Q-Q Plot

Bill Foote

December 2, 2017

What's a Q-Q plot?

Any quantile-to-quantile plot will plot on the x-axis the quantiles of one variable and on the y-axis the quantiles of the other variable. In the case of ordinary least squares, we want to know the answer to this question:

> Are the error terms (residuals of the regression) normally distributed?

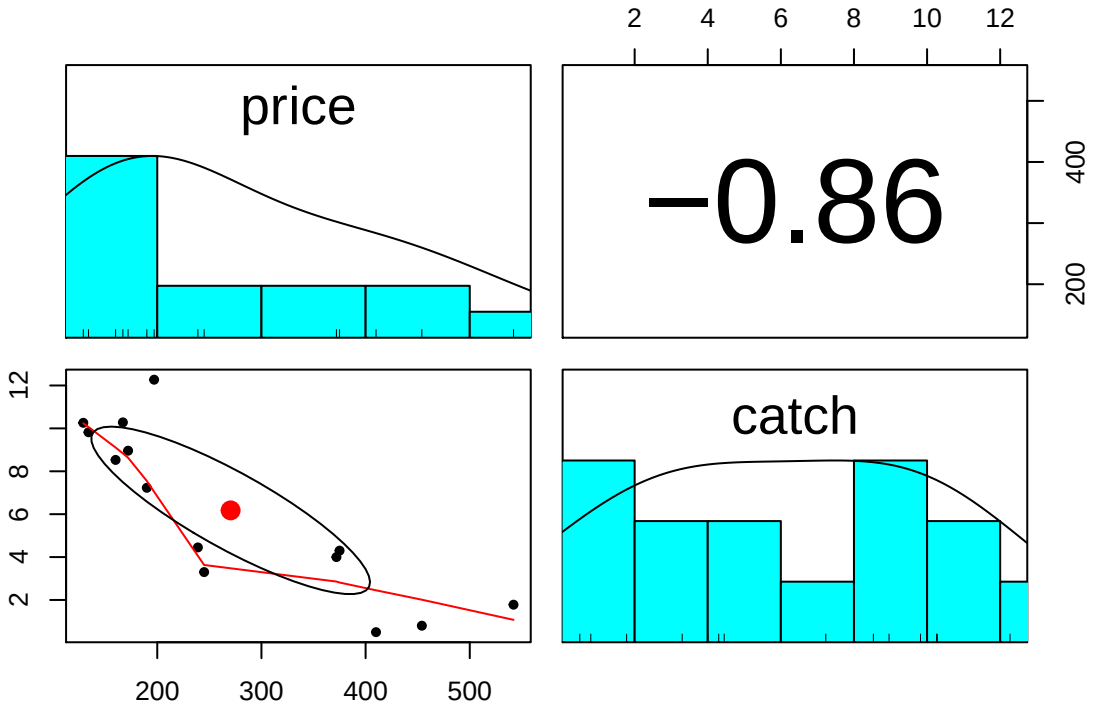
One way to answer this question is to compare the quantiles of the error terms against the quantiles of the normal distribution. If the quantiles of the error terms are near enough to the quantiles of the corresponding values computed from the normal distribution, then we might begin to accept the idea that the error terms are normally distributed.

Other tests based on this idea include the Jarque-Bera test that tries to match jointly under the null hypothesis the skewness and kurtosis of sample error terms to the normal distribution. If the data rejects the null hypothesis at an α level of significance, then the inference of regression parameters and of the regression itself may be called into statistical question.

Let's read some data

The data is a data frame of Peruvian anchovie prices in USD per metric ton (Y) and size of catch in millions of metric tons (X). Histograms, scatterplots, and a correlation are displayed in a scatter plot matrix.

```
require(psych)
XY_df <- read.csv("peruvian_anchovies.csv")
pairs.panels(XY_df)
```



We see a clear "kink" in the relationship where catch behavior seems to switch to a smaller slope at a higher level of catch. This signals a possible non-normal response of price to the size of the catch.

Run a regression

Let's now run a simple linear regression using the assumption that error terms are normally distributed. Without the normal distributional assumption we can describe the relationship using sample slope and intercept parameters b_0 and b_1 as

```
lm_fit <- lm(price ~ catch, data = XY_df)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = price ~ catch, data = XY_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -110.00  -38.29  -19.03   34.57  142.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  451.989     36.794  12.284 3.72e-08 ***
## catch        -29.392      5.087  -5.778 8.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.63 on 12 degrees of freedom
## Multiple R-squared:  0.7356, Adjusted R-squared:  0.7136
## F-statistic: 33.39 on 1 and 12 DF,  p-value: 8.766e-05
```

So far everything looks singificantly different from zero. Ridiculously low *p-values* indicate reject of null

hypotheses that either / or intercepts and slope parameters are zero at least at the 1% level of significance or far less. The sign of the slope parameter is negative indicating a typical demand curve.

Let's look at normality

The scatter plot clearly indicates that price and catch are not very normal. We can calculate also

```
# require(moments) will capture kurtosis
(price_skew <- skew(XY_df$price))
```

```
## [1] 0.6371672
```

```
(price_kurt <- kurtosis(XY_df$price))
```

```
## [1] 2.15725
```

```
(catch_skew <- skew(XY_df$catch))
```

```
## [1] -0.02634619
```

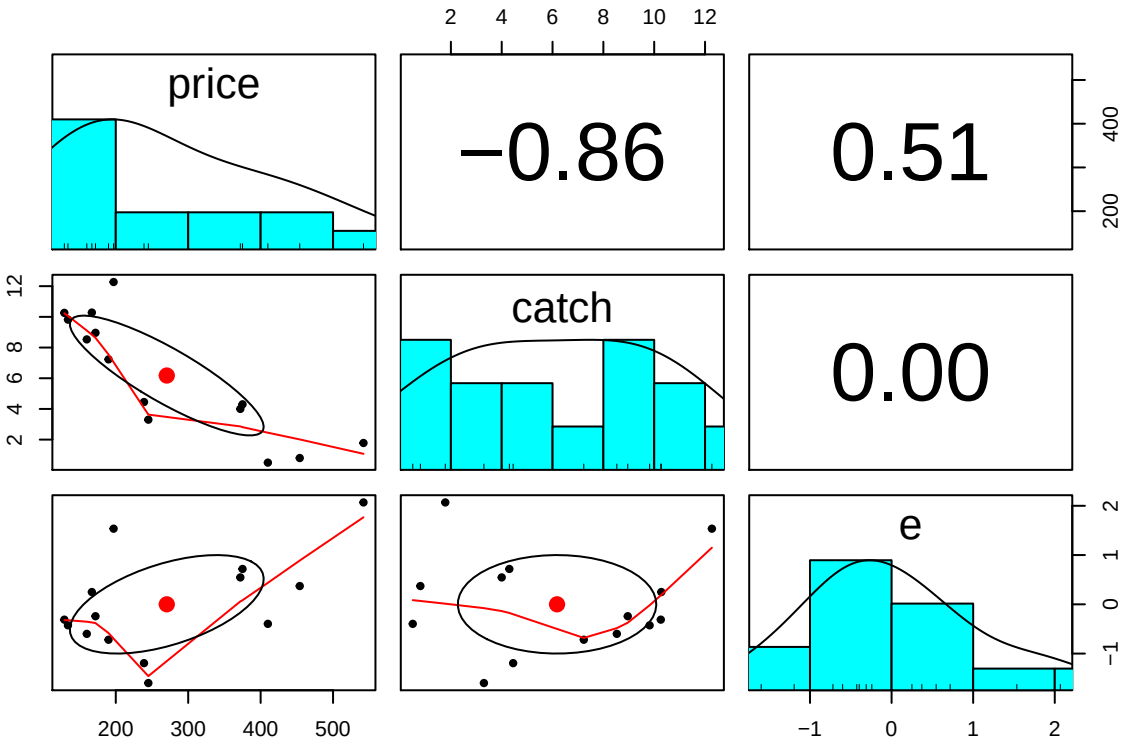
```
(catch_kurt <- kurtosis(XY_df$catch))
```

```
## [1] 1.626601
```

Price and catch have kurtosis different from 3.0, the normal distribution kurtosis. Their skewness is nearly zero, meaning they are almost symmetric. Thick tails then may indicate non-normal variates in this relationship.

Let's also calculate the error terms.

```
b0 <- lm_fit$coefficients[1]
b1 <- lm_fit$coefficients[2]
e <- XY_df$price - b0 - b1*XY_df$catch
e <- e / sd(e)
XY_df$e <- e
pairs.panels(XY_df)
```




```
summary(XY_df)
```

```
##      price           catch              e
##  Min.   :129.0   Min.   : 0.500   Min.   :-1.5983
##  1st Qu.:168.2   1st Qu.: 3.475   1st Qu.:-0.5564
##  Median :218.0   Median : 5.840   Median :-0.2765
##  Mean   :270.4   Mean   : 6.177   Mean   : 0.0000
##  3rd Qu.:374.2   3rd Qu.: 9.605   3rd Qu.: 0.5023
##  Max.   :542.0   Max.   :12.270   Max.   : 2.0682
```

```
(e_skew <- skew(XY_df$e))
```

```
## [1] 0.4646875
```

```
(e_kurt <- kurtosis(XY_df$e))
```

```
## [1] 2.755021
```

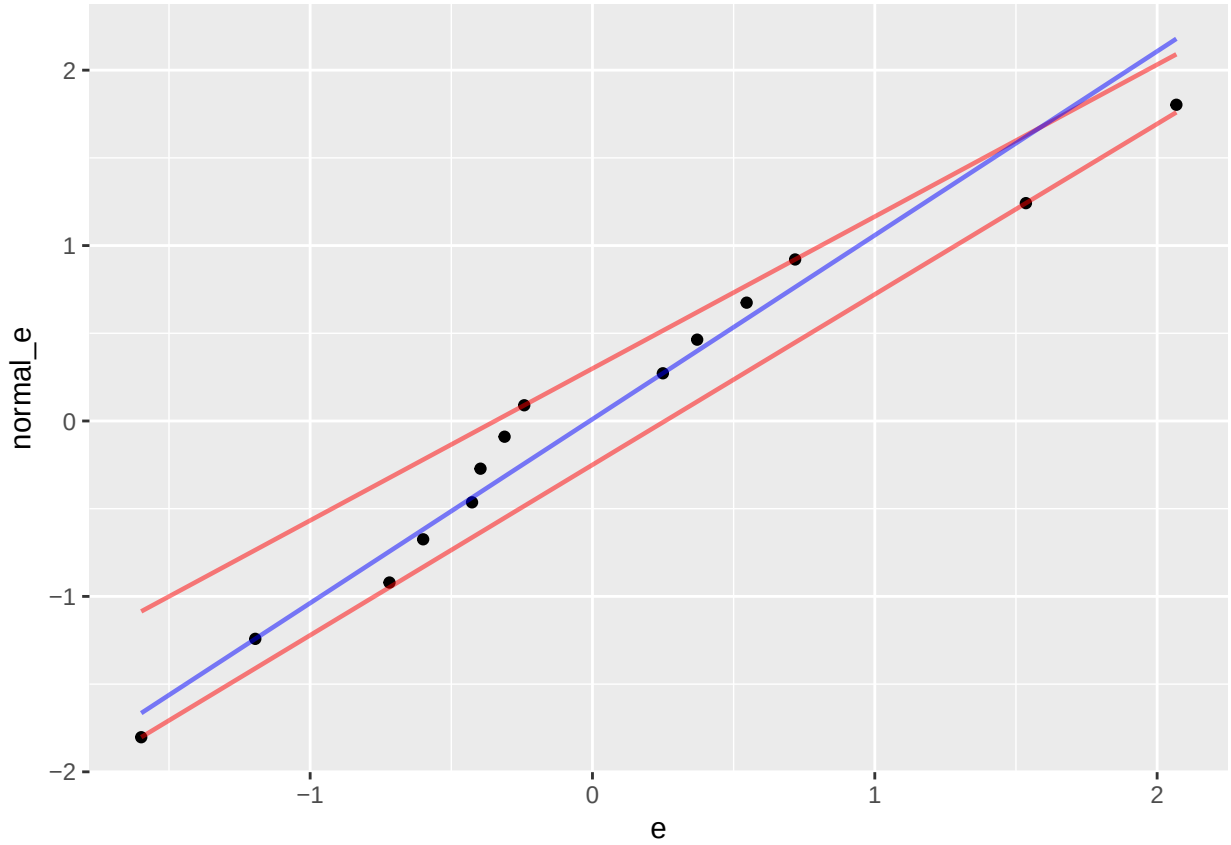
By definition of `e` the mean is zero. We have a small skewness and a kurtosis close to 3.0, the normal distribution kurtosis. But how close?

Eyeballing `e` we see a slightly right-skewed distribution with potentially thick tails: not normally distributed. Let's overlay a normal distribution onto the empirical `e` distribution.

```
# require(ggplot2) will capture suite of ggplot + geom calls
n <- length(XY_df$e)          # number of observations
r <- order(order(XY_df$e))    # order of values, i.e. ranks without averaged ties
p <- (r - 1/2) / n       # assign to ranks using Blom's method
y <- qnorm(p)            # theoretical standard normal quantiles for p values
XY_df$normal_e <- y      # append column of theoretical normal quantiles
plt <- ggplot(XY_df, aes(x = e, y = normal_e)) # set up canvas of x and y values
plt <- plt + geom_point()                      # plot points
plt <- plt + stat_quantile(colour = "red", size = 0.8, alpha = 0.5, geom = "quantile", quantiles = c(0.
plt <- plt + stat_quantile(colour = "blue", size = 0.8, alpha = 0.5, geom = "quantile", quantiles = 0.5
```

```
plt
```

```
## Smoothing formula not specified. Using: y ~ x
## Smoothing formula not specified. Using: y ~ x
```



The median line is through the middle of the scatter plot of `normal_e` versus actual `e`. There are no significant deviations from the 95% confidence lines that would have led us to believe that `e` is not necessarily normal. Thus our inferences using the ordinary least squares approach assuming normal error terms should be acceptable.

Here are more precise calculations using "quantile" regression. In these regressions, instead of using regression of the independent variable X against the arithmetic mean of the dependent variable Y we are using a quantile of the dependent variable Y. So, instead of using the arithemetic mean we are using quantiles of 50% (the median), 5% (lower bound) and 95% (upper bound) of the dependent variable.

```
# require(quantreg) brings in the lm version of quantile regression rq
rq_fit_05 <- rq(normal_e ~ e, data = XY_df, tau = 0.05) # lower 5%
rq_fit_50 <- rq(normal_e ~ e, data = XY_df, tau = 0.50) # median 50%
rq_fit_95 <- rq(normal_e ~ e, data = XY_df, tau = 0.95) # upper 95%
summary(rq_fit_05, se = "boot")
```

```
##
## Call: rq(formula = normal_e ~ e, tau = 0.05, data = XY_df)
##
## tau: [1] 0.05
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept) -0.24980  0.08143   -3.06756  0.00976
```

```
## e            0.97160  0.09489   10.23931  0.00000
```

```
summary(rq_fit_50, se = "boot")
```

```
##
## Call: rq(formula = normal_e ~ e, tau = 0.5, data = XY_df)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 0.01038 0.07313    0.14193 0.88949
## e           1.04847 0.12672    8.27407 0.00000
```

```
summary(rq_fit_95, se = "boot")
```

```
##
## Call: rq(formula = normal_e ~ e, tau = 0.95, data = XY_df)
##
## tau: [1] 0.95
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 0.29902 0.08256    3.62180 0.00350
## e           0.86626 0.20785    4.16765 0.00130
```

At the lower 5% bound, the relationship does not seem to fit normal distribution expectations to some extent. A null hypothesis of the slope equal to 1 would seem to be probably acceptable lending credence to normally distributed error terms. A null hypothesis of an intercept equal to zero is not. We have some little bit of evidence of non-normality in the lower range of error terms.

At the median 50%, the relationship does seem to fit normal distribution expectations to some extent as well. A null hypothesis of the slope equal to 1 would seem to be probably acceptable. However a null hypothesis of an intercept equal to zero is not. But the number of standard deviations measured by the intercept 0.010138 is indeed small.

As we meander out into the 95% bound, the relationship moves positively from a zero intercept, while maintaining a slope that is different from the lower bound and median slopes. We have some non-normality in the upper range of the support of error terms.

Overall, we should accept the null hypothesis that the error terms are normally distributed.

One other test, the Jarque-Bera test, looks at skewness and excess kurtosis and calculates a sample test statistic that follows a χ^2 distribution with 2 degrees of freedom. Using this test we calculate

```
# require(tseries) will enable the Jarque-Bera test
n_sim <- length(XY_df$price)
x <- rnorm(n_sim, mean = 0, sd = sd(XY_df$e))  # under the null hypothesis e~ N(0, se^2)
jarque.bera.test(x)
```

```
##
##  Jarque Bera Test
##
## data:  x
## X-squared = 0.8206, df = 2, p-value = 0.6635
```

```
x <- e  # alternative
jarque.bera.test(x)
```

```
##
##  Jarque Bera Test
##
## data:  x
## X-squared = 0.6643, df = 2, p-value = 0.7174
```

The χ^2 statistic with 2 degrees of freedom is 0.6643. The *p-value* of 71% is greater than a 1% significance level. This indicates an acceptance of the null hypothesis that the `e` series is normally distributed.

The quantile regressions do indicate some marginal non-normality while the Jarque-Bera test unequivocally does not. To tread carefully, we should err on the side of the (slim) poosibility of non-normal error terms. More data would be helpful to resolve any further issues with heteroskedasticity in the error terms.

How to fix?

We could transform variables. We could also simply split the regression sample into two parts as indicated by the scatter matrix. We see a switch point at a level of catch and price. The "price" we pay with this very small sample is the power (type II error) of the regression results with such small sub-sample sizes. The answer is to use median regression that is robust with respect to potentially non-normally distributed error terms. We have indeed used this approach to describe the upper and lower bounds on the 95% interval that confirms the Jarque-Bera test on the one hand, and provides a nuanced analysis of the distribution of error terms on the other hand.

References

Jarque, Carlos M. and Bera, Anil K. (1980). "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". Economics Letters. 6 (3): 255–259.

Koenker, Roger (2005). Quantile Regression. Cambridge University Press.