Homework 3

1. The linear model can be written as . Each term has a corresponding dimension, and interpretation:

Y:

The y term is an n x 1, random column vector of observable terms.

X:

The X term is an n x p matrix of observed values for our regressors/predictors. It can also include a columns of one 1, which represent the intercept.

β:

The β term is a p x 1 fixed, non-random column vector, which represents the weights of each regressor in the linear combination equation . These are unknown, and are determined by solving the normal equations.

ε:

The ε represents the error in the model, and is a random column vector of dimension n x 1.

2.1 a)

summary(fit)$r.squared

[1] 0.5267234

b)

max(residuals(fit)) = 94.2522

c)

median(residuals(fit)) = -1.451392

mean(residuals(fit)) = 0

d)

cor(x = residuals(fit), y = fitted(fit)) = 0

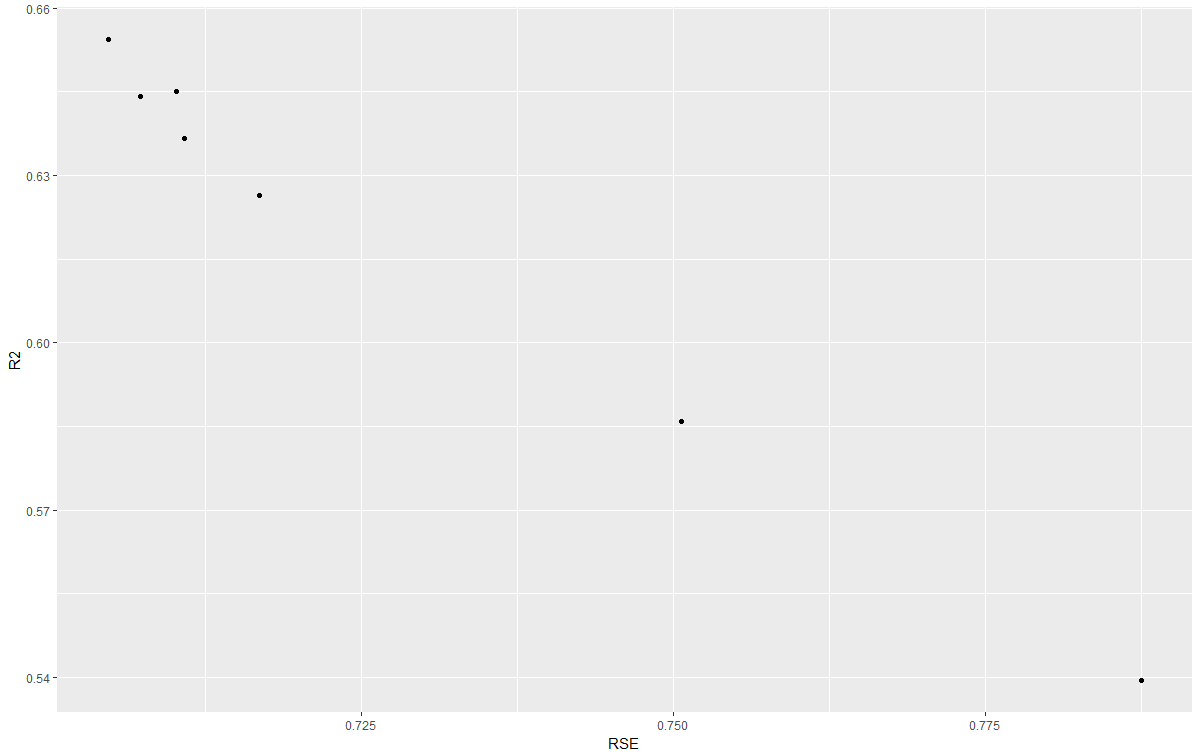
e)

cor(x = residuals(fit), y = teengamb$income)) = 0

2.3

The method for direct calculation fails with a degree six polynomial.

2.4



2.6

a)

> cheese <- lm(taste ~ ., data = cheddar)

> cheese$coefficients

(Intercept) Acetic H2S Lactic

-28.8767696 0.3277413 3.9118411 19.6705434

b)

> correl <- cor(x = fitted(cheese), y = cheddar$taste)

> correl^2

[1] 0.6517747

c)

> no.int <- lm(taste ~ . - 1, data = cheddar)

> summary(no.int)$r.squared

[1] 0.8877059

A better metric for goodness of fit would be the correlation of the fitted values with the observed values squared.

> cor(fitted(no.int), cheddar$taste)^2

[1] 0.6244075