M358K - Homework 3 (Linear regression basics)

posted by: Tuesday, October 10th at 12AM, 2017

due by: Wednesday October 25th at 11.59PM, 2017

Instructions and grading scheme

Submission instructions. Please submit three files on Canvas with the following names:

- **homework3-writeup**: this is a word file containing your answers. All figures, tables etc must be inside this file! For how to save R figures, see instructions on Canvas.
- **homework3-code-final.R**: this is a text file containing the R codes you used to compute the numbers, tables, and figures CONTAINED IN your report ONLY. This file IS graded, see below.
- **homework3-code-draft.R**: this is a text file containing the R codes you used to explore the data. This file should give an idea on how you explored the dataset, and how you arrived at code-final.R. It can tables/figures which are EXCLUDED from your report, comments, etc. For instance, the R history of your various sessions is enough.

Acceptable word files are: anything that can be opened by LibreOffice, OpenOffice, Microsoft Word, or equivalent software. Example file extensions are .doc, .docx, .odt.

Grading scheme. For the write up: on each question you can earn 0/1/2 points.

- 2 = correct answer

- 1 = partially correct answer, or correct answer but with muddy or missing justification
- 0 = incorrect answer, unreasonable answer

For code-final.R: you can earn 0/5 points.

- 5 = code-final.R runs on the given dataset, gives all the tables and figures included in your report.
- 0 = no code file OR code file does not run at all OR code file does not produce the reports' tables/figures OR code file raises a plagiarism red flag, etc

For code-draft.R: you can earn 0/5 points.

- 5 = The grader is sufficiently convinced that you explored this dataset on your own.
- 0 = no code file OR code file raises a plagiarism red flag OR code file is irrelvant, does not give an indication on how you arrived at code-final.R

Bonus points for presentation:
Write-up: +2 points for nice report(grammatically correct sentences, no rambling discussions, discussions exceed expectations, extra analysis of the dataset)
R code: +2 points for neat layout. (code adequately commented, clearly laid out, easy to understand).

Bonus questions
These are extra challenging questions, and are additional opportunities to score points. On each bonus question you can get 0/4 points.

Questions

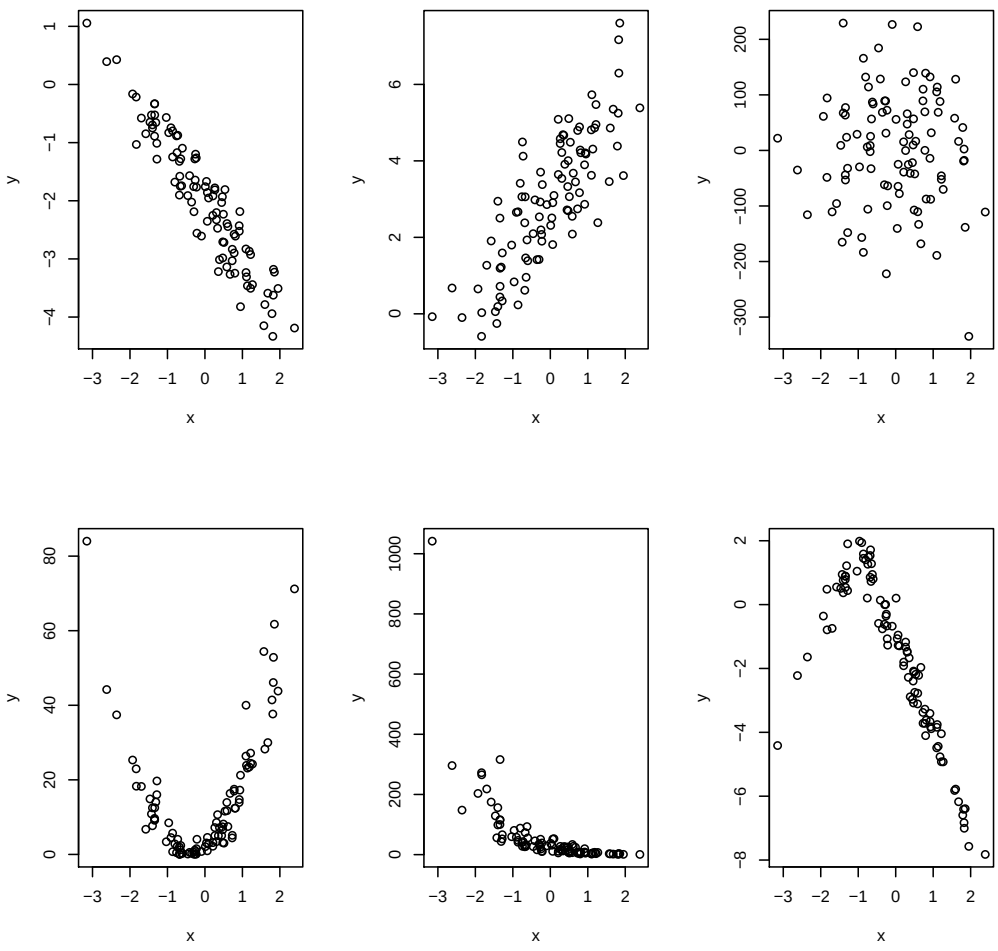
1 Interpretations

1.1 Descriptives with scatterplots (2 points total)

For each of the following plots

1. Identify the strength of the relationship of the two variables in the data (eg: weak, moderate or strong, linear or not, any obvious trends).

2. Would it be reasonable to fit a linear model? Why or why not?



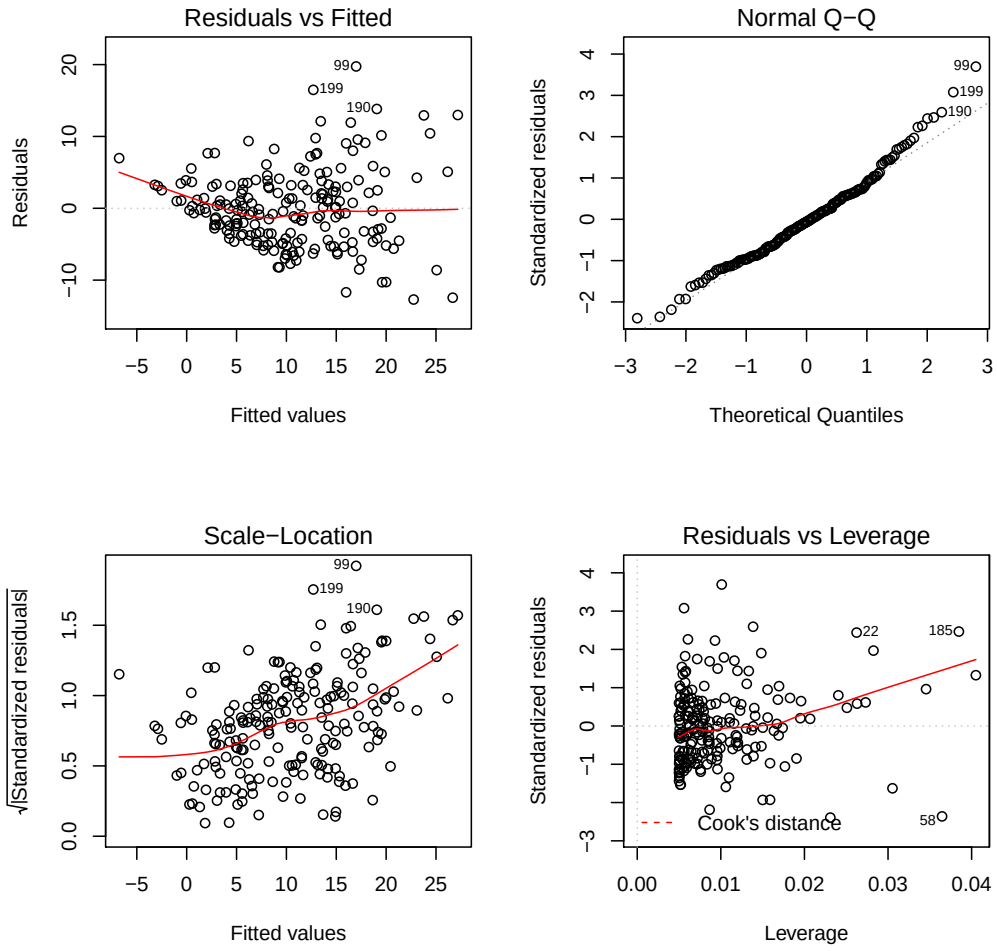
1.2 Correlations (2 points total)

Match the following calculated correlations to the corresponding six scatterplots above: $-0.7, -0.7, 0.77, 0.3, -0.9, -0.01$. Give a brief justification for your choices.

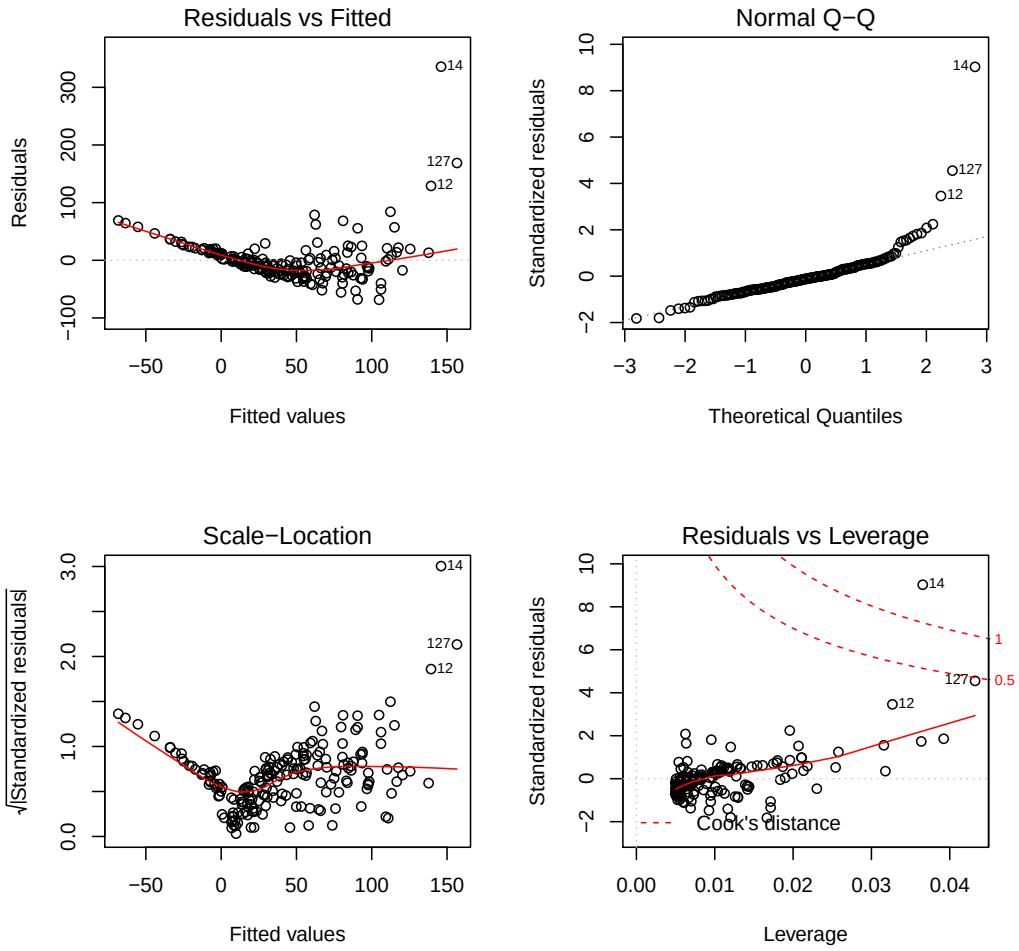
1.3 Residuals (2 points per model)

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

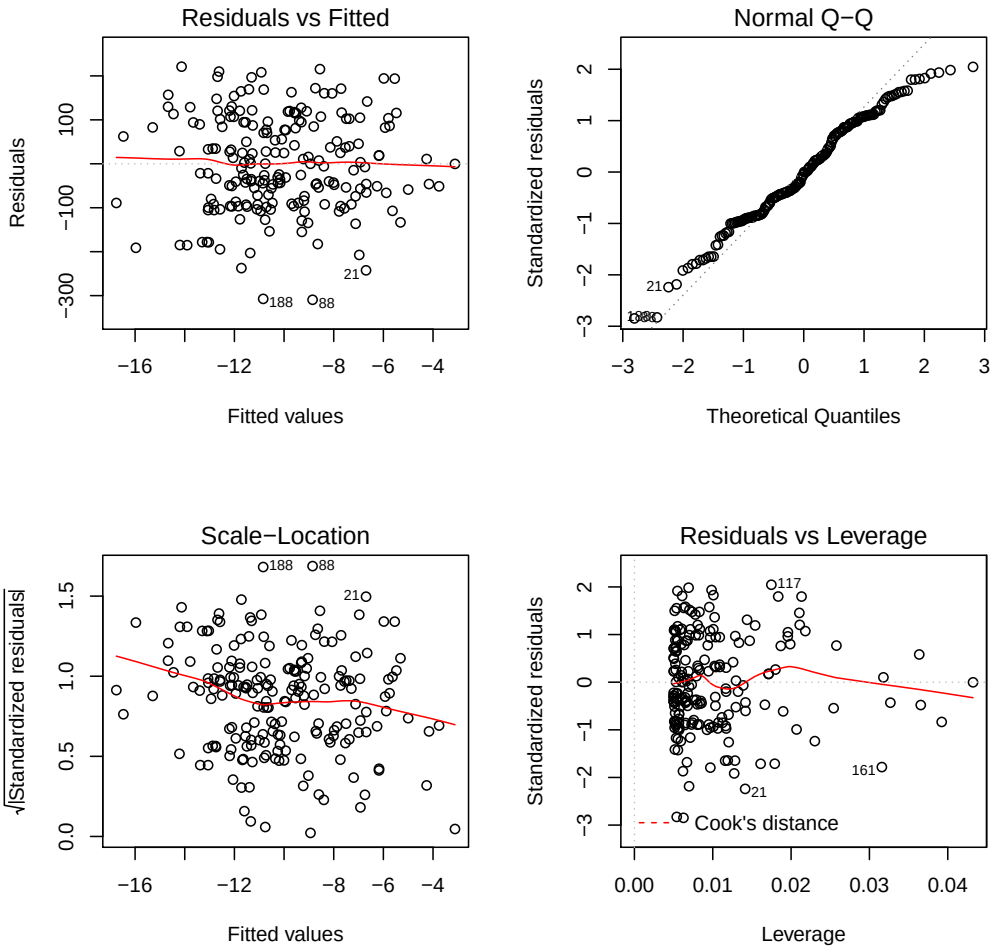
residual plots say which, if any, of the following assumptions are violated: (a) no mean trend, (b) normal distribution and (c) constant variance. Based on your answer, which models are good fits?



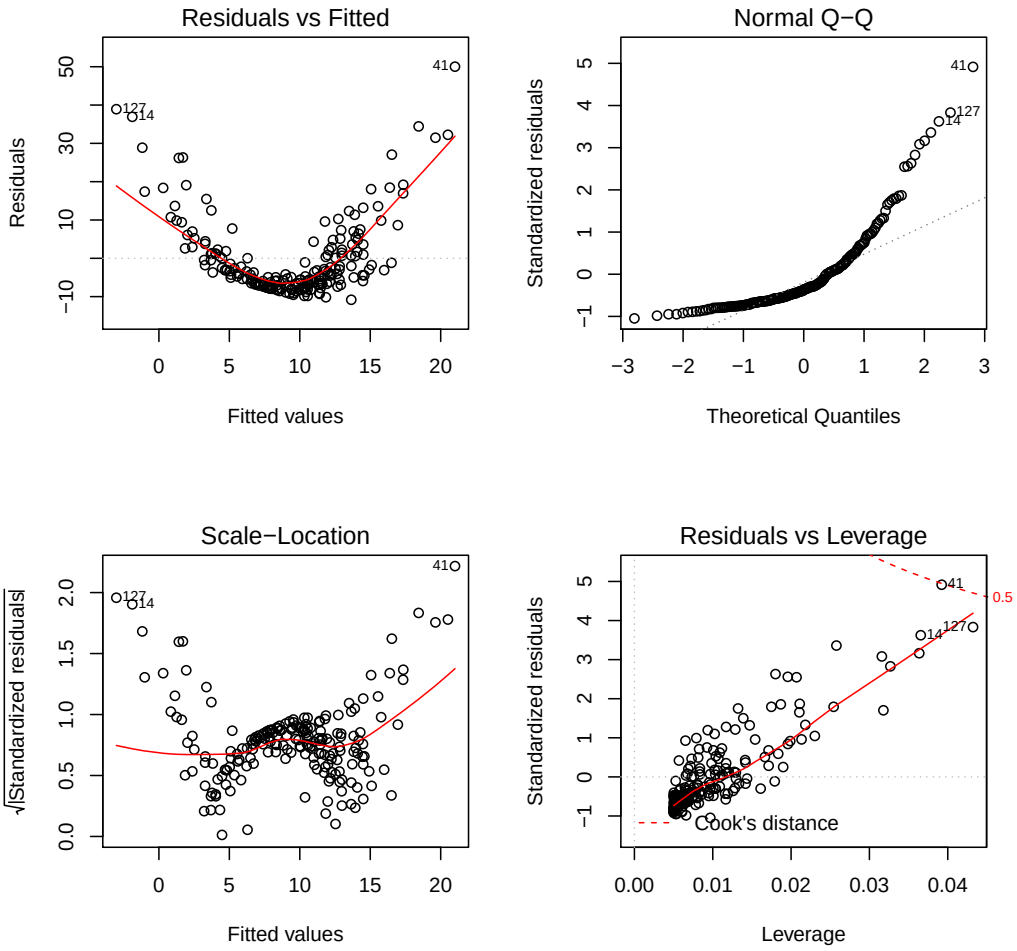
Model 1


Residuals vs Fitted
Residuals
Fitted values
14
127
12
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1
0.5

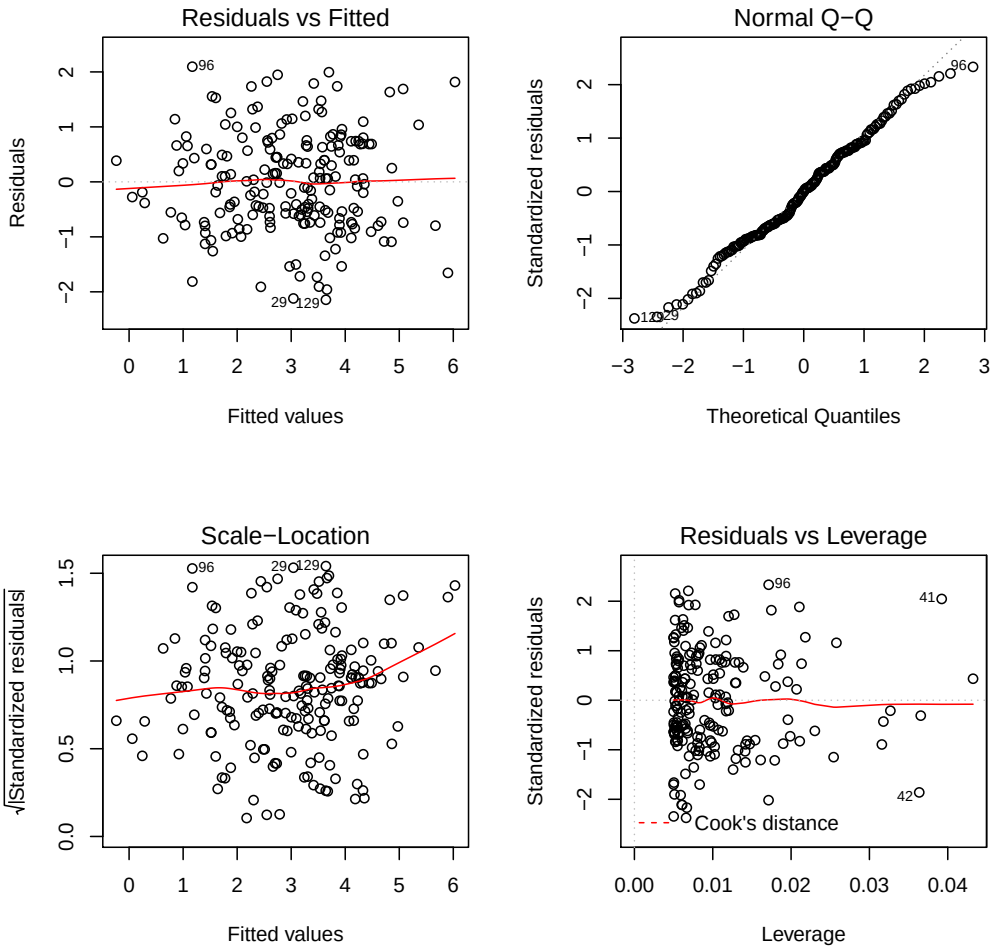
Model 2


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

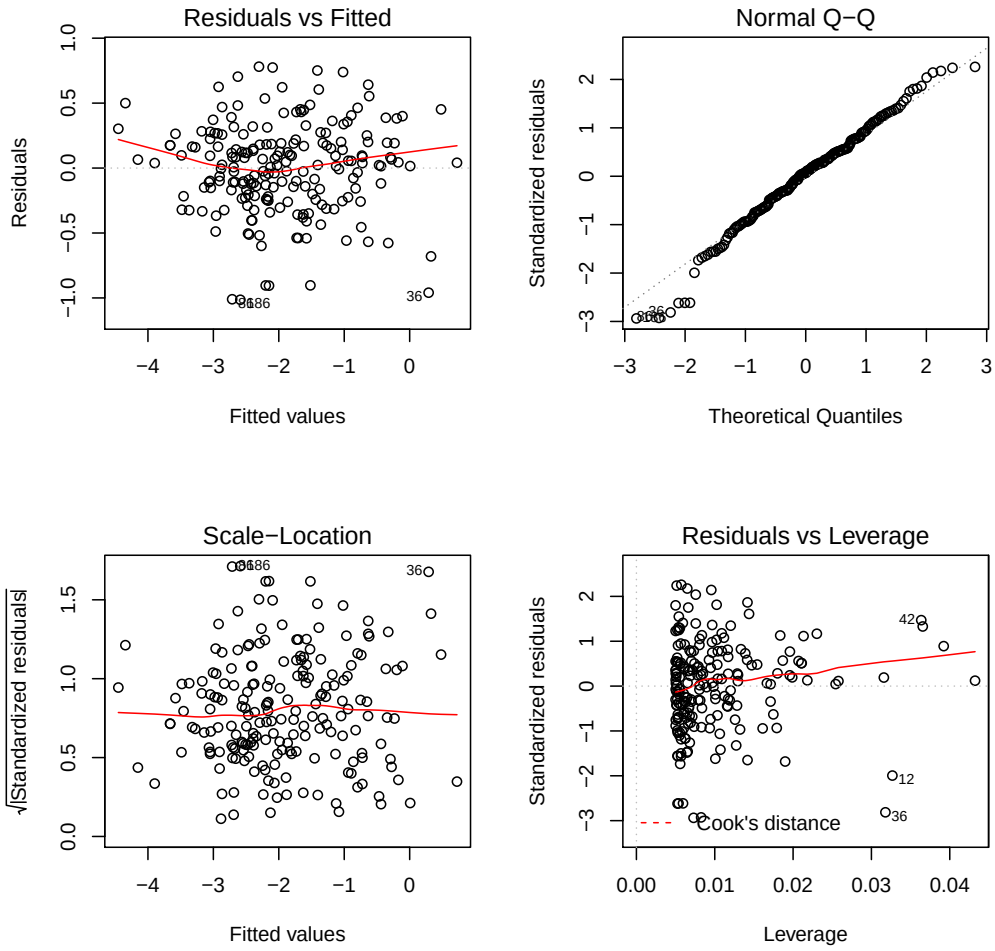
Model 3


Residuals vs Fitted
Residuals
Fitted values
41
127
14
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Leverage

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
Cook's distance
96
41
42

Model 5



Model 6

2 Linear regression in R

Consider the hsb2 dataset.

1. Do a scatter plot of social studies (socst) scores vs math scores (math). Report the correlation. Describe the strength in relationship between these two variables

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.

Bonus question: perform the same analysis for read scores vs math scores. Provide a scatterplot overlaid with the regression line, and a paragraph that summarizes of the relationship between these two variables like question 2.7.