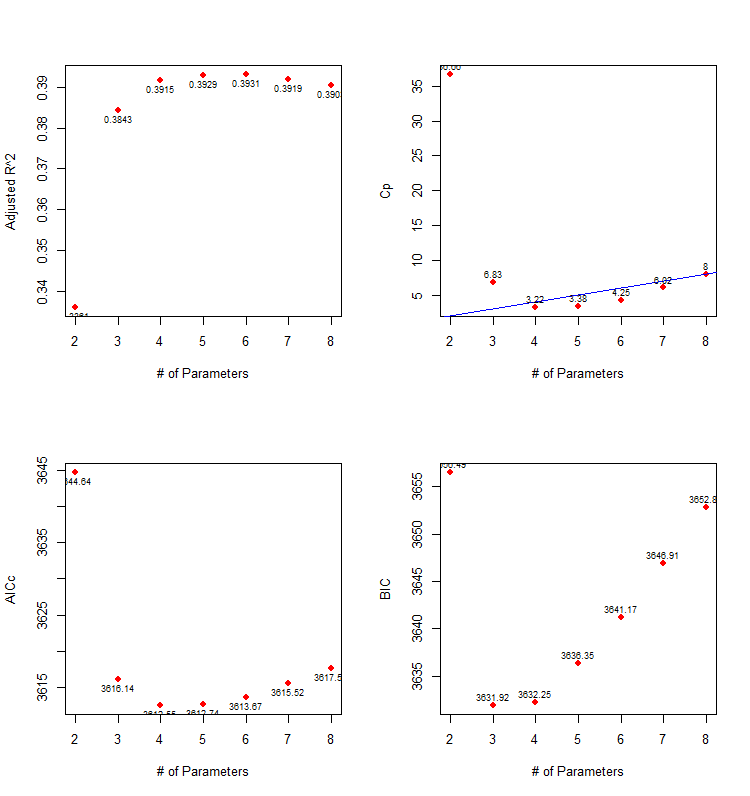
Alyssa Semerdjian

Stat 510

29 Jan 2018

**Homework 2**

**Figure 1:** Graphs correspond to questions below. Top left corresponds to question a, top right to b, bottom left to c, and bottom right to d.

a) The equation for the ‘best’ estimated model is: glucose = 65.5 + 0.2423(diastolic) +0.1372(insulin) + 0.6036(age). See figure 1 for graph.

b) The models with 4, 5, 6, 7 and 8 variables all have very similar adjusted R2 values, however the ‘best’ model is the one with 6 parameters (5 x-variables; insulin, age, diastolic, diabetes, and bmi) because it has highest adjusted R2. See figure 1 for graph.

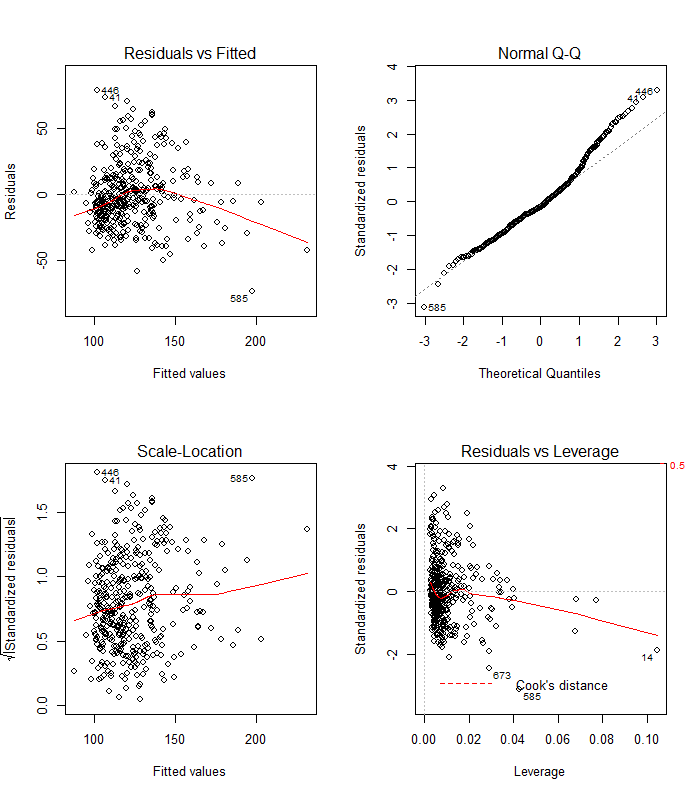
c) According to Mallows’ Cp the ‘best model’ has 7 parameters (6 x-variables; insulin, age, diastolic, diabetes, bmi, and triceps). See figure 1 for graph.

d) n/(p+1) in this case is greater than 40 so you’d want to use AIC for this data. However, since the question asks which model is best using AICc, the model with the lowest AICc has 4 parameters (3 x-variables; insulin, age, diastolic). See figure 1 for graph.

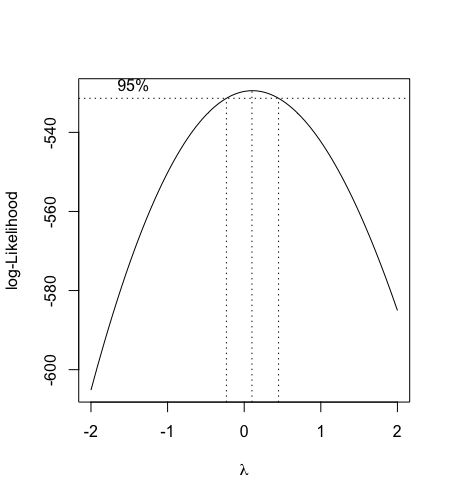
e) The model with the lowest BIC (and therefore the ‘best’ model) has 3 parameters (2 x-variables; insulin, age). See figure 1 for graph.

f) The equation AICc = AIC + = . Loglikelihood for model 5 as calculated in R is -1799.6988, p is 6 and n is 392. So AIC = = = **.**

g) The best model according to AIC is also the best according to AICc and has 4 parameters (3 x-variables; insulin, age, diastolic). The normal QQ plot shows us that model seems to have fairly normal distribution, except for some of the higher residuals. The residuals vs fitted plot shows that there maybe some problems with variance homogeneity. The points do not seem to be evenly scattered around 0 and instead are clustered more to the right. Overall is model does not seem to be ideal in terms of meeting the assumptions.



h) According to the box-cox routine, the recommended transformation of y is log(y) because lamda~0. This transformation should make the model’s variance more homogeneous.



1. When R^2 is close to 1 vif will be very big. When R^2 is close to 0 vif will get closer to 1.

R^2 = 1/(1-R^2) = 1/(1-0.1255) = 1.1435

> model\_99=lm(age~insulin+diastolic,data=data2)

> summary(model\_99)$r.sq

[1] 0.1255335

> 1/(1-summary(model\_99)$r.sq)

[1] 1.143554

> vif(model\_3)

insulin age diastolic

1.050805 1.143554 1.100343

j) There are 18 positive outliers and they are #’s 41, 111, 121, 176, 259, 261, 360, 428, 446, 499, 507, 516, 562, 596, 648, 697, 733, and 749. There are 3 negative outliers and they are #’s 255, 585 and 673.

