STT 465 Optional HW

Nate Davis

11/6/2017

Question 1

1.1 & 1.3

```
gout <- read.table("gout.txt")
colnames(gout) <- c("sex", "race", "age", "serum_urate", "gout")

gout$sex <- ifelse(gout$sex == "F", 1, 0)
gout$race <- ifelse(gout$race == "W", 1, 0)

model <- lm(serum_urate ~ sex + age + race, data = gout)

summary(model)
```

```
##
## Call:
## lm(formula = serum_urate ~ sex + age + race, data = gout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4843 -0.9717 -0.1829  0.8276  5.4296
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.84828    0.82314   7.105 5.64e-12 ***
## sex         -1.52853    0.14306 -10.684  < 2e-16 ***
## age          0.02674    0.01299   2.058   0.0402 *
## race        -0.78212    0.16932  -4.619 5.22e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 396 degrees of freedom
## Multiple R-squared:  0.2504, Adjusted R-squared:  0.2447
## F-statistic: 44.09 on 3 and 396 DF,  p-value: < 2.2e-16
```

1.2

From above we see that all coefficients are significant at the 95%. Being female results in an expected decrease of urate by 0.78 and being white results in an expected decrease of urate by about 1.52. Each year one ages has an expected increase in urate by about 0.02. By the intercept, a newborn black male would have an expected urate of about 5.84.

Question 2

2.1 & 2.2

```
X <- model.matrix(~sex + age + race, data = gout)
Y <- gout$serum_urate

XY <- t(X) %*% Y
XtX <- t(X) %*% X

bhat <- solve(XtX, XY)

error <- Y - X %*% bhat

error_var <- sum(error^2/(nrow(gout)-2))

se <- sqrt(diag(solve(XtX))*error_var)

tstat <- bhat / se

pvalues <- 2 * pt(-abs(tstat), nrow(gout) - 1)

new_summary <- cbind(bhat, se, tstat, pvalues)

colnames(new_summary) <- c("Coef", "SE", "t-stat", "pvalue")

new_summary
```

```
##                    Coef         SE     t-stat       pvalue
## (Intercept)  5.84828010 0.82107010   7.122754 4.970191e-12
## sex         -1.52852797 0.14270205 -10.711325 1.059436e-23
## age          0.02673734 0.01295701   2.063543 3.970731e-02
## race        -0.78211876 0.16889349  -4.630840 4.935380e-06
```

These results are almost identitcal to those above and have the same interpretations. The largest difference is the pvalue for the age variable, which is slightly more significant (smaller pvalue) than in the lm function results. There are minor differences in some of the values at the fourth or fifth decimal place.

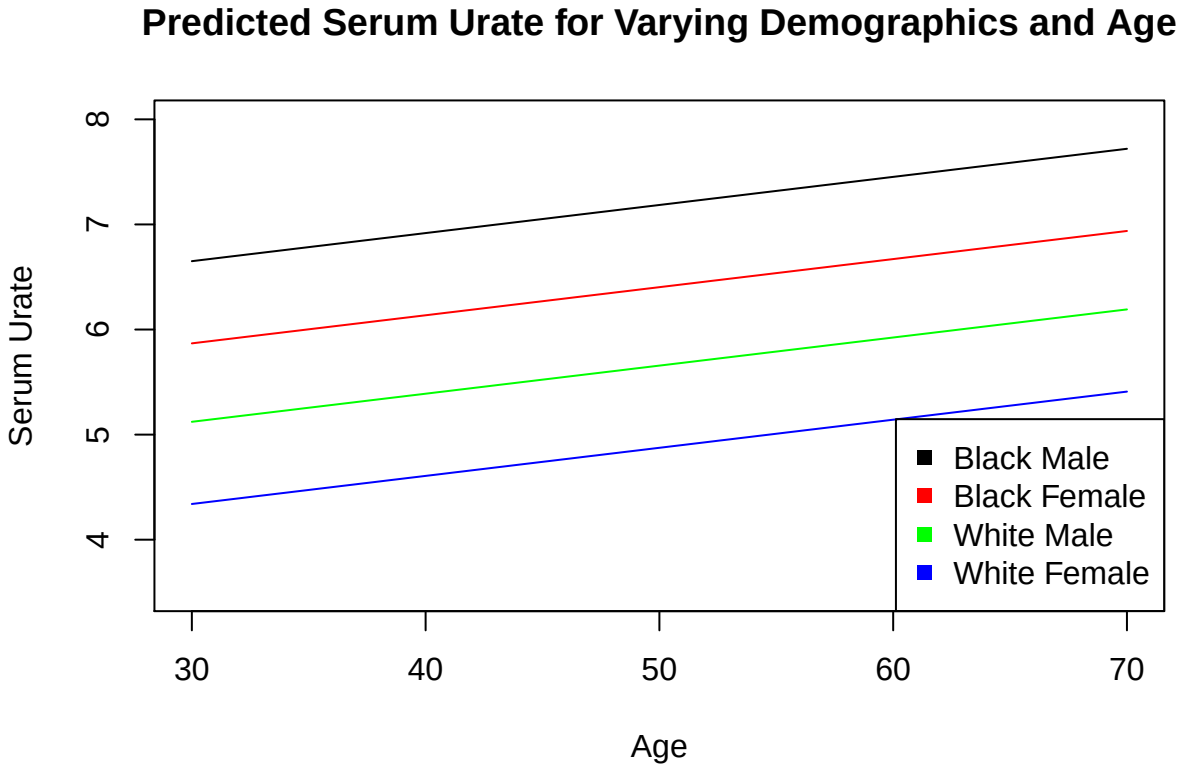
Question 3

```
age <- rep(seq(30, 70, length.out = 100), 4)
race <- rep(c(0, 1), each = 200)
sex <- rep(c(0, 1, 0, 1), each = 100)

pred_data <- data.frame(cbind(sex, age, race))

pred_data <- cbind(pred_data, predict(model, pred_data))

plot(pred_data[1:100,2], pred_data[1:100,4], type = "l", ylim = c(3.5,8), xlab = "Age",
     ylab = "Serum Urate", main = "Predicted Serum Urate for Varying Demographics and Age")
lines(pred_data[101:200,2], pred_data[101:200,4], col = "green")
lines(pred_data[201:300,2], pred_data[201:300,4], col = "red")
lines(pred_data[301:400,2], pred_data[301:400,4], col = "blue")
legend("bottomright", legend = c("Black Male", "Black Female", "White Male", "White Female"),
                                col = c("black", "red", "green", "blue"),
                                pch = 15)
```


Predicted Serum Urate for Varying Demographics and Age
Serum Urate
4
5
6
7
8
30
40
50
60
70
Age
Black Male
Black Female
White Male
White Female