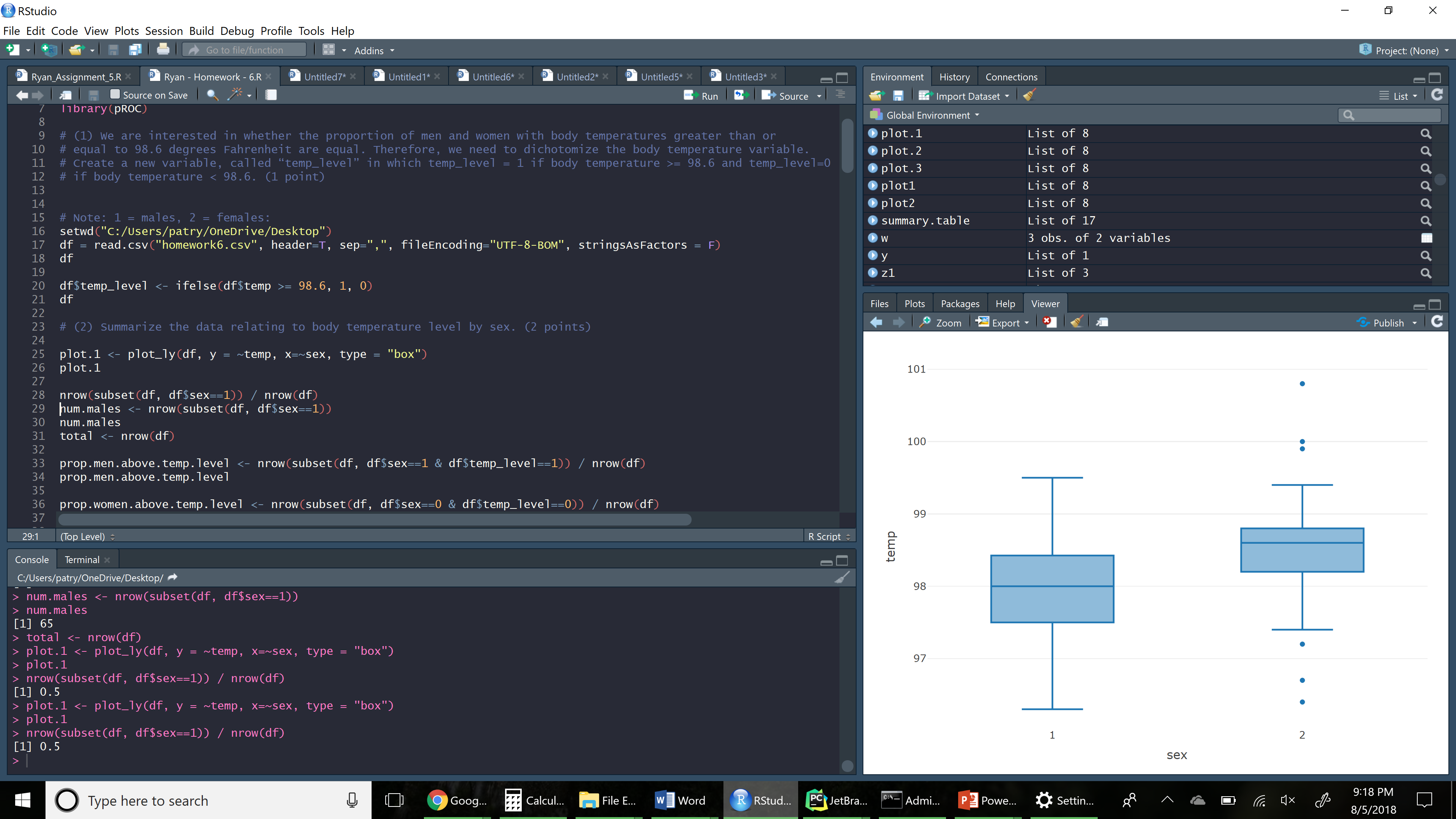
(1) We are interested in whether the proportion of men and women with body temperatures greater than or equal to 98.6 degrees Fahrenheit are equal. Therefore, we need to dichotomize the body temperature variable. Create a new variable, called “temp\_level” in which temp\_level = 1 if body temperature >= 98.6 and temp\_level=0 if body temperature < 98.6. (1 point)

|  |
| --- |
| # Note: 1 = males, 2 = females:  setwd("C:/Users/patry/OneDrive/Desktop")  df = read.csv("homework6.csv", header=T, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df  df$temp\_level <- ifelse(df$temp >= 98.6, 1, 0)  df |

(2) Summarize the data relating to body temperature level by sex. (2 points)

|  |
| --- |
| plot.1 <- plot\_ly(df, y = ~temp, x=~sex, type = "box")  plot.1  nrow(subset(df, df$sex==1)) / nrow(df) # 0.5  num.males <- nrow(subset(df, df$sex==1))  num.males # 65  total <- nrow(df)  total # 130  prop.men.above.temp.level <- nrow(subset(df, df$sex==1 & df$temp\_level==1)) / nrow(df)  prop.men.above.temp.level  prop.women.above.temp.level <- nrow(subset(df, df$sex==2 & df$temp\_level==1)) / nrow(df)  prop.women.above.temp.level |

[Figure 1: Boxplot for temp per sex group]



(3) Calculate the risk difference. Formally test (at the α=.05 level) whether the proportion of people with higher body temperatures (greater than or equal to 98.6) is the same across men and women, based on this effect measure. Do females have higher body temperatures than males? (4.5 points)

|  |
| --- |
| n.1 <- sum(df$sex==1)  n.1  n.2 <- sum(df$sex==2)  n.2  successes.1 <- nrow(subset(df, df$sex ==1 & df$temp\_level == 1 ))  successes.1  failures.1 <- nrow(subset(df, df$sex ==1 & df$temp\_level == 0 ))  failures.1  successes.2 <- nrow(subset(df, df$sex == 2 & df$temp\_level == 1 ))  successes.2  failures.2 <- nrow(subset(df, df$sex ==2 & df$temp\_level == 0 ))  failures.2  p.hat.1 <- successes.1 / n.1  p.hat.1  p.hat.2 <- successes.2 / n.2  p.hat.2  # risk difference:  p.hat.1 - p.hat.2 # - 0.3230769  # i.e. the risk of a temp above 98.6 is 32.307% higher in females than males  # or is 32.307% lower in females than in males  # Formally test at a alpha = 0.05 level:  # (3.1)  # H0: p.1 == p.2 (the population proportions are equal )  # H1: p.1 != p.2 (the population proportions are not equal )  # alpha = 0.05  # (3.2)  # Select the appropriate test statistic:  # z <- (p.hat.2 - p.hat.1) / sqrt ( p.hat.1 \* (1 - p.hat.1) \* ( (1/n.1) + (1/n.2) ) )  # (3.3)  # Deterine the decision rule:  # Reject H0 if |z| >= 1.960  # i.e. qnorm(0.975) == 1.960  # (3.4) COmpute the test statistic  z <- (p.hat.1 - p.hat.2) / sqrt ( p.hat.1 \* (1 - p.hat.1) \* ( (1/n.1) + (1/n.2) ) )  z  # (3.5)  # Reject H0, since |-4.480349| > 1.960 |

**Answer:**

* Yes, women do have a higher temperature than men.

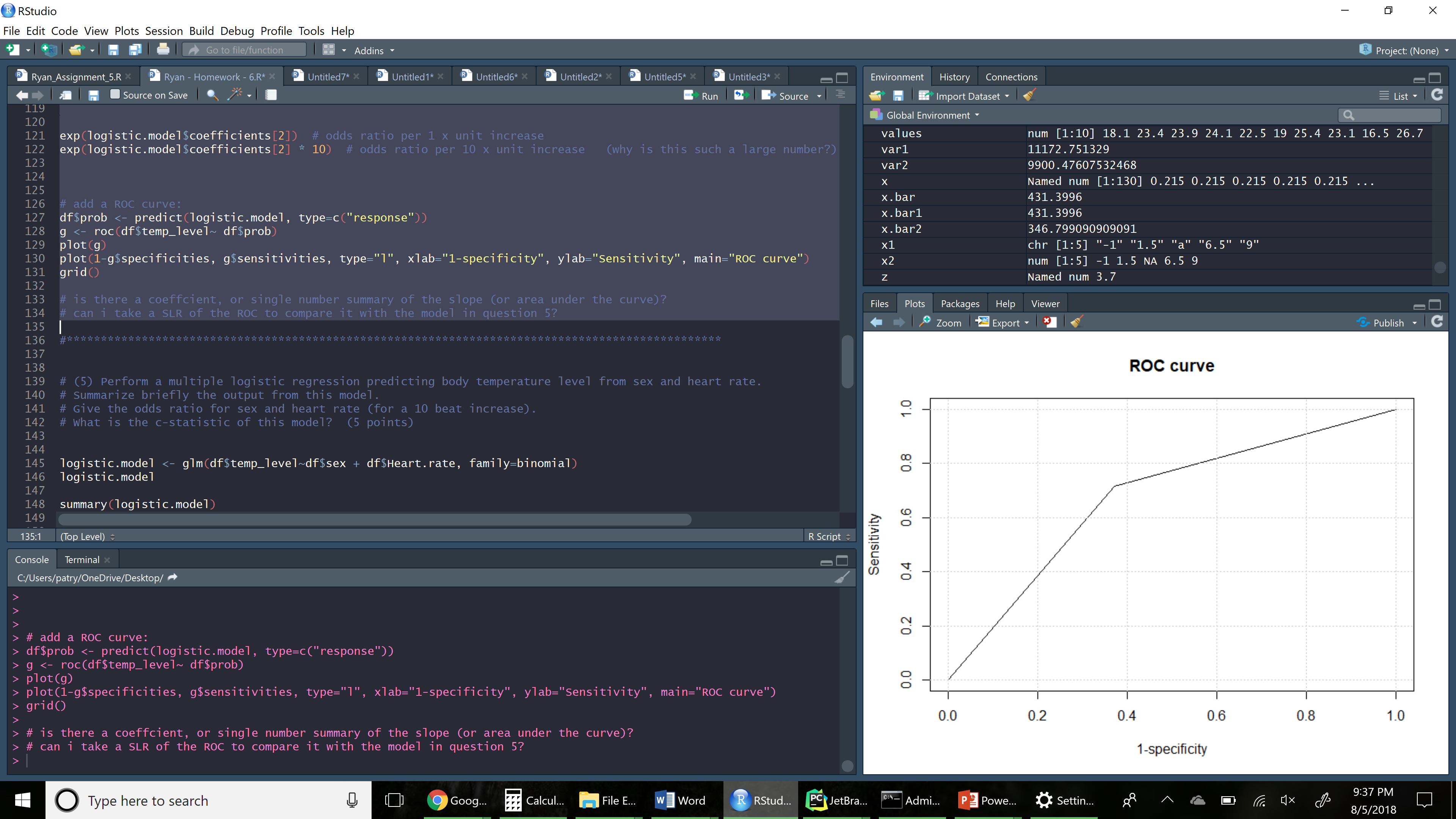
(4) Perform a logistic regression with sex as the only explanatory variable. Formally test (at the α=.05 level) if the odds of having a temperature greater than or equal to 98.6 is the same between males and females. Include the odds ratio for sex and the associated 95% confidence interval based on the model in your summary and interpret this value. What is the c-statistic for this model? (5.5 points)

|  |
| --- |
| logistic.model <- glm(df$temp\_level~df$is.girl, family=binomial)  summary.table <- summary(logistic.model)  summary.table  #(4.1)  # H0: p.hat(m.temp >= 98.6) == p.hat(f.temp >= 98.6) Male probabilty of having a temp of 98.6 == female probabilty of having a temp of 98.6  # H0: p.hat(m.temp >= 98.6) != p.hat(f.temp >= 98.6) Male probabilty of having a temp of 98.6 != female probabilty of having a temp of 98.6  #(4.2) # Select the appropriate test statistic:  # z = beta.1 / SE(beta.1)  # (4.3) # State the decision Rule  # Reject H0 if |z| >= 1.959964; or reject H0 if p <- alpha; otherwise do not reject H0  qnorm(0.975, lower.tail = TRUE) #1.959964  # (4.4) Compute the test statistic:  z <- logistic.model$coefficients[2] / summary.table$coefficients[4]  z # 3.699833  pnorm(z, lower.tail = FALSE)\* 2  # Reject H0 because 3.699833 is >= 0.0002157413  # The population probabilty for males and females having a temp >= 98.6 is not equal  # x <- predict(logistic.model, type=c("response"))  # plot(x)  # Confidence interval on Odds Ratio:  # Odds ratio:  or <- exp(cbind(OR = coef(logistic.model), confint.default(logistic.model)))  or[2] #4.25  lower <- exp((logistic.model$coefficients[2] - qnorm(0.975) \* summary(logistic.model)$coefficients[2,2]))  lower # 1.974712  upper <- exp((logistic.model$coefficients[2] + qnorm(0.975) \* summary(logistic.model)$coefficients[2,2]))  upper # 9.146904  exp(logistic.model$coefficients[2]) # odds ratio per 1 x unit increase  # add a ROC curve:  df$prob <- predict(logistic.model, type=c("response"))  g.1 <- roc(df$temp\_level~ df$prob)  plot(g.1)  plot(1-g.1$specificities, g.1$sensitivities, type="l", xlab="1-specificity", ylab="Sensitivity", main="ROC curve")  grid()  # Associated C-Statistic:  g.1 # 0.672 |

**Confidence Interval:** 1.974712 - 9.146904

**C-Statistic:** 0.672

[Figure 2: ROC Curve for temp\_level ~ is.girl]



(5) Perform a multiple logistic regression predicting body temperature level from sex and heart rate. Summarize briefly the output from this model. Give the odds ratio for sex and heart rate (for a 10 beat increase). What is the c-statistic of this model? (5 points)

|  |
| --- |
| logistic.model <- glm(df$temp\_level ~ df$is.girl + df$Heart.rate, family=binomial)  logistic.model  summary(logistic.model)  # Odds ratio  or <- exp(cbind(OR = coef(logistic.model), confint.default(logistic.model)))  or  lower <- exp((logistic.model$coefficients[3] - qnorm(0.975, lower.tail = TRUE) \* summary(logistic.model)$coefficients[3,2]\*10))  lower # 0.6094185  upper <- exp((logistic.model$coefficients[3] + qnorm(0.975, lower.tail = TRUE) \* summary(logistic.model)$coefficients[3,2]\*10))  upper # 1.862621  # plot as ROC curve  df$prob <- predict(logistic.model, type=c("response"))  g.2 <- roc(df$temp\_level~ df$prob)  plot(g.2)  plot(1-g.2$specificities, g.2$sensitivities, type="l", xlab="1-specificity", ylab="Sensitivity", main="ROC curve")  grid()  g.2  # C-statistic  g.2 # 0.7289 |

**Odds Ratio:**

OR 2.5 % 97.5 %

(Intercept) 0.002584129 3.895706e-05 0.1714124

df$is.girl 4.011585330 1.836365e+00 8.7634090

df$Heart.rate 1.065418154 1.007534e+00 1.1266280

**Summary:**

* Heart rate’s slope is 0.0633, and SE is 0.398
* Gender (is.girl)’s slope is 1.389 and SE is 0.02850

Call:

glm(formula = df$temp\_level ~ df$is.girl + df$Heart.rate, family = binomial)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.6524 -0.8639 -0.6103 1.0489 2.0480

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -5.95837 2.14018 -2.784 0.005369 \*\*

df$is.girl 1.38919 0.39868 3.484 0.000493 \*\*\*

df$Heart.rate 0.06337 0.02850 2.223 0.026195 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

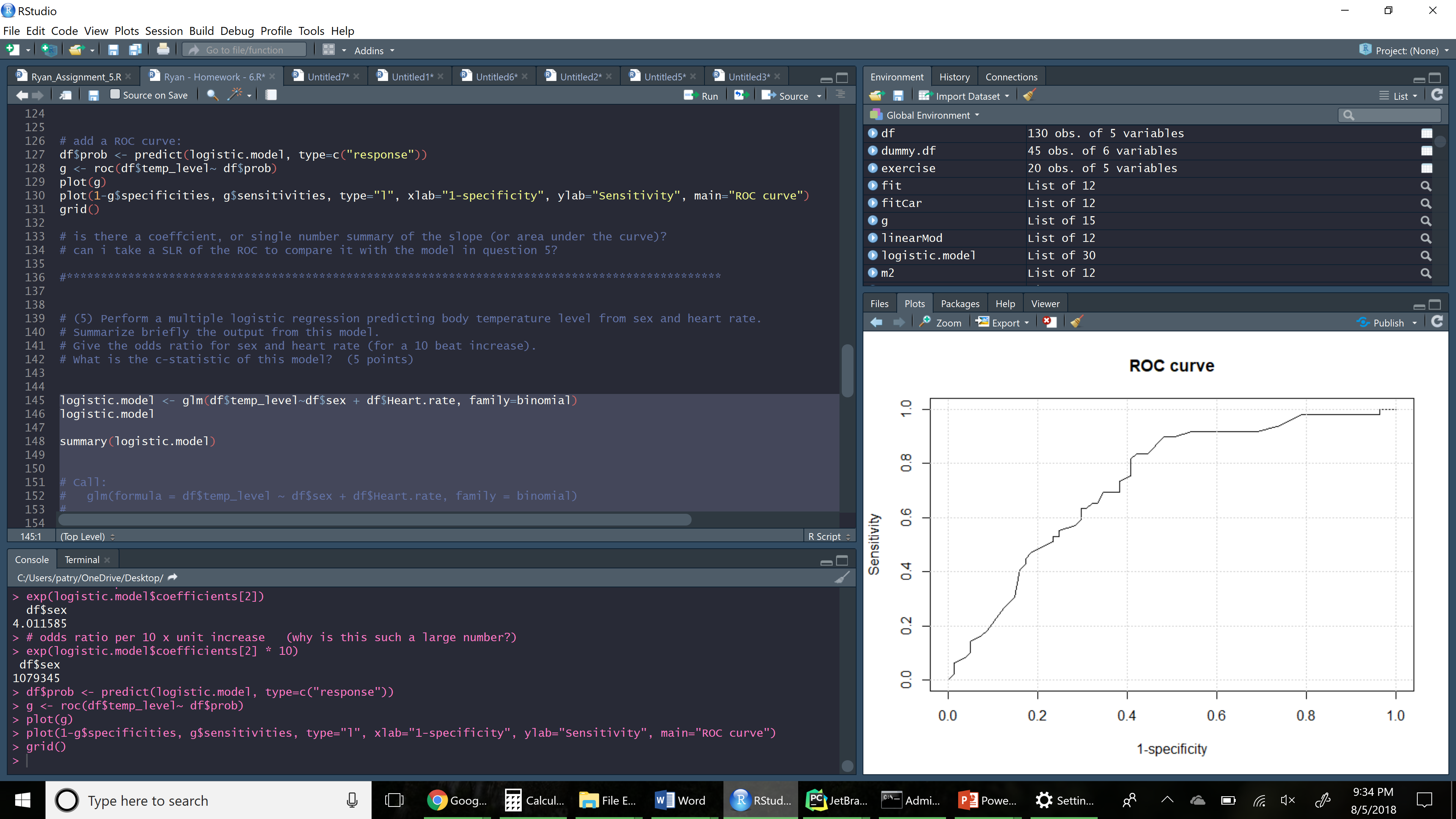
Null deviance: 172.26 on 129 degrees of freedom

Residual deviance: 152.24 on 127 degrees of freedom

AIC: 158.24

**C-Statistic**: 0.7289

[Figure 3: ROC Curve for temp\_level ~ is.girl + Heart.rate]



(6) Which model fit the data better? Support your response with evidence from your output. Present the ROC curve for the model you choose. (2 points)

**Decision:**

* The second model fit the data better.

**Model 1’s C-Statistic:** 0.672

**Model 2’s C-Statistic:** 0.7289

[Figure 4: ROC Curve for Superior Fit Curve (temp\_level ~ is.girl + Heart.rate)]

