1
RefresheR

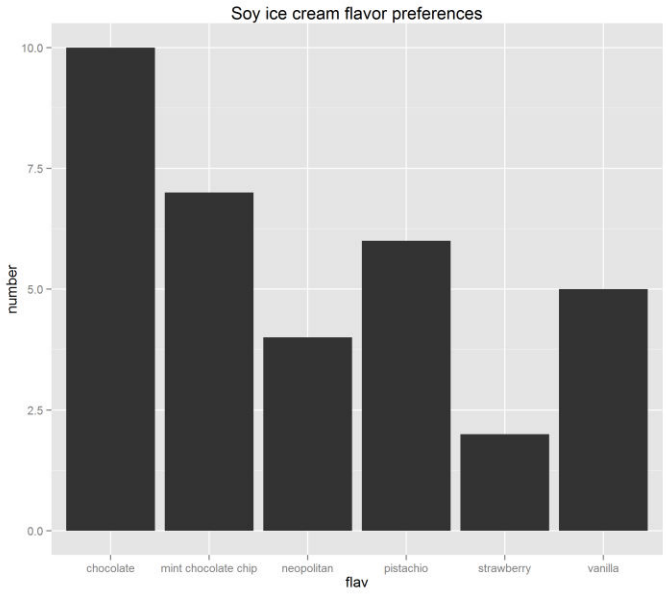


Figure 1.1: Soy ice cream flavor preferences

2
The Shape of Data

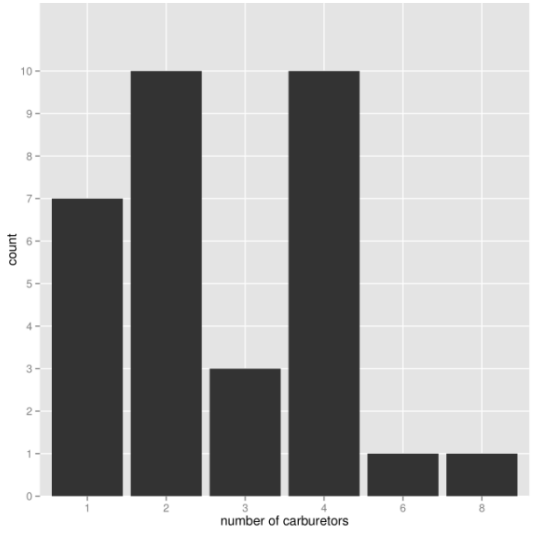


Figure 2.1: Frequency distribution of number of carburetors in mtcars dataset

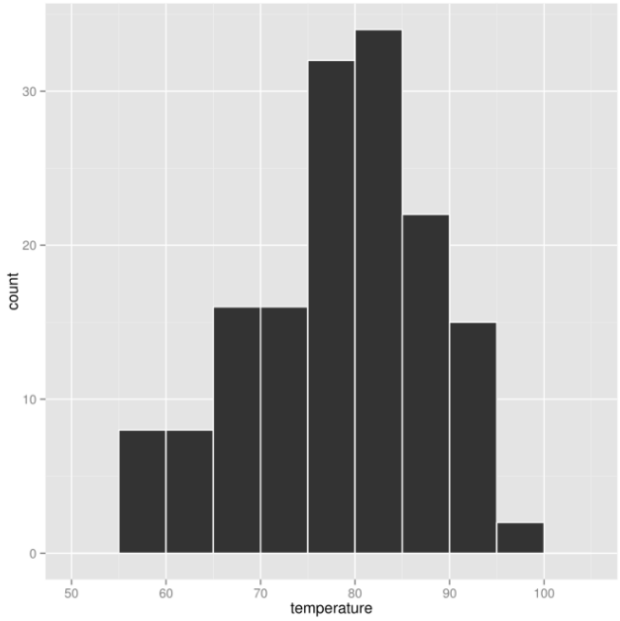


Figure 2.2: Daily temperature measurements from May to September in NYC

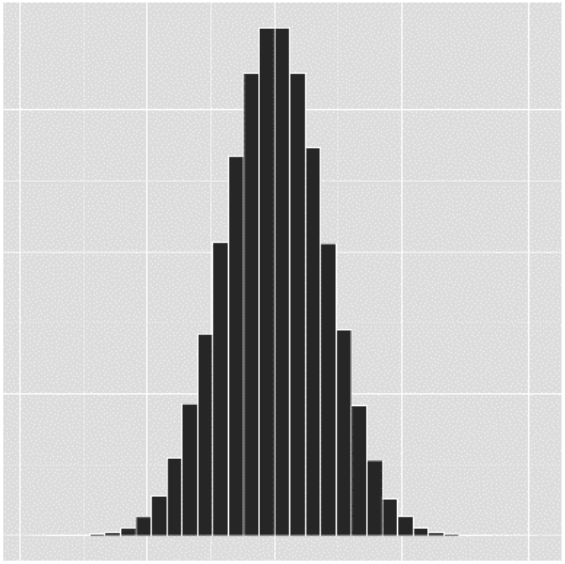


Figure 2.3: A normal distribution

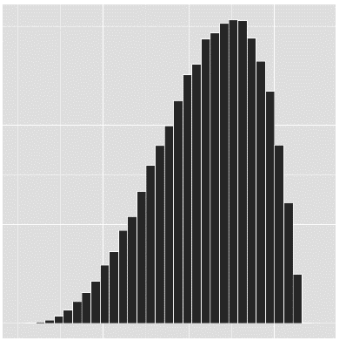


Figure 2.4a: A negatively skewed distribution

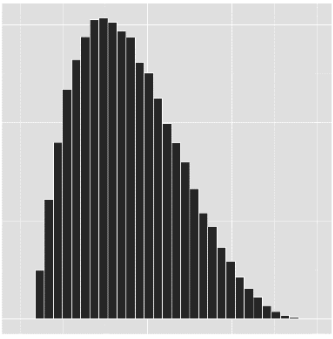


Figure 2.4b: A positively skewed distribution

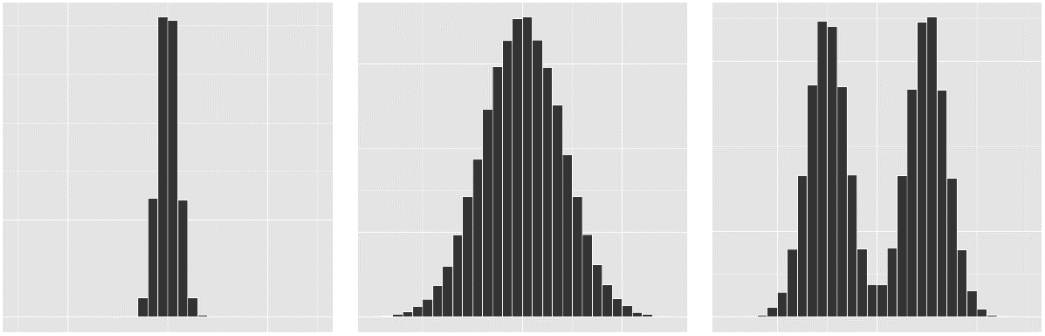


Figure 2.5: three distributions with the same mean and median

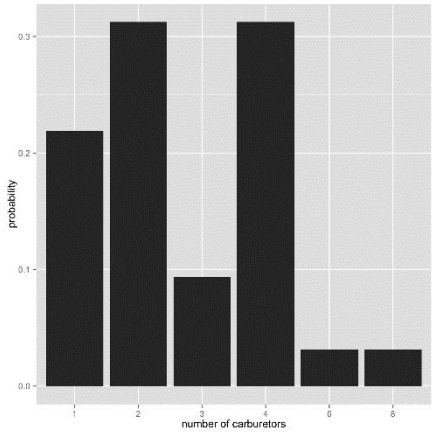


Figure 2.6a: Probability mass function of number of carburetors

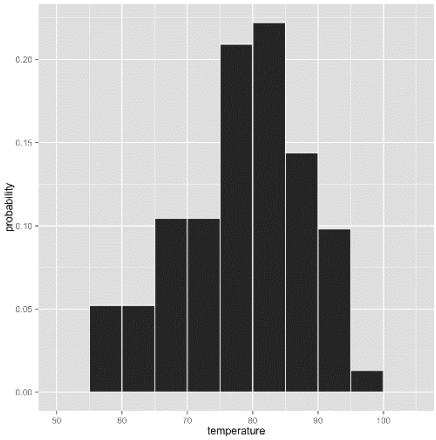


Figure 2.6b: Probability mass function of daily temperature measurements from May to September in NY

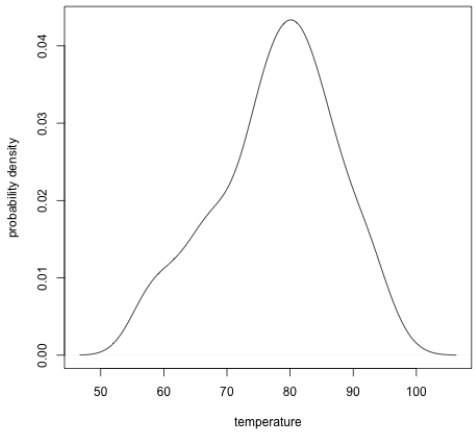


Figure 2.7: Three distributions with the same mean and median

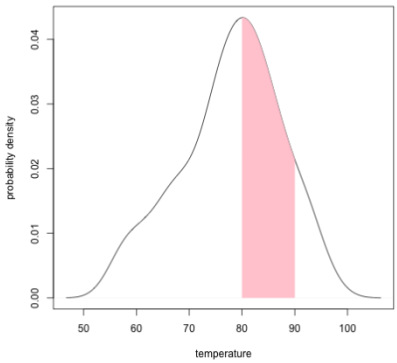


Figure 2.8: PDF with highlighted interval

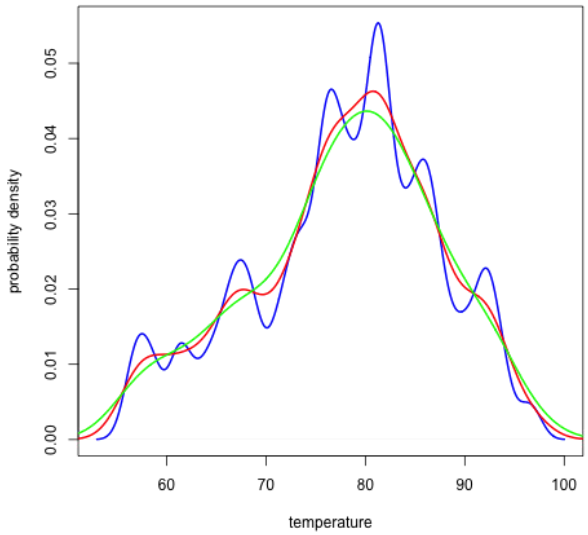


Figure 2.9: Three different bandwidths used on the same data.

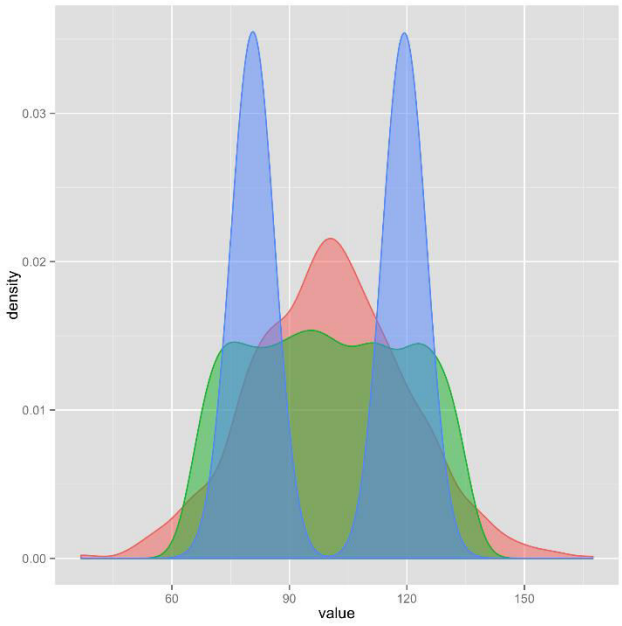


Figure 2.10: Three PDFs with the same mean, median, and standard deviation

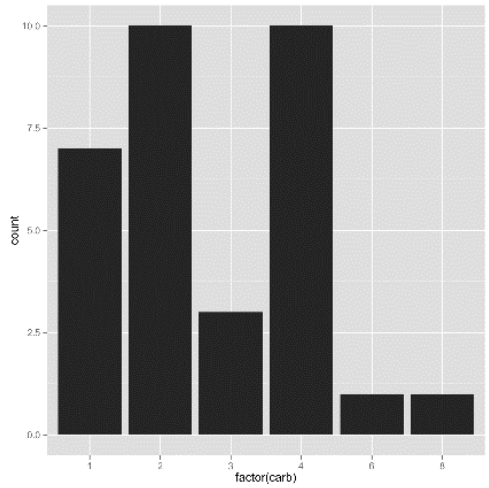


Figure 2.11: Frequency distribution of the number of carburetors

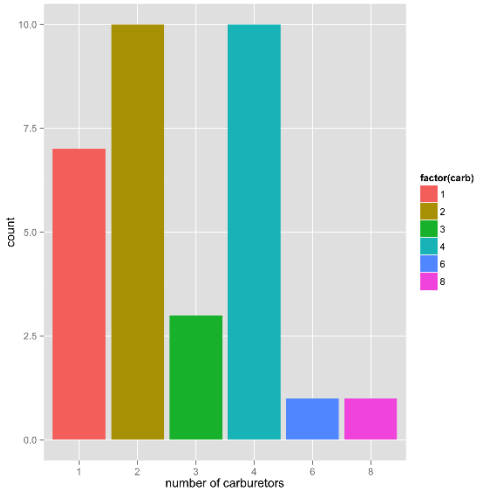


Figure 2.12: With color and label modification

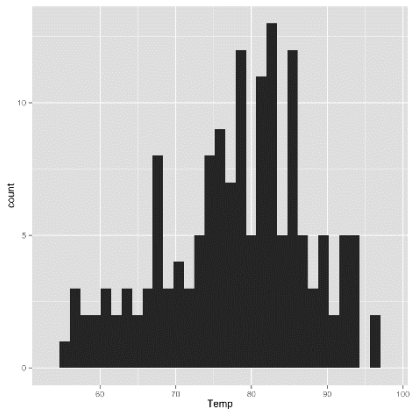


Figure 2.13: Histogram of temperature data

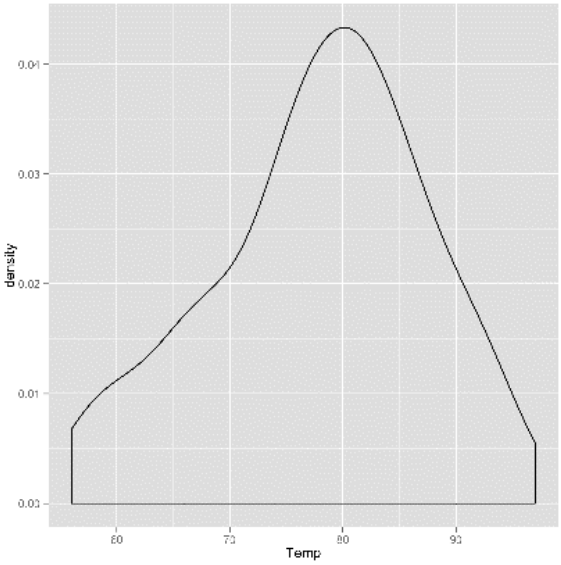


Figure 2.14: PDF of temperature data

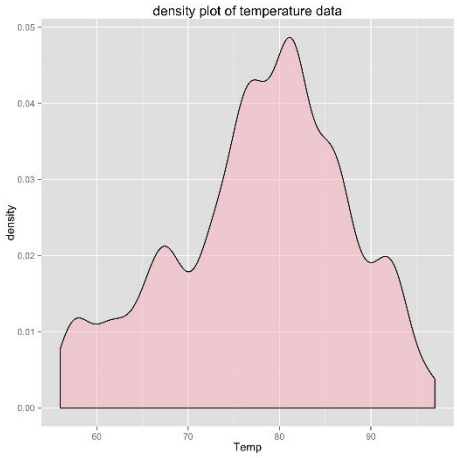


Figure 2.15: Figure 2.14 with modifications

3
Describing Relationships

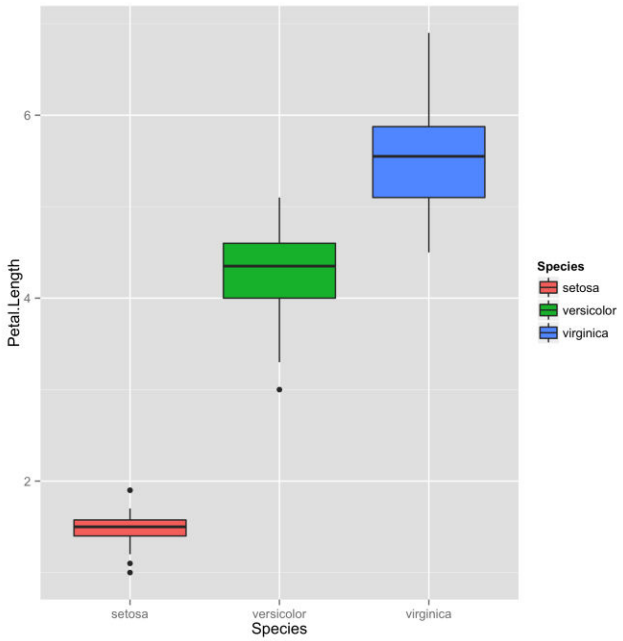


Figure 3.1: A box-and-whisker plot depicting the relationship between the petal lengths of the different iris species in iris dataset

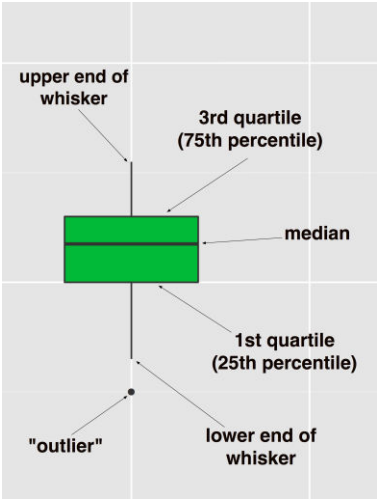


Figure 3.2: The anatomy of a box plot

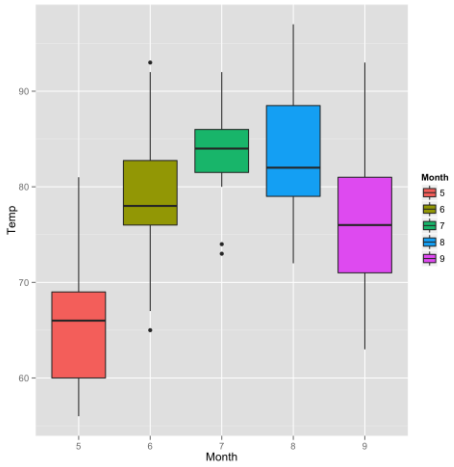


Figure 3.3: A Box plot of NYC temperatures across months (May to September)

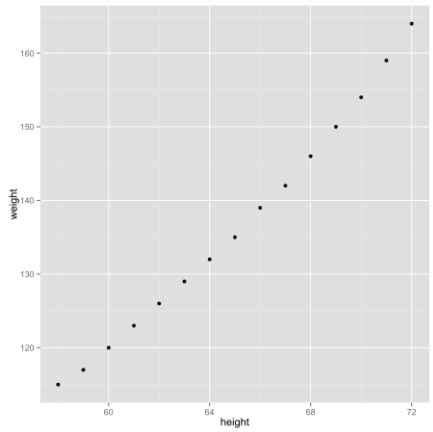


Figure 3.4: Scatterplot of women's heights and weights

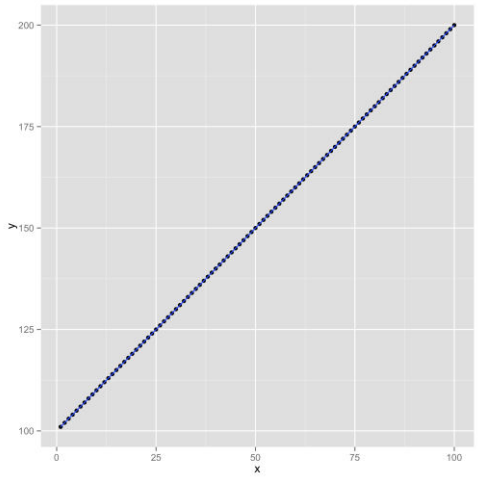


Figure 3.5: Scatterplot of y=x + 100 with regression line. r and rho are both 1

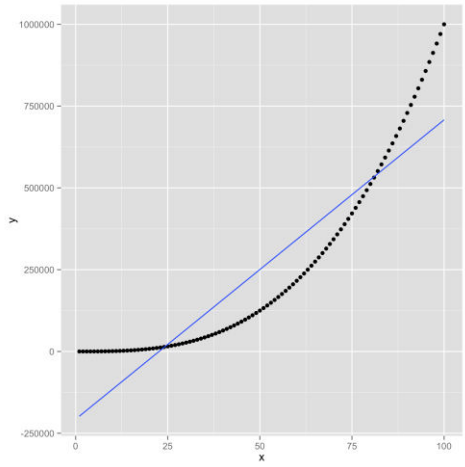
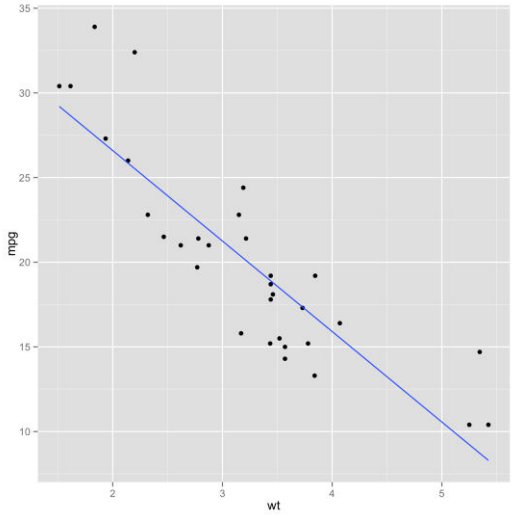


Figure 3.6: Scatterplot of $y = x^3$ with regression line. r is .92, but rho is 1



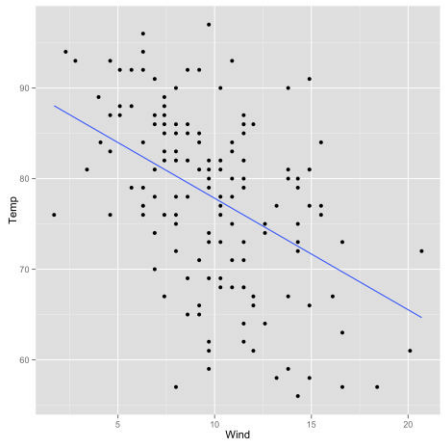


Figure 3.8: Scatterplot of the relationship between wind speed and temperature

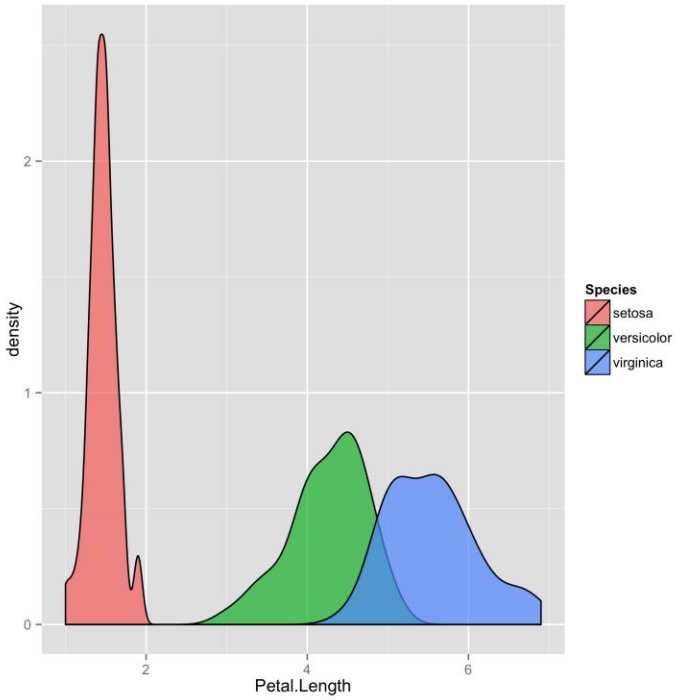


Figure 3.9: Overlapping density plot of petal length of iris flowers across species

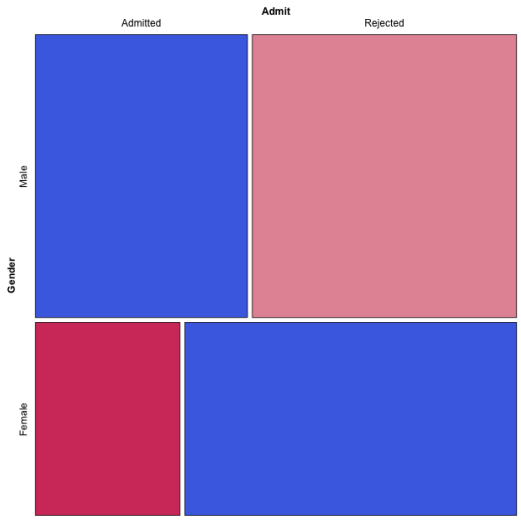


Figure 3.10: A mosaic plot of the UCBAdmissions dataset (across all departments)

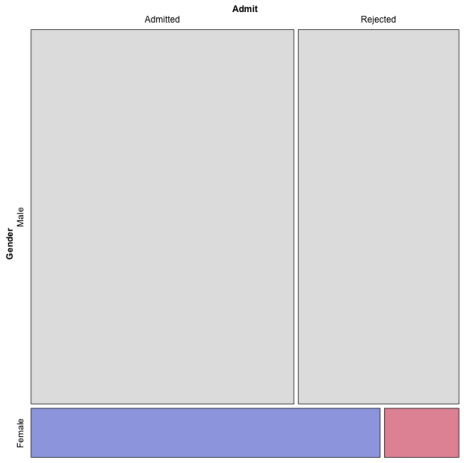


Figure 3.11: A mosaic plot of the UCBAdmissions dataset for department A

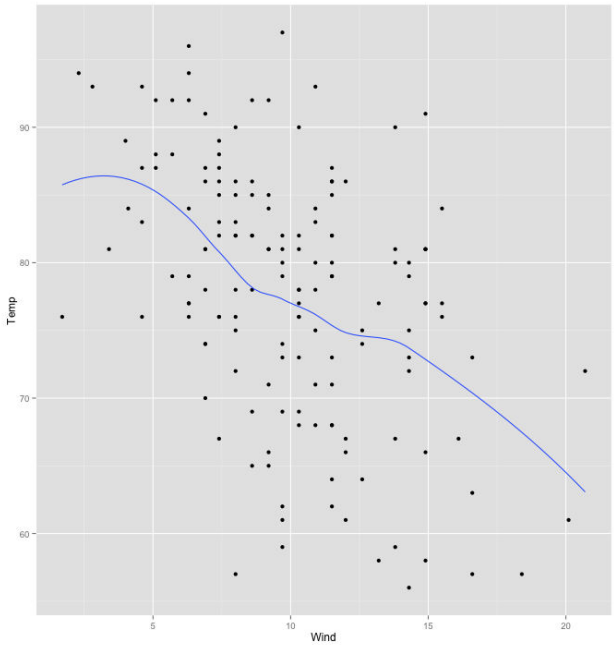


Figure 3.12: A scatterplot of the relationship between the weight of a car, and its miles per gallon, and a trend-line smoothed with LOESS

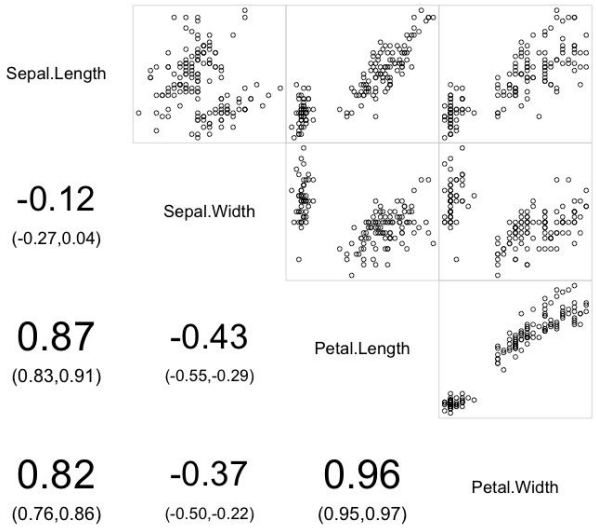


Figure 3.13: A corrgram of the iris data set's continuous variables

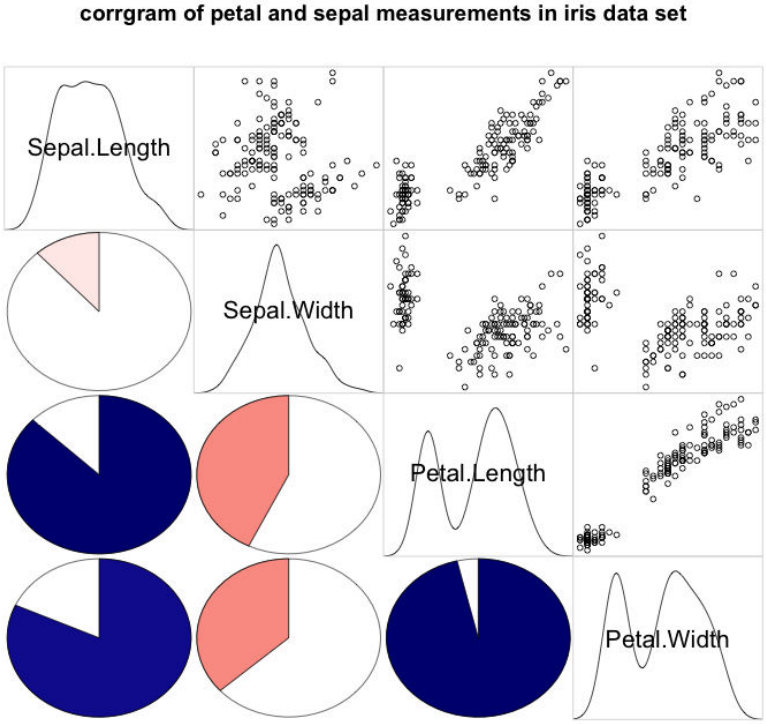


Figure 3.14: Another corrgram of the iris data set's continuous variables

4
Probability

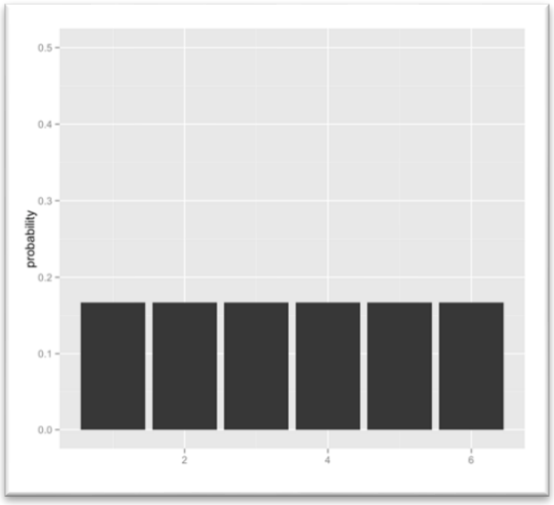


Figure 4.1: Probability distribution of outcomes of a die roll

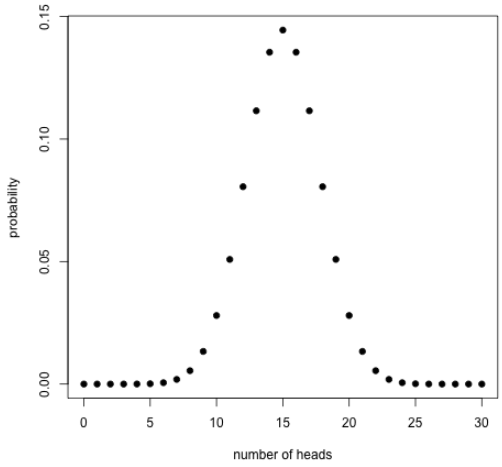


Figure 4.2: A binomial distribution (n=30, p=0.5)

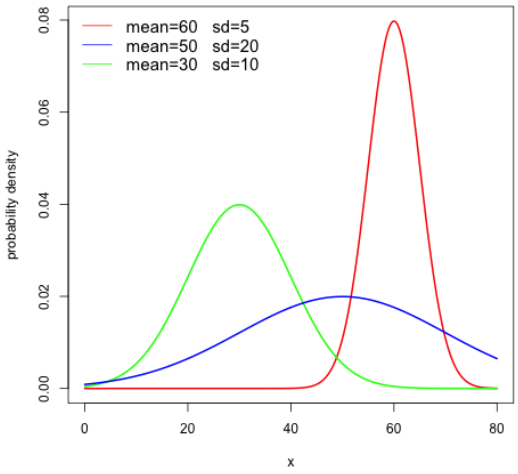


Figure 4.3: Normal distributions with different parameters

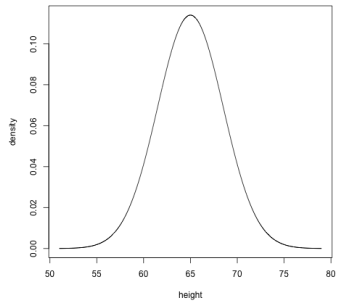


Figure 4.4: Normal distributions with different parameters

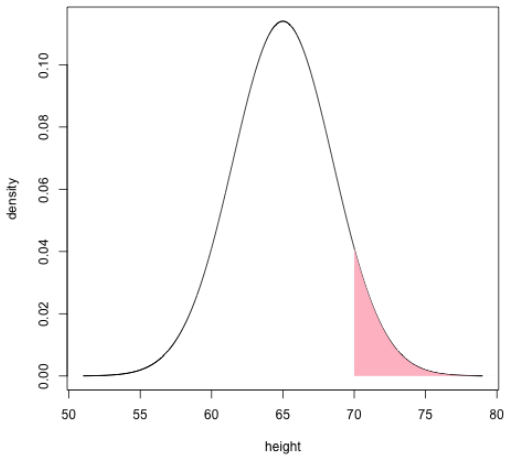


Figure 4.5: Area under the curve of the height distribution from 70 inches to positive infinity

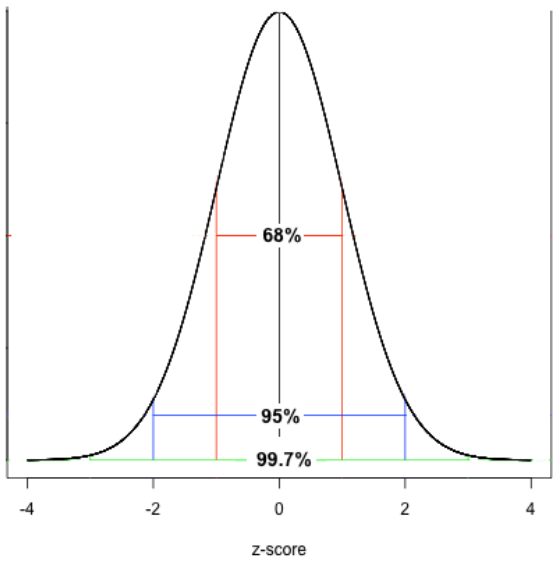


Figure 4.6: The three-sigma rule

5

Using Data to Reason About The World

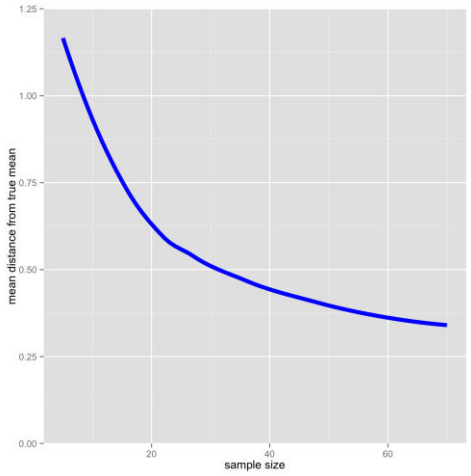


Figure 5.1: Accuracy of sample means as a function of sample size

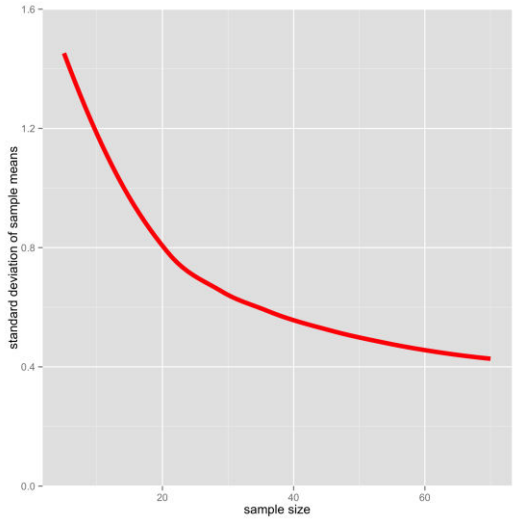


Figure 5.2: The variability of sample means as a function of sample size

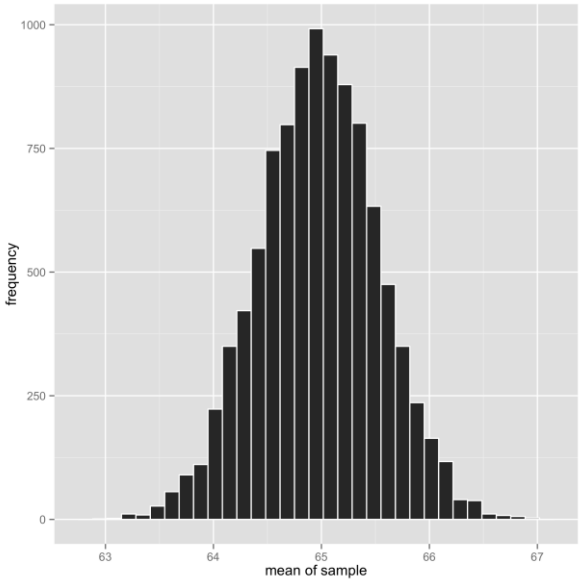


Figure 5.3: The sampling distribution of sample means

This frequency distribution is called a *sampling distribution*. In particular, since we used sample means as the value of interest, this is called the sampling distribution of the

sample means (whew!!). You can create a sampling distribution of any statistic (median, variance, and so on), but when we refer to sampling distributions throughout this chapter,

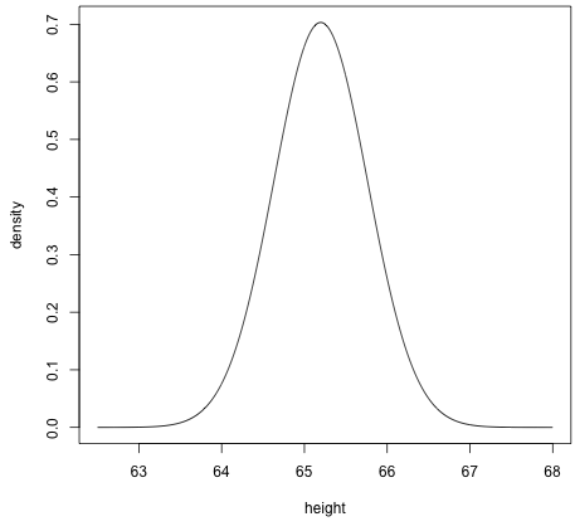


Figure 5.4: Estimated sampling distribution of sample means based on one sample

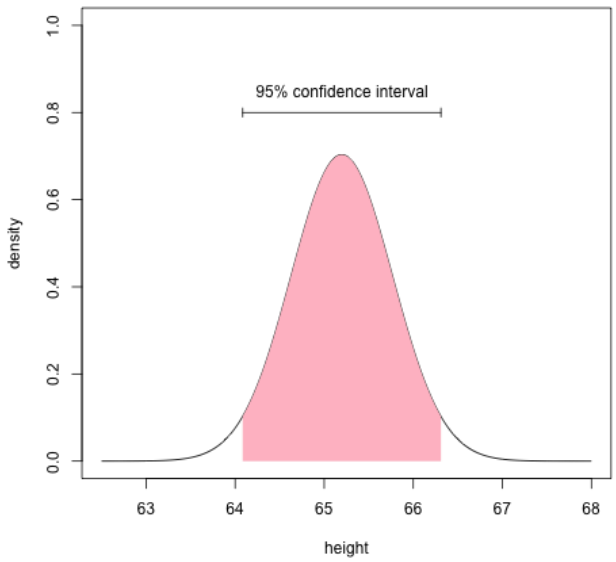


Figure 5.5:The 95% confidence interval of our estimate of the sample mean (64.085 to 66.31) covers 95% of the area in the our estimated sampling distribution

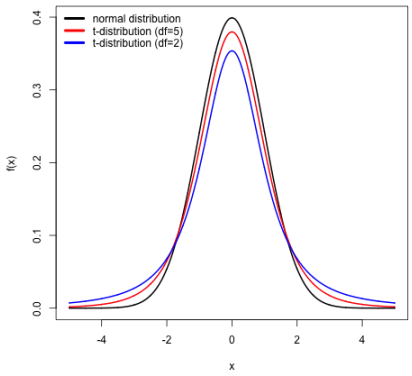


Figure 5.6: The normal distribution, and two t-distributions with different degrees of freedom

6
Testing Hypotheses

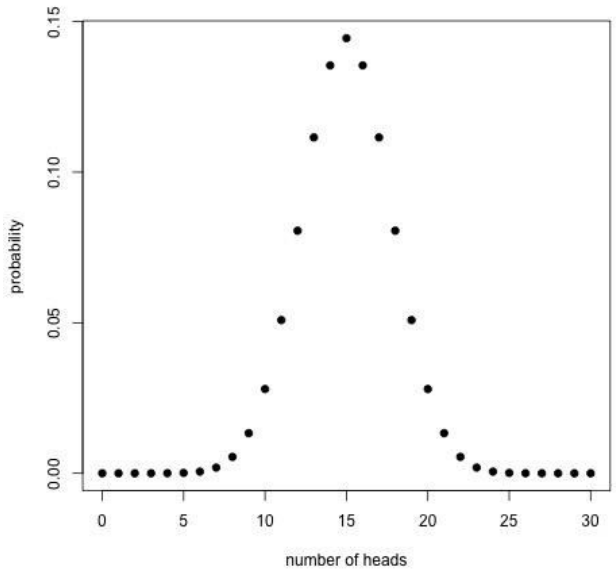


Figure 6.1: The sampling distribution of our coin-flip test statistic (the number of heads)

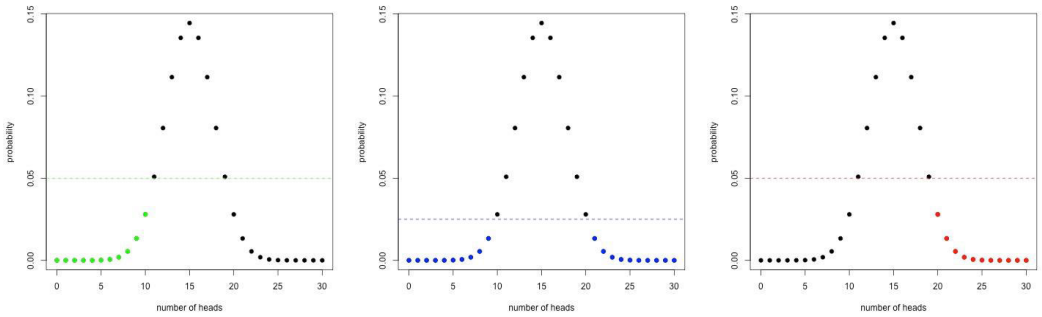


Figure 6.2: The three panels, from left to right, depict the critical regions of the left ("lesser") one-tailed, two-tailed, and right ("greater") alternative hypotheses. The dashed horizontal line is meant to show that, for the two-tailed tests, the critical region starts below p=.025, because it is being split between two tails. For the one-tailed tests, the critical region is below the dashed horizontal line at p=.05.

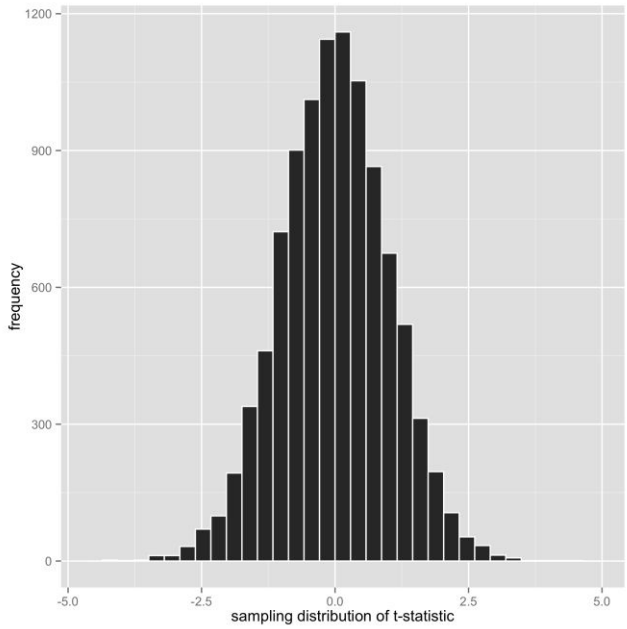


Figure 6.3: The sampling distribution of the t-statistic

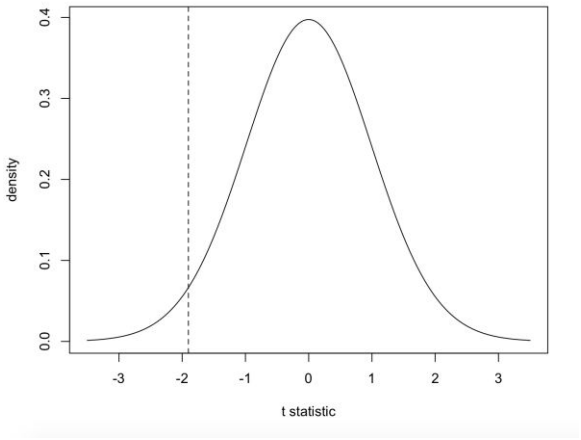


Figure 6.4: The t-distribution with 69 degrees of freedom. The t-statistic of our sample is shown as the dashed line

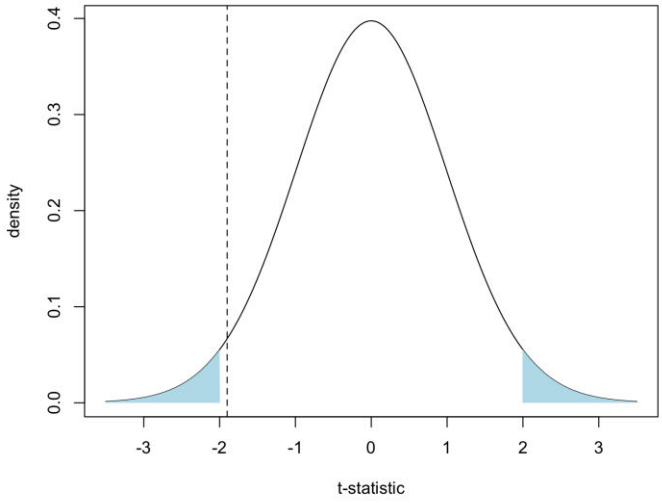


Figure 6.5: The previous figure with the critical region for non-directional hypothesis highlighted

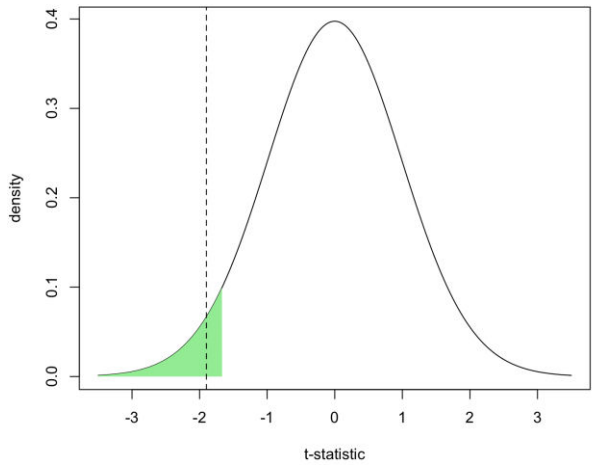


Figure 6.6: figure 6.4 with directional critical region highlighted

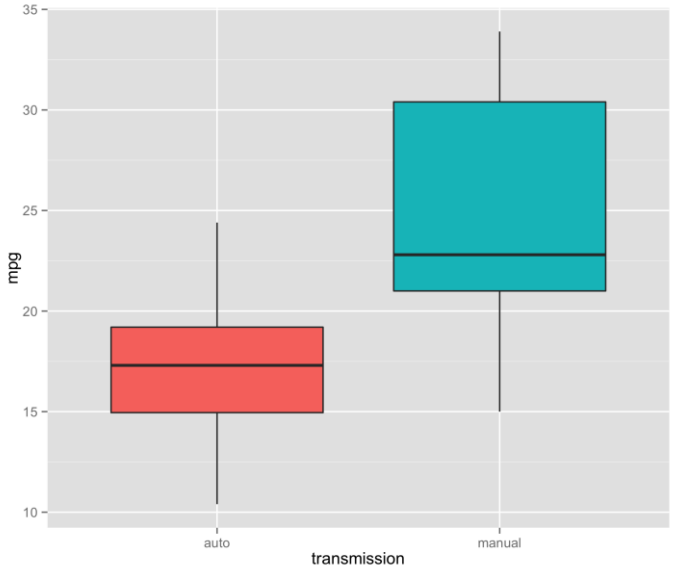


Figure 6.7: Boxplot of the miles per gallon ratings for automatic cars and cars with manual transmission

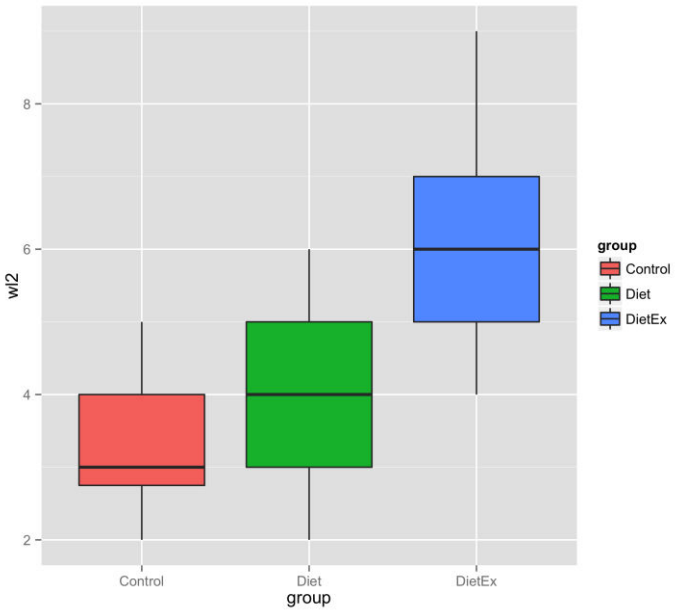


Figure 6.8: Boxplot of weight lost in week 2 of trial for three groups: control, diet, and diet & exercise

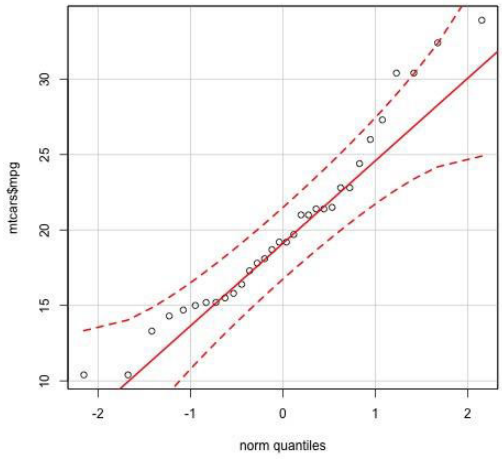


Figure 6.9: A QQ-plot of the mile per gallon vector in mtcars

7
Bayesian Methods

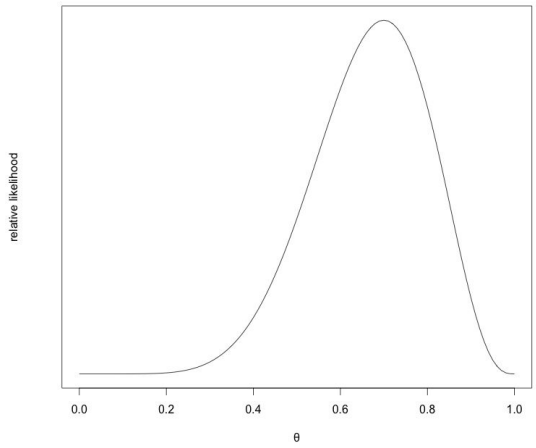


Figure 7.1: The likelihood function of theta for 20 out of 30 successful Bernoulli trials.

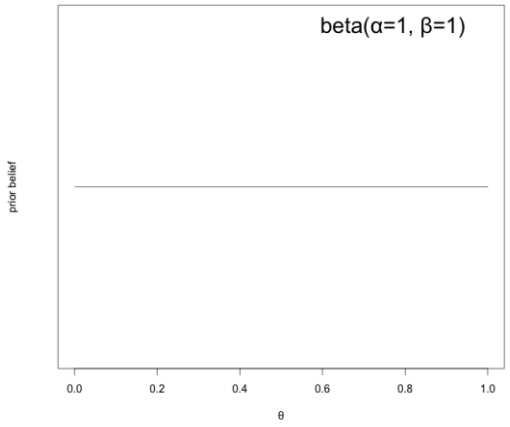


Figure 7.2: A flat prior on the value of theta. This beta distribution, with alpha and beta = 1 confers an equal level of credibility to all possible values of theta, our parameter of interest.

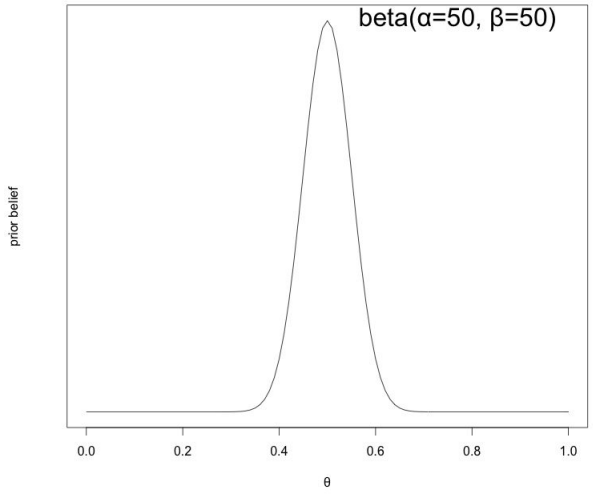


Figure 7.3: A skeptic's prior

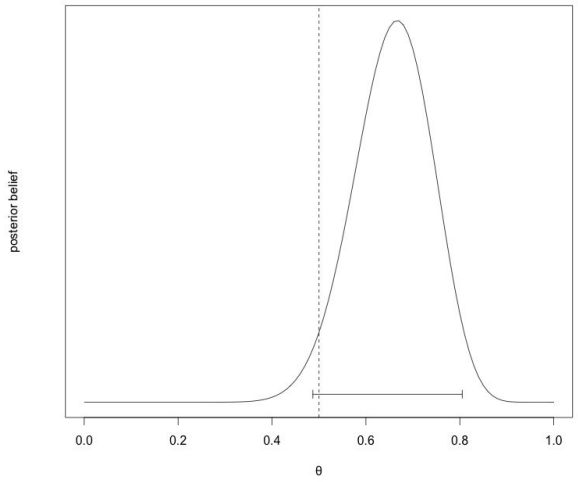


Figure 7.4: The result of the Bayesian update of the evidence and prior #1. The interval depicts the 95% credible interval (the densest 95% of the area under the posterior distribution). This interval overlaps slightly with theta = 0.5.

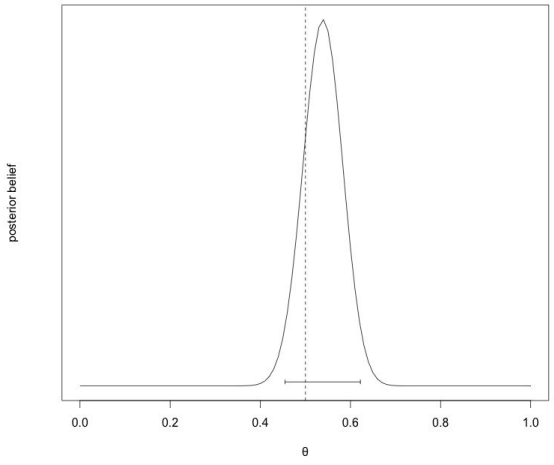


Figure 7.5: Posterior distribution of theta using prior #2

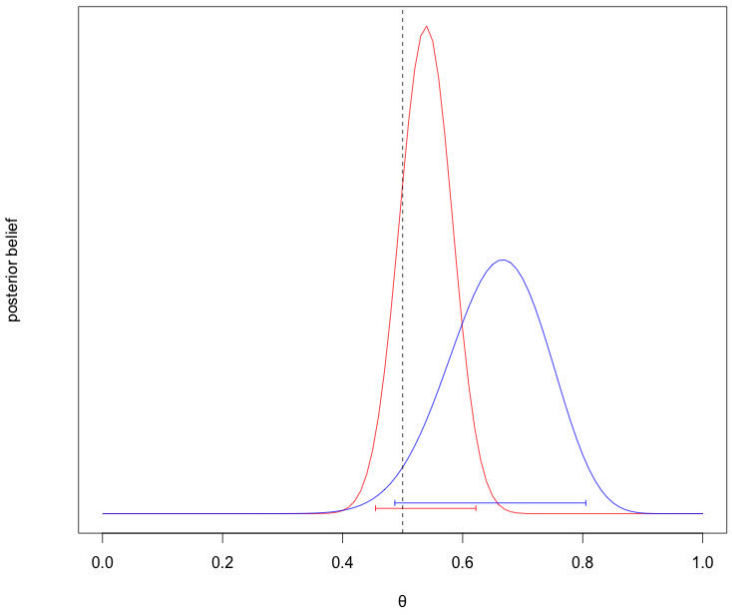


Figure 7.6: The posterior distributions from prior #1 and #2

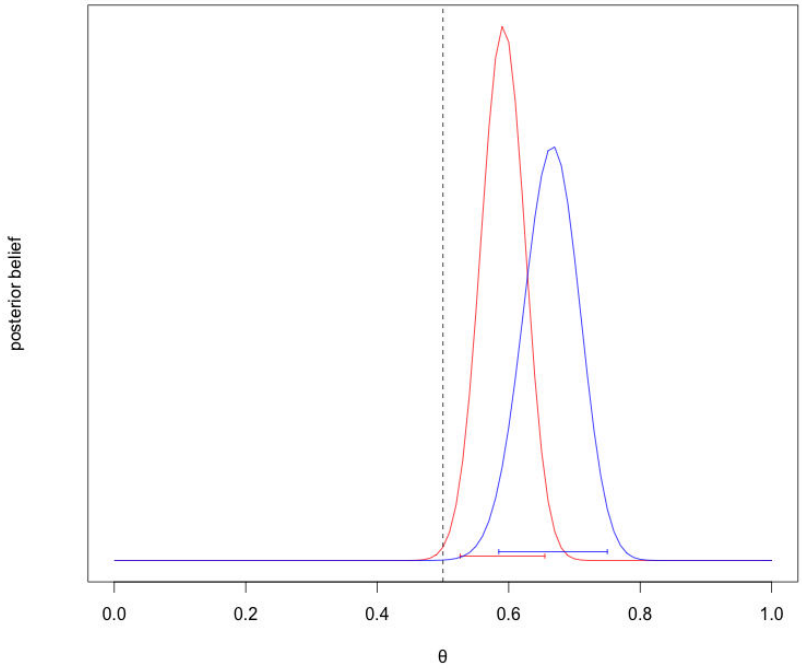


Figure 7.7: The posterior distributions from prior #1 and #2 with more evidence

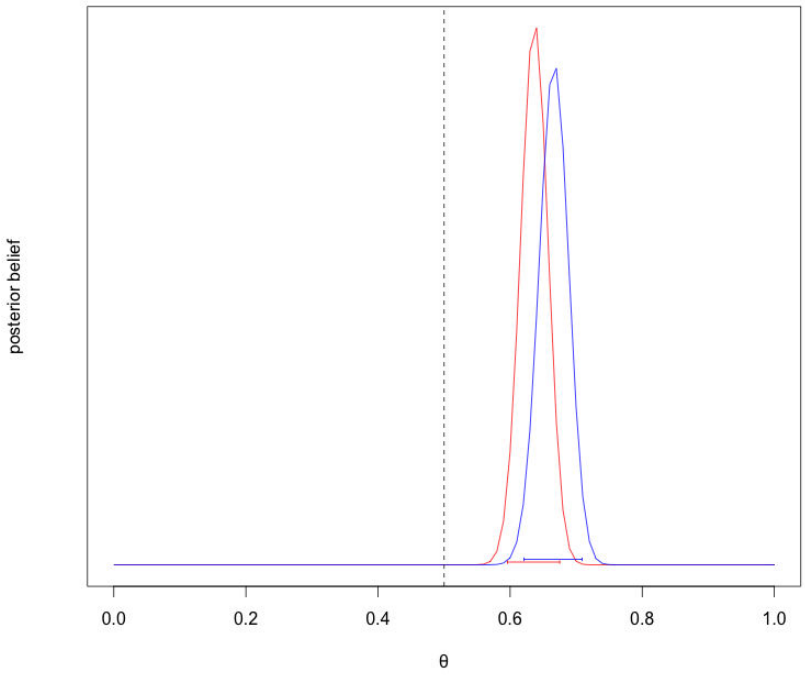


Figure 7.8: The posterior distributions from prior #1 and #2 with even more evidence

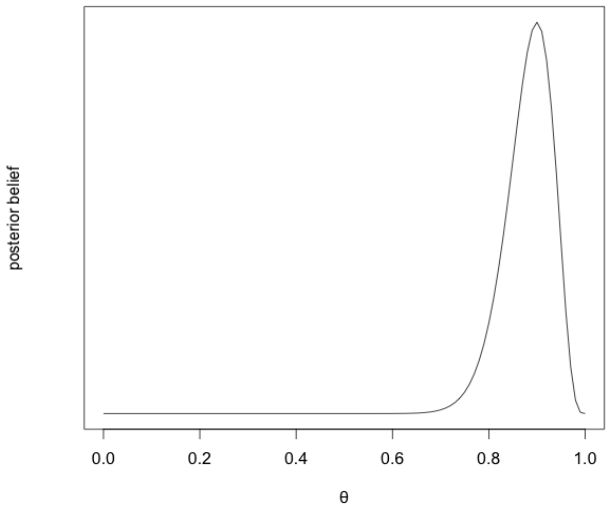


Figure 7.9: The posterior distribution of the effectiveness of my recommendations using a uniform prior

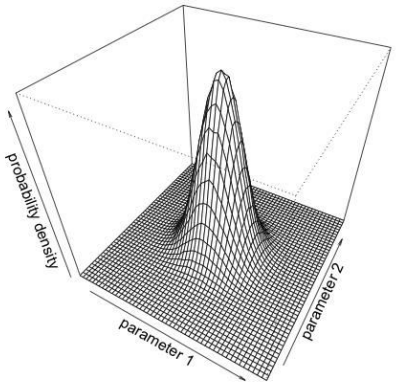


Figure 7.10: A bivariate normal distribution

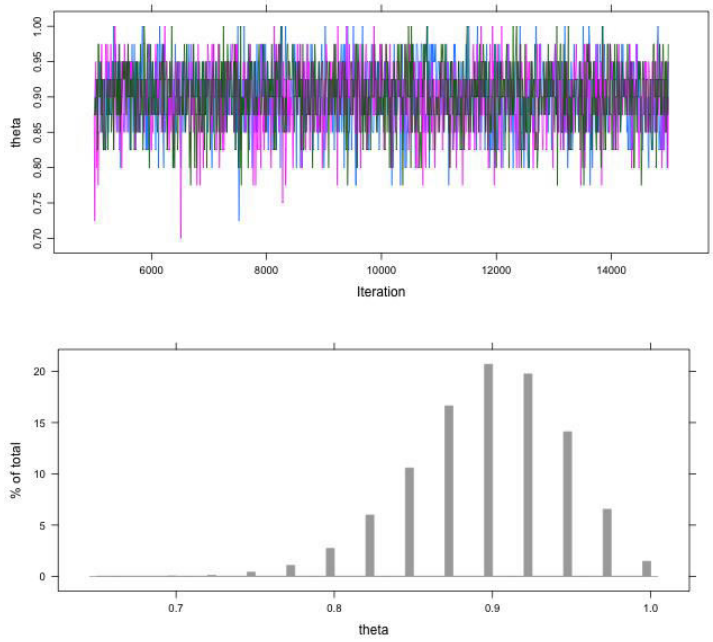


Figure 7.11: Output plots from the MCMC results. The top is a trace plot of theta values along the chain's length. The bottom is a bar plot depicting the relative credibility of different theta values.

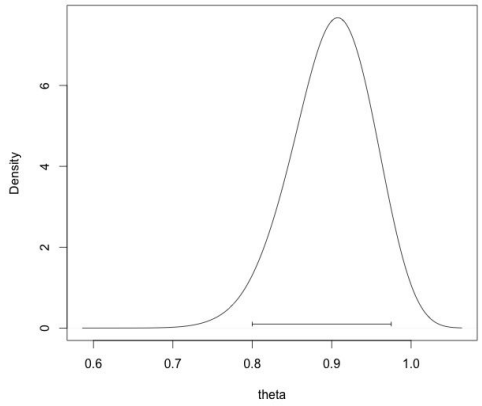


Figure 7.12: Density plot of the posterior distribution. Note that the x-axis starts here at 0.6

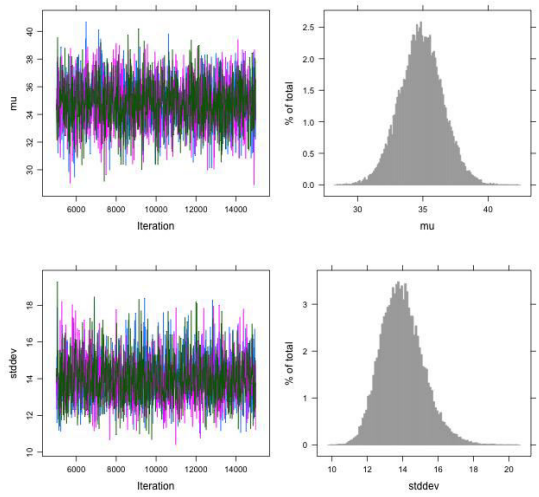


Figure 7.13: Output plots from the MCMC result of fitting a normal curve to the built-in precipitation data set.

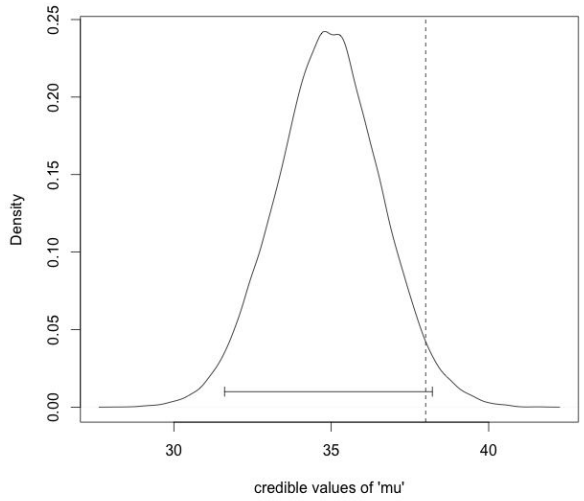


Figure 7.14: Marginal distribution of posterior for parameter ‘mu’. Dashed line shows hypothetical population mean within 95% credible interval

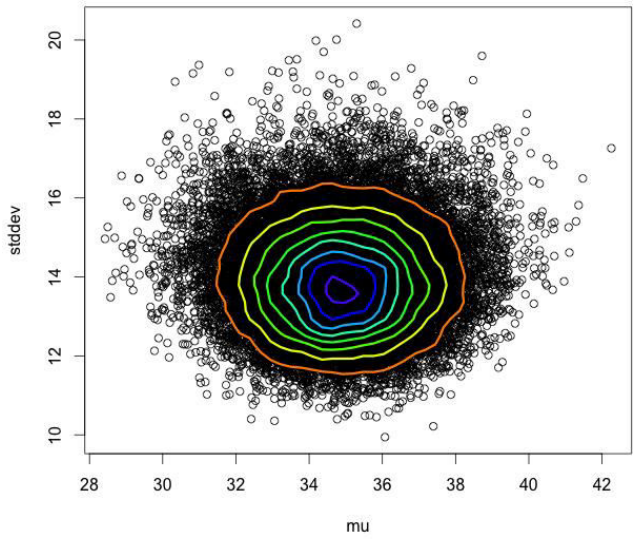


Figure 7.15: Contour plot of the joint posterior distribution. The purple contour corresponds to the region with the highest probability density.

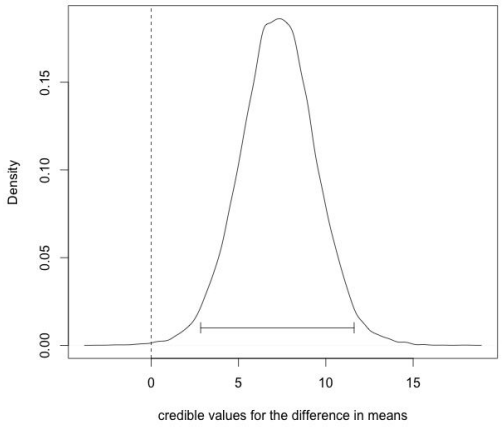


Figure 7.16: Credible values for the difference in means of the gas mileage between automatic and manual cars. The dashed line is at a difference of zero.

8

Predicting Continuous Variables

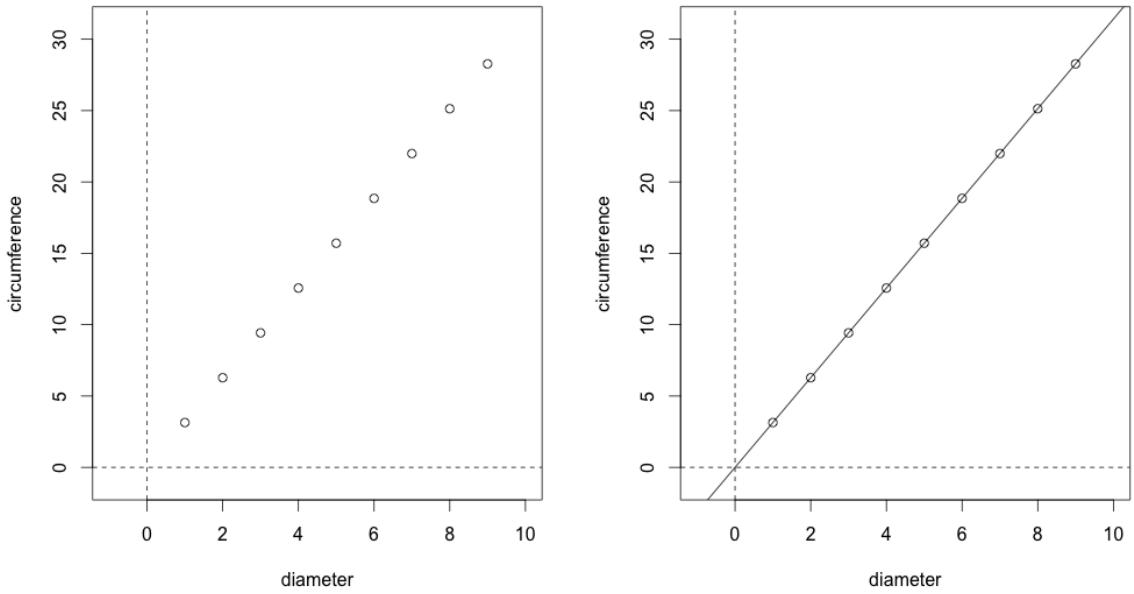


Figure 8.1: (left) A scatterplot of diameters and circumferences of No Scone Unturned's cookies; (right) the same plot with a best fit regression line plotted over the data points

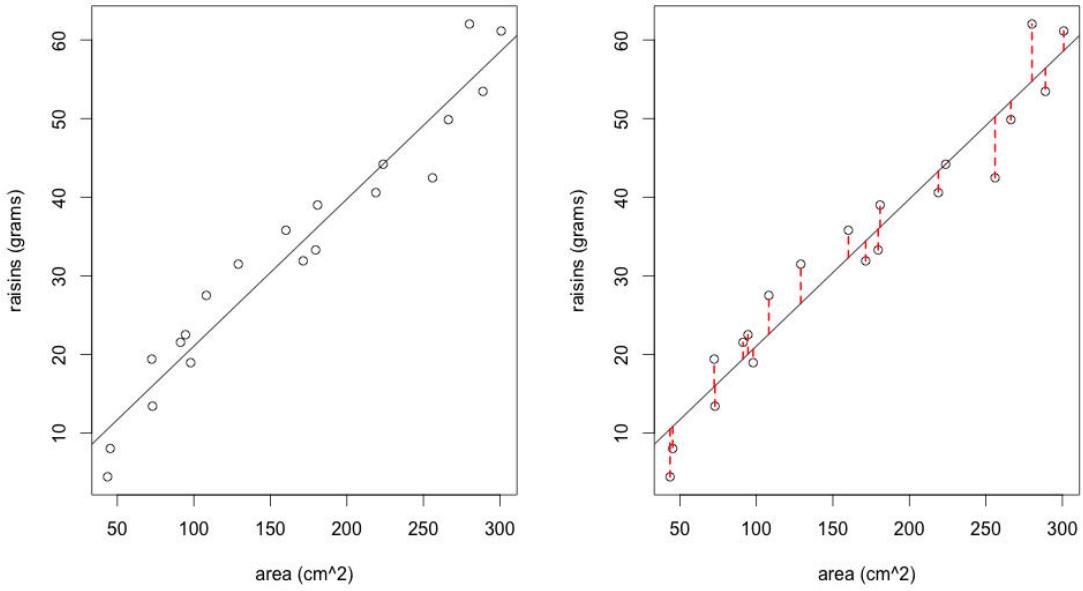


Figure 8.2: (left) A scatterplot of areas and grams of raisins in No Scone Unturned's cookies with a best-fit regression line; (right) the same plot with highlighted residuals.

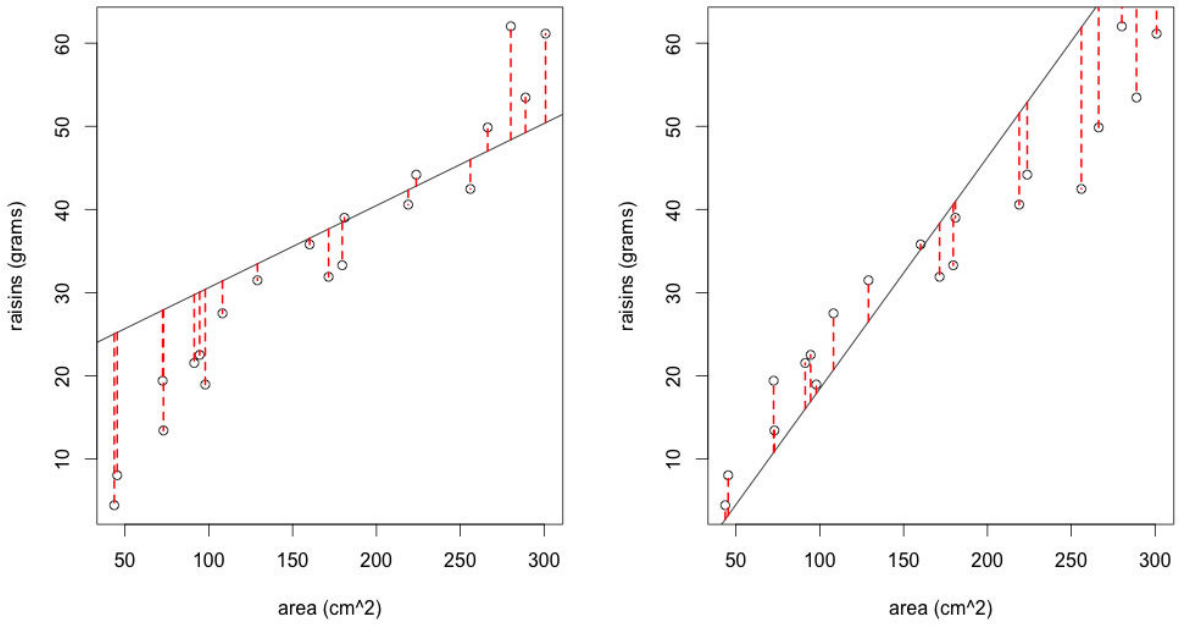


Figure 8.3: Two regression lines that do not minimize the RSS

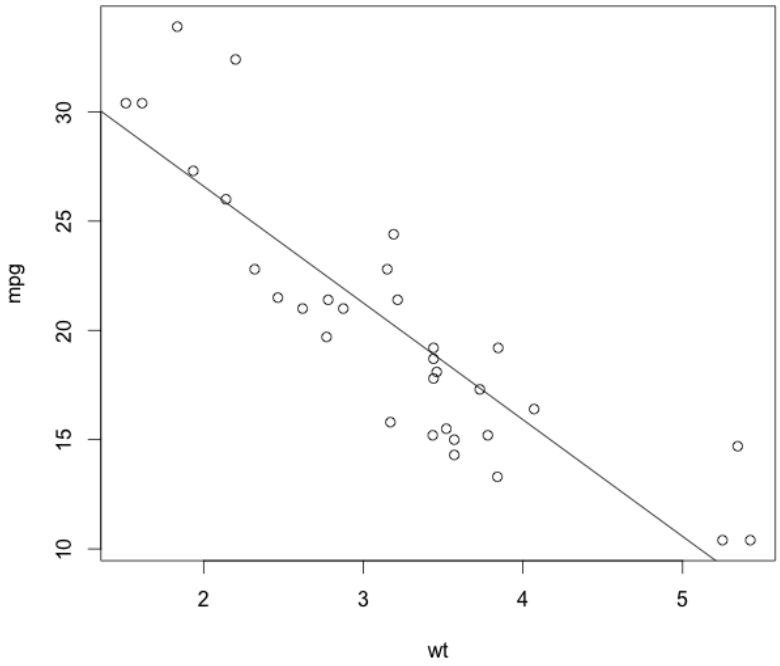


Figure 8.4: The result of plotting output from `lm`

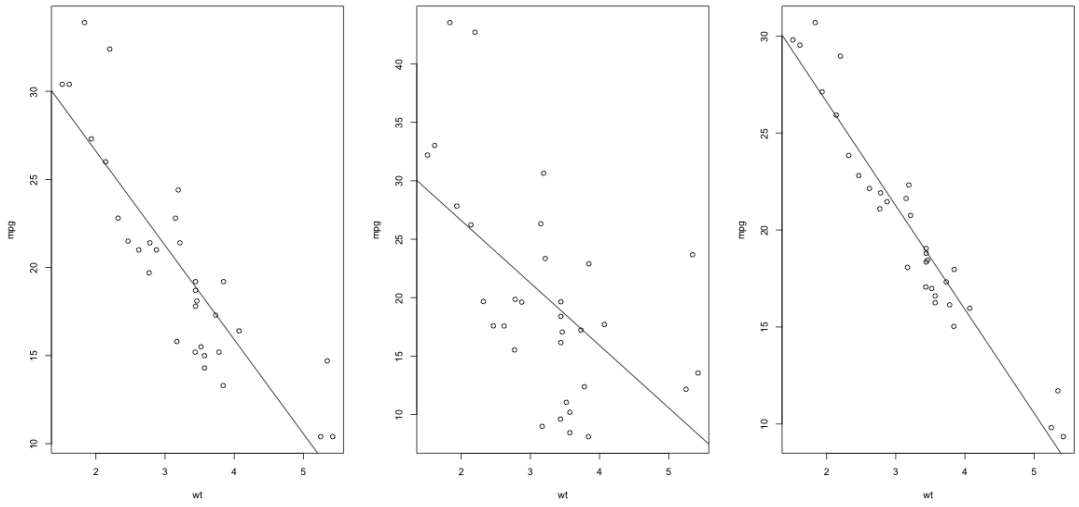


Figure 8.5: Linear models (from left to right) with R^2s of 0.75, 0.33, and 0.92

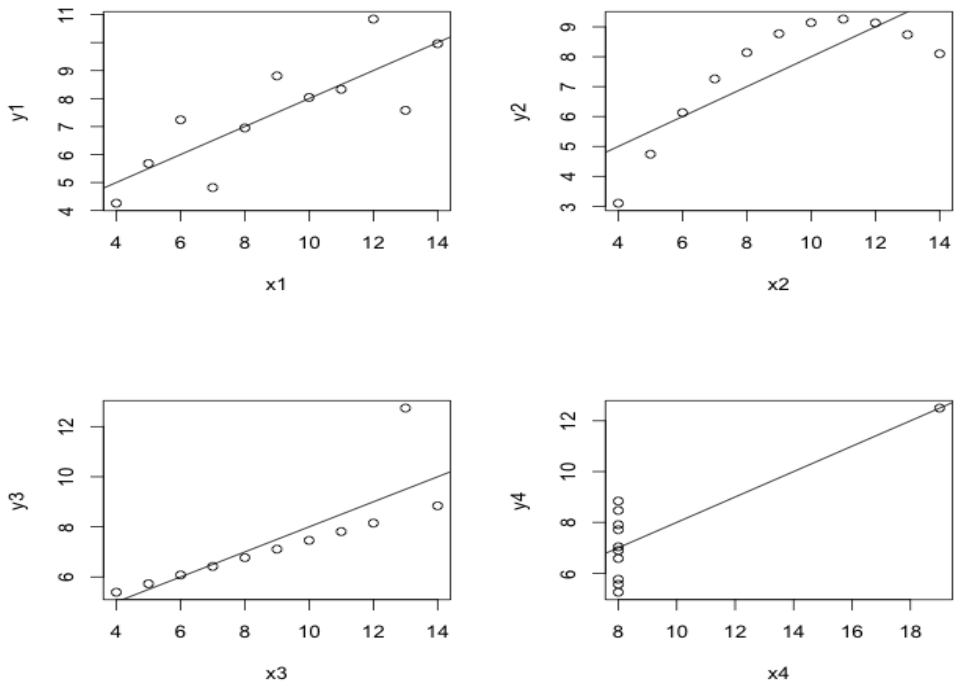


Figure 8.6: Four datasets with identical means, standard deviations, regression coefficients, and R^2

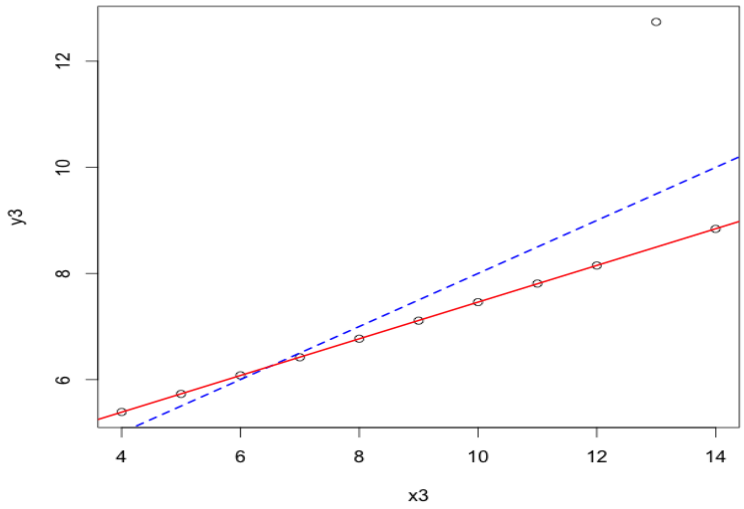


Figure 8.7: The difference between linear regression fit with OLS and a robust linear regression fitted with IWLS

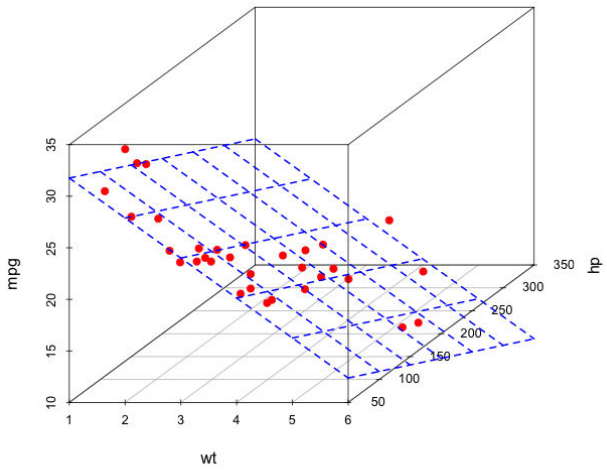


Figure 8.8: The prediction region that is formed by a two-predictor linear model is a plane

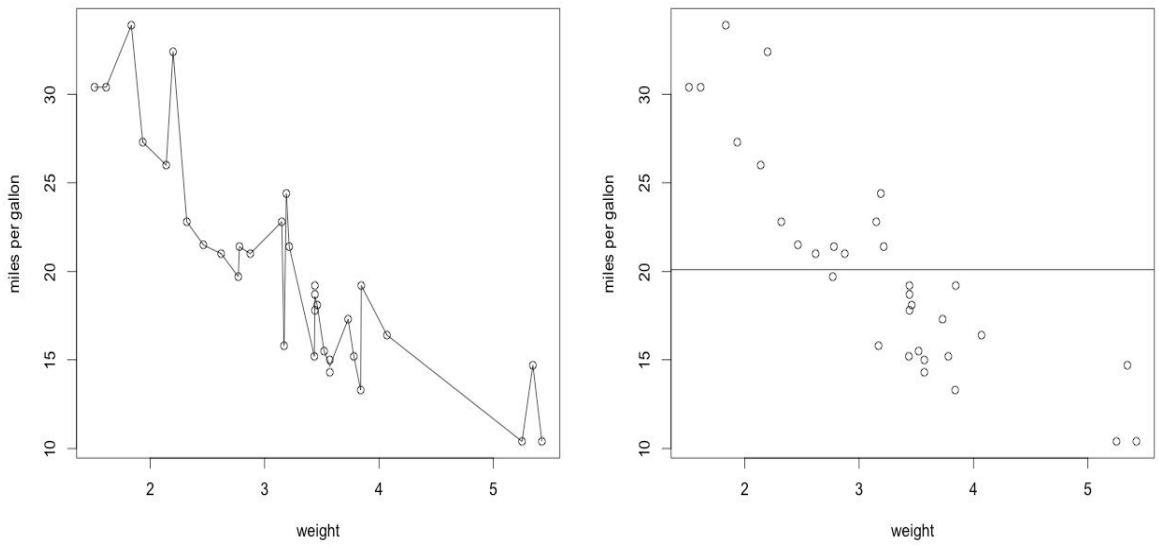


Figure 8.9: The two extremes of the bias-variance tradeoff. (left) a (complicated) model with essentially zero bias (on training data) but enormous variance. (right) a simple model with high bias but virtually no variance

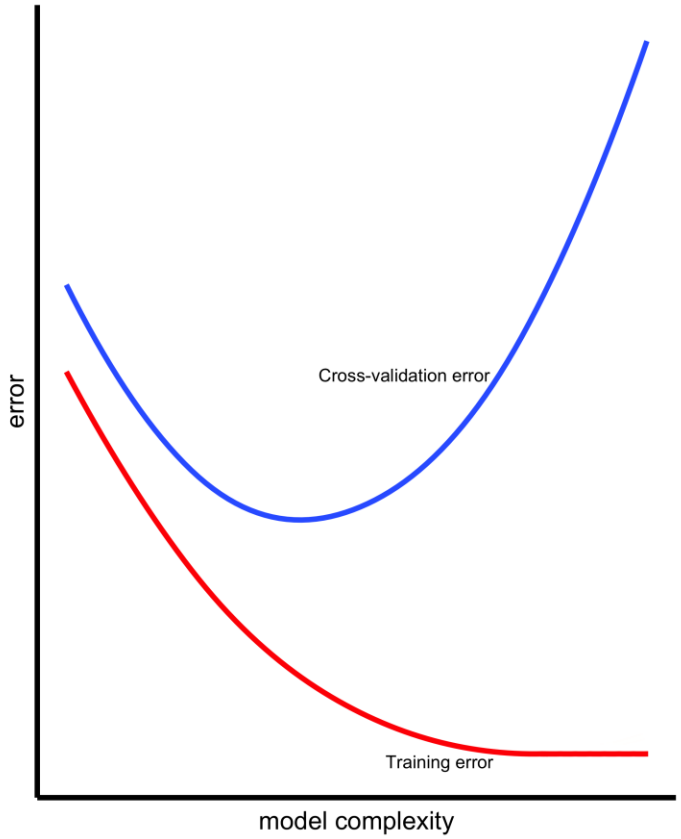


Figure 8.10: As model complexity/flexibility increases, training error (bias) tends to be reduced. Up to a certain point, the cross-validation error decreases as well. After that point, the cross-validation error starts to go up again, even as the model's bias continues to decrease. After this point, the model is too flexible and overfits.

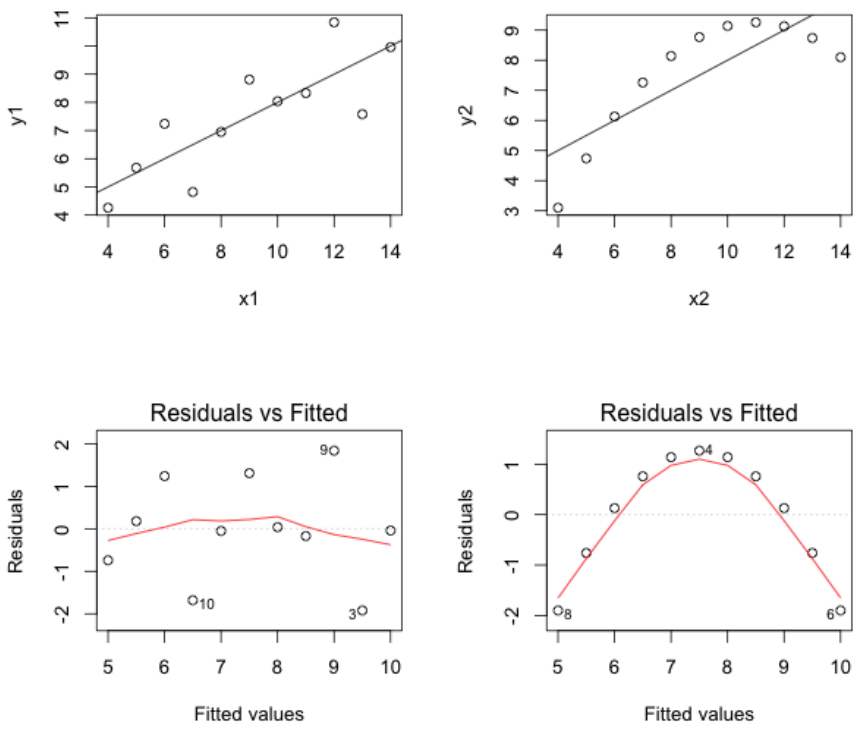


Figure 8.11: The top two panels show the first and second relationships of Anscombe's quartet, respectively. The bottom two panels depict each top panel's respective residual-fitted plot

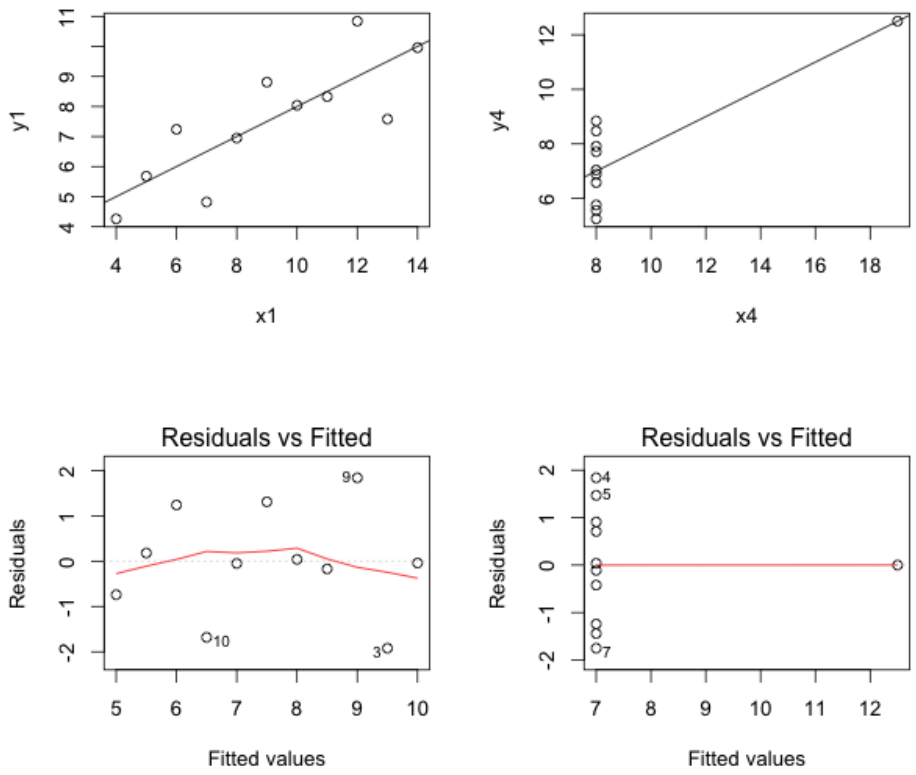


Figure 8.12: The first and the fourth Anscombe relationships and their respective residual-fitted plots.

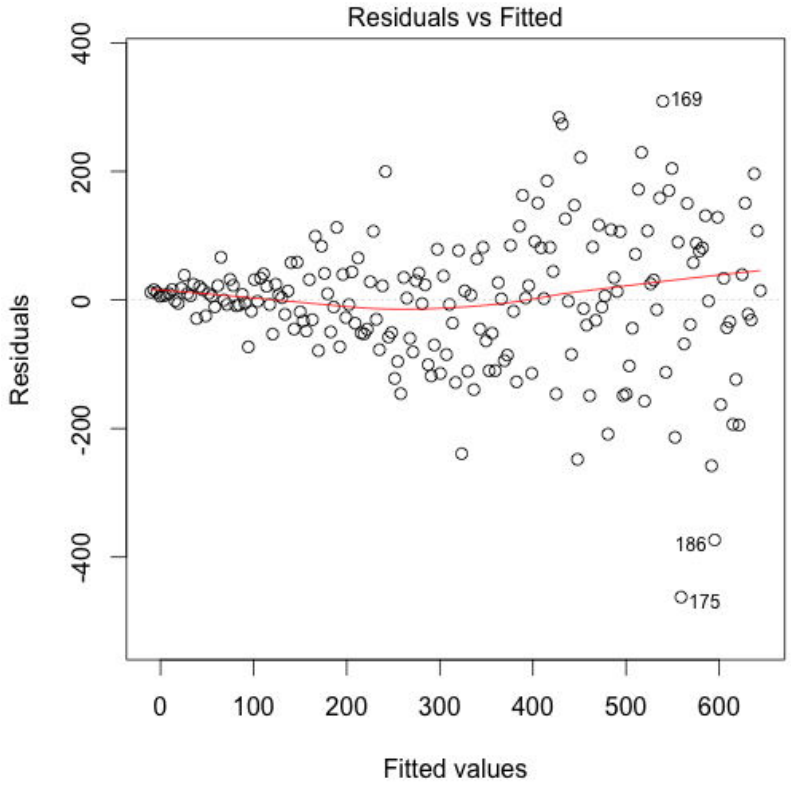


Figure 8.13: A paradigmatic depiction of the residual-fitted plot of a regression model for which the assumption of homogeneity of variance is violated

9

Predicting Categorical Variables

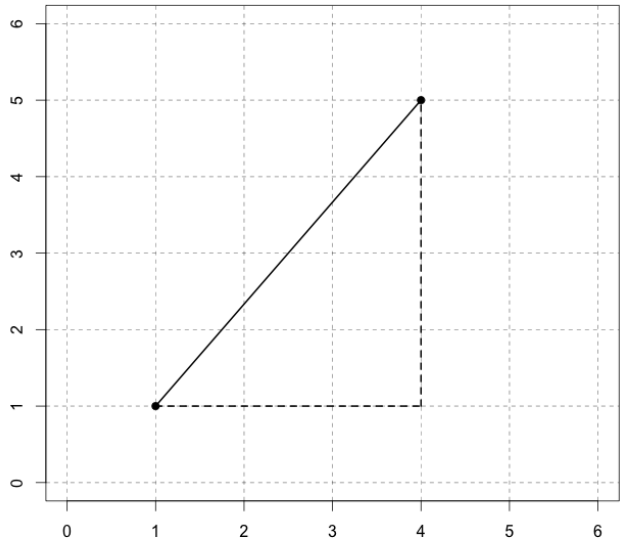


Figure 9.1: Two points on a Cartesian plane. Their Euclidean distance is 5. Their Manhattan distance is 3+4=7

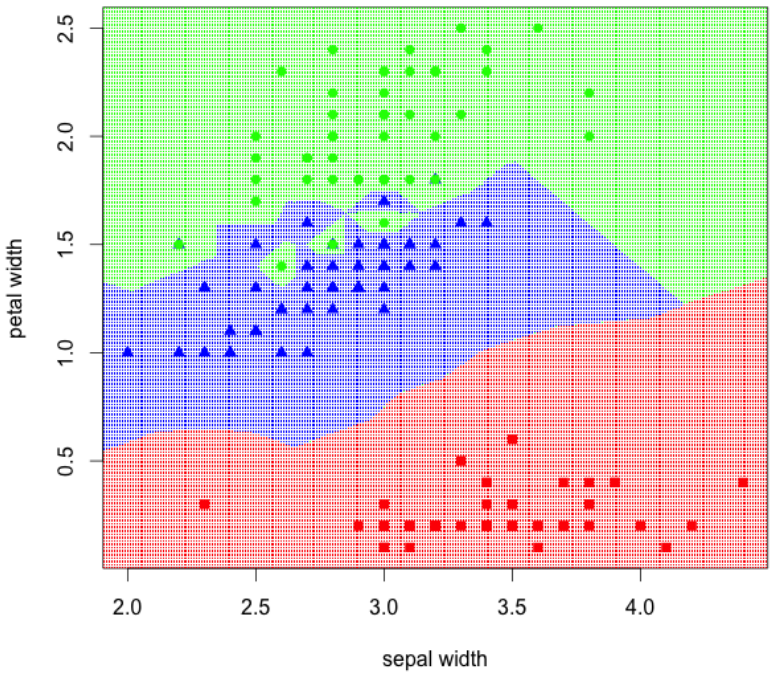


Figure 9.2: The species classification regions of the iris data set using 1-NN.

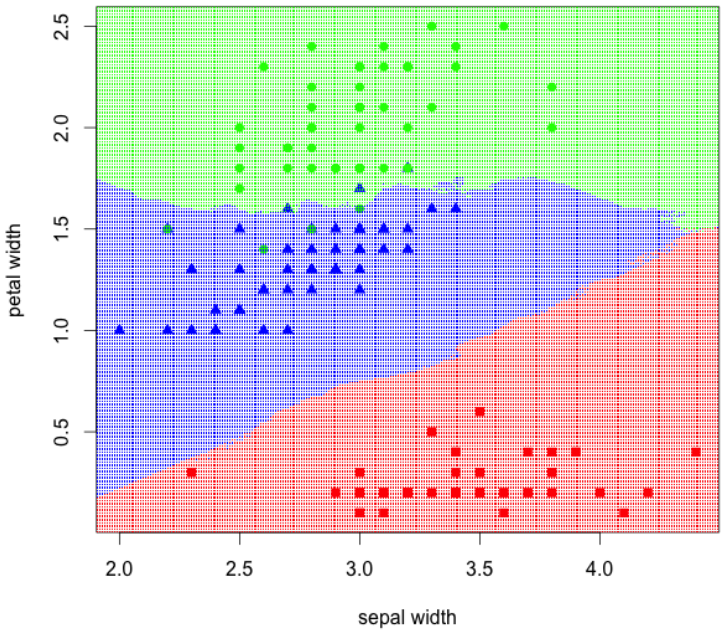


Figure 9.3: The species classification regions of the iris data set using 15-NN. The boundaries between the classification regions are now smoother and less overfit

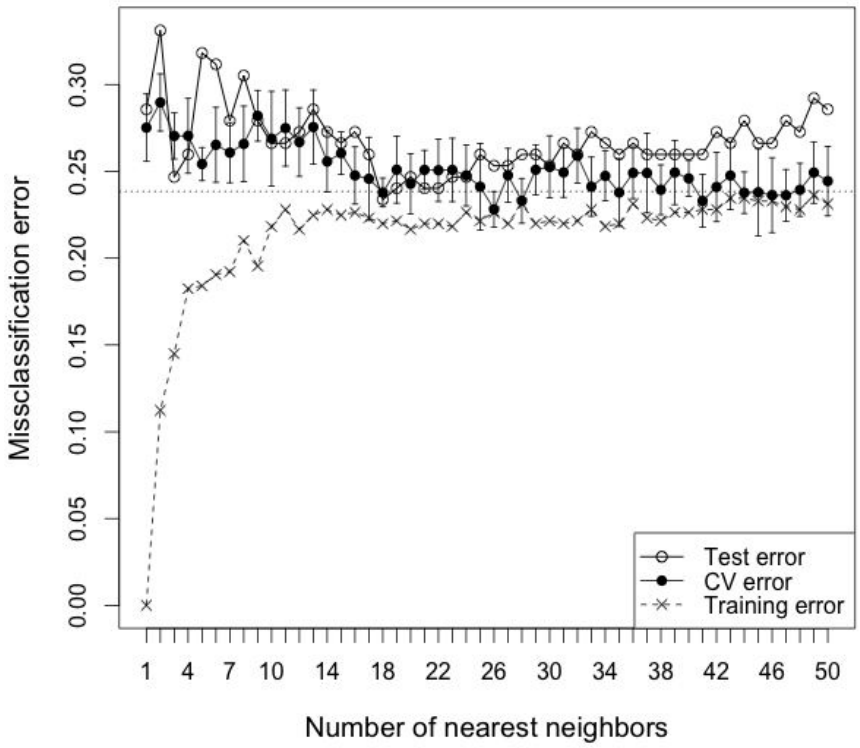


Figure 9.4: A plot illustrating test set error, cross-validated error, and training set error as a function of k in k-NN. After about k=15, the test and CV error doesn't appear to change much.

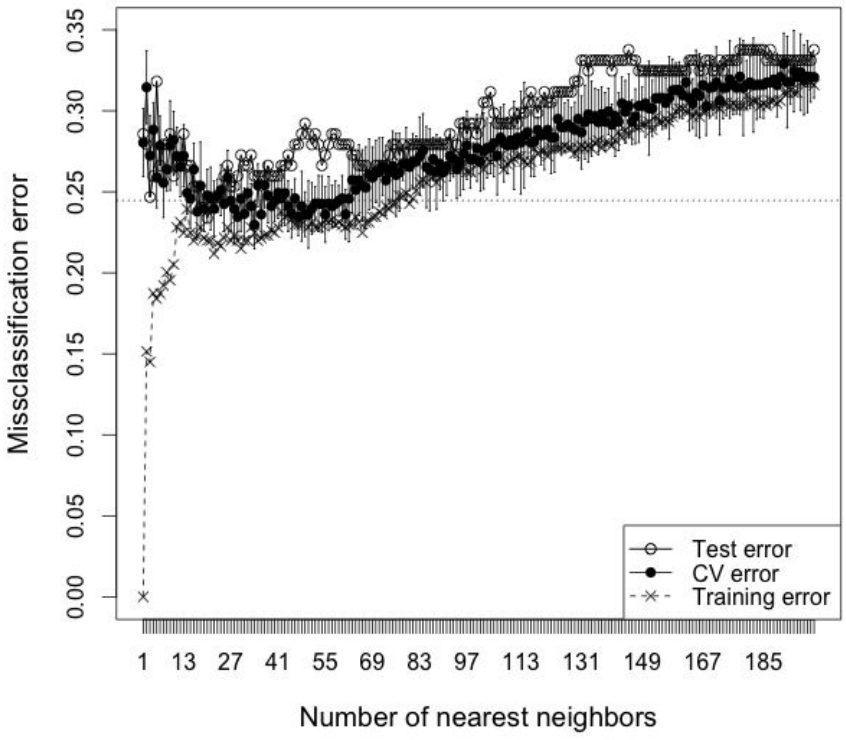


Figure 9.5: A plot illustrating test set error, cross-validated error, and training set error and a function of k in k-NN up to k=200. Notice how error increases as the number of neighbors becomes too large and causes the classifier to overfit.

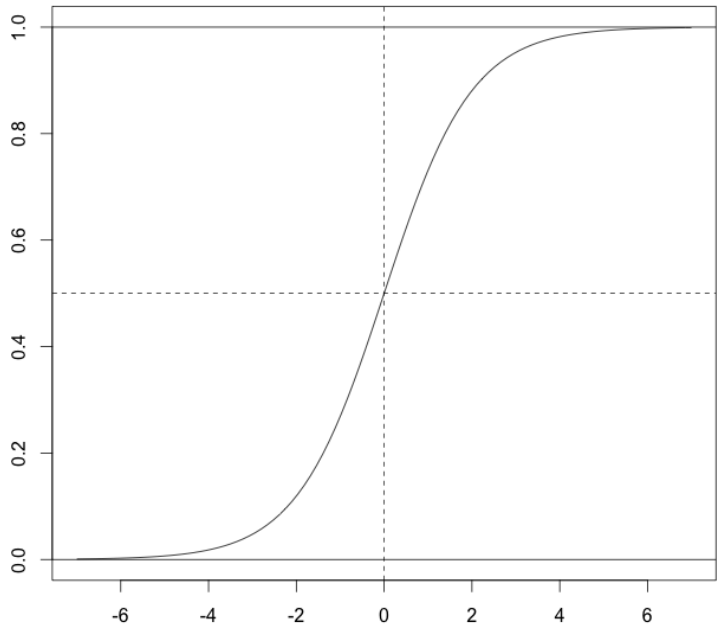


Figure 9.6: The logistic function

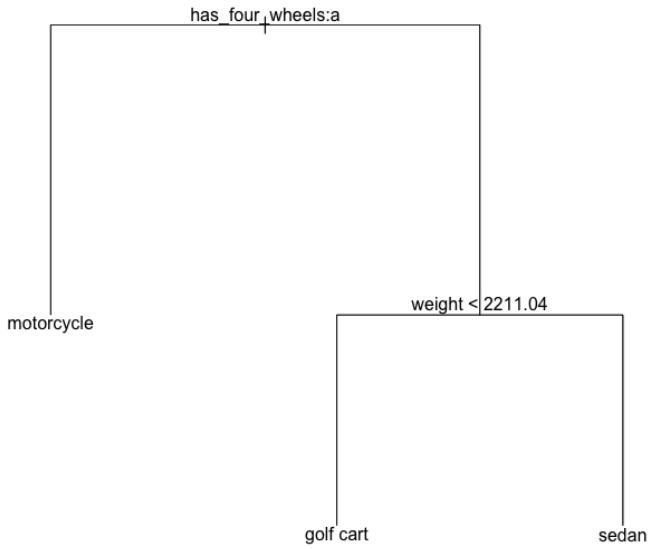


Figure 9.7: A simple and illustrative decision tree that classifies motor vehicles into either motorcycles, golf carts, and sedans

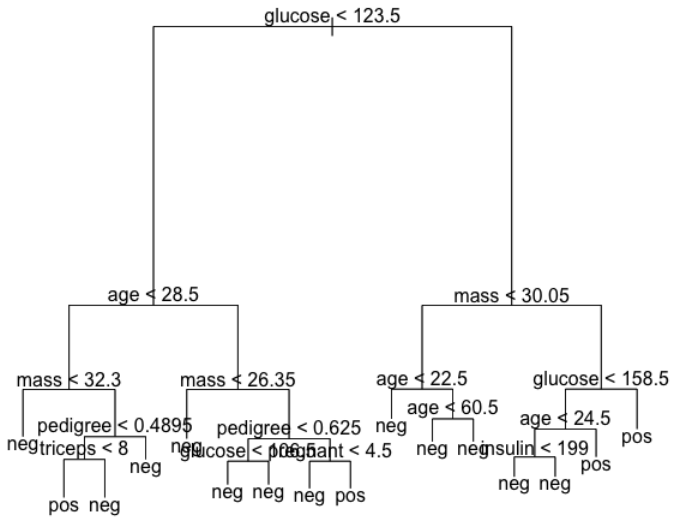


Figure 9.8: An unpruned and complex decision tree

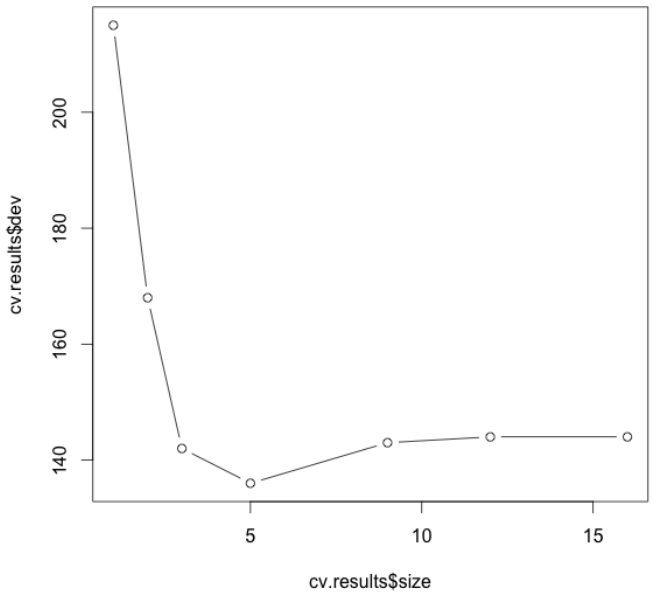


Figure 9.9: A plot cross-validated misclassification error as a function of tree size. Observe that tree of size one performs terribly, and that the error rate steeply declines before rising slightly as the tree is overfit and large sizes

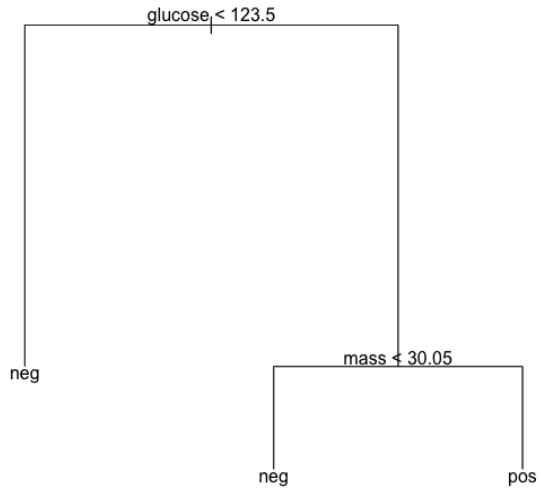


Figure 9.10: Simpler decision tree with the same testing set performance as the tree in Figure 9.8

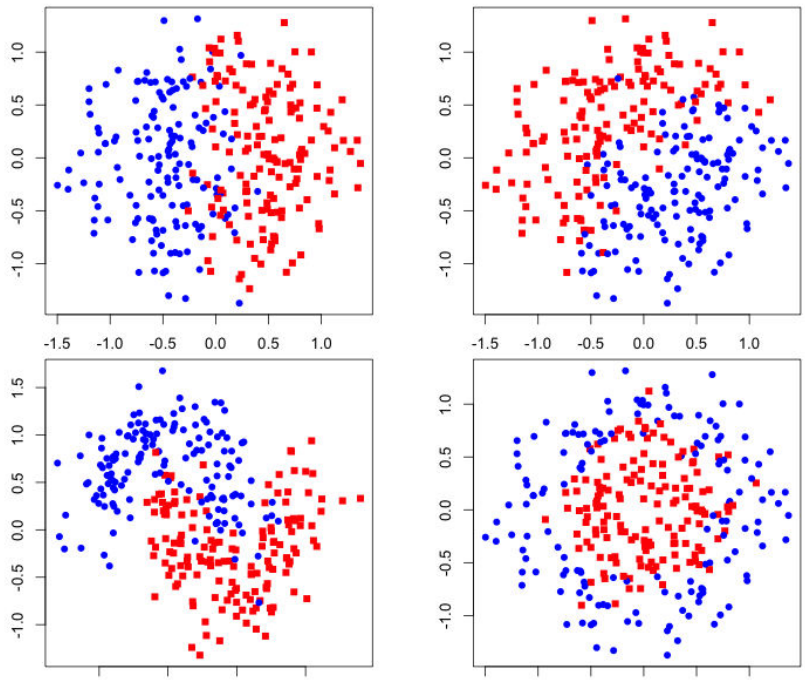


Figure 9.11: A plot depicting the class patterns of our four illustrative and contrived data sets

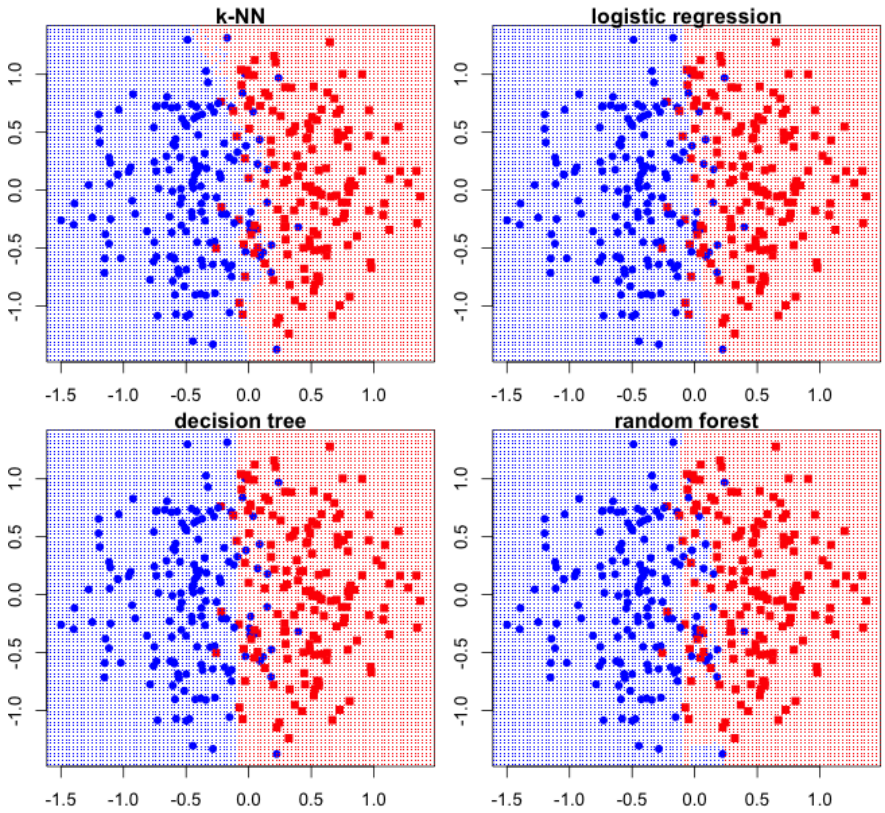


Figure 9.12: A plot of the decision boundaries of our four classifiers on our first contrived dataset.

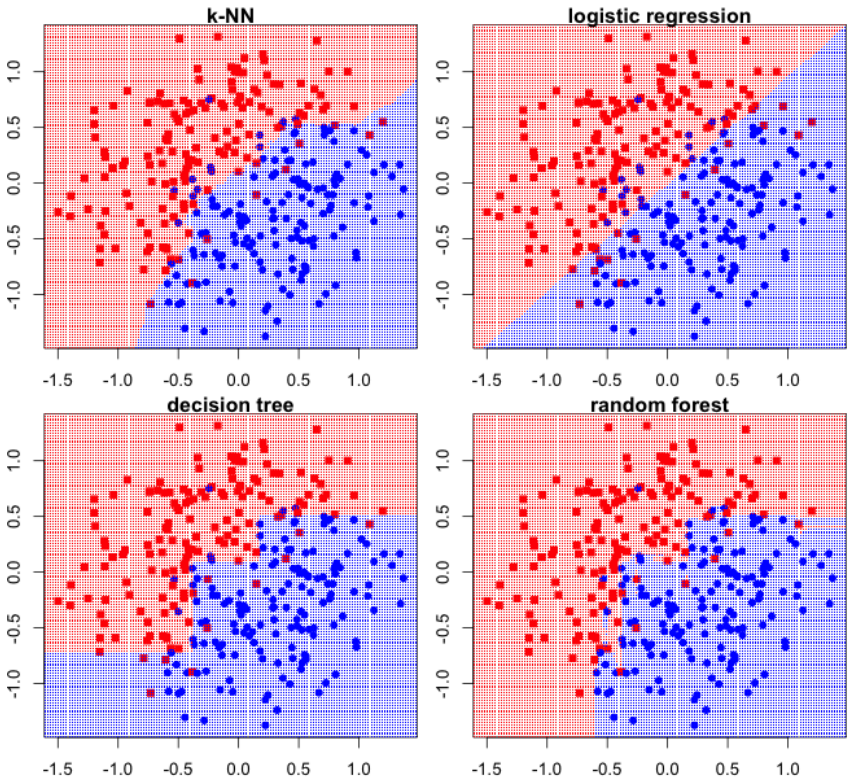


Figure 9.13: A plot of the decision boundaries of our four classifiers on our second contrived dataset.

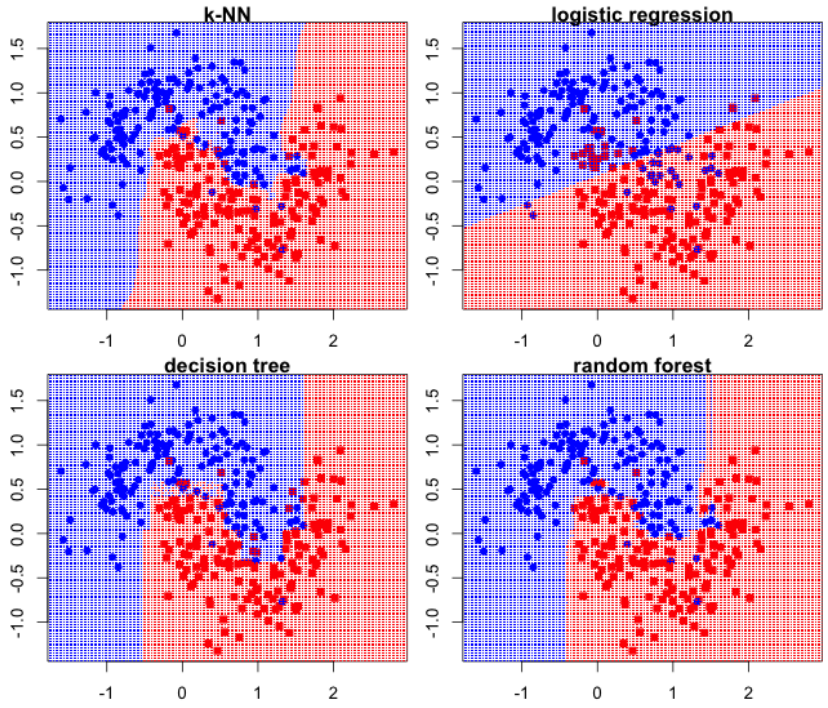


Figure 9.14: A plot of the decision boundaries of our four classifiers on our third contrived dataset.

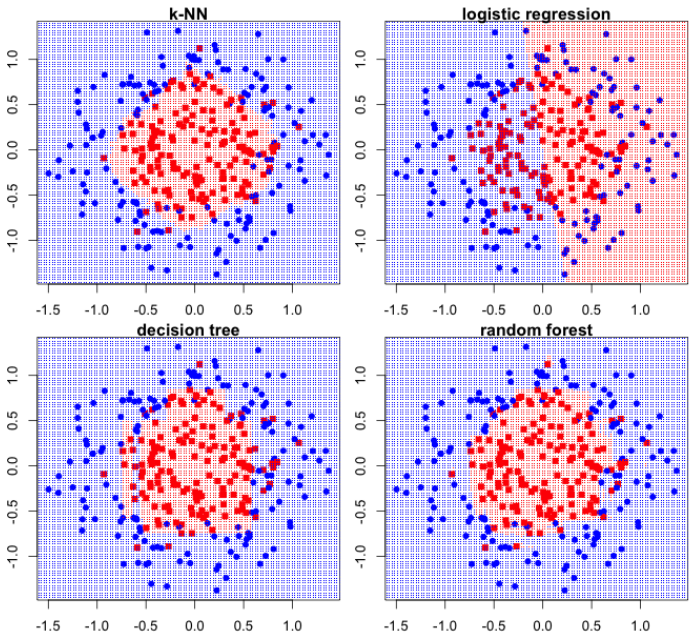


Figure 9.15: A plot of the decision boundaries of our four classifiers on our fourth contrived dataset.

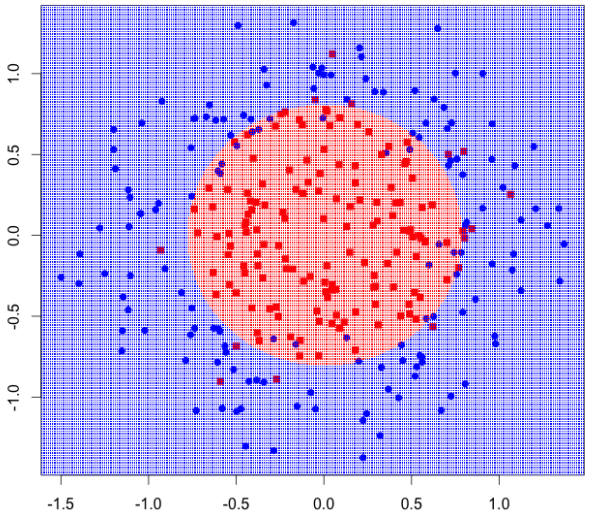


Figure 9.16: A second-order (quadratic) logistic regression decision boundary

10
Sources of Data

No pictures in this chapter

11

Dealing with Messy Data

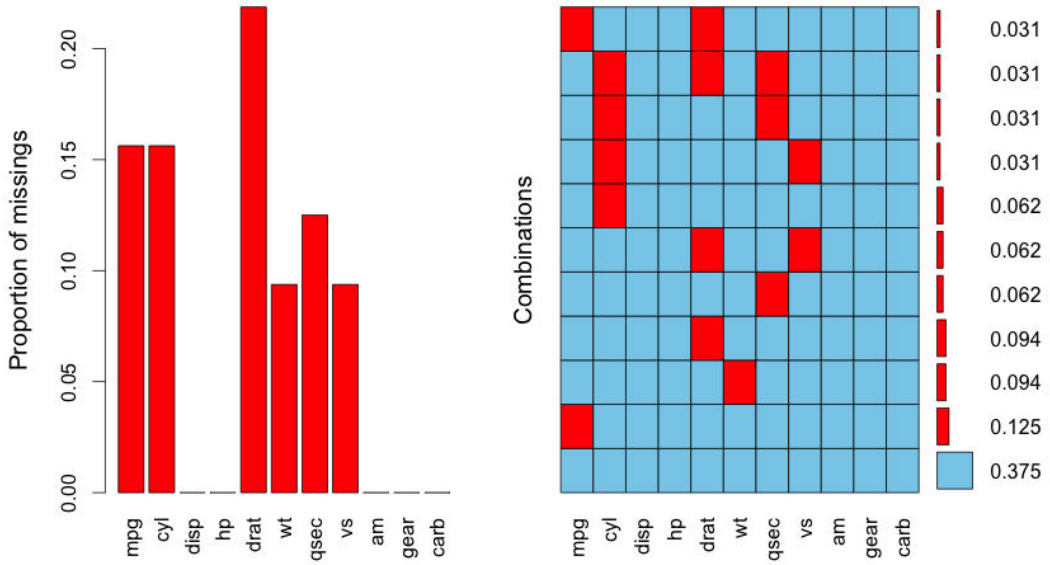


Figure 11.1: The output of VIM's visual aggregation of missing data. The left plot shows the proportion on missing values for each column. The right plot depicts the prevalence of row-wise missing data patterns, like md.pattern

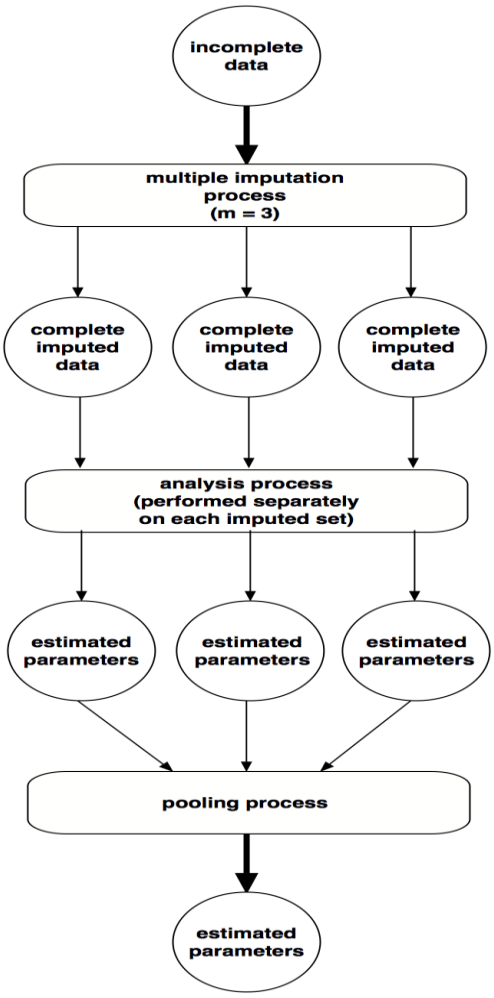


Figure 11.2: Multiple imputation in a nutshell

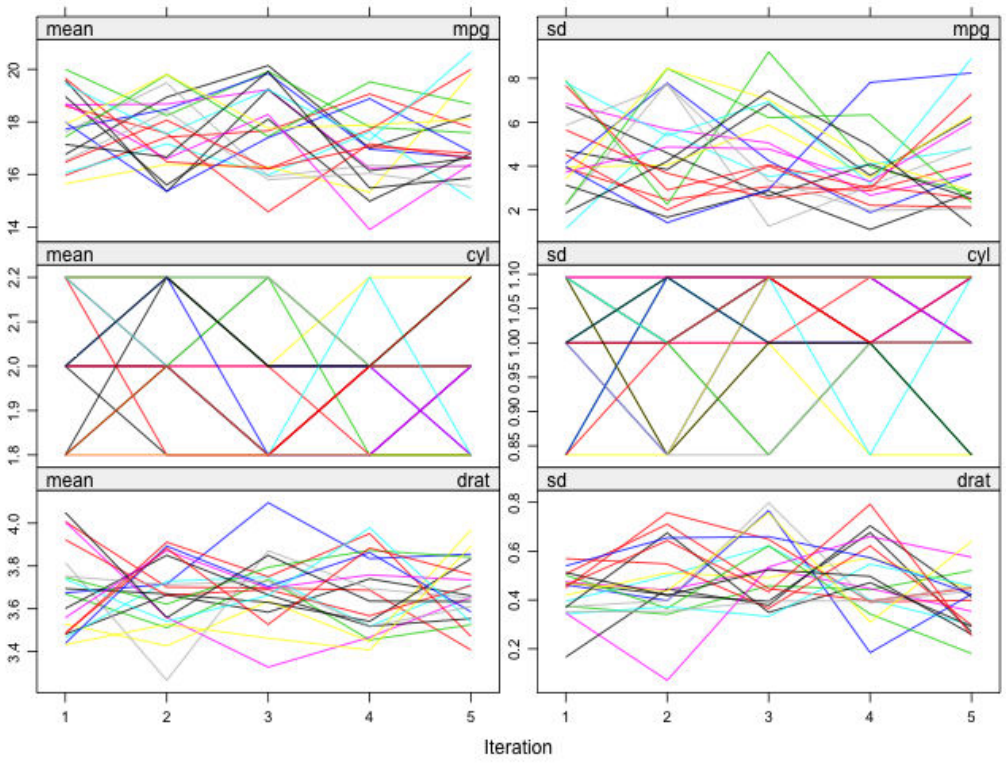


Figure 11.3: A subset of the trace plots produced by plotting an object returned by a mice imputation

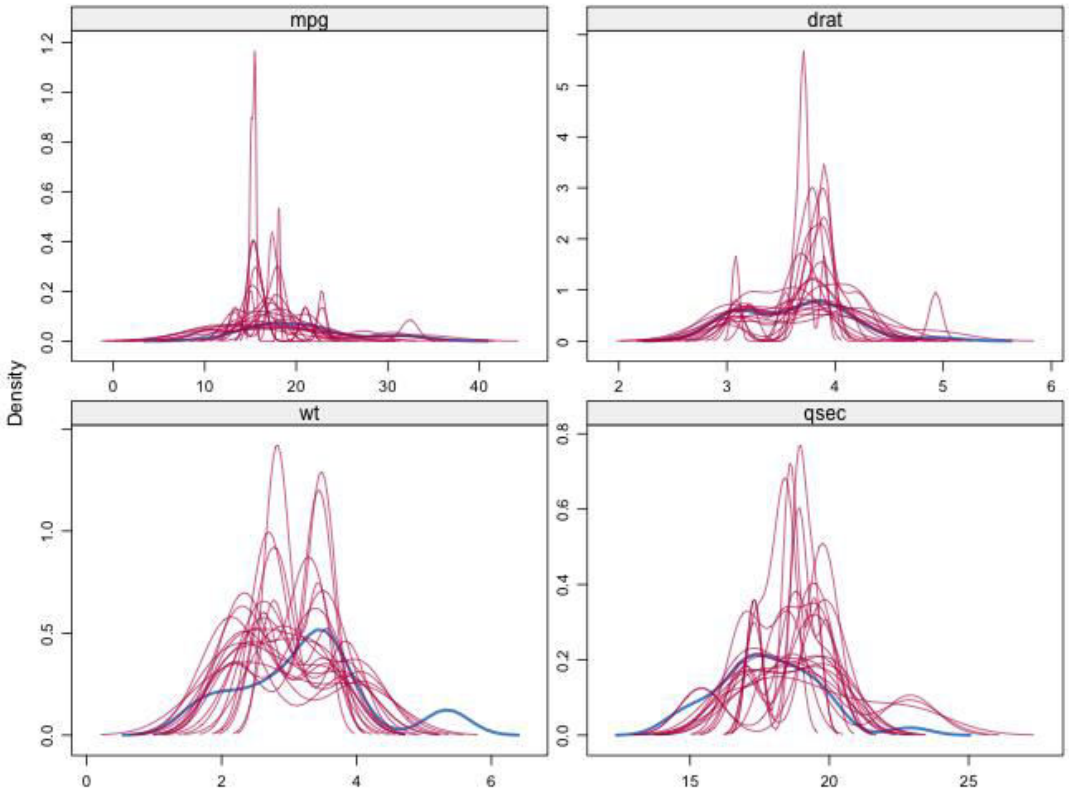


Figure 11.4: Density plots of all the imputed values for `mpg`, `drat`, `wt`, and `qsec`. Each imputation has its own density curve in each quadrant

12
Dealing with Large Data

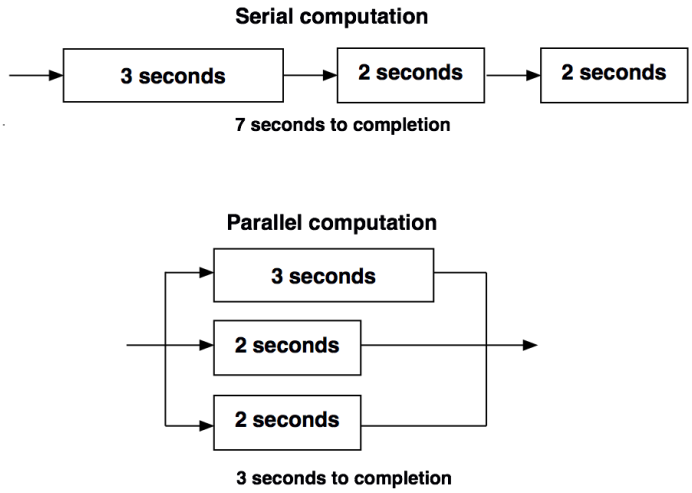


Figure 12.1: diagram of the parallelization and the resultant reduced time to completion

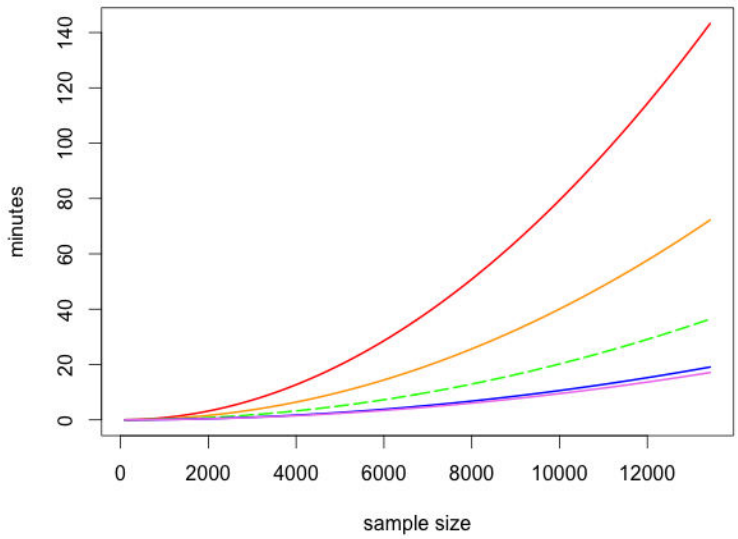


Figure 12.2: The running times for the average-distance-between-all-airports task at different sample sizes for 1, 2, 4, 8, and 16 cores. For reference, the dashed line is the 4 core performance curve, the top most curve is the single core performance curve, and the bottom most curve is the 16 core curve.

13

Reproducibility and Best Practices

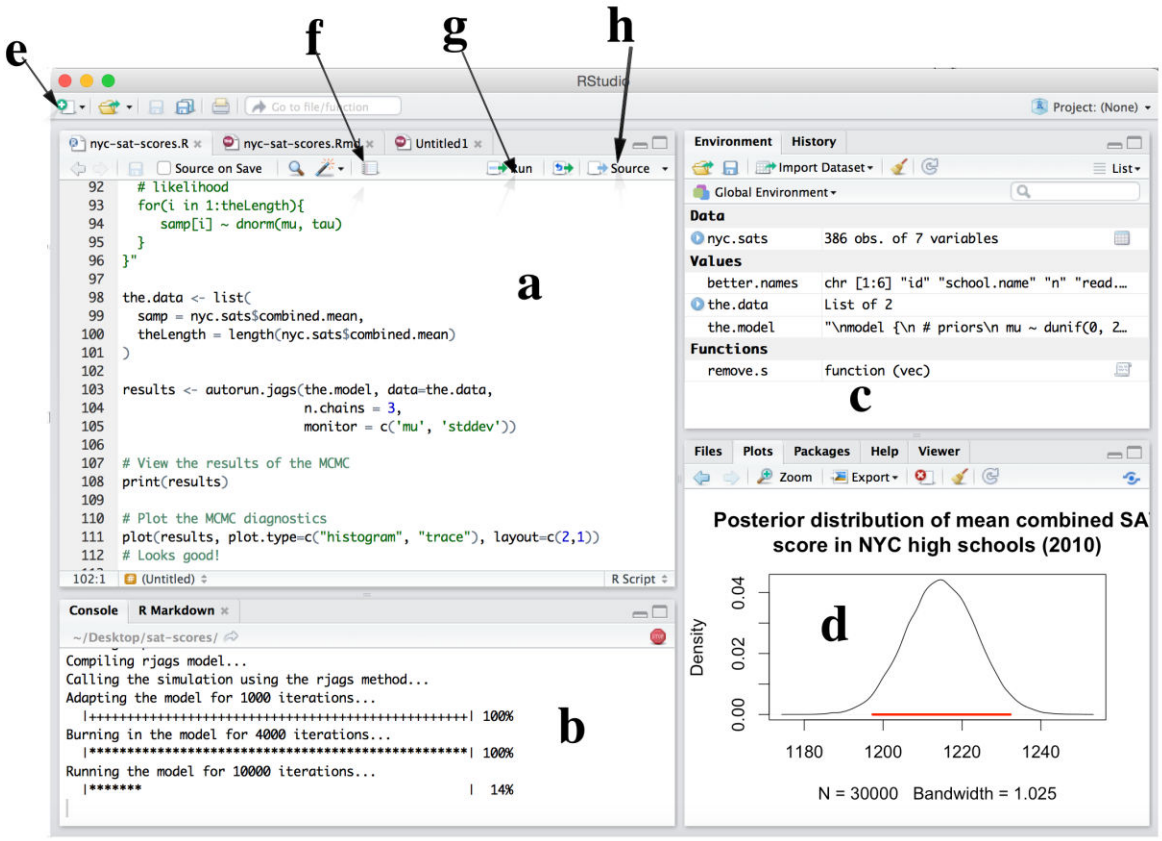


Figure 13.1: RStudio's four-panel interface in Mac OS X (version 0.99.486)

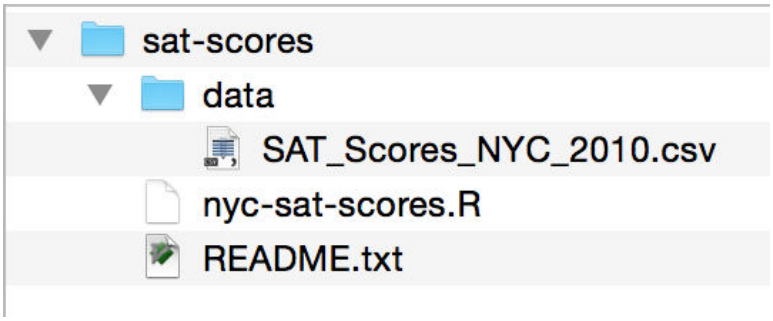


Figure 13.2: A sample file/folder hierarchy for an R analysis project

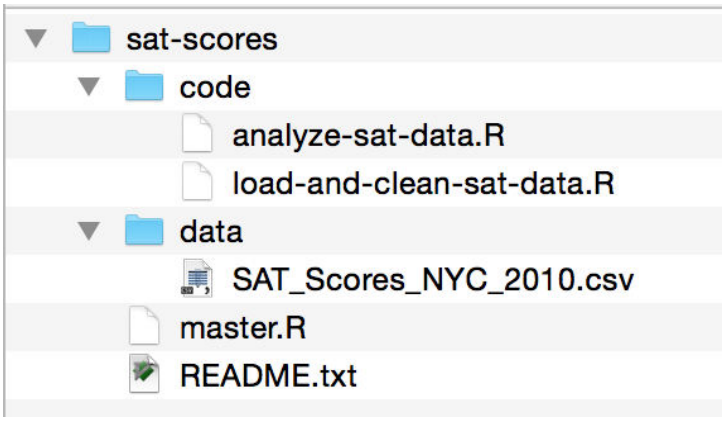


Figure 13.3: A sample file/folder hierarchy for a multiscript R analysis project

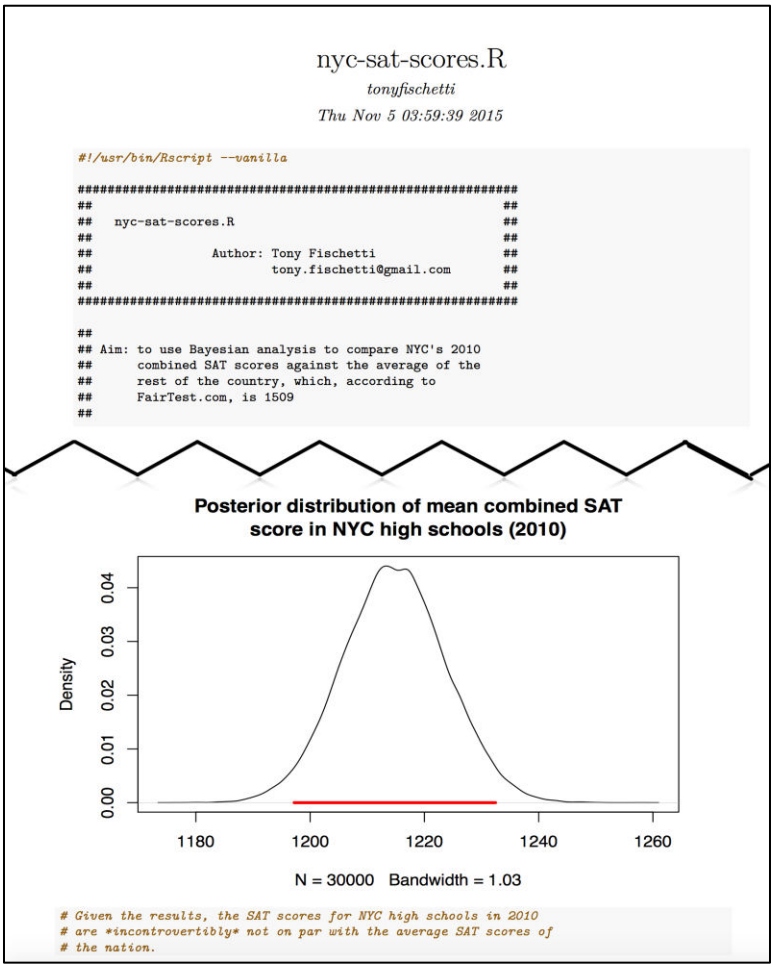


Figure 13.4: An excerpt from the output of Compile Notebook on our example script

NYC SAT Scores Analysis

Tony Fischetti

November 1, 2015

Aim:

To use Bayesian analysis to compare NYC's 2010 combined SAT scores against the average of the rest of the country, which, according to FairTest.com, is 1509

We are going to use the `assertr` and `runjags` packages for data checking and MCMC, respectively.

```
# libraries
library(assertr)   # for data checking
library(runjags)   # for MCMC
```

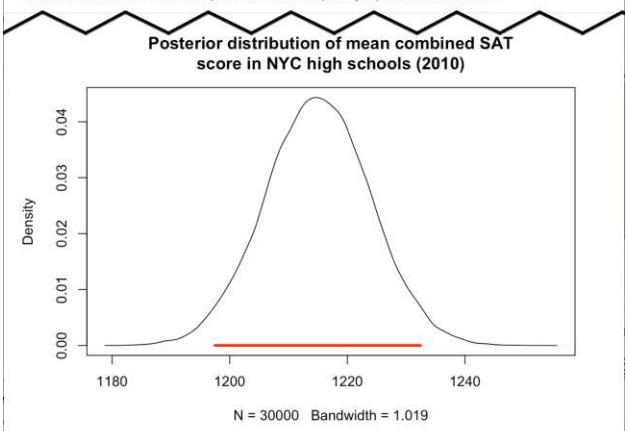
Let's make sure everything is all set with JAGS!

```
testjags()
```

```
## You are using R version 3.2.1 (2015-06-18) on a unix machine, with
## the X11 GUI
## The rjags package is installed
## JAGS version 3.4.0 found successfully using the command
## '/usr/local/bin/jags'
```

Great!

This data was found in the NYC Open Data Portal: https://nycopendata.socrata.com



Given the results, the SAT scores for NYC high schools in 2010 are **incontrovertibly** not on par with the average SAT scores of the nation.

This is some session information for reproducibility:

```
devtools::session_info()
```

Figure 13.5: An excerpt from the output of Knit HTML on our example R Markdown document.