Vladimir Kustov

Regression Analysis

Fall 2010

Student Project

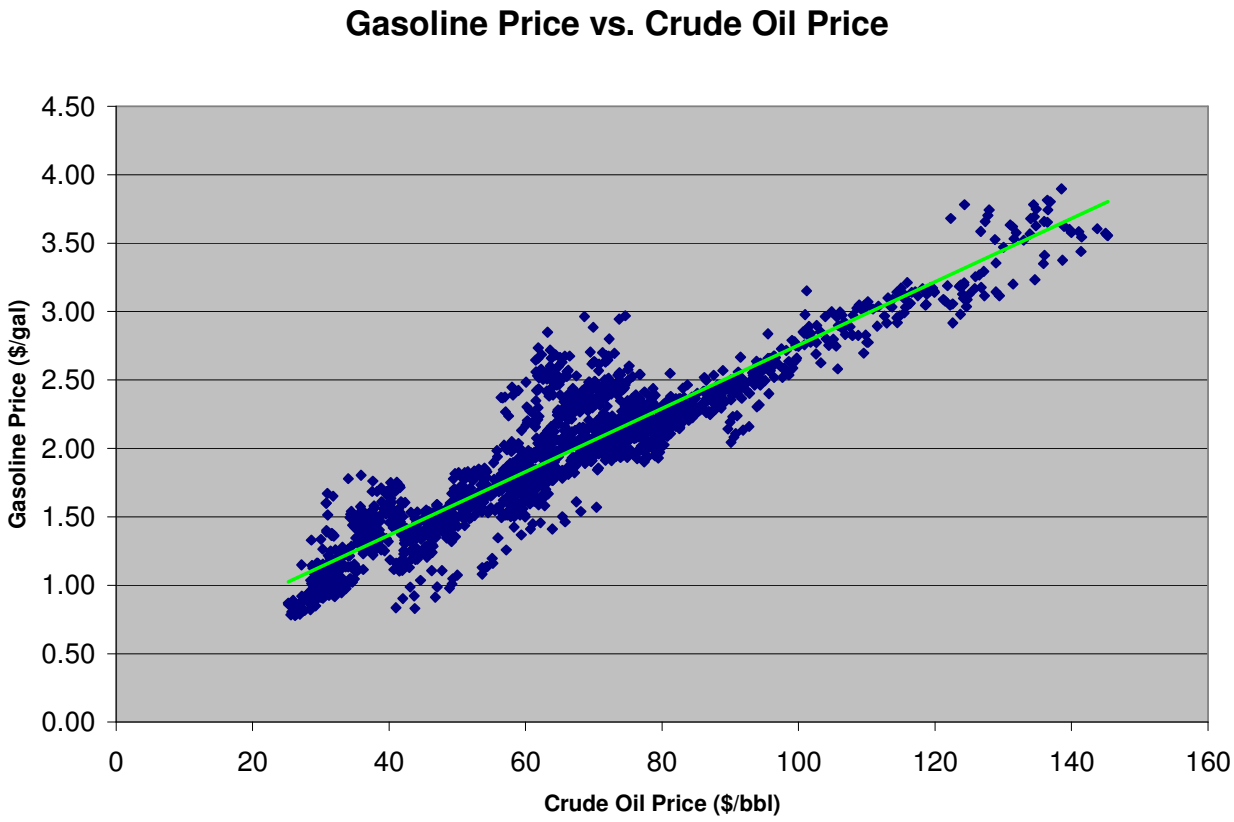
Crude Oil and Gasoline Futures

In this project I compare price of gasoline futures and price of crude oil, both traded on New York Mercantile Exchange, using simple linear regression. While it is obvious that the prices are positively correlated, it is not obvious in advance what model would be appropriate.

The data have been obtained at http://www.eia.doe.gov/dnav/pet/pet_pri_spt_s1_d.htm (US Energy Information Administration) and consist of two datasets available in the Excel format: daily prices of West Texas intermediate (WTI), a type of crude oil, and daily prices of Reformulated Gasoline Blendstock for Oxygen Blending (RBOB) Gasoline Futures over a period from 2003 to 2011.

For this project I used simple linear regression implemented in SAS.

The following scatter plot is constructed in Excel and represents gasoline futures and crude oil prices, along with a regression line. The scatter plot gives an overall impression that a linear model might be quite appropriate here.



SAS Code:

```
/* reading data --> */

data dog (drop=dt);
format date date9.;
infile 'F:/regdata/data_dog.csv' dsd missover;
input dt $ oil gas;
date=input(dt,mmddyy10.);
run;

/* end of reading data */


proc reg data=dog;
model gas=oil; /* simple linear regression: gas price vs. oil price */
output out=res_ds p=yhat r=yresid student=student;
run;
quit;

goptions reset; /* plotting residuals against predicted values */
symbol1 v=circle;
proc gplot data=res_ds;
plot student*yhat/vref=0;
run;

proc univariate data=res_ds; /* histogram and q-q plot for residuals */
histogram student;
qqplot student /normal (mu=est sigma=est);
run;
```

SAS Output:

The REG Procedure
Model: MODEL1
Dependent Variable: gas

Number of Observations Read	1968
Number of Observations Used	1968

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	591.34997	591.34997	12874.8	<.0001
Error	1966	90.29963	0.04593		
Corrected Total	1967	681.64959			

Root MSE	0.21431	**R-Square**	0.8675
Dependent Mean	1.93131	**Adj R-Sq**	0.8675
Coeff Var	11.09683		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	0.44030	0.01400	31.45	<.0001
oil	**1**	0.02314	0.00020392	113.47	<.0001

Results:

$$\text{gas\_price} = 0.440 + 0.023 * \text{oil\_price},$$

where oil price is in $ per barrel and gas price is in $ per gallon.

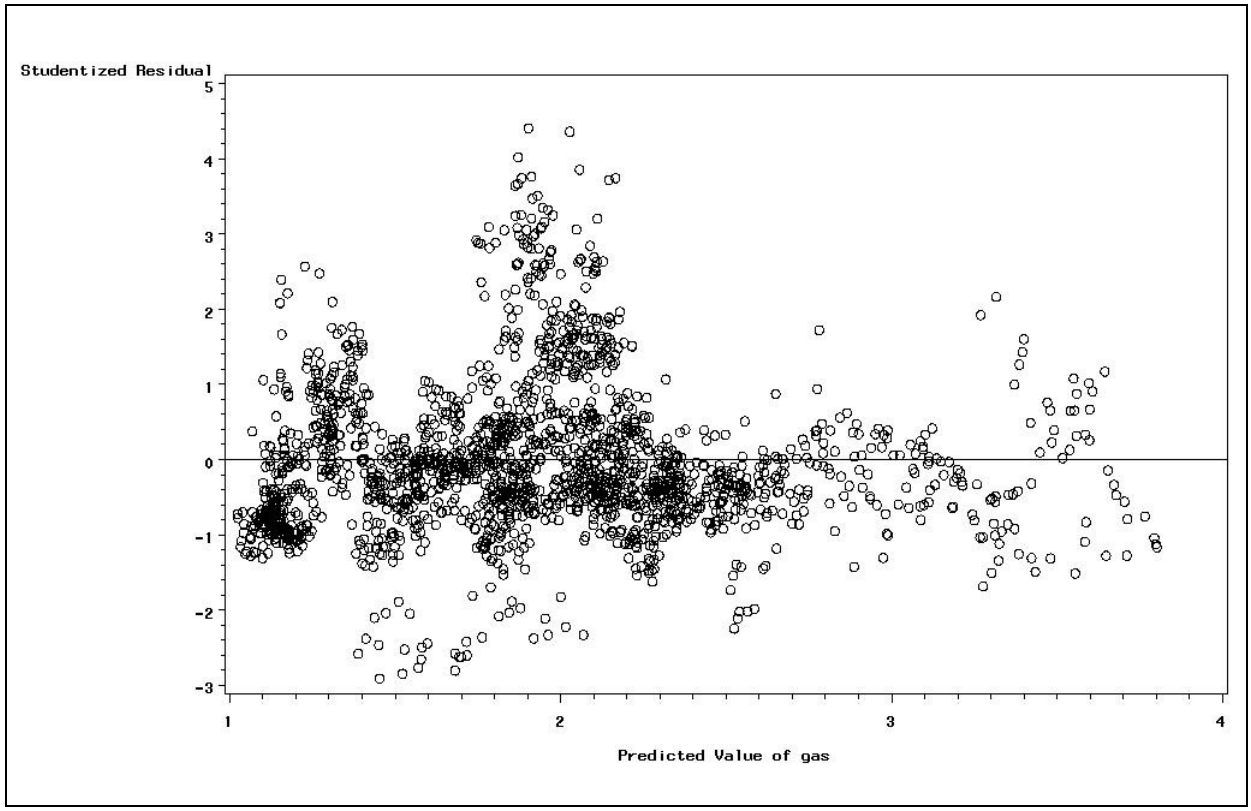
Since 1 oil barrel is equivalent to 42 gallons, the equation can be rewritten as

$$gas\_price = 0.440 + 0.966*oil\_price,$$

where oil price is measured in $/gallon. This means that for each gallon futures price of gasoline increases by 96.6 cents per each $1 increase in price of oil. This almost gallon-per-gallon equivalence is not something to be expected in advance.
The small p-values strongly reject the null hypothesis that the regression coefficients might be zero. The value of $R^2 = 87\%$ also seems very optimistic.

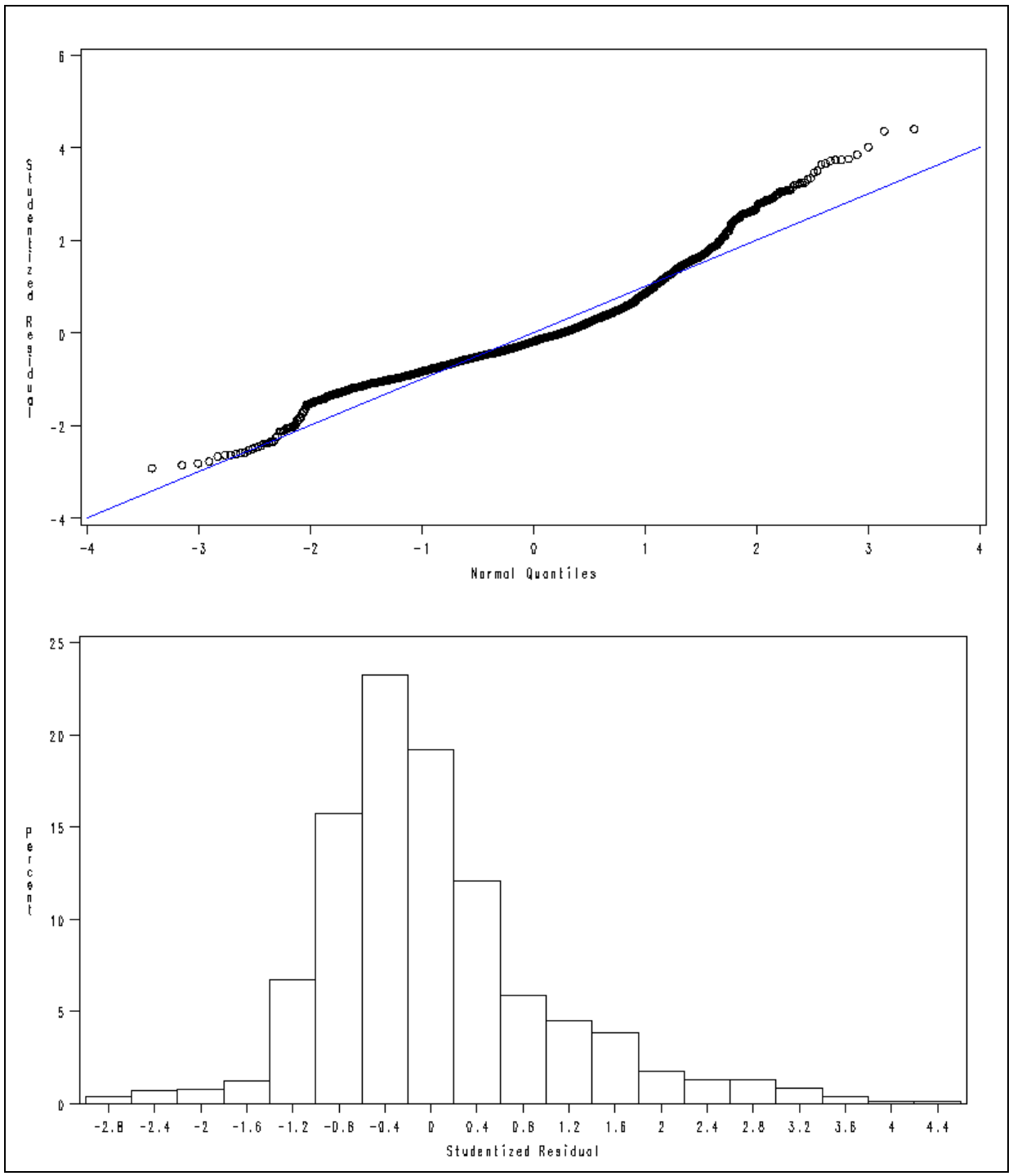
Next, we will analyze distribution of the residuals.

Studentized residuals vs. predicted value:

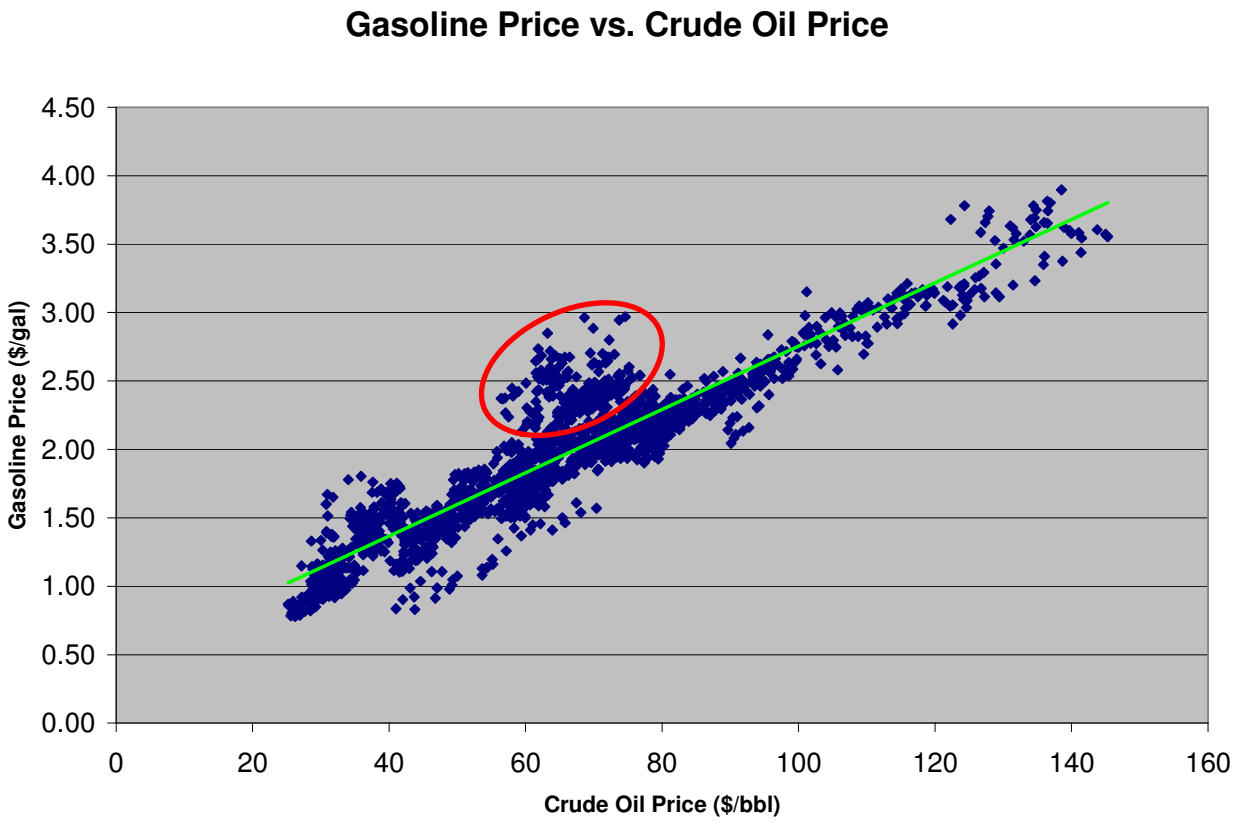


The range in which the residuals fluctuate is not quite consistent: residuals seem to vary more when predicted values of gasoline price are around 1.3 and 2 ($/gallon).

Q-Q plot and histogram for studentized residuals:

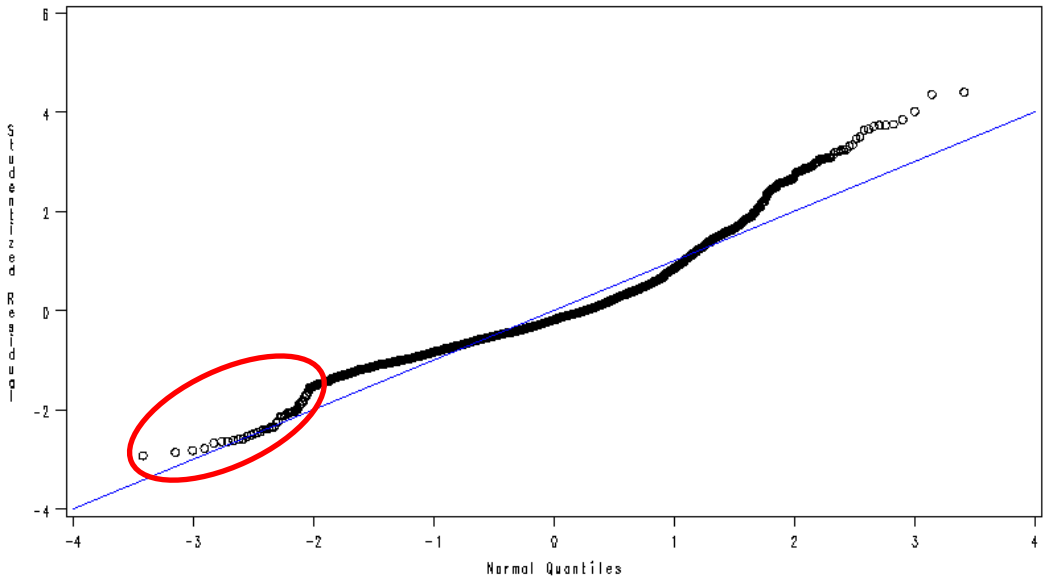


As we can see, the residuals have a heavy right tail. This can be noticed on the original graph of gasoline price vs. oil price: the central region contains many data points that are far above the regression line, while the points below the line tend to be closer to it (see picture below).



As can be seen from the q-q plot and residual plot, there is no consistent distribution of residuals throughout the range of predicted values which makes it difficult to suggest alternative models. Interestingly, many points in the circle region come from May 2007, when the price of crude oil was anomalously lower from world oil prices. This could move some points of this group to the right moving thus reducing their vertical distance from the regression line.

Let us again at the q-q plot again. Without the most left piece (see picture below), the curve looks like a U-shape which suggests a distribution with a long right tail, so we could try to use a generalized linear model with gamma-distributed errors.



Conclusions: A simple linear model seems to be an appropriate simple solution when comparing gasoline futures and crude oil prices. However such a linear model is not the most efficient solution among other possible alternatives. Using variable transformations or a generalized linear model does not seem adequate here since there is no consistent pattern of how residual errors are distributed throughout the data.