Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2020

Answers for Assignment 8

Sample answer for Exercise 1:

1. (5 pts) Carefully state the null hypothesis. The null hypothesis is that the probability that the World Series lasts m games is

$$P(m) = \binom{m-1}{m-4}\Big(p^4(1-p)^{m-4} + (1-p)^4 p^{m-4)}\Big),$$

where p is the probability that the better team wins. We have estimated that to be $\hat{p} = 0.5906$, so the null hypothesis we will test is that the multinomial probabilities for lengths of $4, \dots, 7$ are

$$H_0 \colon \boldsymbol{p} = \boldsymbol{p}_0 = \begin{pmatrix} 0.150 & 0.266 & 0.302 & 0.283 \end{pmatrix}$$

versus the alternative hypothesis

$$H_1 \colon \boldsymbol{p} \neq \boldsymbol{p}_0.$$

2. (15 pts) Write out by hand the formula for the test statistic. (Hint: All the numbers you need are in the two tables above.) What is the value of the test statistic? (You may use a computer/calculator to evaluate the formula.)

The test statistic D is the sum of the squared (expected minus actual) number of games divided by the expected number of gaes, which is

$$\frac{(16.62-21)^2}{16.62} + \frac{(29.48-26)^2}{29.48} + \frac{(33.51-42)^2}{33.51} + \frac{(31.38-40)^2}{31.38} = 6.63.$$

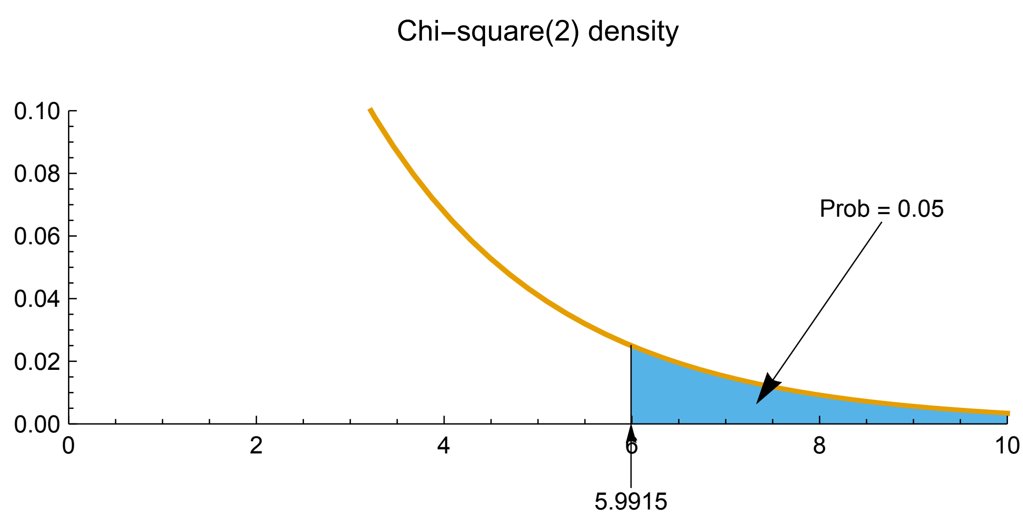
3. (5 pts) Should you use a two-sided test or a one-sided test? Why?

I would use a one-sided test. The test statistic is larger the worse the model fits the data. I do not care if the fit is too good, because it it is hard to imagine that over the course of more than a century the teams could conspire to make the data too close to the model.

4. (10 pts) Explain how many degrees of freedom you should use. (Remember, p was estimated by MLE.) What is the critical value of the test statistic? Why?

> There are four categories for the multinomial (lengths of 4, 5, 6, and 7), but we estimated one parameter $\hat{p}$ by MLE, so the test staistic D should have a χ^2-distribution with $(4-1)-1=2$ degrees of freedom.
>
> To test at the 5% level of significance, the critical value $\chi^2_{2,.95} = 5.99$

5. (5 pts) Draw a rough sketch of the pdf to illustrate the critical value for a test at the $\alpha = 0.05$ level of significance.



6. (5 pts) What is the p-value of the test statistic you computed?

> The p-value is 0.036.

7. (5 pts) Do you reject or fail to reject the null hypothesis at the $\alpha = 0.05$ level of significance?

> The null hypothesis is **rejected**. Bummer.

◇

Sample answer for Exercise 2:

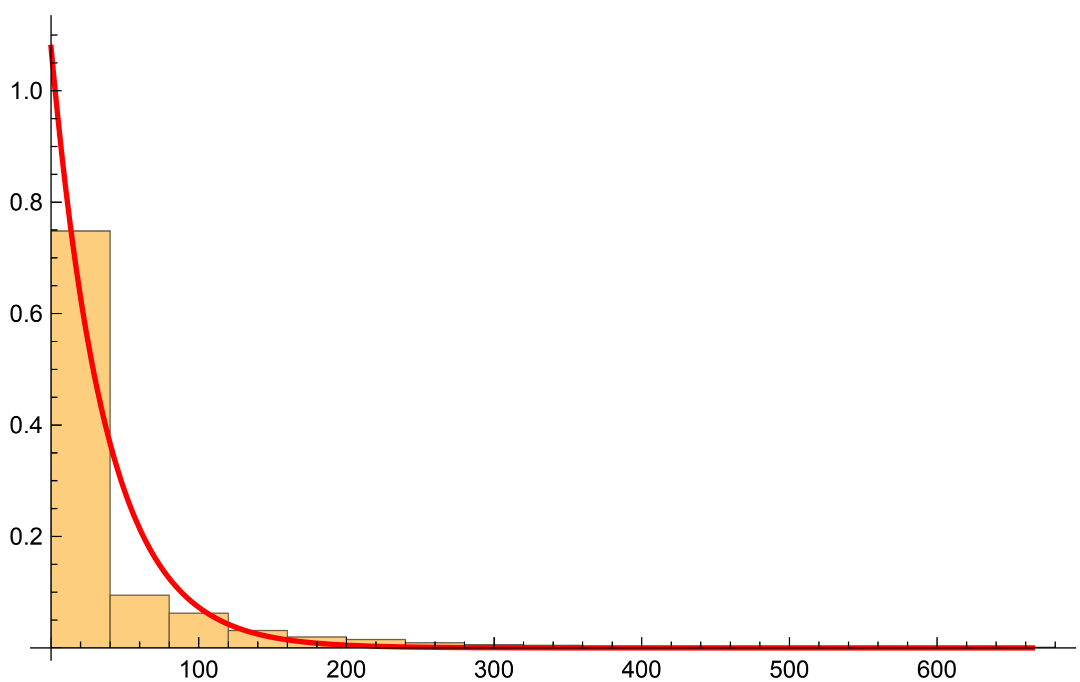
1. (10 pts) What is the relationship between the mean and the standard deviation of an exponential distribution?

The exponential distribution with parameter λ has density $f(x) = \lambda e^{-\lambda x}$. The mean μ and standard deviation σ satisfy

$$\mu = \sigma = 1/\lambda,$$

so they are equal.

2. (10 pts) Create a histogram of the inter-arrival times.



3. (10 pts) Find the mean and standard deviation of the inter-arrival times. Do they come close to satisfying the relationship in part 1?

The mean waiting time is 37.1 days, and the standard deviation is 74.7 days. These do not seem close, but I should really construct confidence intervals.

4. (10 pts) What is the log-likelihood function for a sample $x_1, \dots, x_n$ drawn from an Exponential(λ) distribution?

The exponential density with parameter λ is $f(x) = \lambda e^{-\lambda x}$, so the log likelihood function for a sample $x_1, \dots, x_n$ is just

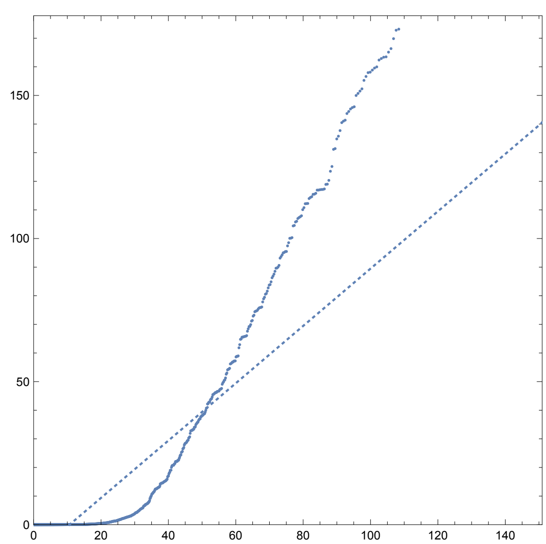
$$n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

5. (10 pts) Assuming the earthquake inter-arrival times are exponentially distributed with parameter λ, what is the maximum likelihood estimate of λ?

The MLE of the parameter λ is gotten by maximizing the log likelihhod function, and the result is that $\hat{\lambda}_{\text{MLE}} = 1/\bar{x}$, where $\bar{x}$ is the sample mean. In this case

$$\hat{\lambda}_{\text{MLE}} = 0.026946 \quad \text{M4.5+ earthquakes per day.}$$

6. (10 pts) Create a Q-Q plot of the quantile of the empirical cdf vs the quantiles of an Exponential distribution with parameter $\hat{\lambda}_{\text{MLE}}$. *Do not create a Normal Q-Q plot.* How does it look?



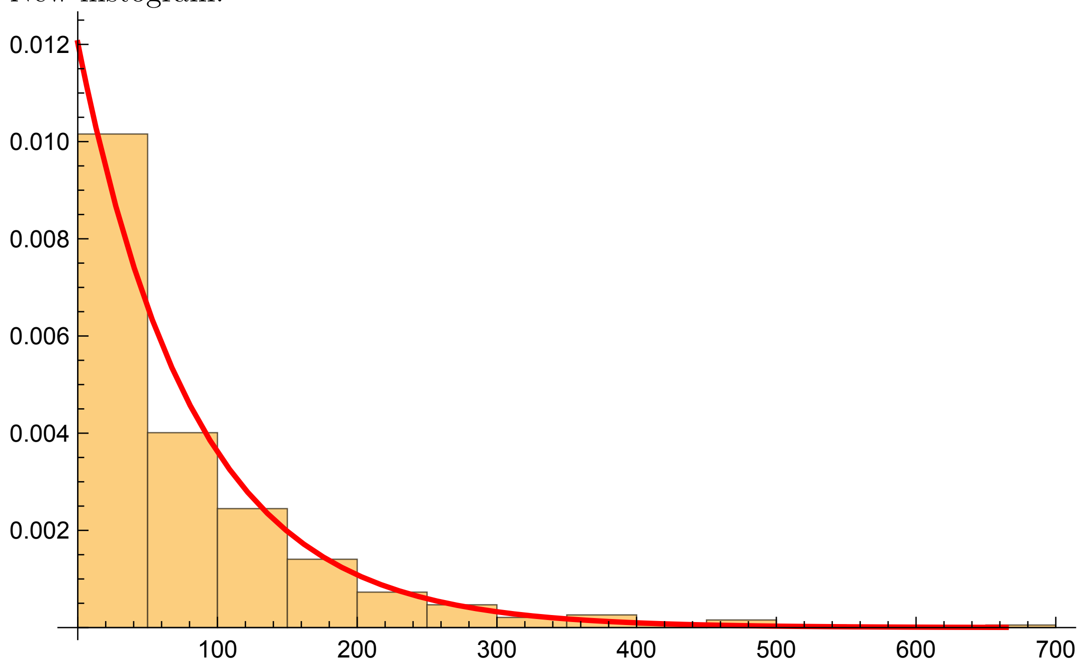
The fit is terrible

7. (10 pts) Use a Kolmogorov–Smirnov test to test the null hypothesis that your data are exponentially distributed with parameter $\hat{\lambda}_{\text{MLE}}$ versus the "two-sided" alternative hypothesis that the distributions are different. Does it agree with your visual assessment?

Mathematica 12 reports the Kolmogorov–Smirnov test statistic has a value of 0.462703, which has a p-value of essentially zero ($\approx 1.5 \times 10^{-162}$). So, yeah, the K–S test agrees with my visual assessment.

8. (30 pts) Redo parts (2)–(7) with the smaller dataset obtained by simply discarding all inter-arrival times less than four days. *Make sure to recompute your means and standard deviations, and your estimate of* λ*!*

New histogram:

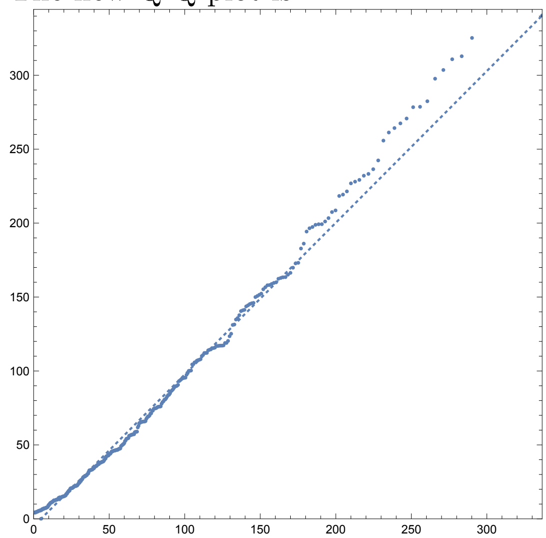


The new mean waiting time is 83.1 days, and the new standard deviation is 93.8 days, which are much closer to being equal.

The new

$$\hat{\lambda}_{\text{MLE}} = 0.012$$

The new Q-Q plot is



and the fit looks much better.

The new K–S statistic takes on a value 0.064, which has a reported p-value of 0.084. So we do not reject the null hypothesis that the inter-earthquake waiting times are exponentially distributed at the conventional 5% level of significance.

9. (10 pts) There is no real justification for the four-day minimum above. Suggest a more intelligent, but more time-consuming, approach to deciding which are aftershocks and foreshocks. (Hint: Look at the list of references.)

A more intelligent idea would be to consult the geology literature on foreshocks and aftershocks. For instance, Gardner and Knopoff[a] suggest a "windowing" procedure where windows in time and space are constructed around each quake, and quakes in that window are classified as foreshocks or aftershock of the larger quake.

We have the necessary data to do this, but automating such a procedure would require more programming than the current Ma 3 instructor is willing to impose on the class.

[a] Gardner and Knopoff, 1974, Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bulletin of the Seismological Society of America* 64(5):1363–1367. bssa.geoscienceworld.org/content/64/5/1363.full.pdf+html

10. (10 pts) How long should we expect to wait for the next magnitude 4.5+ quake?

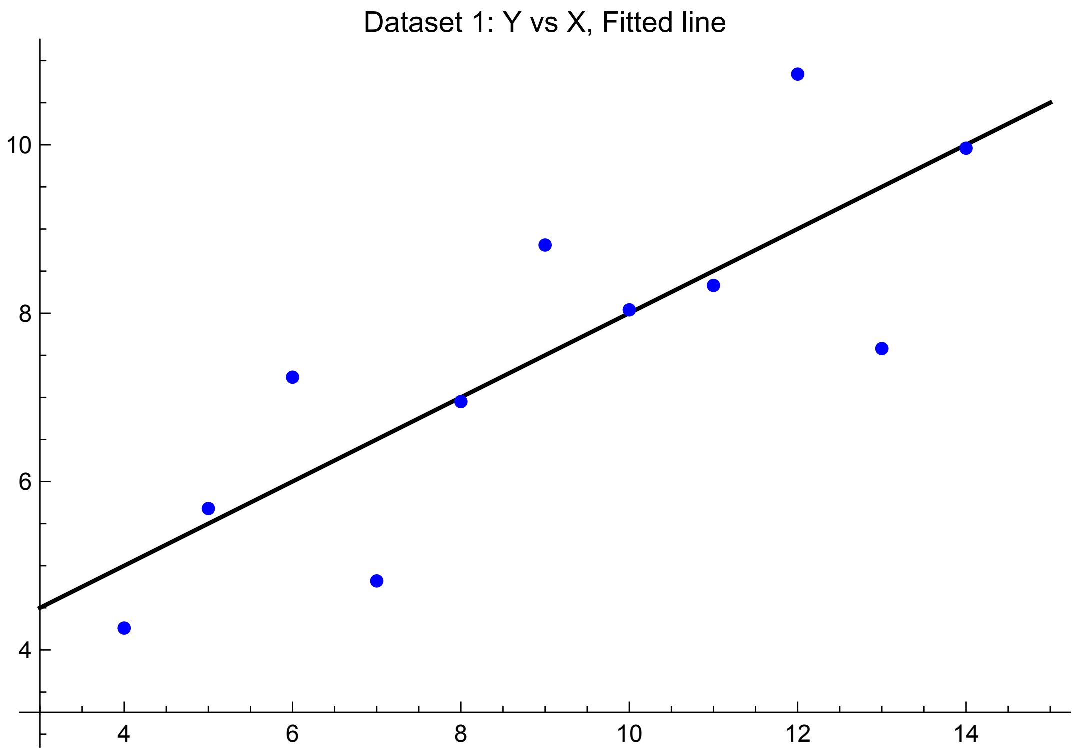
The expected waiting time between magnitude 4.5+ mainshocks is estimated to be 83.1 days, so that is how long we should expect to wait for the next one, somewhere in Southern California.

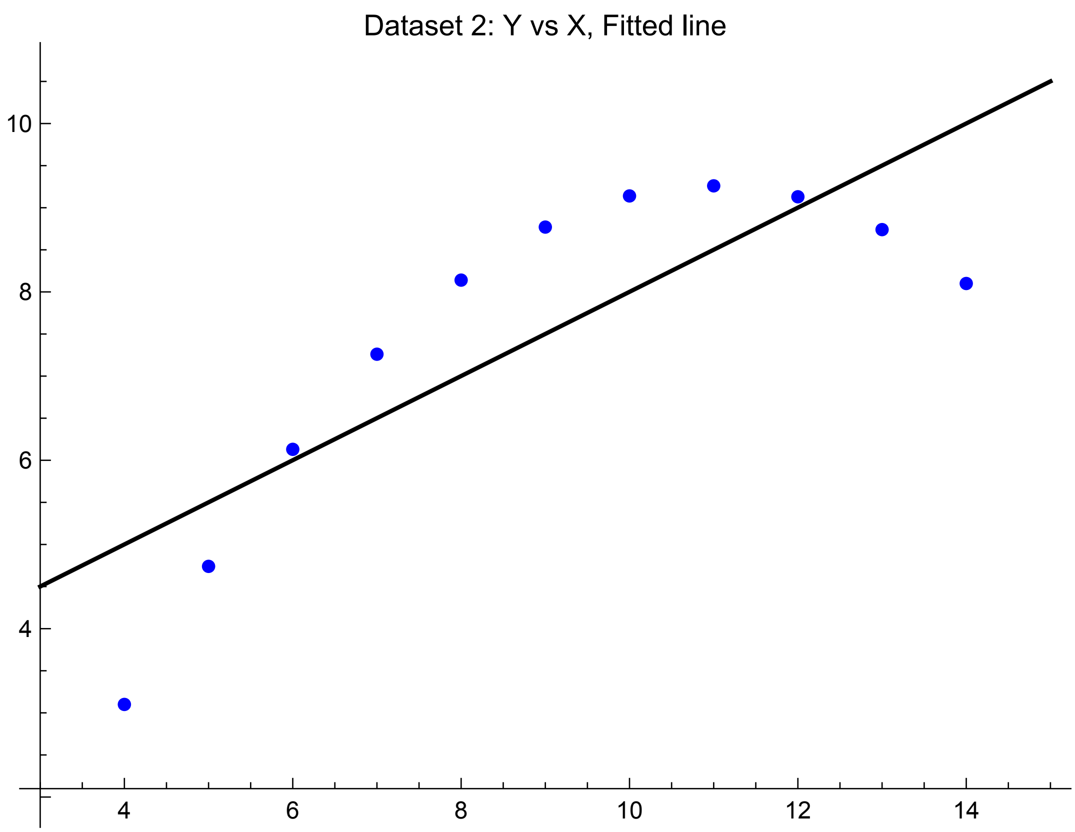
Sample answer for Exercise 3: Here are the (rounded) results reported by MATHEMATICA 12 from the analysis of the four datasets.

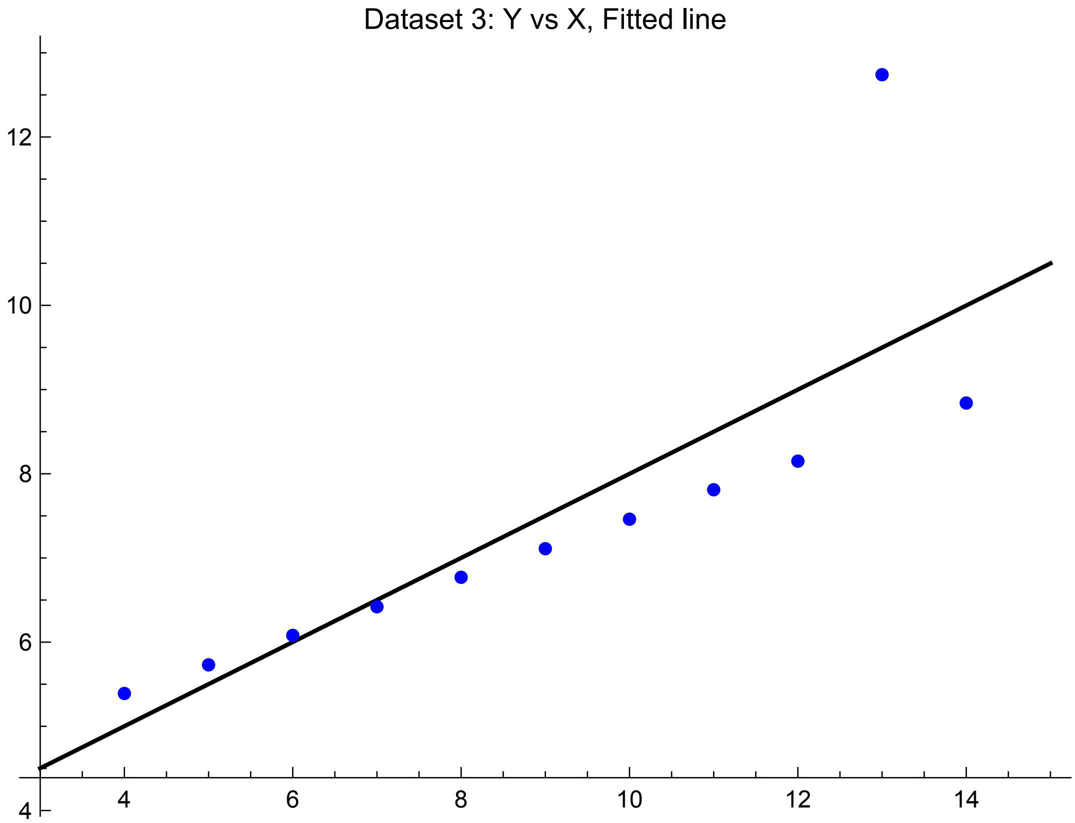
	Dataset 1	Dataset 2	Dataset 3	Datadatasetset 4
Mean of X	9.0	9.0	9.0	9.0
Std. Dev. of X	3.32	3.32	3.32	3.32
Mean of Y	7.5	7.5	7.5	7.5
Std. Dev. of Y	2.03	2.03	2.03	2.03
$\hat{\beta}_0$	3.0	3.0	3.0	3.0
t-statistic for $\hat{\beta}_0$	2.67	2.67	2.67	2.67
p-value of t-statistic for $\hat{\beta}_0$	0.026	0.026	0.026	0.026
Reject $H_0\colon \beta_0 = 0$? Y/N	Y	Y	Y	Y
$\hat{\beta}_1$	0.50	0.50	0.50	0.50
t-statistic for $\hat{\beta}_1$	4.24	4.24	4.24	4.24
p-value of t-statistic for $\hat{\beta}_1$	0.002	0.002	0.002	0.002
Reject $H_0\colon \beta_1 = 0$? Y/N	Y	Y	Y	Y
R^2	0.67	0.67	0.67	0.67
Adjusted $\bar{R}^2$	0.63	0.63	0.63	0.63
F-statistic for the regression	18.0	18.0	18.0	18.0
p-value of F-statistic	0.002	0.002	0.002	0.002
Sum of squared residuals	13.76	13.78	13.76	13.74
Reject $H_0\colon$ Normal residuals? Y/N	N	N	N	N

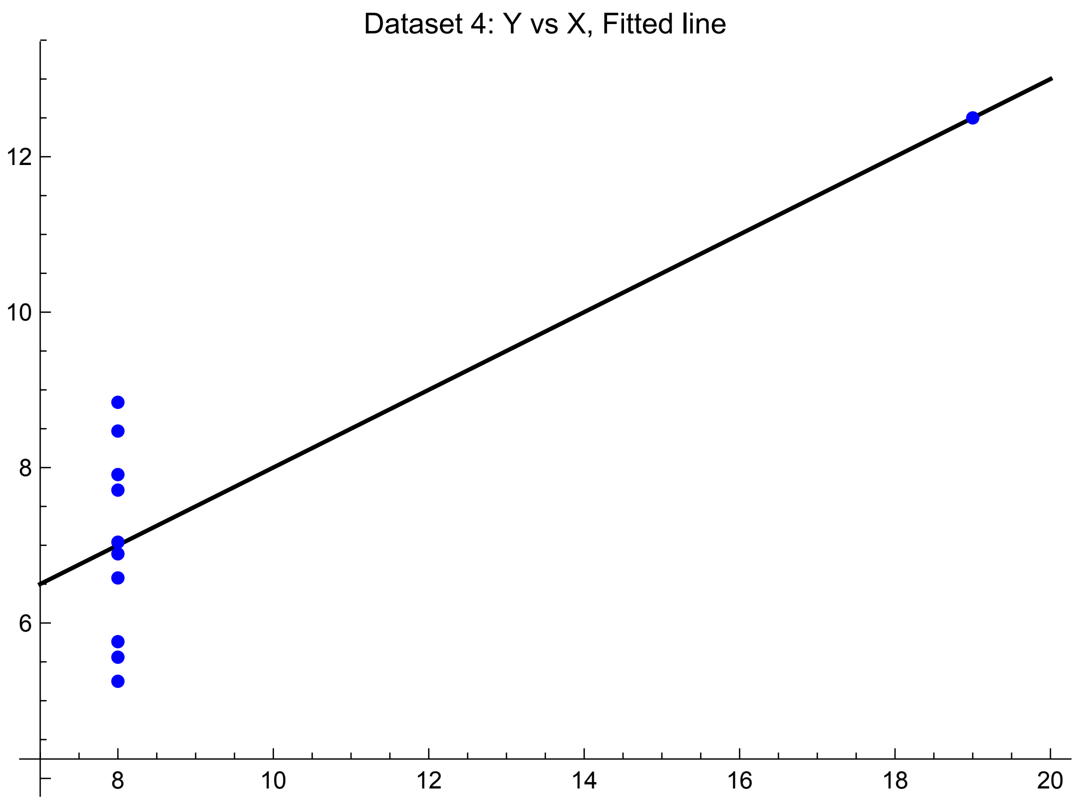
The summaries of the analysis of the four datasets are nearly identical. The only noticeable differences are in the total sum of squared residuals, and even they are very close.

Here are the X-Y scatter plots:

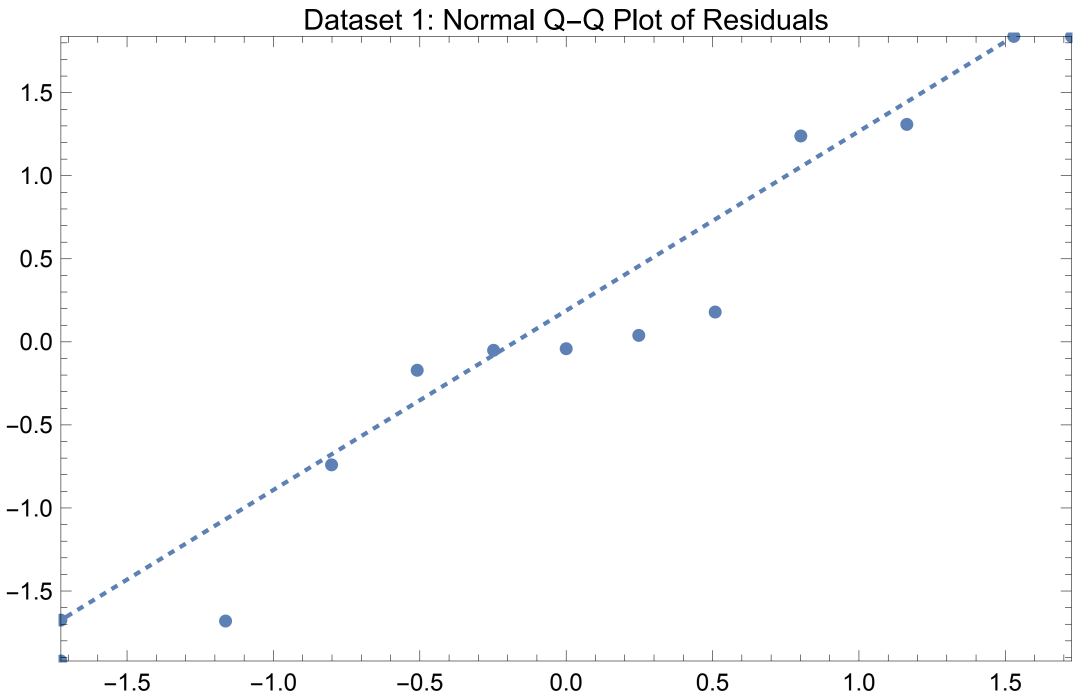


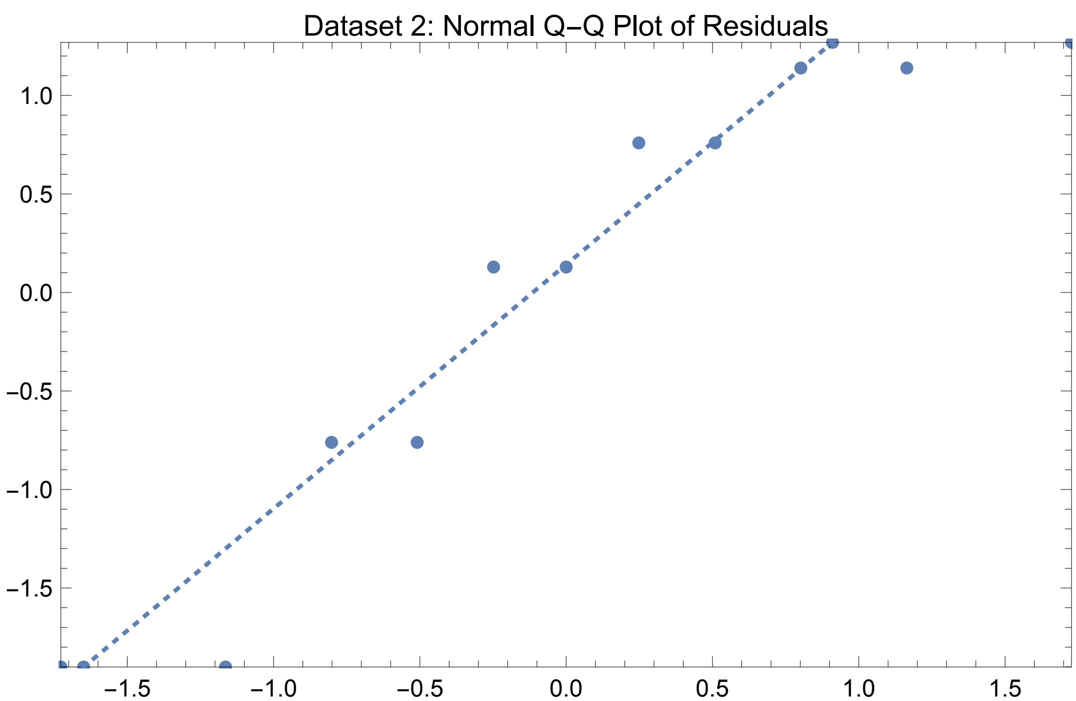


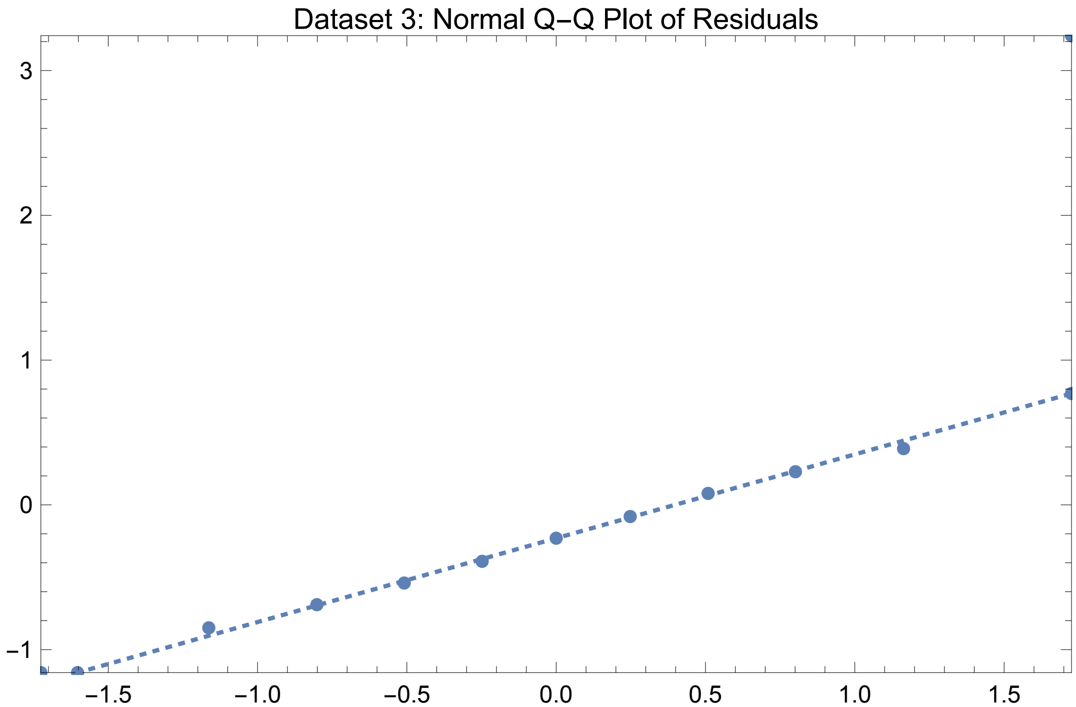

Dataset 3: Y vs X, Fitted line
12
10
8
6
4
4
6
8
10
12
14

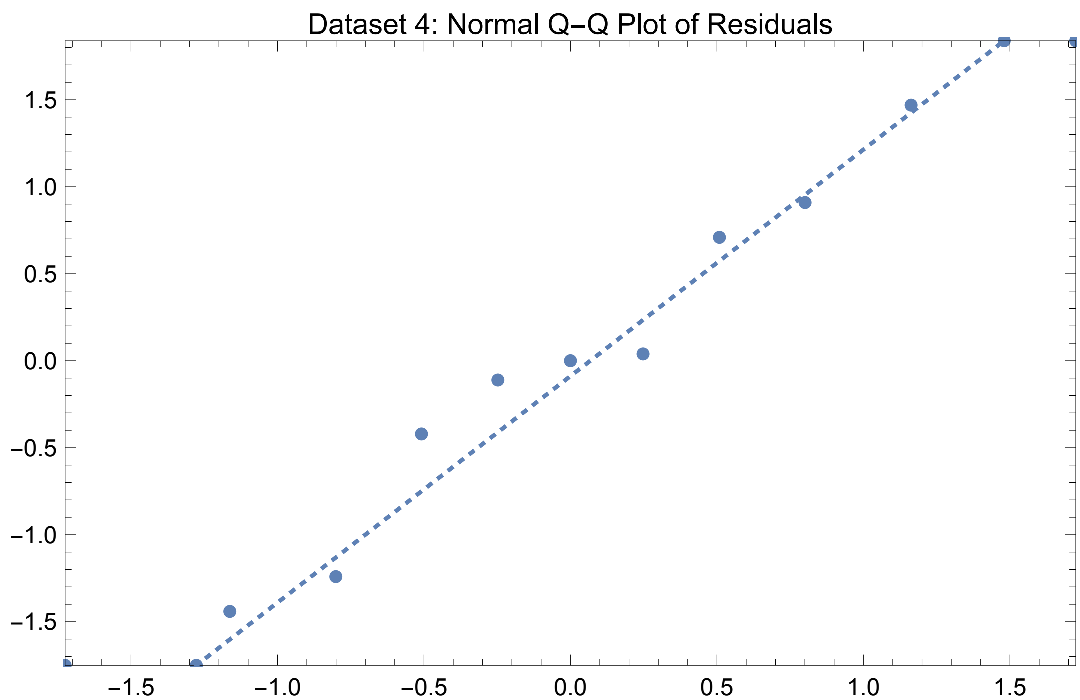

Dataset 4: Y vs X, Fitted line
12
10
8
6
8
10
12
14
16
18
20

And here are the Normal Q-Q plots of the residuals.






Dataset 3: Normal Q–Q Plot of Residuals
3
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Dataset 4: Normal Q–Q Plot of Residuals
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

The X-Y scatter plots are the only way to make it clear that these are very different datasets.

- Dataset 1 seems like the best instance of the standard linear model. The fit is loosely linear and there is no apparent pattern to the residuals.
- For dataset 2, it is clear from the scatter plot that a much better model would be that Y is a quadratic function of X, not linear.
- Dataset 3 looks like a linear relationship, but the model may have misestimated the slope and intercept, because of a single outlying datum.
- Dataset 4 is unusual in that there are only two distinct X-values. The slope is entirely determined by a single datum, and so is really more unreliably estimated than in dataset 1.

◇