Paper 5620-2016

Quantile Regression versus Ordinary Least Squares Regression

Ruth Croxford, Institute for Clinical Evaluative Sciences

ABSTRACT

Regression is used to examine the relationship between one or more explanatory (independent) variables and an outcome (dependent) variable. Ordinary least squares regression models the effect of explanatory variables on the average value of the outcome. Sometimes, we are more interested in modeling the median value or some other quantile (for example, the 10th or 90th percentile). Or, we might wonder if a relationship between explanatory variables and outcome variable is the same for the entire range of the outcome: Perhaps the relationship at small values of the outcome is different from the relationship at large values of the outcome. This presentation illustrates quantile regression, comparing it to ordinary least squares regression. Topics covered will include: a graphical explanation of quantile regression, its assumptions and advantages, using the SAS® QUANTREG procedure, and interpretation of the procedure's output.

INTRODUCTION

Regression is used to learn about the associations between one or more predictors (the independent variables) and the value of an outcome (the dependent variable). The value of the outcome varies from one observation to another, and you'd like to understand why, or you'd like to be able to predict the value of the outcome for a new observation.

The generic linear regression equation relates the values of an observation's predictors to the value of its outcome:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

The ε's, or error terms, allow two observations with identical values of the x's to nevertheless have different values for their outcomes.

The goals of regression analysis include obtaining numeric estimates for each of the coefficients (the β's). The value of a coefficient estimates how much a 1-unit change in the value of the corresponding predictor will cause the value of the outcome to change. This allows the results to be used to predict the value of the outcome for a new observation, given that observation's characteristics. As well as estimating the value of each coefficient, the regression analysis provides its standard error and p-value, allowing inferences to be made about the association between the predictor and the outcome.

For continuous outcome variables, probably the most common type of regression is ordinary least squares (OLS) regression. The estimates of the coefficients provided by OLS regression are those which minimize the sum of the squared residuals (for each observation, the residual is the difference between the actual value of the outcome and the value that is predicted using the regression equation). SAS provides a number of procedures for OLS regression, including the GLM and REG procedures.

Figure 1 shows a graph of two variables: the predictor, X, on the horizontal axis and the outcome, Y, on the vertical axis. The appearance of the graph is sufficient to justify using a straight line to model the data: that is, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

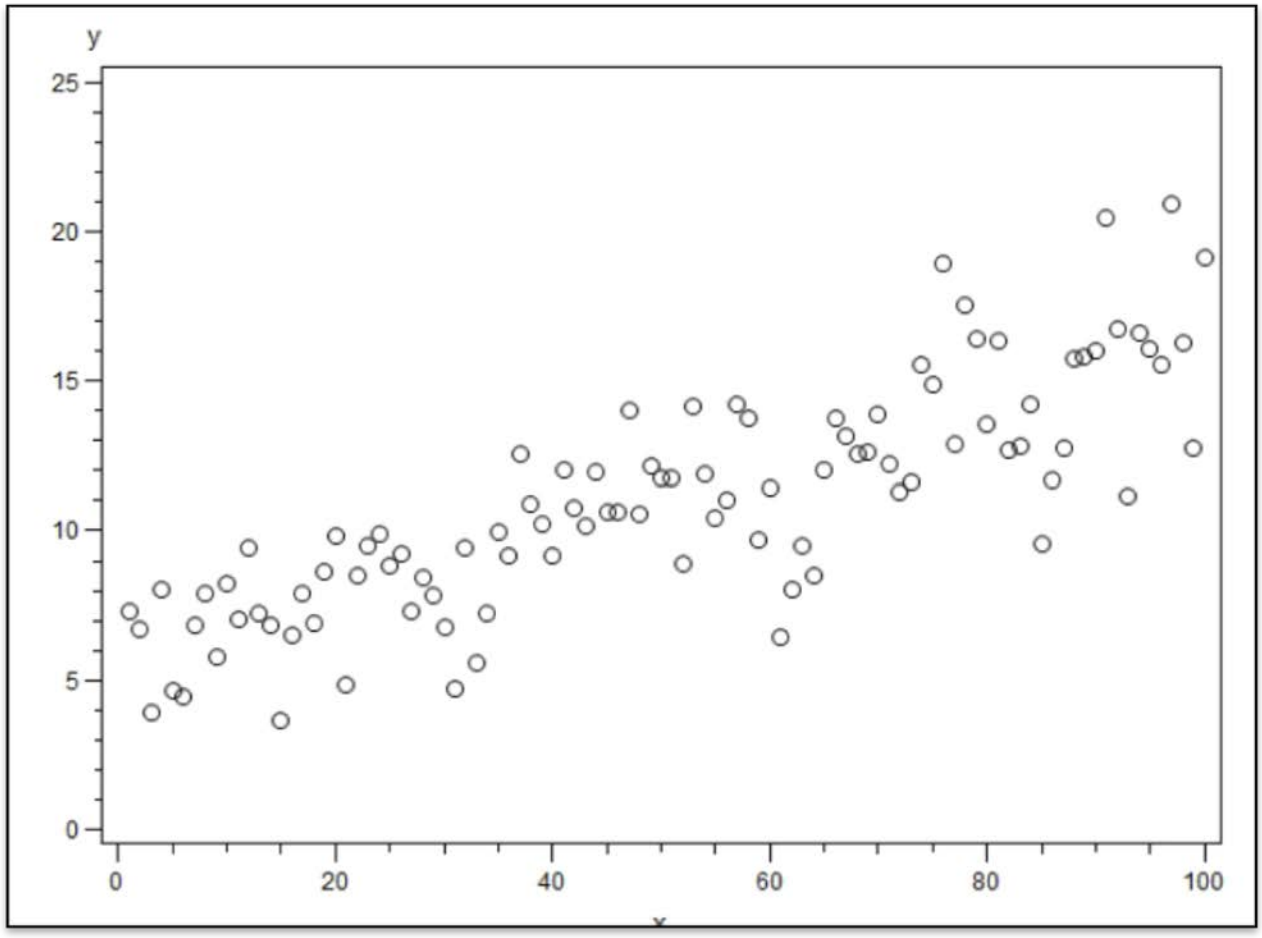

Figure 1. Scatterplot of a linear relationship, with normally distributed errors of constant variance

The analysis, using PROC REG, tells me that the estimated value of β_0 (the intercept) is 5.6 and the estimated value of β_1 (the slope) is 0.107 (with a standard error of 0.007, $p < 0.0001$). I conclude that there is evidence that, the slope, is not equal to 0, and that each 1-unit change in X will result in an estimated change of 0.107 in the value of Y. The standard error the slope can be used to construct its confidence interval.

Figure 2, created using the SGPLOT procedure, shows the same data, overlaid with the regression line. The line represents the value of Y which is predicted for each value of X. The shaded area on the graph shows the 95% confidence limits for the mean value of Y for each value of X, while the dashed lines are the 95% confidence limits for individual predicted values.

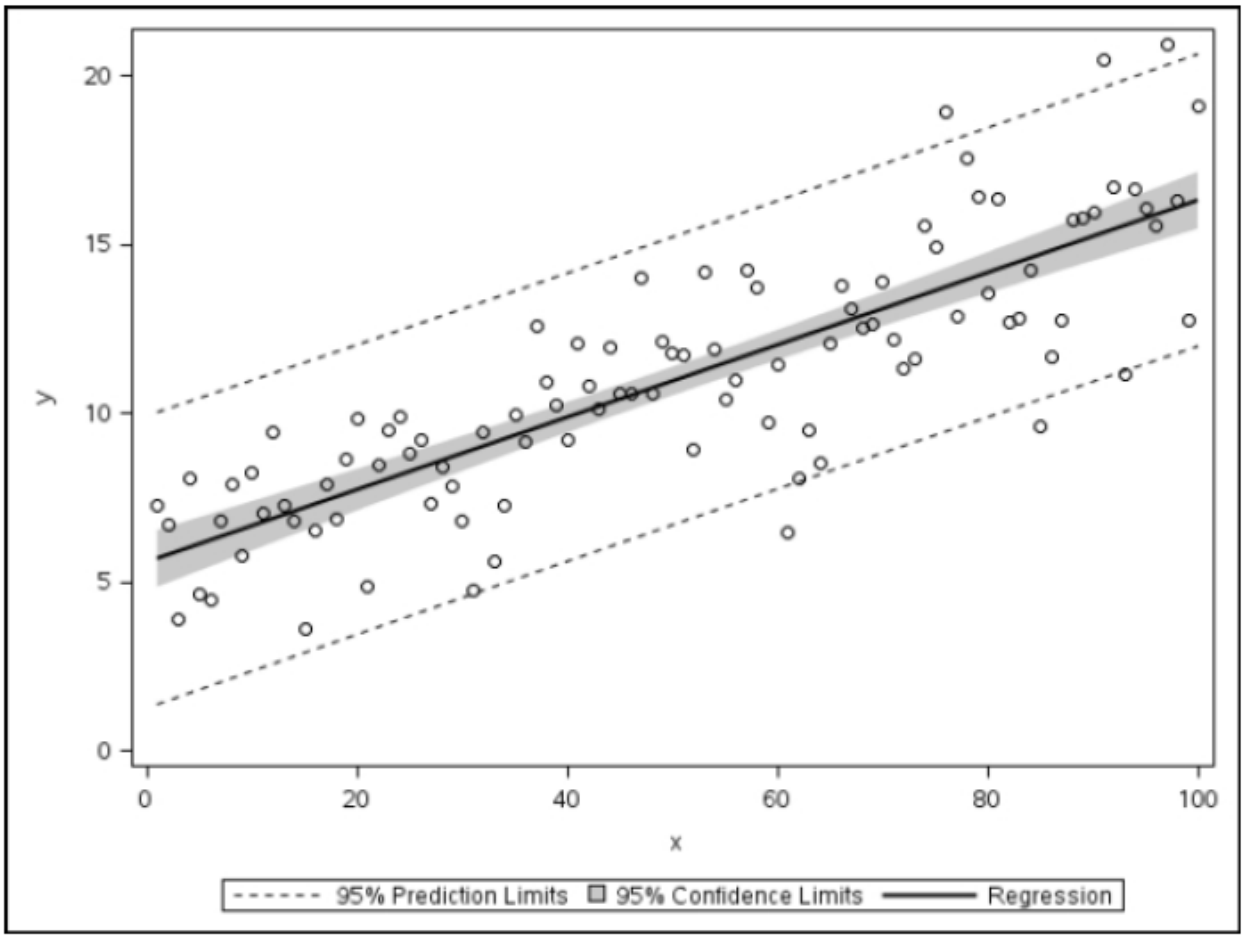

Figure 2. Scatterplot, overlaid with the regression line, confidence limits for the mean, and confidence limits for individual predicted values

The residual, or error, for each of the 100 observations in the dataset is the vertical distance from the point to the regression line. Because the residuals sum to 0, the coefficient estimates are unbiased – their expected values are equal to the true values for the population from which the data were sampled. Furthermore, because the errors have constant variance, that is, the spread of the data points around the regression line is the same over the range of X values, and they are normally distributed, any inferences I draw from the regression are likely to be valid.

From this short, simplified, and somewhat incomplete discussion, note two things about OLS regression. The first is that my ability to draw inferences from the regression analysis depends on the characteristics of the residuals. If the errors in the data are not normally distributed, with constant variance, those inferences may be misleading. Many interesting relationships don't meet the criteria of normal errors with constant variance. Biological concentrations, and money are among the types of measurements which tend to have highly skewed distributions with non-constant variance – as their values get larger, so does the amount of variability.

The second thing to note is that the regression equation estimates only how the conditional mean of the outcome is related to the values of the predictors. **Fi**gure 2 illustrates this for the case of a single predictor: the predicted mean value of the outcome for each value of X can be read from the regression line (or calculated using the estimates of the intercept and slope), but the regression results do not provide any information about values other than the mean. If you are interested in predicting a value other than the mean, or you are interested in finding out whether and how the impact of a predictor changes over the range of outcome values, OLS regression won't help.

Quantile regression addresses both the need to model data from skewed or heteroskedastic distributions

and an interest in modeling values of the outcome variable other than its mean.

MODELING SKEWED DATA USING MEDIAN REGRESSION

Figure 3 again shows two variables: the predictor, X, and the outcome, Y. The simulated data were generated using normally distributed errors, but clearly the errors do not have constant variance. Instead the amount of variation in Y increases as the value of X increases. This pattern, in which the spread of the data increases as the values increase, is common in many applications.

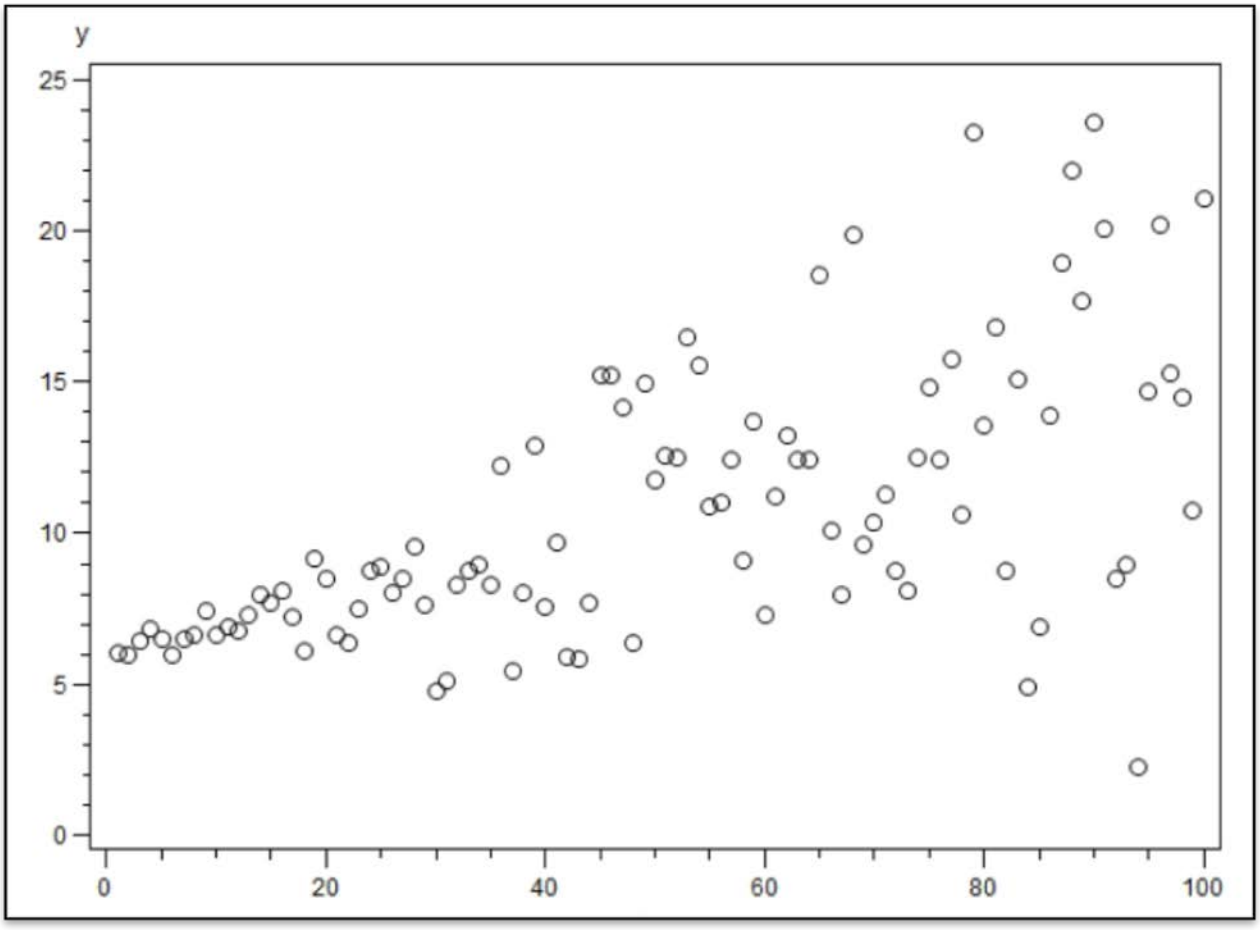

Figure 3. Scatterplot of a linear relatinship, with non-constant variance.

You can still use OLS regression to estimate the relationship between X and Y, and to obtain predictions of the value of Y for each value of X. **Figu**re 4 shows the results. The estimated regression line, with an intercept of 6.05 and a slope of 0.082, does indeed capture the relationship that was used to create the simulated observations. The results of the regression analysis can be used for the purpose of prediction. However, as you can see from **Figure** 4, the results of the regression are misleading if the goal of the analysis is to make inferences. The estimated variance, used to form the confidence intervals, is misleading. For small values of X, the confidence interval is too wide, overestimating the amount of uncertainty in the predictions. For large values of X, the confidence interval is clearly too narrow, seriously underestimating the uncertainty in any predictions that might be made.

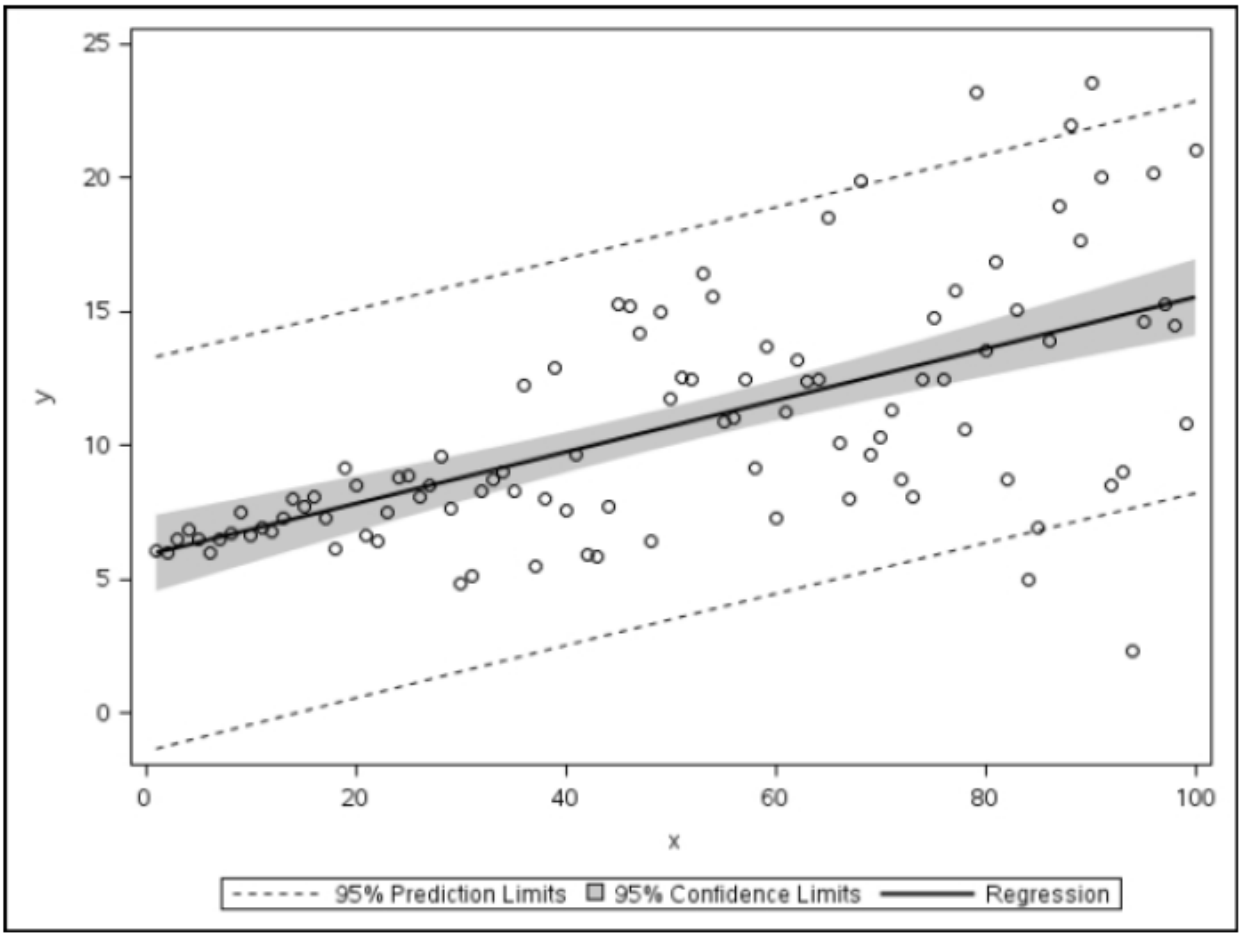

Figure 4. Scatterplot of data with non-constant variance, overlaid with the regression line, confidence limits for the mean, and confidence limits for individual predicted values

These data are easily analyzed using quantile regression in place of OLS regression. Rather than a model predicting the mean value of Y for each value of X, quantile regression provides a model which predicts another measure of central tendency, the median value of Y for each value of X. Unlike OLS regression, median regression does not assume either normality or constant variance in the residuals.

At its simplest, an invocation of the QUANTREG procedure looks the same as a call to PROC GLM or PROC REG. In the sample code, I have explicitly specified 'quantile = 0.5', but this was unnecessary; if no quantile is specified, the QUANTREG procedure will fit a median regression:

```
ods graphics on;
proc quantreg data = non_constant ci = resampling;
   model y = x / quantiles = 0.5 plot = fitplot (showlimits);
run;
ods graphics off;
```

A portion of the output produced by this analysis is shown in **Figure 5**. There are several things to note. The first is that the output looks reassuringly familiar. It provides an estimate for the intercept and the slope, in a format similar to that provided by other SAS regression procedures. The interpretation of these estimates is also familiar: the intercept estimates the median (not the mean) value of Y when X is equal to 0 and the parameter estimate for X estimates the effect of a 1-unit change in X on the median (not the mean) value of Y. The confidence interval and p-value allow you to make inferences about the association between X

and Y.

The requested plot is similar to the plots shown in **Figure 2** and F**igure 4** but now shows the 95% confidence limits for the median value of Y at each value of X. Most importantly, the 95% confidence interval for the median captures both the small amount of variation in the value of Y at small values of X and the high amount of variation in the value of Y at high values of X.

Quantile Level and Objective Function	
Quantile Level	0.5
Objective Function	124.9306
Predicted Value at Mean	10.5631

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept	1	5.9175	0.2302	5.4608	6.3743	25.71	<.0001
x	1	0.0920	0.0059	0.0803	0.1037	15.65	<.0001

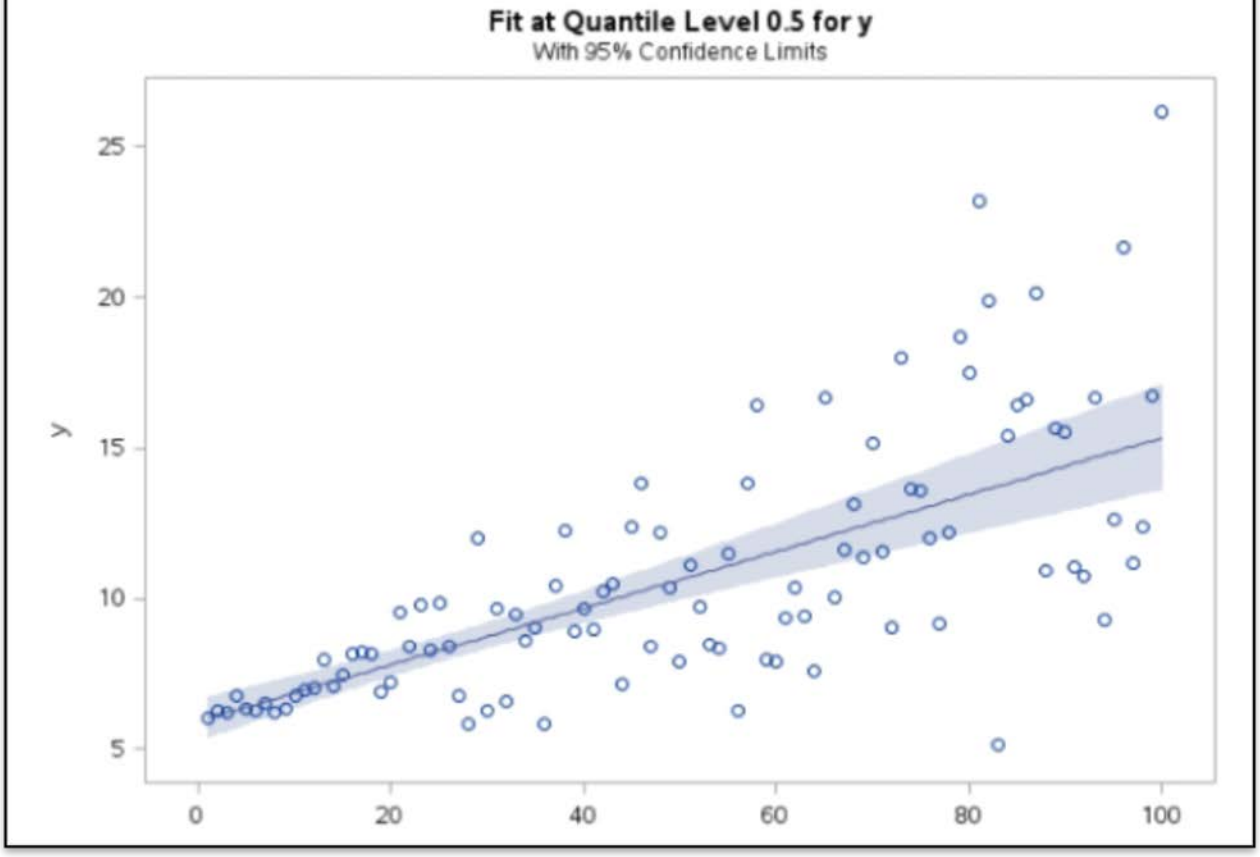

Figure 5. Partial output from a median regression using the QUANTREG procedure, showing the parameter estimates and plot of the median regression line

INVESTIGATING A RANGE OF QUANTILES

The median is, of course, not the only quantile, and there are several reasons to be interested in modeling more than just the median. In healthcare, for example, the 25th or 75th percentiles are sometimes used as benchmarks, and the 90th percentile is sometimes used as performance indicator (e.g., hospitals may be rated on the 90th percentile wait in the Emergency Department or the 90th percentile wait for a consultation). Thus, there may be interest in modeling a percentile other than the median. As well, association can be understood in greater depth by examining the impact of predictors on a range of the percentiles of the outcome variable, rather than on a single percentile.

The "Recommended Reading" section at the end of this paper lists several studies from the Institute for Clinical Evaluative Sciences which used quantile regression to look at the impact of a predictor over a range of outcome values. For illustration, I'll use a fictitious simulated dataset which contains the following variables, describing the health care services used over a one-year period:

- total_cost: total annual coat of health care use (the outcome variable)
- age
- sex: a categorical variable with the values M and F
- num_comorb: number of chronic conditions

While I stress that these data are totally fictitious, there is a lot of interest in predicting people who use a lot of health care services. In Ontario, in 2007/08, the top 1% of health care users accounted for one-third of health care spending (Rosella et al., 2014) and similar results have been reported for other jurisdictions in North America.

The data were analyzed using the following SAS code:

```
ods graphics on;
proc quantreg data = fake_cost ci = sparsity
     algorithm = interior (tolerance = 1.e-4);
   class sex;
   model total_cost = age age*age sex num_comorb /
         quantile = 0.1 to 0.9 by 0.1
         plot = quantplot;
run;
```

The code demonstrates PROC QUANTREG's ability to handle categorical data, using the same parameterization used by PROC GLM. It also demonstrates that quantile can accommodate interactions, which are specified for QUANTREG in the same way they are specified for PROC GLM. In other words, the possibilities for the model statement are the same as those for OLS regression.

The output contains parameter estimates for each of the quantiles specified in the model statement. To save space, only the estimates for the 10th and 90th percentiles are reproduced below, in **Figure 6**. The interpretation of the parameter estimates for this regression is the same as their interpretation for any other regression. The parameters quantify the effect of each predictor on the value of the 10th percentile and 90th percentile of total health care spending. In other words, quantile regression assesses how the predictors relate to the outcome for individuals whose outcome tends to be low (e.g., 10th percentile), high (e.g., 90th percentile) or anywhere in between.

As for any regression, the intercept estimates the value of the quantile for an individual with a value of 0 for all of the predictors (in this case, a newborn (age 0) male with no chronic conditions). The parameter value for female sex estimates the difference in health care utilization for females relative to males at each percentile. For individuals whose total spending puts them at the 10th percentile, the difference is estimated to be $35, while for individuals at the 90th percentile, the difference is an estimated $523. Similarly, the effect of an additional chronic condition, at the 10th percentile level of spending, is to increase spending by an estimated $95, whereas by the 90th percentile, this has increased to $3,064.

The quantile plots shown in **F**igure 7 summarize the parameter estimates for each predictor over the entire range of quantiles. For example, the plot for female sex shows that costs for females are always higher than costs for males, that the difference increases as health care utilization increases, and that the difference is particularly pronounced in the upper tail. The same pattern is seen for the effect of the number of chronic conditions. As expected, spending increases with each additional chronic condition, but the effect is particularly large in the upper tail.

As is the case for median regression, quantile regression does not assume either normality or constant variance for the residuals.

Quantile Level and Objective Function	
Quantile Level	0.1
Objective Function	123872821.28
Predicted Value at Mean	98.2685

Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept		1	6.7810	0.6668	5.4741	8.0879	10.17	<.0001
age		1	-4.4253	0.0567	-4.5364	-4.3141	-78.01	<.0001
age*age		1	0.1146	0.0011	0.1125	0.1168	103.45	<.0001
sex	F	1	35.8649	0.2983	35.2803	36.4496	120.23	<.0001
sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
num_comorb		1	96.0854	0.9104	94.3010	97.8699	105.54	<.0001

Quantile Level and Objective Function	
Quantile Level	0.9
Objective Function	641345359.05
Predicted Value at Mean	3674.4846

Parameter Estimates

Parameter		DF	Estimate	Standard Error	95% Confidence Limits		t Value	Pr > |t|
Intercept		1	1434.899	16.9392	1401.6986	1468.0991	84.71	<.0001
age		1	-94.3678	1.6275	-97.5577	-91.1779	-57.98	<.0001
age*age		1	2.3975	0.0303	2.3382	2.4569	79.22	<.0001
sex	F	1	571.9905	14.4792	543.6117	600.3693	39.50	<.0001
sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
num_comorb		1	3263.120	44.3547	3176.1864	3350.0539	73.57	<.0001

Figure 6. Partial output from the QUANTREG procedure, showing the parameter estimates for the 10th and 90th percentiles of total cost (simulated data)

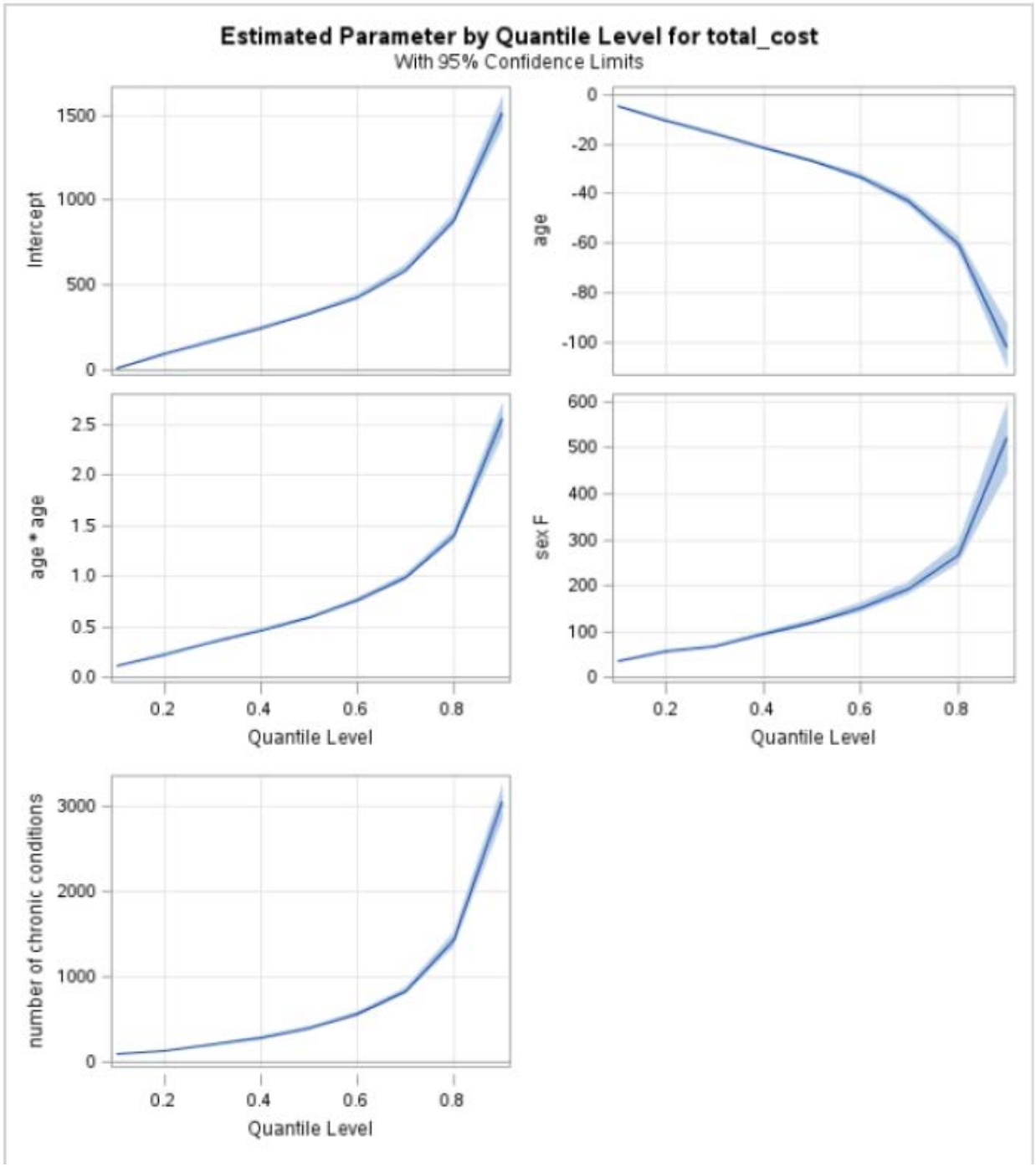

Figure 7. Quantile plots of the parameter estimates, with 95% confidence intervals (simulated data)

CONCLUSION

Researchers are often faced with data that do not meet the assumptions necessary in order to draw valid inferences using ordinary least squares regression, namely that the errors are normally distributed with constant variance. When the assumptions of OLS regression are not met, a number of tools are available, including transformation of the dependent or independent variables and weighted regression.

Quantile regression is another tool, one which is easy to use and which produces easily interpreted results. Quantile regression is more robust in the presence of outliers, and is not affected by violations of the assumptions of normality and constant variance.

In addition, quantile regression provides the ability to explore the association between predictors and an outcome variable in greater depth. Using quantile regression, it is possible to explore whether and how the effects of predictors vary across the entire distribution of the outcome.

REFERENCES

Rosella, L. C., Fitzpatrick, T, Wodchis, W. P., Calzavara, A., Manson, H., & Goel, V. (2014). High-cost health care users in Ontario, Canada: demographic, socio-economic, and health status characteristics. *BMC Health Services Research, 14* (532). Available at http://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-014-0532-2.

RECOMMENDED READING

- *Atzema CL, Austin PC, Tu JV, Schull MJ. Emergency department triage of acute myocardial infarction patients and the effect on outcomes. Ann Emerg Med. 2009; 53(6): 736-45.*
- Austin PC, Schull MJ. Quantile regression: A statistical tool for out-of-hospital research. Acad Emerg Med. 2003; 10(7): 789-97. Available at *http://onlinelibrary.wiley.com/doi/10.1197/aemj.10.7.789/epdf*
- *Austin PC, Tu JV, Daly PA, Alter DA. The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy. Stat Med. 2005; 24(5): 791-816.*
- *Jiang L, Gilbert J, Langley H, Moineddin R, Groome PA. Effect of specialized diagnostic assessment units on the time to diagnosis in screen-detected breast cancer patients. Br J Cancer. 2015; 112(11): 1744-50.*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ruth Croxford
Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada
ruth.croxford@ices.on.ca
www.ices.on.ca