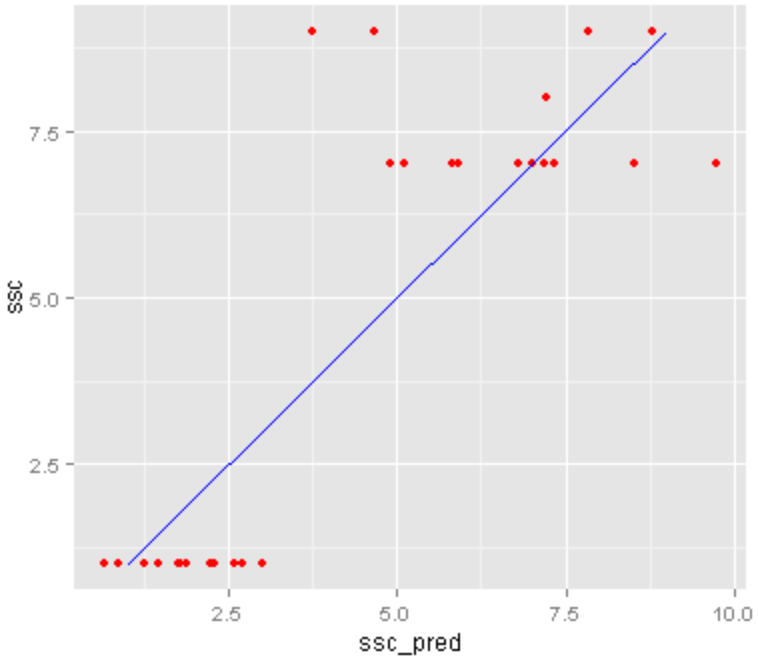
Code and Results

```
> # Load libraries, set wd, and load/attach data
> library(MASS)
> setwd("C:\\Users\\Kyle\\Dropbox\\CS\\CS3654\\R\\Inclass9")
> load("fdata.rdata")
> # Create/attach training and test subset as demonstrated in text
> final$gp <- runif(dim(final)[1])
> test <- subset(final, final$gp <= 0.1)
> train <- subset(final, final$gp > 0.1)
> attach(train)
> # Remove columns used for creating subsets
> train["gp"] <- NULL
> test["gp"] <- NULL
> # Fit linear regression with all features
> fit <- lm(ssc ~ age + gender + location + ethnicity + coder + som1 + som2 + som3 + som4
+ + som5 + som6 + som7 + som8 + som9 + som10 + som11 + som12 + som13 + som14)
> summary(fit)
> step <- stepAIC(fit, direction = "both")
> step
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
ssc ~ age + gender + location + ethnicity + coder + som1 + som2 +
    som3 + som4 + som5 + som6 + som7 + som8 + som9 + som10 +
    som11 + som12 + som13 + som14

Final Model:
```

```
ssc ~ age + location + ethnicity + coder + som1 + som2 + som3 +
    som4 + som5 + som9 + som10 + som11 + som12 + som13 + som14



     Step Df  Deviance Resid. Df Resid. Dev      AIC
1                         265   539.3376 237.9931
2   - som8  1 0.2053721      266   539.5429 236.1054
3   - som6  1 0.2107394      267   539.7537 234.2206
4   - som7  1 0.2348345      268   539.9885 232.3489
5 - gender  1 0.8016397       269   540.7902 230.7865
> rm(step)
> rm(fit)
> # Build model with only retained variables
> fit1 <- lm( ssc ~ age + location + ethnicity + coder + som1 + som2 + som3 +
+ som4 + som5 + som10 + som11 + som12 + som13 + som14)
> summary(fit1)
> # Remove training set (no longer needed)
> detach(train)
> rm(train)
> # Now predict using the test set
> test$ssc_pred <- predict(fit1, newdata = test)
> rm(fit1)
> #See how predicted ssc compares to actual ssc values
> library(ggplot2)
> ggplot(data = test, aes(x = ssc_pred, y = ssc)) +
+ geom_point(color = "red") +
+ geom_line(aes(x = ssc, y = ssc), color = "blue")
```


7.5
5.0
2.5
ssc
2.5
5.0
7.5
10.0
ssc_pred

Interpretation

The actual SSC scores versus the predicted scores in this case are relatively inaccurate. While the relationship between the two is linear with a slope of near 1, we can see from looking at the plot that there are many outliers, and that a given SSC score can lead to a prediction within +/- 5 of the actual score. As a result, I would say that this is not an accurate model for predicting SSC scores.