ASSESSING NORMALITY

DAVAR KHOSHNEVISAN

1. Histograms

Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The pressing question is, "is it true that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$"?

To answer this, consider the "residuals,"

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

If $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ then one would like to think that the histogram of the $\widehat{\varepsilon}_i$'s should look like a normal pdf with mean 0 and variance σ^2 (why?). How close is close? It helps to think more generally.

Consider a sample $U_1, \ldots, U_n$ (e.g., $U_i = \widehat{\varepsilon}_i$). We wish to know where the U_i's are coming from a normal distribution. Again, the first thing to do is to plot the histogram. In R you type,

```
hist(u,nclass=n)
```

where `u` denotes the vector of the samples $U_1, \ldots, U_n$ and `n` denotes the number of bins in the histogram.

For instance, consider the following exam data:

```
16.8 9.2 0.0 17.6 15.2 0.0 0.0 10.4 10.4 14.0 11.2 13.6 12.4
14.8 13.2 17.6 9.2 7.6 9.2 14.4 14.8 15.6 14.4 4.4 14.0 14.4 0.0
0.0 10.8 16.8 0.0 15.2 12.8 14.4 14.0 17.2 0.0 14.4 17.2 0.0 0.0
0.0 14.0 5.6 0.0 0.0 13.2 17.6 16.0 16.0 0.0 12.0 0.0 13.6 16.0
8.4 11.6 0.0 10.4 0.0 14.4 0.0 18.4 17.2 14.8 16.0 16.0 0.0 10.0
13.6 12.0 15.2
```

The command `f1.dat,hist(nclass=15)` produces Figure 1(a).[1]

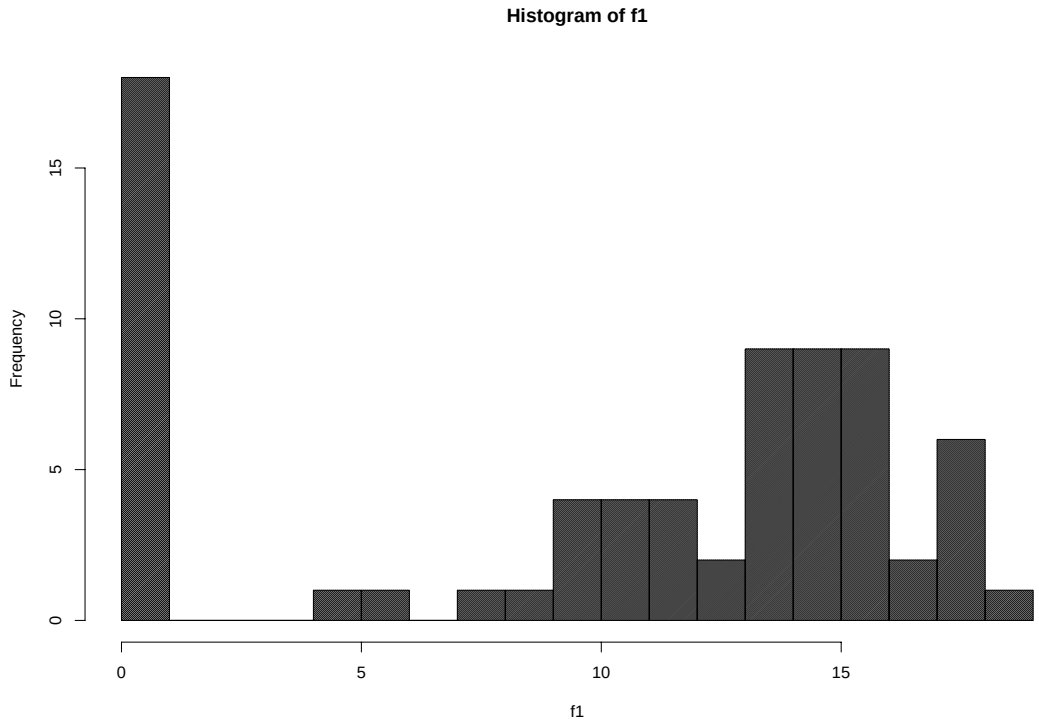
Try this for different values of `nclass` to see what types of hitograms you can obtain. You should always ask, "which one represents the truth the best"? Is there a unique answer?

Now the data $U_1, \ldots, U_n$ is probably not coming from a normal distribution if the histogram does not have the "right" shape. Ideally, it would be symmetric, and the tails of the distribution taper off rapidly.

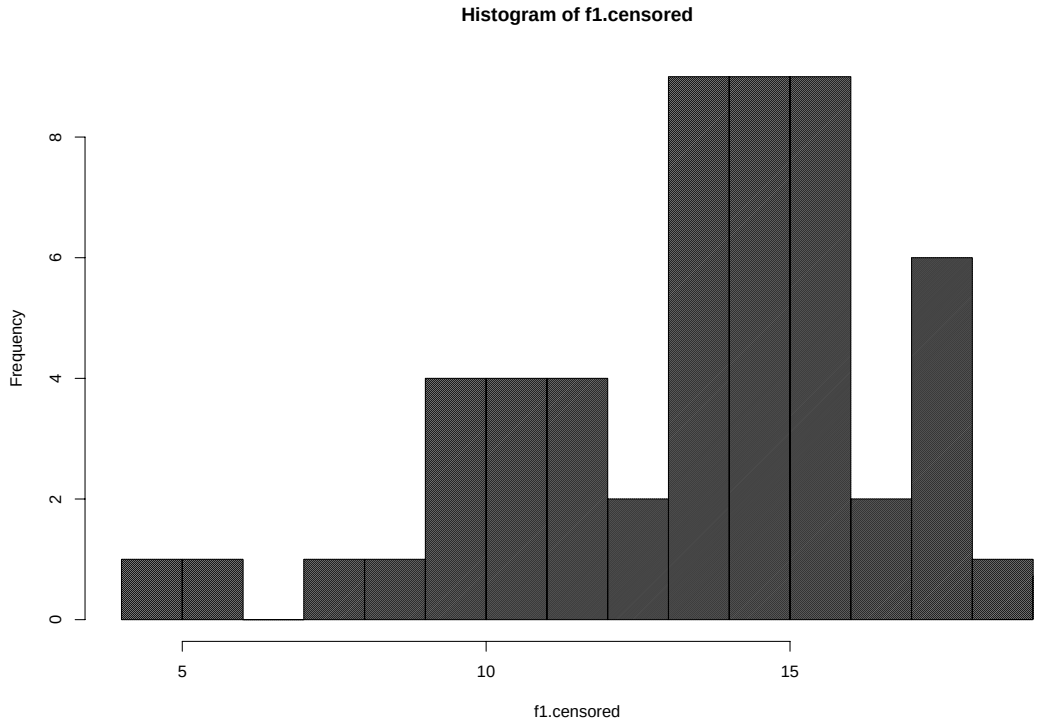
In Figure 1(a), there were many students who did not take the exam in question. They received a '0' but this grade should probably not contribute to our knowledge of the distribution of all such grades. Figure 1(b) shows

Date: September 1, 2004.

[1]You can obtain this data freely from the website below:
http://www.math.utah.edu/~davar/math6010/2004/notes/f1.dat.



(a) Grades



(b) Censored Grades

the histogram of the same data set when the zeroes are removed. [This histogram is closer to a normal density.]

2. QQ-Plots

QQ-plots are a better way to assess how closely a sample follows a certain distribution.

To understand the basic idea note that if $U_1, \dots, U_n$ is a sample from a normal distribution with mean μ and variance σ^2, then about 68.3% of the sample points should fall in $[\mu-\sigma, \mu+\sigma]$, 95.4% should fall in $[\mu-2\sigma, \mu+2\sigma]$, etc.

Now let us be more careful still. Let $U_{(1)} \le \cdots \le U_{(n)}$ denote the order statistics of $U_1, \dots, U_n$. Then no matter how you make things precise, the fraction of data "below" $U_{(j)}$ is $(j \pm 1)/n$. So we make a continuity correction and *define* the fraction of date below U_j to be $(j - \frac{1}{2})/n$. If the U_j's were approximately $N(0,1)$, then we would expect the fraction of data below U_j to be q_j. This is defined to be the "quantile,"

$$P\{N(0,1) \le q_j\} = \frac{j-\frac{1}{2}}{n}; \qquad \text{i.e., } q_j = \Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right).$$

If $U_j \sim N(\mu, \sigma^2)$, then $(U_j - \mu)/\sigma \sim N(0,1)$, so we would expect the fraction below U_j to be $\sigma q_j + \mu$ (work this out!).

Therefore, even if we do not know μ and σ^2, then we would expect the scatterplot of (q_j, U_j) to follows closely a line. [The slope and intercept are σ and μ, respectively.]

QQ-plots are simply the plots of the $N(0,1)$-quantiles $q_1, \dots, q_n$ versus the order statistics $U_{(1)}, \dots, U)(n)$. To draw the qqplot of a vector u in R, you simply type

$$\mathsf{qqnorm}(u).$$

Figure 1(c) contains the qq-plot of the exam data we have been studying here.

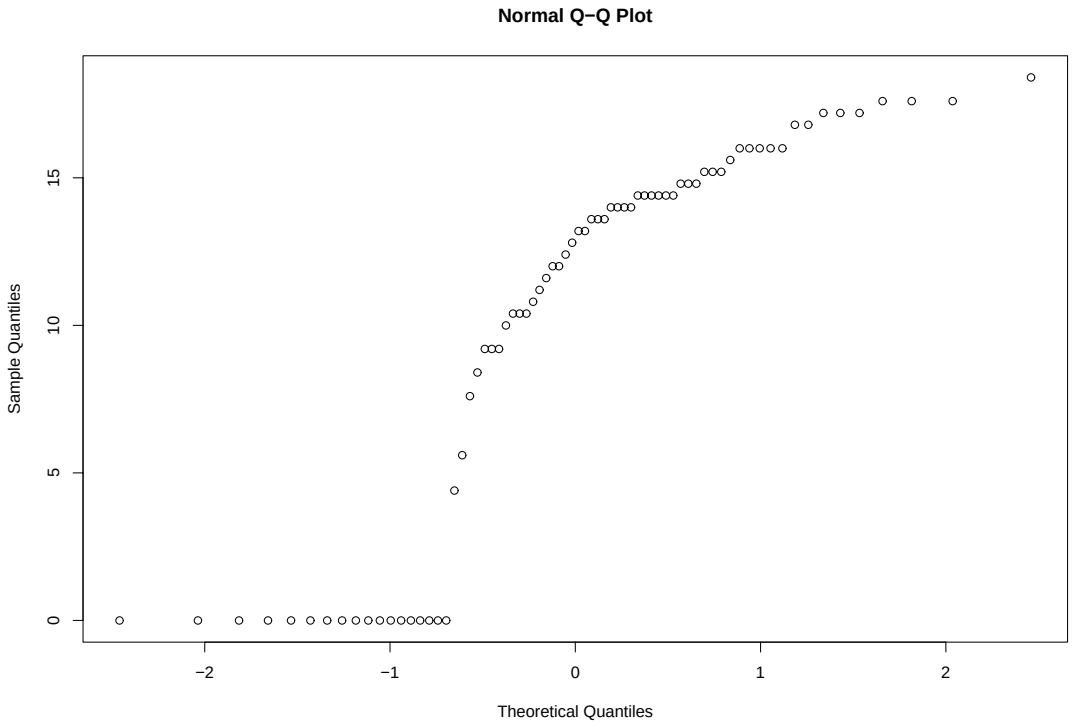
3. The Correlation Coefficient of the QQ-Plot

In its complete form, the R-command qqnorm has the following syntax:

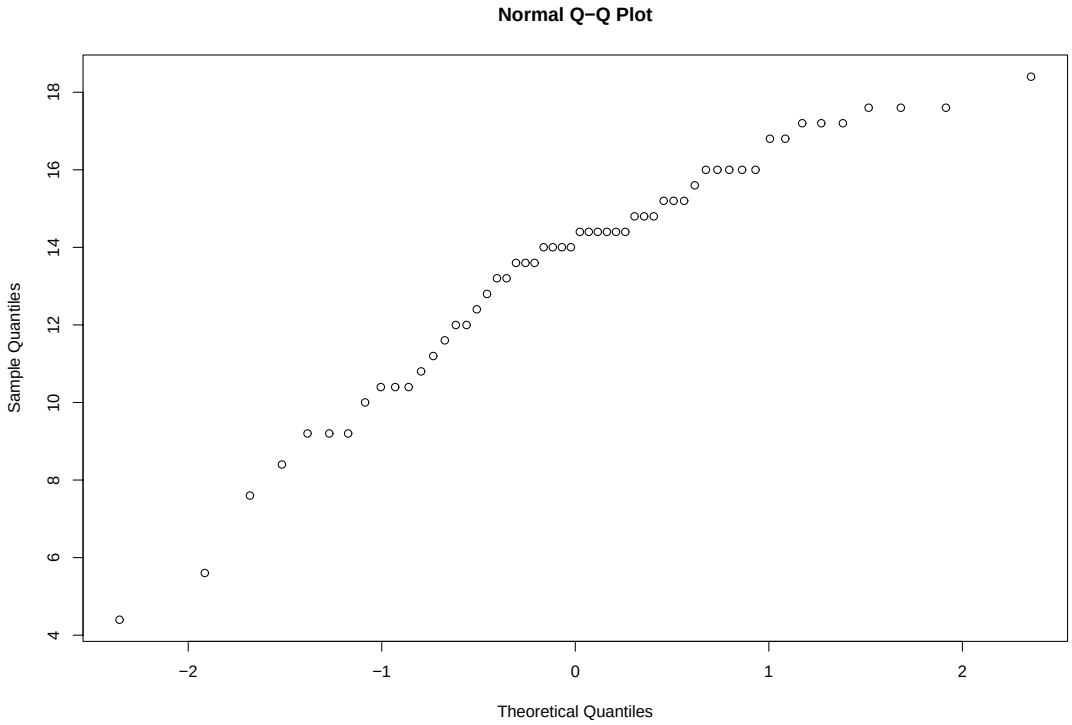
$$\mathsf{qqnorm}(\mathsf{u}, \mathsf{datax} = \mathsf{FALSE}, \mathsf{plot} = \mathsf{TRUE}).$$

The parameter u denotes the data; datax is "FALSE" if the data values are drawn on the y-axis (default). It is "TRUE" if you wish to plot $(U_{(j)}, q_j)$ instead of the more traditional $(q_j, U_{(j)})$. The option plot=TRUE (default) tells R to plot the qq-plot, whereas plot=FALSE produces a vector. So for instance, try

$$\mathsf{V} = \mathsf{qqnorm}(\mathsf{u}, \mathsf{plot} = \mathsf{FALSE}).$$



(c) QQ-plot of grades



(d) QQ-plot of censored grades

This creates two vectors: V\$$x$ and V\$$y$. The first contains the values of all q_j's, and the second all of the $U_{(j)}$'s. So now you can compute the correlation coefficient of the qq-plot by typing:

$$\mathsf{V} = \mathsf{qqnorm}(\mathsf{u}, \mathsf{plot} = \mathsf{FALSE})$$
$$\mathsf{cor}(\mathsf{V}\$x, \mathsf{V}\$y).$$

If you do this for the qq-plot of the grade data, then you will find a correlation of ≈ 0.910. After censoring out the no-show exams, we obtain a correlation of ≈ 0.971. This produces a noticeable difference, and shows that the grades are indeed normal.

In fact, one can analyse this procedure statistically, as we shall do later on.