M358K - Homework 3 (Linear regression basics)

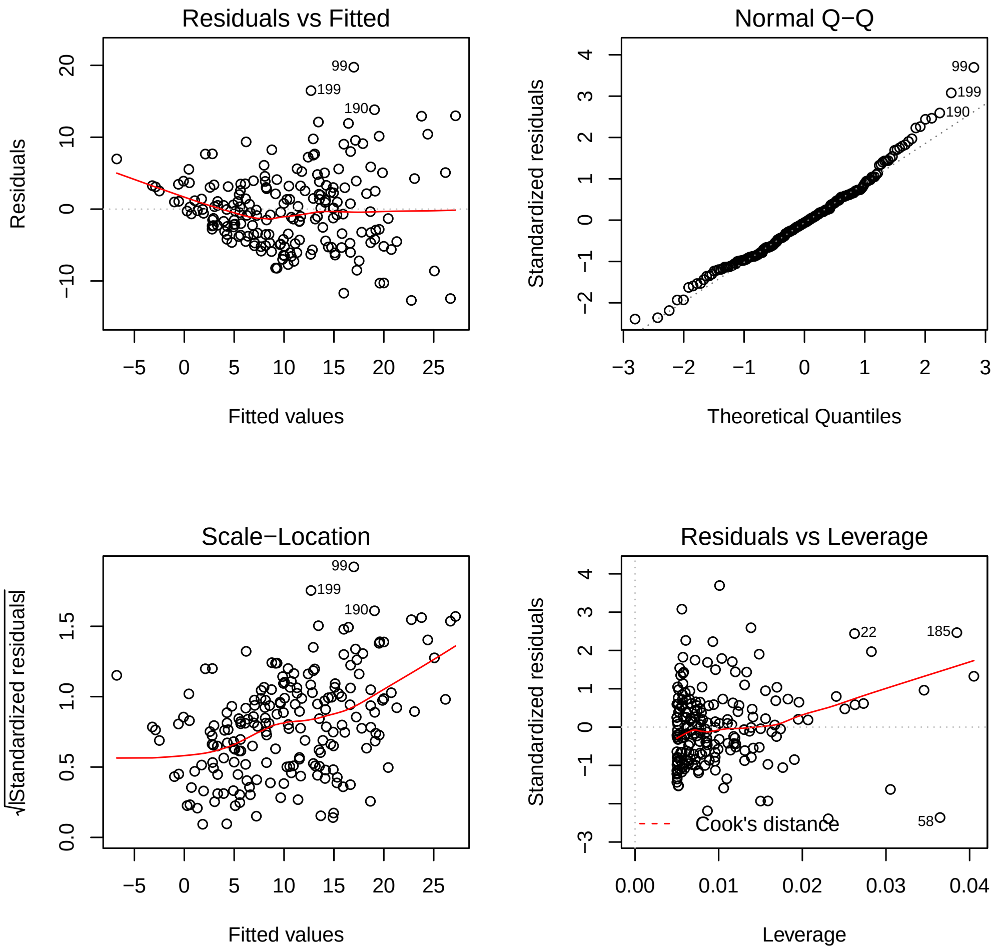
posted by: Monday October 22nd, 2018

due by: Monday November 5th at 11.59PM, 2018

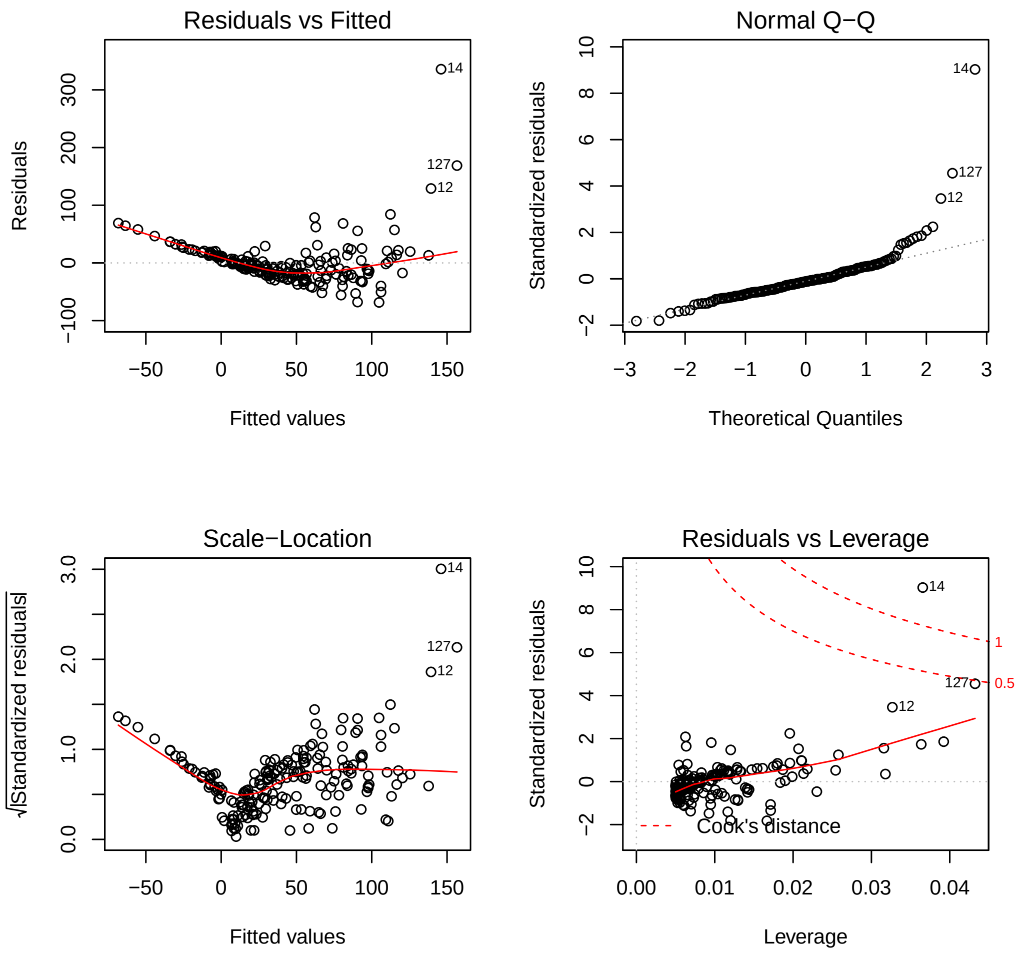
Number of questions in this homework: 3.
Maximal points possible: 6 from writeup, 2 from code, 4 from presentation. This gives a total of **12 points**.

Question 1. Residuals in linear regression models

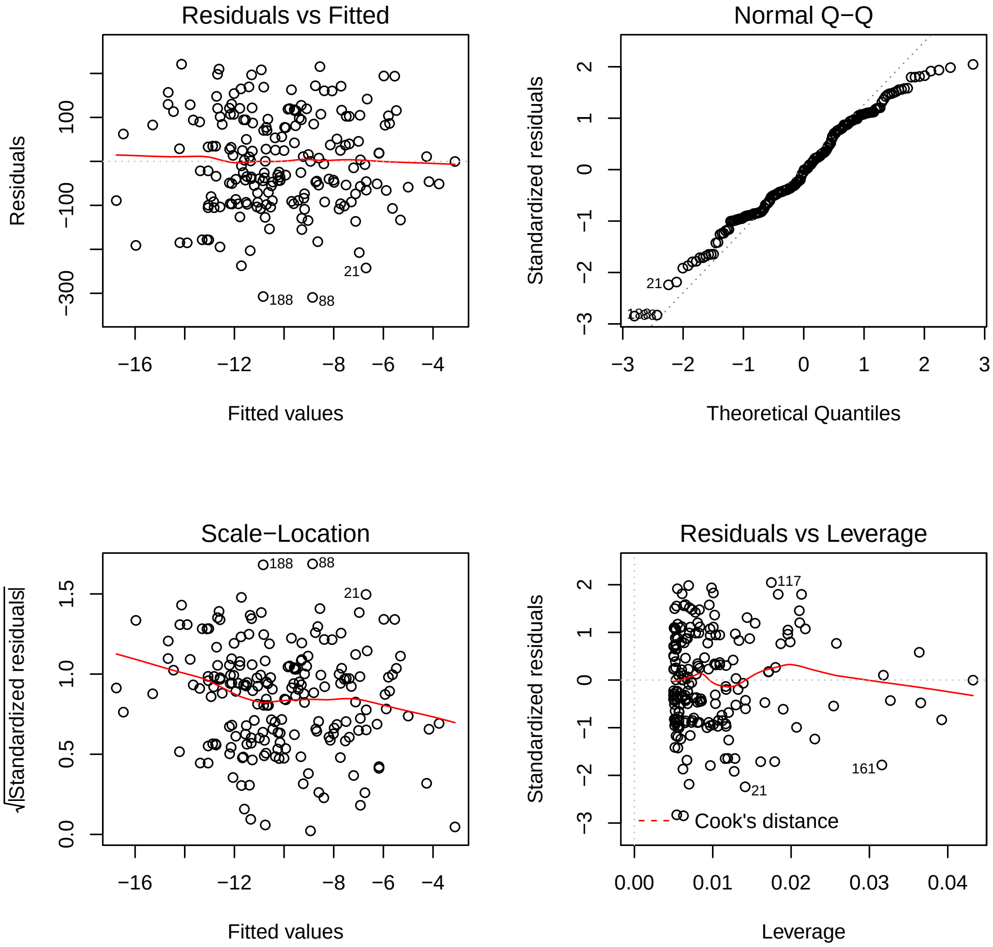
A statistician fitted several regression models to different variable pairs. For each model, she ran diagnotic plots for the residuals. For each of the following residual plots say which, if any, of the following assumptions are violated: (a) no mean trend, (b) normal distribution and (c) constant variance.
Based on your answer, which models are good fits?


Residuals vs Fitted
Residuals
20
10
0
−10
−5
0
5
10
15
20
25
Fitted values
99
199
190
Normal Q−Q
Standardized residuals
4
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
99
199
190
Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−5
0
5
10
15
20
25
Fitted values
99
199
190
Residuals vs Leverage
Standardized residuals
4
3
2
1
0
−1
−3
0.00
0.01
0.02
0.03
0.04
Leverage
22
185
58
Cook's distance

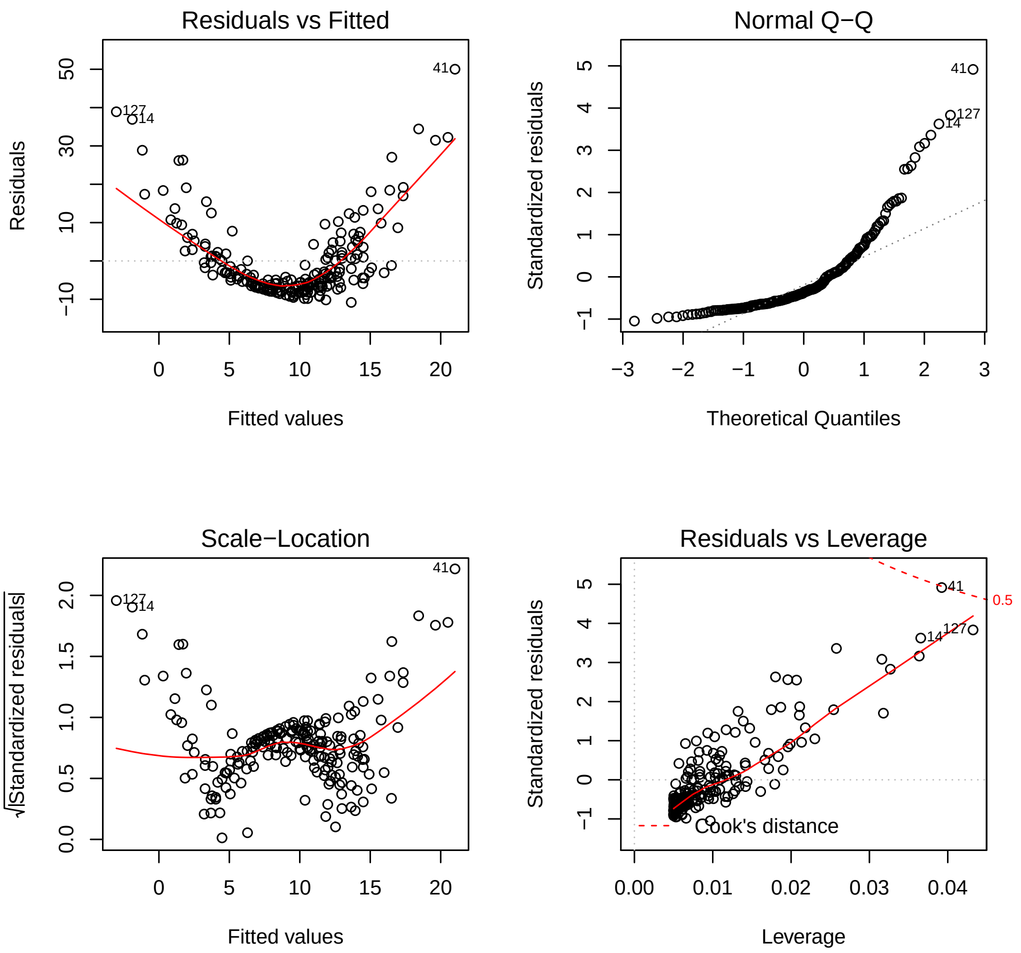
Model 1


Residuals vs Fitted
Residuals
Fitted values
−100
0
100
200
300
−50
0
50
100
150
14
127
12
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−2
0
2
4
6
8
10
−3
−2
−1
0
1
2
3
14
127
12
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
2.0
3.0
−50
0
50
100
150
14
127
12
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
4
6
8
10
0.00
0.01
0.02
0.03
0.04
14
127
12
1
0.5
Cook's distance

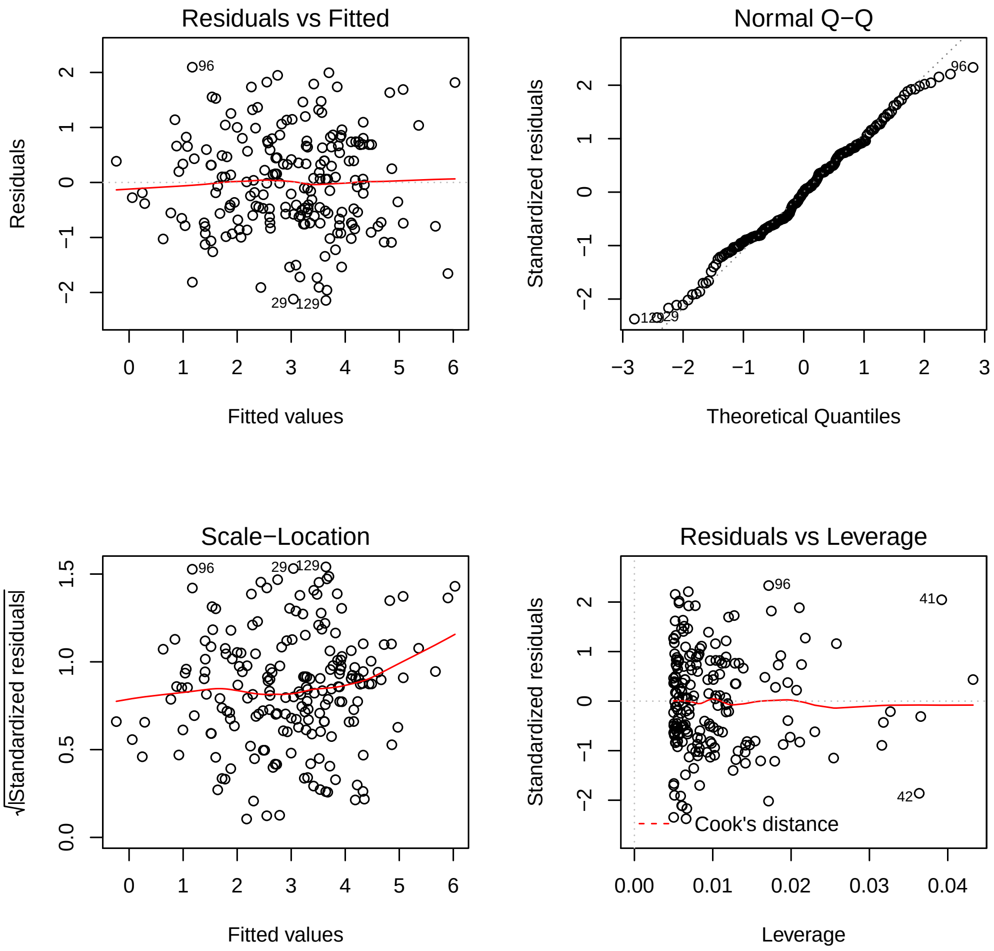
Model 2


Residuals vs Fitted
Residuals
100
−100
−300
−16
−12
−8
−6
−4
Fitted values
21
188
88
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
Theoretical Quantiles
21
188
Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
−16
−12
−8
−6
−4
Fitted values
188
88
21
Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
−3
0.00
0.01
0.02
0.03
0.04
Leverage
117
161
21
Cook's distance

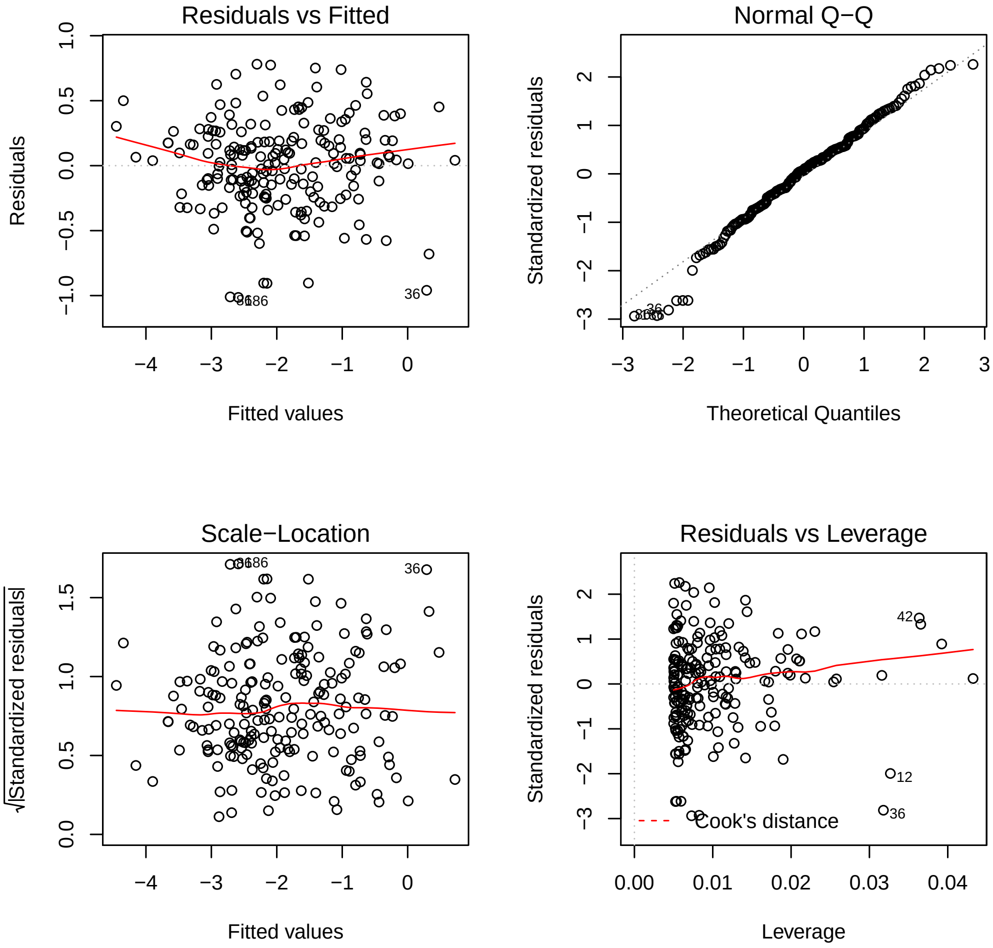
Model 3


Residuals vs Fitted
Residuals
50
30
10
−10
0
5
10
15
20
Fitted values
127
14
41
Normal Q−Q
Standardized residuals
5
4
3
2
1
0
−1
−3
−2
−1
0
1
2
3
Theoretical Quantiles
41
127
14
Scale−Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
0
5
10
15
20
Fitted values
127
14
41
Residuals vs Leverage
Standardized residuals
5
4
3
2
1
0
−1
0.00
0.01
0.02
0.03
0.04
Leverage
41
0.5
14
127
Cook's distance

Model 4


Residuals vs Fitted
Residuals
Fitted values
96
29
129
Normal Q−Q
Standardized residuals
Theoretical Quantiles
96
129
29
Scale−Location
√|Standardized residuals|
Fitted values
96
29
129
Residuals vs Leverage
Standardized residuals
Leverage
96
41
42
Cook's distance
0.00
0.01
0.02
0.03
0.04
−3
−2
−1
0
1
2
3
4
5
6
0.0
0.5
1.0
1.5

Model 5



Model 6

Question 2. Simple linear regression

Consider the `hsb2` dataset.

1. Do a scatter plot of social studies (socst) scores vs math scores (math). Report the correlation.

2. Fit a linear regression model that can be used to predict social studies scores based on math scores. Write down the model equation that R gives you. Overlay this regression line on top of the scatterplot.

3. Interpret the intercept and the slope. Is the intercept meaningful? Why or why not?

4. R reports the p-value of the slope and the intercept. What is the meaning of these p-values? What are the null and alternative hypotheses in these tests? For the p-values in your model, what can you conclude?

5. Do a diagnostic plot for your model. Say which, if any, of the (a) no mean trend, (b) normal distribution and (c) constant variance assumptions are violated. Based on your answer, is the model a good fit? That is, is there a linear relationship between social studies scores and math scores?

6. Report the R^2 value of your model. What is the meaning of this value?

7. Summarize the relationship between social studies scores and math scores in a paragraph that utilizes all of the numbers in the previous questions.

Question 3. Regression with categorical predictors

This question concerns a study on obesity and qualification. Reference: Hebl, M. R., & Mannix, L. M. (2003). The weight of obesity in evaluating others: A mere proximity effect. Personality and Social Psychology Bulletin, 29, 28-38.

People who are obese face a great deal of prejudice and discrimination. But what about people who are somehow associated with an obese person? In the study above, participants had to rate how qualified a particular job applicant was. This applicant was sitting by a woman. The researchers manipulated the following two variables: the weight of the woman and the relationship between the woman and the applicant. The woman was either obese or of average weight. This woman was also portrayed as being the applicant's girlfriend or a woman simply waiting to participate in a different experiment.

Data: `weight.txt` on Canvas. You can read this table in with the `read.table` command (same syntax as `read.csv`)

Variables descriptions:

- Weight: The weight of the woman sitting next to the job applicant.
  1 = obese, 2 = average weight

- Relate: type of relationship between the job application and the woman seated next to him
  1 = girlfriend, 2 = acquaintance (waiting for another experiment)

- Qualified: a numerical. Larger numbers represent higher professional qualification ratings

1. Do a plot of qualified vs weight, qualified vs relate, and qualified vs weight:relate. For each plot, briefly describe what you see. Based on these plots alone, do you think sitting next to an obese woman has an adverse effect on the applicant's qualification ratings?

2. Run three regression models with the following definitions:

   - `weights.model1`: qualified vs weight + relate.
   - `weights.model2`: qualified vs weight:relate
   - `weights.model3`: qualified vs weight + relate + weight:relate

   For each model, clearly show the R command that you use, and include the R's model summary. Write down the equation that R gives you. Interpret all the coefficients and the *p*-values associated with the coefficients.

3. Do a diagnostic plot for each of your models. Say which, if any, of the (a) independence (no mean trend) (b) normal distribution and (c) constant variance assumptions are violated.

4. Why do the coefficients estimates of `weights.model3` differ from those in `weights.model1` and `weights.model2`, even for the same variables?

5. Between these three models, which one is the best? Justify your choice based on the diagnostics and the model's interpretability.