Simple Linear and Quantile Regression

STEM

2022

Bivariate models.

- When we have two variables measured at the same time, we want to know the "law" connecting them.
- The law may take the form of a functional relation, $Y = f(X)$. In Linear Regression, the law is a linear combination of power functions or other functions.
- A simple example of a multilinear regression model is a polynomial ("linear" in the coefficients $\beta_0, \beta_1, \beta_2, \ldots$):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$

- In Simple Linear Regression (SLR) the polynomial has degree 1:

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

The number β_1 is known as the *regression coefficient*.

Bivariate models.

- One of the variables will be regarded as dependent, and the other as in some way determining or explaining the variations in the values of the dependent variables.
- Because the word "independent" has a special meaning within statistics, we use the term *regressor* to describe this explanatory variable.
- Regressors may be controlled, that is, strictly deterministic, or uncontrolled, that is, subject to observation error, or partially controlled, where any observation error in the regressor is small.

Regression model examples:

- The monthly sales of marine radio devices depends on the month. (Controlled regressor.)
- The speed of a fan depends on the input power. (Partially controlled.)
- The hardness of a sample of a steel alloy depends on its vanadium content. (Uncontrolled.)
- The vanadium content of a steel alloy may be found from its hardness. (Uncontrolled.)

Some possible laws.

Some other laws can be reduced to the fundamental SLR relationship

$$Y = \beta_0 + \beta_1 X.$$

- The power law $Y = b_0 X^{\beta_1}$.

  Take logarithms: $\ln Y = \ln b_0 + \beta_1 \ln X.$

  (This is now in the correct form for SLR, with $\ln Y$ in place of Y, $\ln b_0$ in place of β_0, and $\ln X$ in place of X.)

- The exponential law $Y = b_0 e^{\beta_1 X}$.

  Take logarithms: $\ln Y = \ln b_0 + \beta_1 X.$

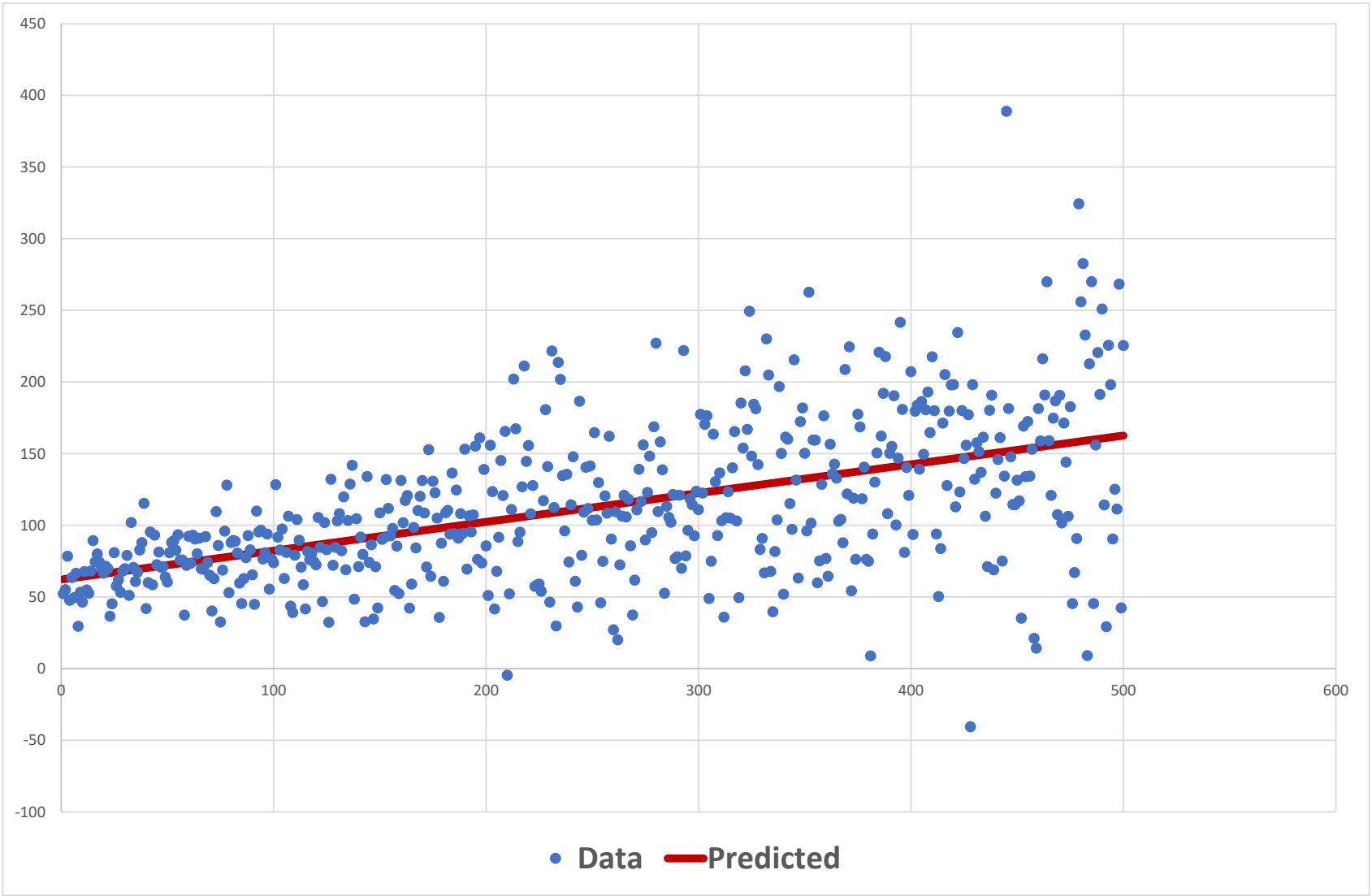
Controlled regressor.

- In the discussion that follows, we will always assume that only Y is subject to measurement errors, that is, assume X is controlled (deterministic), Y is random.
- We will further assume that the observation errors are Gaussian distributed.
- The observations for SLR come in (X, Y) pairs. The aim of regression analysis is to determine the coefficients β_0, β_1, in the model $Y = \beta_0 + \beta_1 X$, in some optimal fashion.

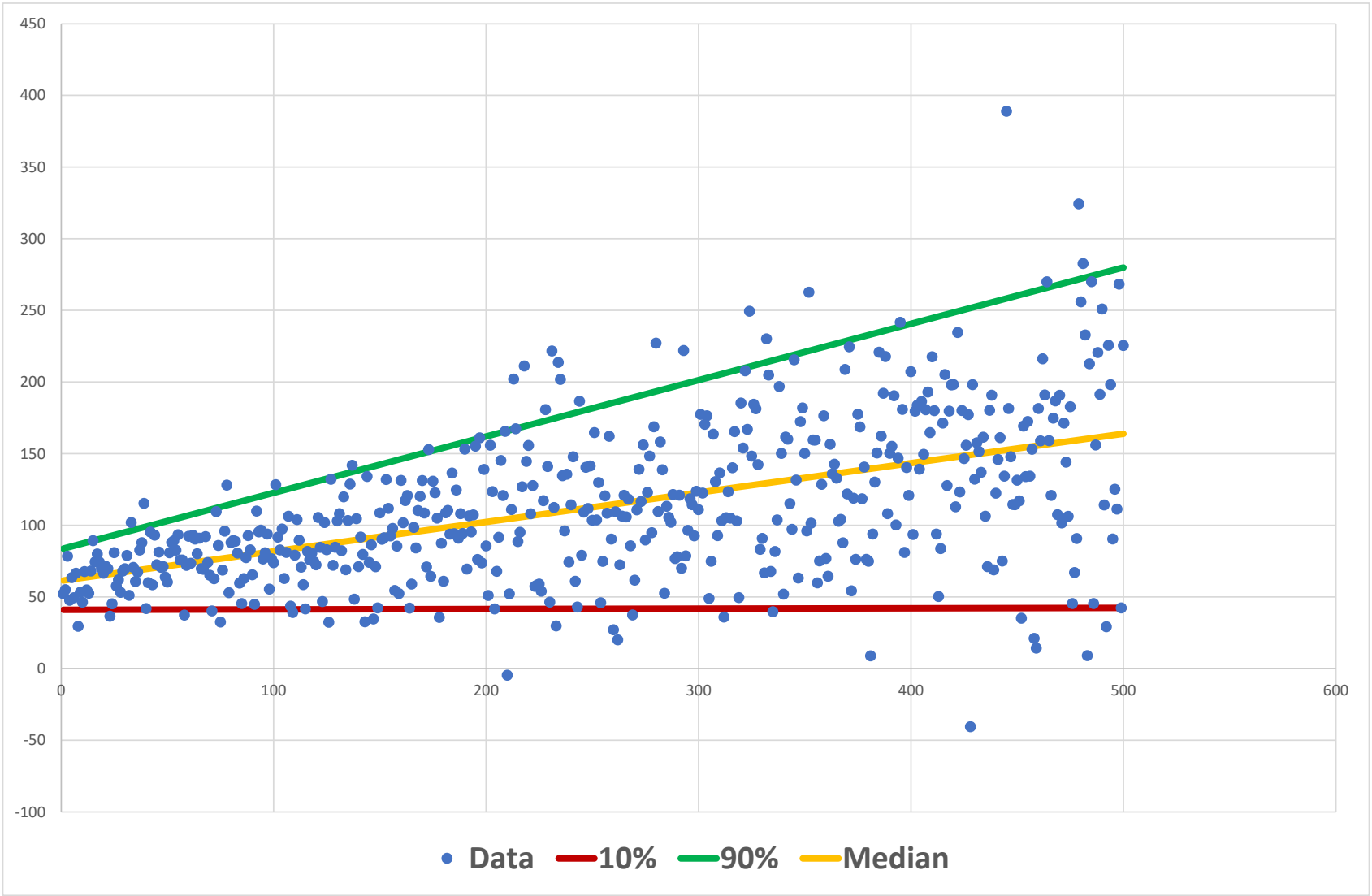
Quantile Regression

- Standard least squares regression models only the conditional mean of the response
- Quantile regression allows one to model the specific conditional quantiles of the response.
- **It does not assume a particular parametric distribution for the response, nor does it assume a constant variance for the response, unlike least squares regression.**

Simple Linear Regression



Quantile Regression



Definition of the Quantile Level

- The Quantile level τ is the probability $Pr[Q_\tau(Y|X)|X]$.
- The conditional quantile $Q_\tau(Y|X)$ is the value of Y below which the proportion of the conditional response population is τ.

Estimating the regressor coefficients for SLR

- The standard regression equation for the expected value of the response is given by
- $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$
- The β_j's are estimated using

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2.$$

Estimating the regressor coefficients for Quantile Regression

For quantile level τ of the response, the goal is to

$$\min_{\beta_0(\tau),\beta_1(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau))^2.$$

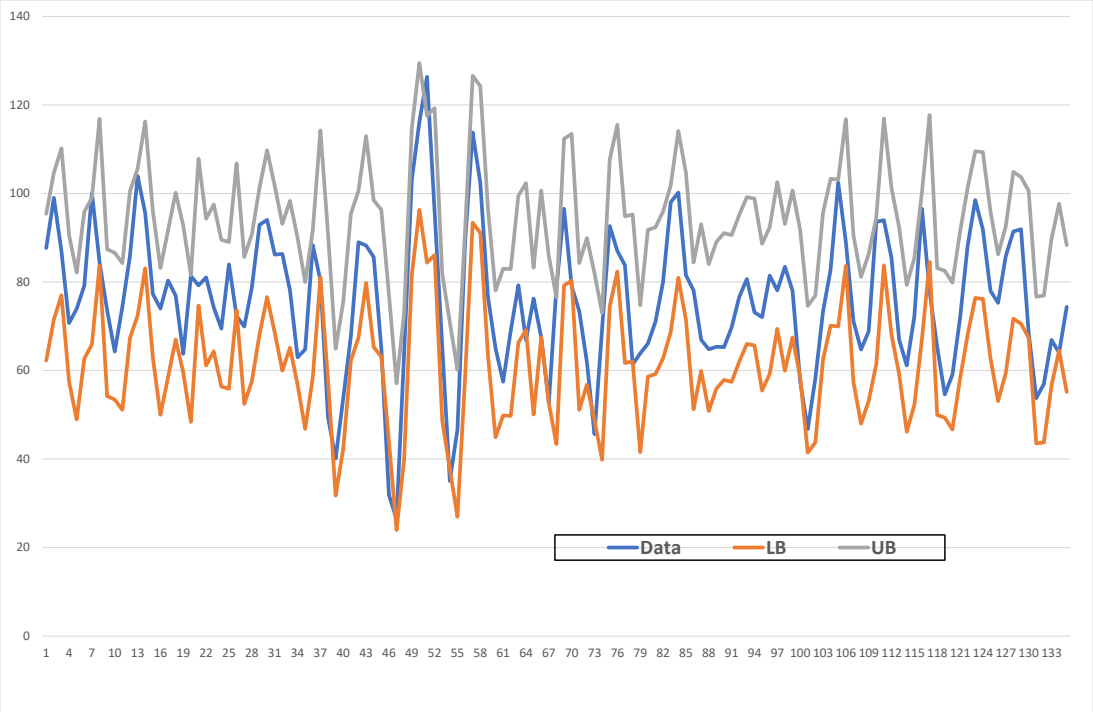
$$\rho_\tau = \tau max(r, 0) + (1 - \tau)max(-r, 0)$$

is the check function. If the error in a single period, r, is positive, then the check function multiplies the error by τ and by $(1 - \tau)$ if negative.

Testing the procedure

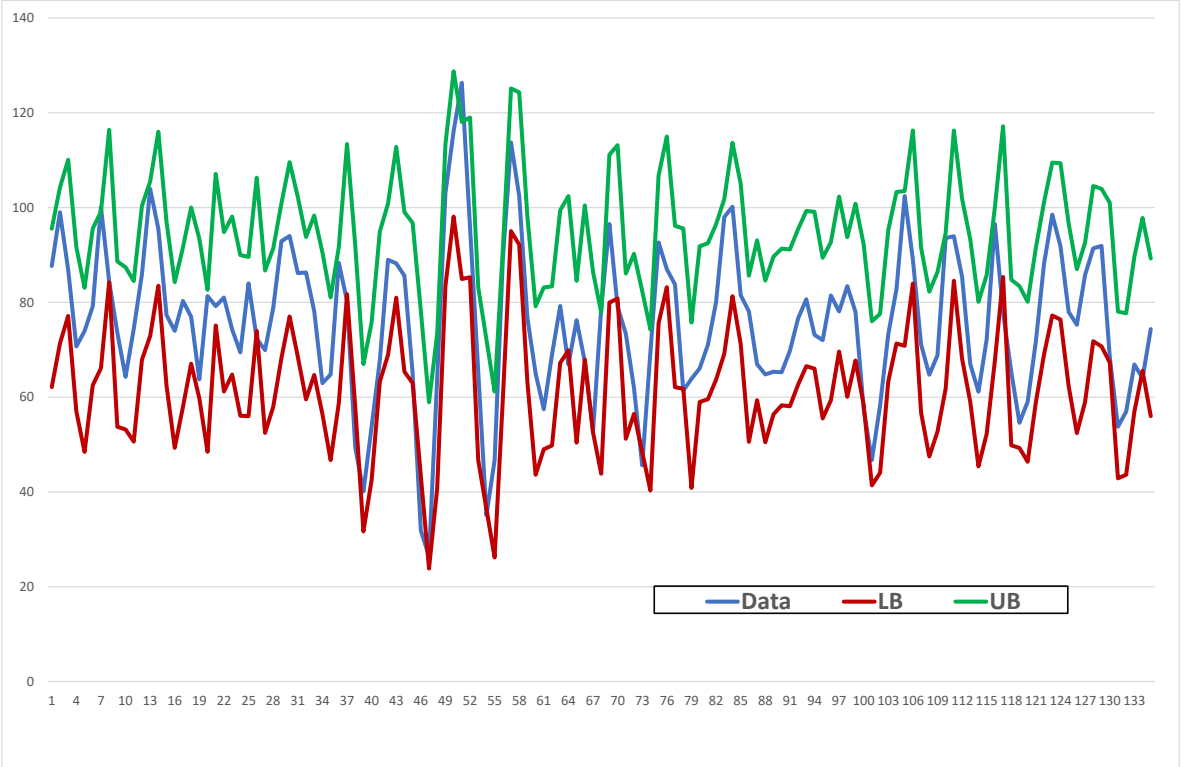
- We will first go through an experiment,with two sets of synthetic data.
- The first set is an $AR(2)$ process, with coefficients $\phi_1 = 1.01, \phi_2 = -0.2$, and noise being $N(10, 13^2)$.
- The second set has the same structure but with the noise following a Gamma distribution, with $\alpha = 0.5, \beta = 20$.
- Subsequently we will see the use of quantile regression with a solar radiation series.

Results assuming normally distributed noise - using Normal Prediction intervals



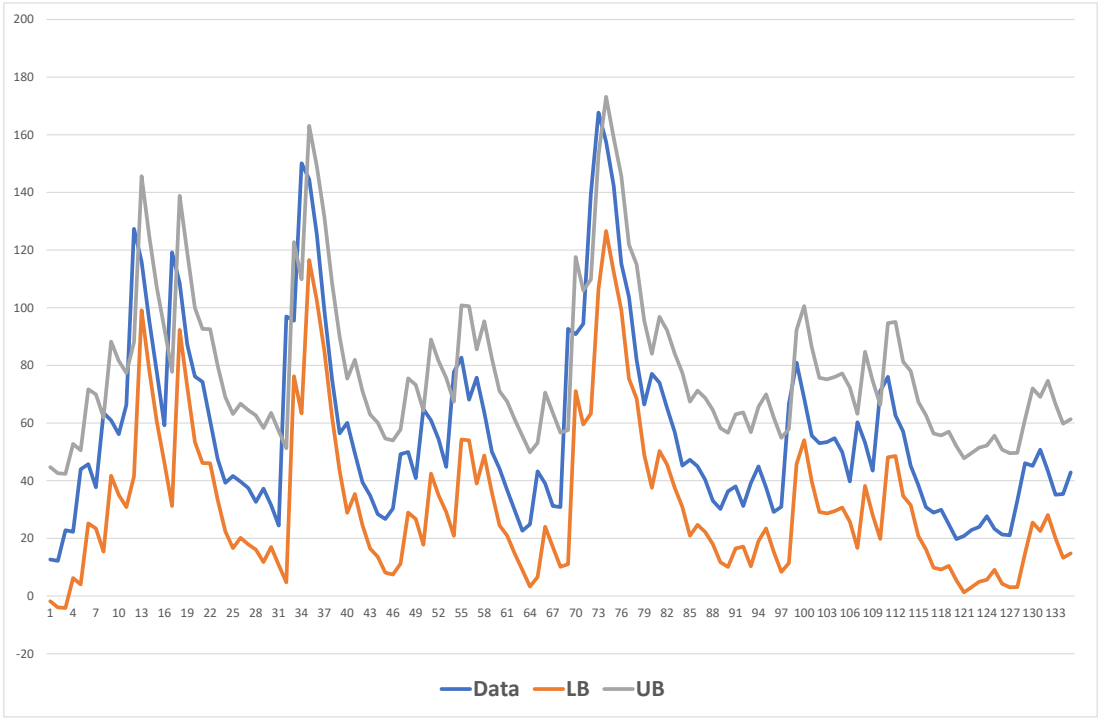
- For a 90% prediction interval, the coverage was 90.2%.
- The average width of the interval (all intervals the same width) was 33.17.

Results assuming normal distributed noise - using Quantile regression



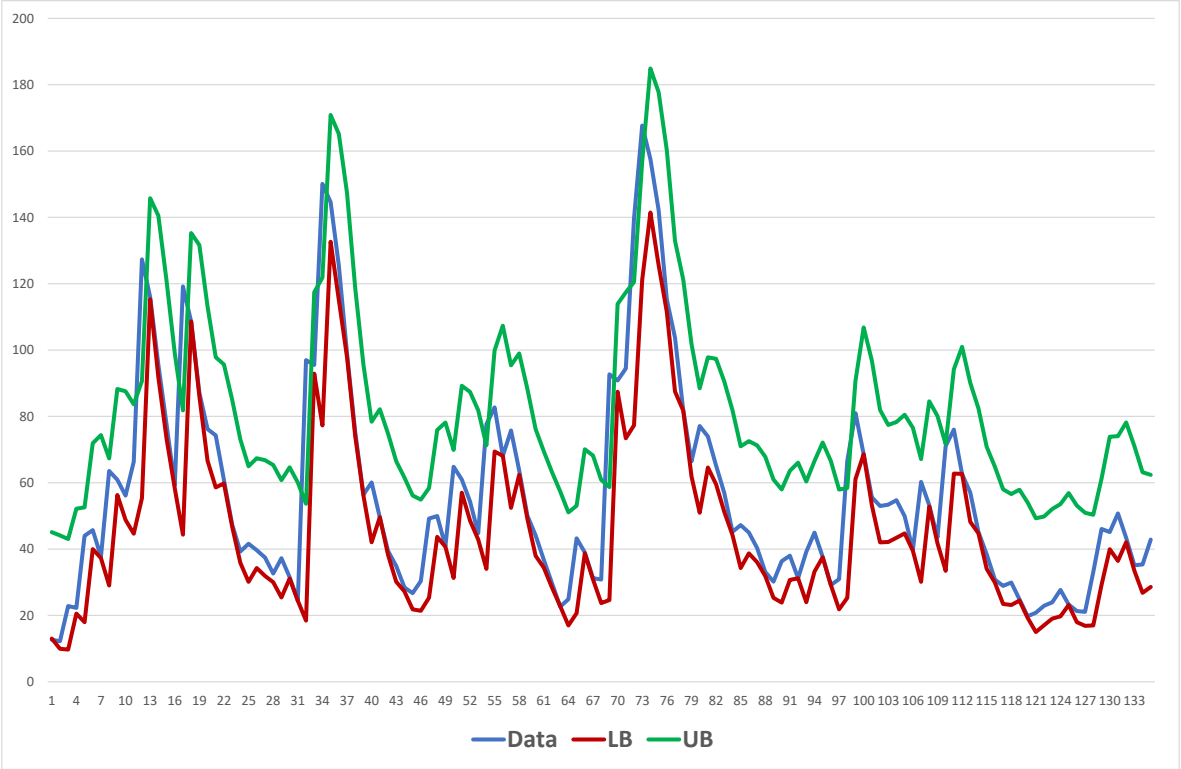
- For a 90% prediction interval, the coverage was 90.6%.
- The average width of the interval was 33.42.

Results assuming skewed noise - using Normal Prediction intervals



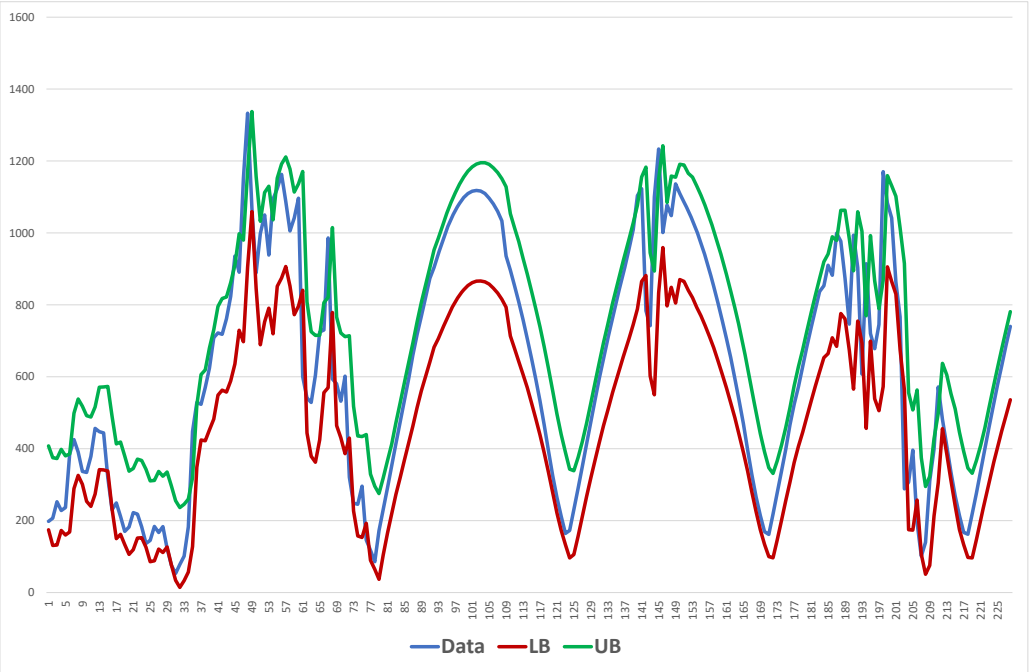
- For a 90% prediction interval, the coverage was 92.7%.
- The average width of the interval (all intervals the same width) was 46.53.

Results assuming skewed noise - using Quantile regression



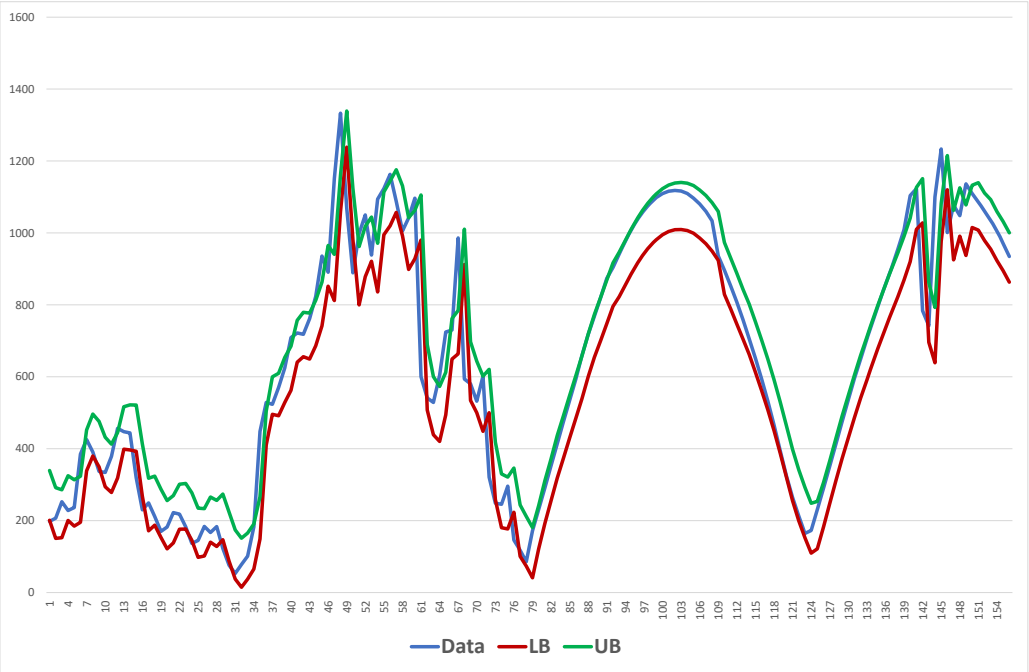
- For a 90% prediction interval, the coverage was 89.1%.
- The average width of the interval was 36.23.

Solar radiation - 90% prediction interval



- For a 90% prediction interval, the coverage was 90.0%.
- The average width of the interval was 266.75.

Solar radiation - 80% prediction interval



- For an 80% prediction interval, the coverage was 79.9%.
- The average width of the interval was 130.22.

Solar Farm Output

- This is a special case since in Australia, almost all solar farms are oversized.
- They have more capacity in terms of the panels than the inverters can process.
- There are various financial reasons for this, as what happens is that on a clear day at any time of the year, the output is constant for a number of hours.
- We will use the Broken Hill solar farm in NSW as an example.

Difference of Solar Farm Output from Solar Energy Profile - Summer

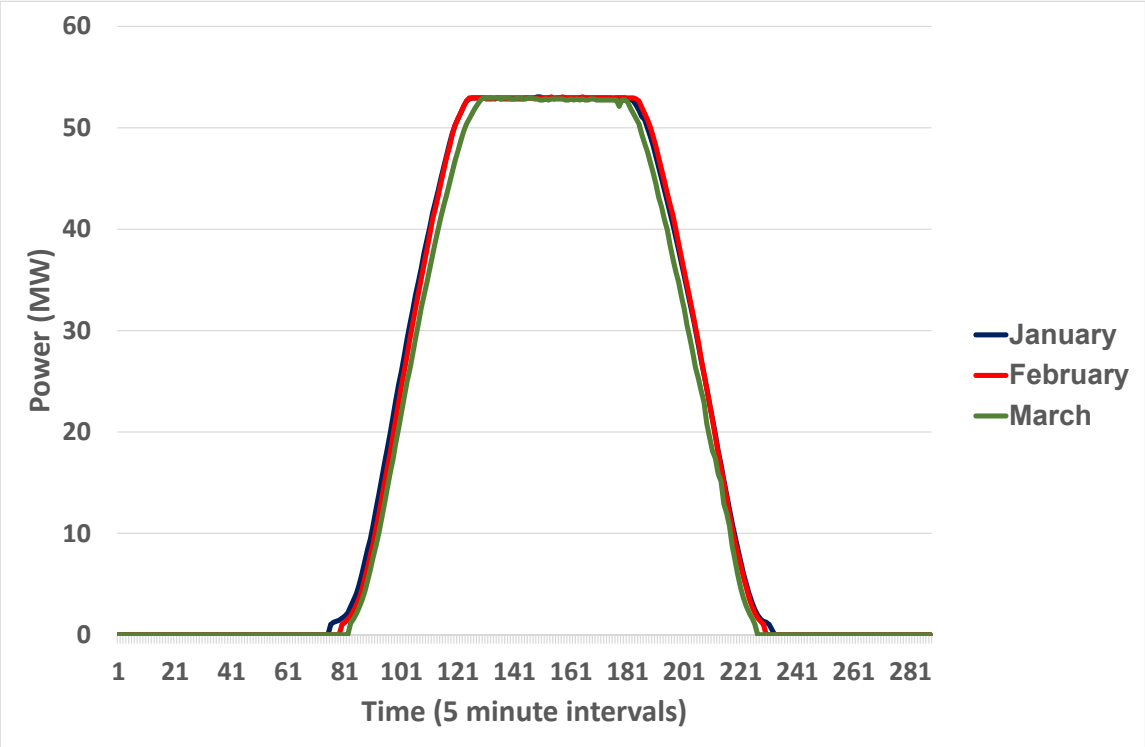


Figure: January-March

Difference of Solar Farm Output from Solar Energy Profile - Winter

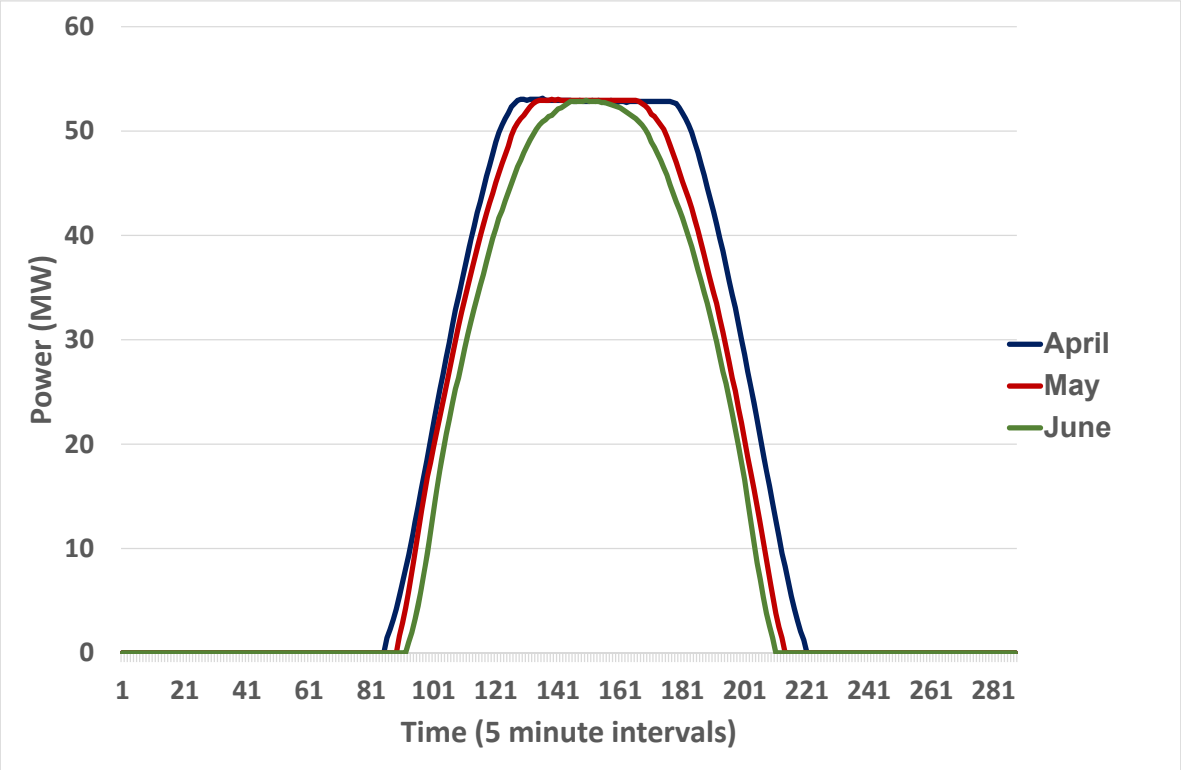


Figure: April-June

Quantile Regression for Broken Hill Solar Farm

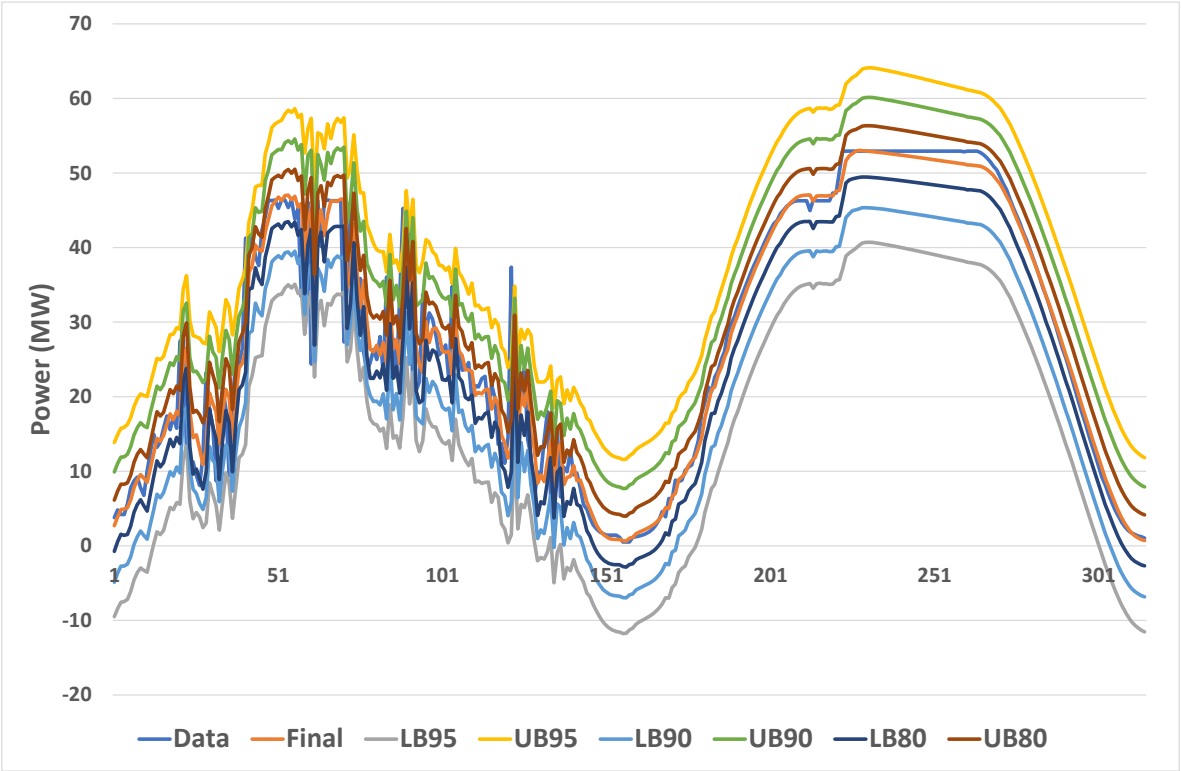


Figure: Output, forecast and 95, 90 and 80% prediction intervals

Quantile Regression for Broken Hill Solar Farm

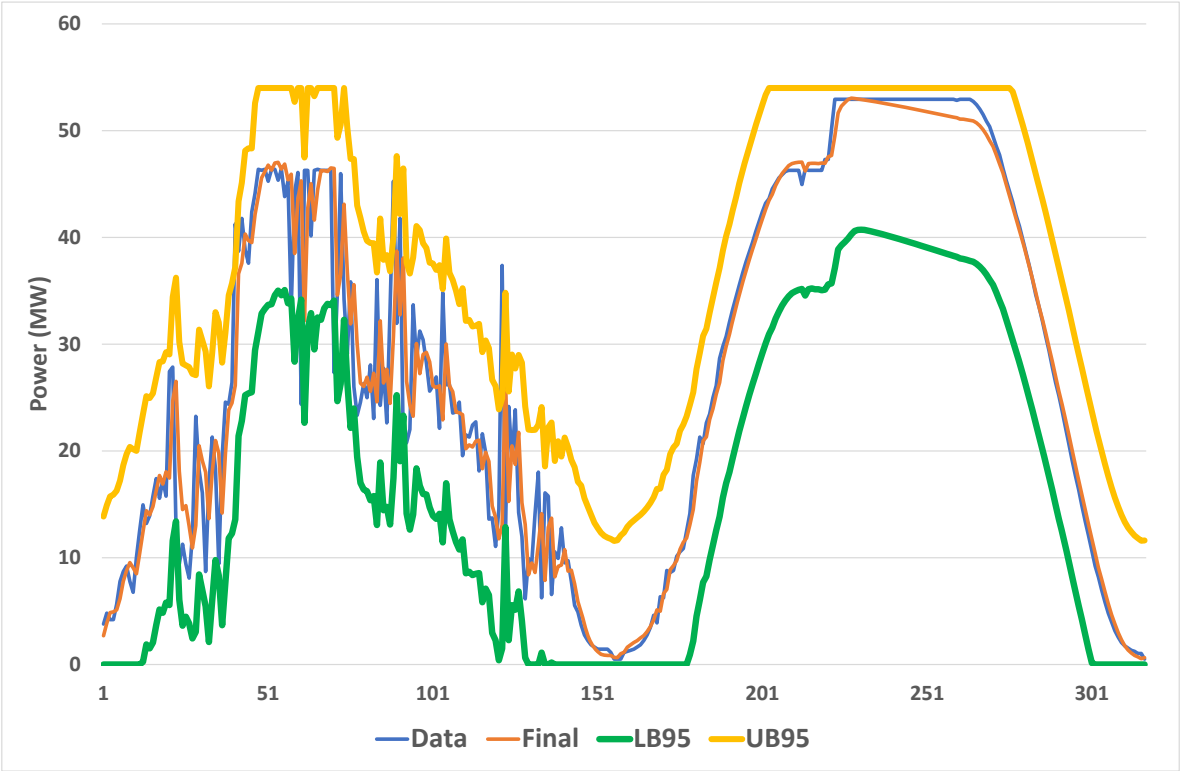


Figure: Output, forecast and 95% prediction interval

Coverage and Interval Width

- Hopefully you noticed that in the first figure, the interval bounds could go below zero, and also above the capacity.
- In the second one, I filtered out any values in the intervals below zero or above 54 MW.
- As well as giving a more realistic appearance, the average interval width was lowered - see the next slide.

Quantile Regression for Broken Hill Solar Farm - Results

Level	Coverage	Width	Width (Filtered)
95%	95.55%	23.38	19.53
90%	90.73%	14.75	12.95
80%	81.83%	6.86	6.44

Figure: Coverage and Interval Width