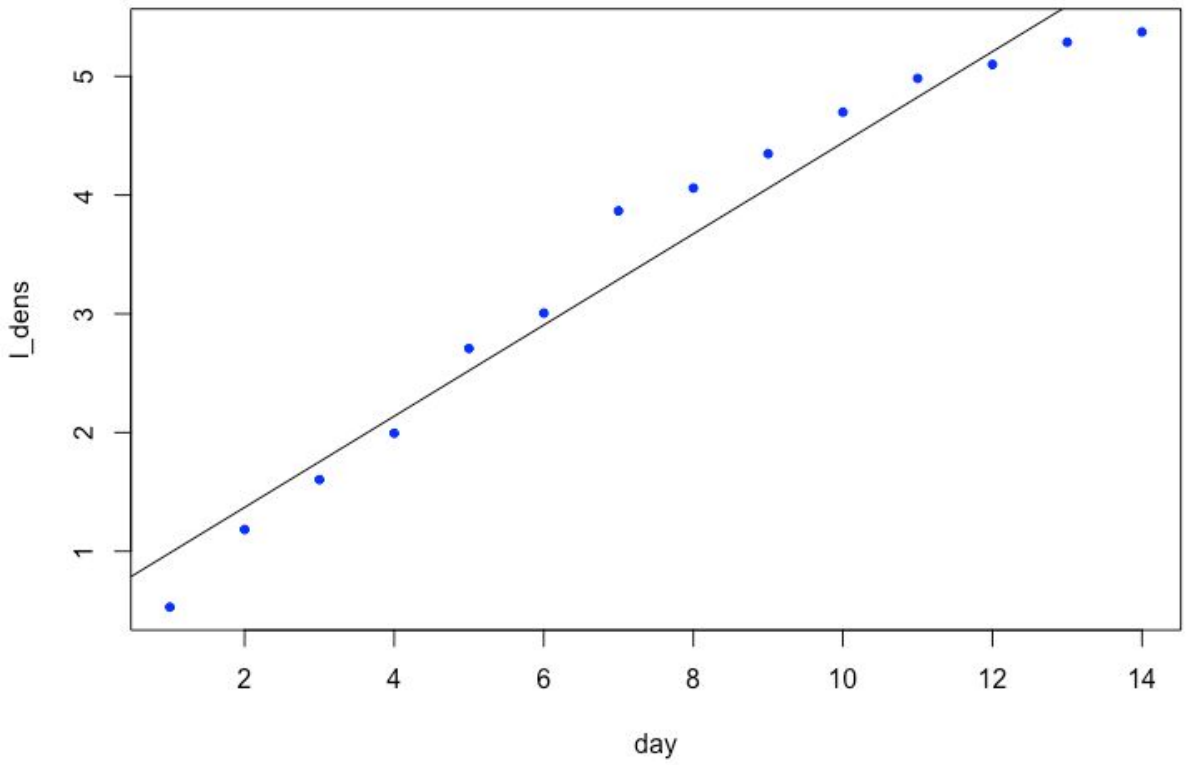
Calvin Todorovich

Homework 3

a) Construct a scatter-plot of log-density vs day



The data suggests a positive linear relationship between days and log density, because as the values of day increase, the values of log density increase too, with symmetry about the fitted line.

b) Use the R-function `cor(x,y)` to obtain the correlation between the variables log-density and day

Correlation coefficient = .9789, this confirms my thoughts about the plot suggesting a linear relationship, due to the value being close to 1, it suggests a strong linear relationship between the two variables.

c) Fit 4 polynomial regression models for log-Density Vs Day : linear, quadratic, cubic, and quartic

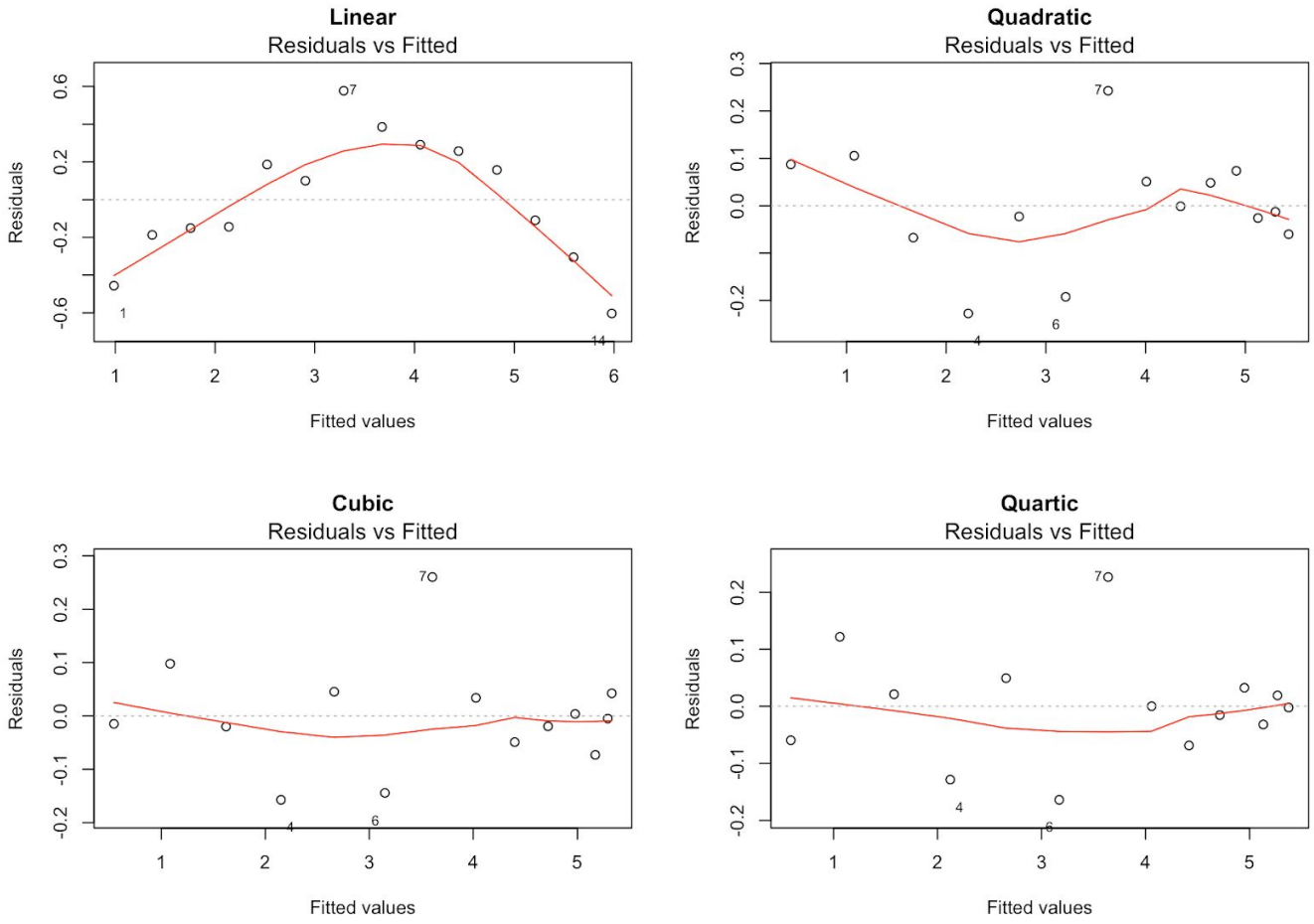
Fit1(linear) #R sq: .9583, adj R sq: .9549
Fit2(quadratic) #R sq: .9947, adj R sq: .9937
fit3(cubic) #R sq: .9961, adj R sq: .9949
Fit4(quartic) #R sq: .9965, adj R sq: .9949

As we add another term to the fit, the R squared and adjusted R squared both rise. This shows that the fitted line is closer to the observed points as you add higher degree terms.

d) What is the highest order polynomial that can be fitted using these data?

N-2, because p-values are unavailable at n-1 degree polynomial, where n is the number of observations, due to the fit precisely fitting to each value on the plot. The ‘perfect fit’ is actually just an over-complicated fit.

e) Construct the residual plots for all the 4 models under consideration and produce a 2 by 2 display



f) Use the ESS test to compare the models: linear vs. quadratic, quadratic vs. cubic, cubic vs. quartic.

Fit1 vs Fit2 #p = 3.04e-6
Fit2 vs Fit3 #p = .08519
Fit3 vs Fit4 #p = .3519

With low p-values for all three comparisons, I reject the null hypothesis that there is no difference between the models in each comparison. This means that as we add higher degree terms, the higher degree model is a better statistical predictor than the simpler model. This matches the statistics in part c), that the fitted line is tighter as you add higher degree terms.

Software: (sorry for not including this section on both of the other assignments)

```
#ST 412 HW 3
#Calvin Todorovich

library(ggplot2)
density

#a
#to my understanding the data is already log transformed

l_dens <- density$log_density
day <- density$day
fit1 <- lm(l_dens~day)
help(plot)
plot(day,l_dens, pch=20, col="blue")
abline(fit1, pch=20, col="black")
plot(fit1, which = 1)

#b
cor(day,l_dens)
# = .9789

#c
#need 4 fits for l_dens, l_dens^2, l_dens^3, l_dens^4

fit2 <- lm(l_dens~ poly(day, 2, raw = TRUE))
fit3 <- lm(l_dens~ poly(day, 3, raw = TRUE))
fit4 <- lm(l_dens~ poly(day, 4, raw = TRUE))

summary(fit1)#R sq: .9583, adj R sq: .9549
summary(fit2)#R sq: .9947, adj R sq: .9937
summary(fit3)#R sq: .9961, adj R sq: .9949
summary(fit4)#R sq: .9965, adj R sq: .9949

#d conceptual

#e
par(mfrow=c(2,2))
plot(fit1,which = 1, main="Linear")
plot(fit2, which = 1, main="Quadratic")
plot(fit3, which = 1, main="Cubic")
plot(fit4, which = 1, main="Quartic")

#f
anova(fit1, fit2) #p = 3.04e-6
anova(fit2, fit3) #p = .8519
```

```
anova(fit3, fit4) #p = .3519
```