BUSINESS ANALYTICS

MODEL BUILDING – Improving Regression Model

Residual = Observed – Predicted

…positive values for the residual (on the y-axis) mean the prediction was too low, and negative values mean the prediction was too high; 0 means the guess was exactly correct.

Treating the outlier

- Outlier is not because of a measurement or data error
- then we have to think about model adjustment
- Do regression with and without outlier
- If the model doesn’t change much no need to worry much.
- If the model changes significantly, we have to decide what to do, that is which is good for the model

Normalising the Variable

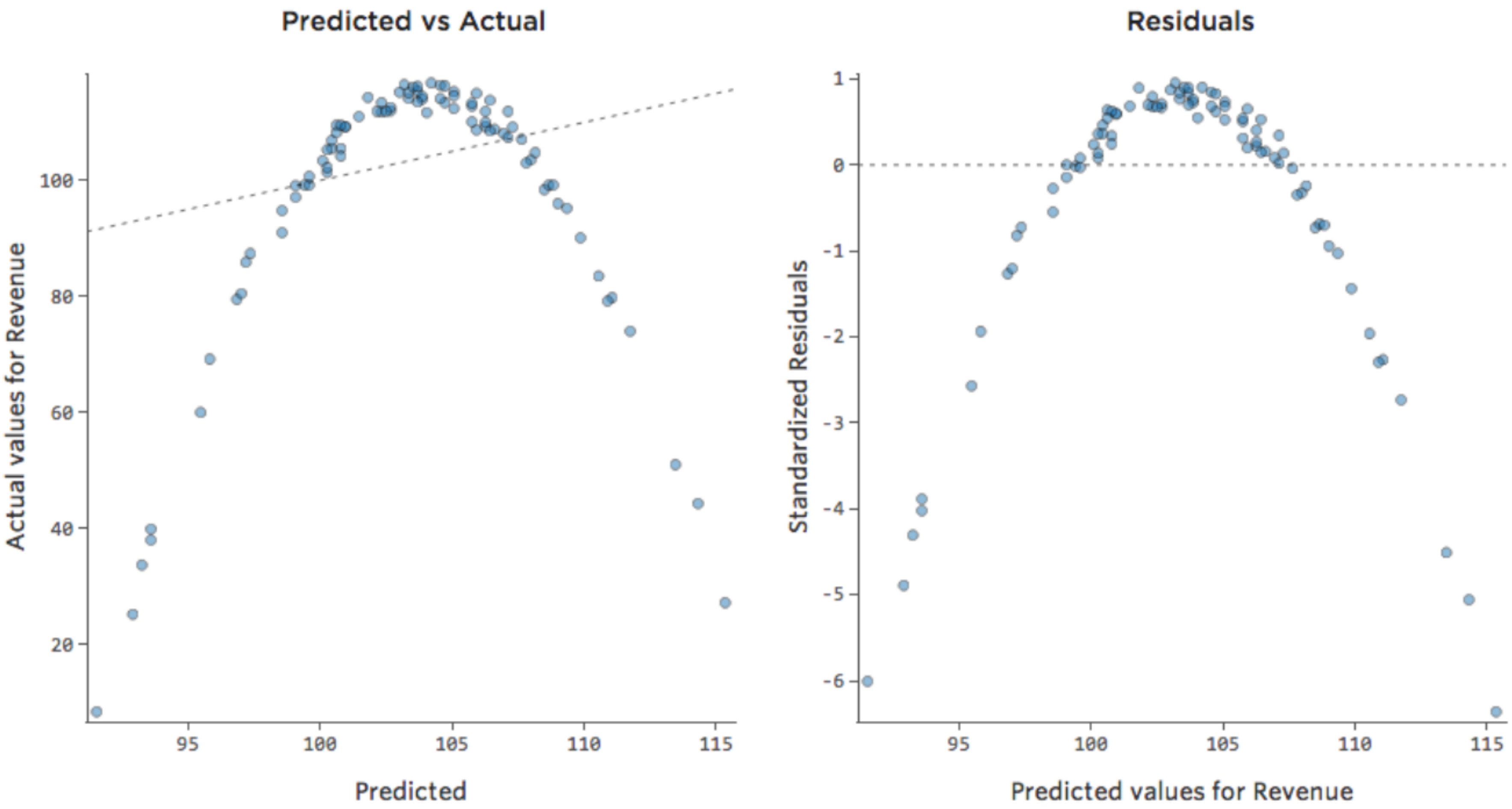
- One of the best methods to improve the models is transforming the variable into log value.
- By doing this transformation there will be a change in shape of its distribution.
- Regression models work better with more symmetrical, bell-shaped curves.

Missing Variable/Adding a New Variable

- Probably the most common reason that a model fails to fit is that not all the right variables are included.
- This particular issue has a lot of possible solutions
- Sometimes the fix is as easy as adding another variable to the model.

For example, Sales vs marketing investment is not giving a good model, so to that model add temperature variable, model becomes good

Fixing Nonlinearity



Fixing Nonlinearity

- The line in the plot above has this formula: y = 1.7x + 51
- But it’s a terrible ft. So if we add an x2 term, our model has a better chance of fitting the curve.
- The above approach can be extended to other kinds of shapes, particularly an S-shaped curve, by adding an x3 term.
- That’s relatively uncommon, though.

Unavailable Omitted Variable

- Quite frequently the relevant variable isn’t available because you don’t know what it is, or it was difficult to collect.
- May be it wasn’t a temperature issues, but instead something like number of competitors in the area that you failed to collect at the time.