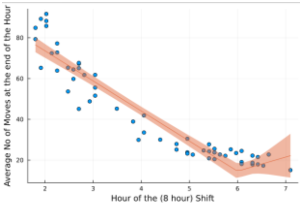
(Don’t) Assume curvy-linear relationships for cranes simulations.



<u>Below are some quick insights from recent shipping industry’s Machine Learning experiment.</u>

Business Problem: Client identified that simulations from curvy linear estimates (used in simulations) were yielding always either higher or lower estimates when validated with real time crane performances.

Business Objective: To provide more better estimates that can current ones (curvy linear estimates) that can be used for inputs into simulations for estimating crane works.

Data and Sample: Took a sample of one year data for a specific port, terminal and Quay Cranes (QCs).

What EDA is telling?

1) In initial EDA, we understood (aligns with research papers available), individual QCs have different performance level (give their location and moves to be placed in the terminal), hence, drilled down sample to a single QC.
2) Made sample independent, that is, as each QC has three shifts and again each shift has its own distribution, drilled sample to a single QC and a single shift.
3) However, understood, different operators (each crane has 3 shifts – 8 hours into 3 shifts per day) has different distributions. Hence, utilized normalized sample (minimum of 1 hour to max of 8 hours) of a shift by same operator for a quarter, below is the sample data of a QC that went into modelling.
4) It has 2 columns; when operator been on same QC, hours in shift being on the same QC for instance 6.64 represents that operator has been 6.64 hours of total 8 hour shift on the same crane. And average number of moves per hour he has done in those 8 hours.

hour_of_the_shift	avg_crane_moves
6.647887	22.71398
5.746479	22.10049
6.535211	17.29095
3.042254	51.60648
1.915493	65.25735
6.422535	21.48684
2.253521	77.2125
3.830986	38.89338
2.140845	72.40132
6.084507	24.48843
7.098592	15.07341
2.704225	65.08333
1.802817	79.32031
2.478873	53.65341
5.408451	20.89323
1.915493	89.27206
2.253521	63.9
4.957746	30.45739
2.704225	67.30208
6.309859	17.90848
4.056338	41.90972
6.422535	17.9057
5.070423	22.68056
6.309859	18.70089
4.957746	23.80114

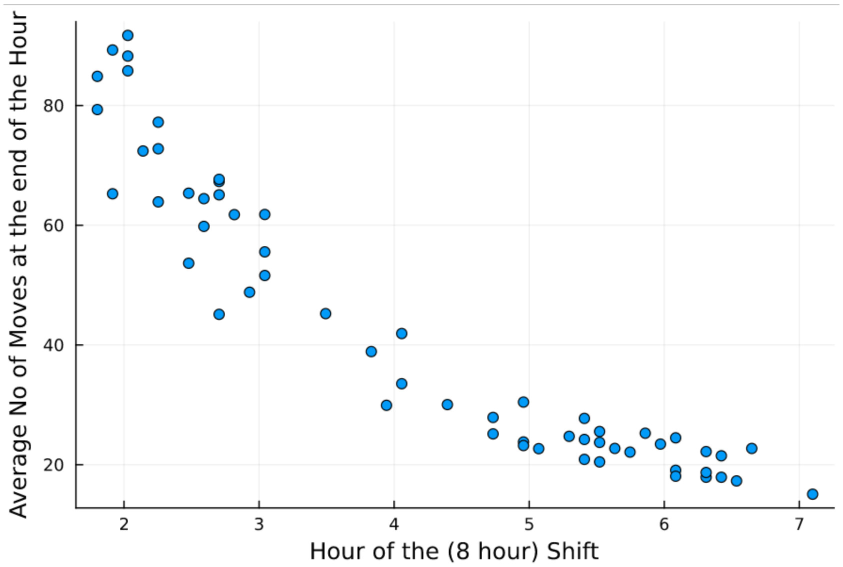
2.591549	59.80978
2.028169	91.70833
4.732394	27.89286
6.084507	19.06481
5.295775	24.7367
2.704225	45.11458
5.859155	25.25962
3.042254	61.7963
2.028169	85.79167
5.408451	27.73438
3.492958	45.23387
5.521127	23.72704
5.408451	24.22135
5.521127	20.46684
2.591549	64.44022
2.253521	72.775
2.704225	67.67188

3.943662	29.92143
4.056338	33.52778
4.957746	23.19602
5.521127	25.53827
5.971831	23.4434
1.802817	84.86719
6.084507	18.0787
2.816901	61.77
3.042254	55.55093
2.028169	88.25694
2.929577	48.8125
2.478873	65.35227
5.633803	22.72
4.732394	25.14583
4.394366	30.03846
6.309859	22.1875

~~~
```{julia}
using CSV, DataFrames, StatsPlots, GLM
cranedata = CSV.read("crane_p_reg_2.csv", DataFrame)

@df cranedata plot(:hour_of_the_shift, :avg_crane_moves, seriestype = :scatter,
legend = :none, xlabel = "Hour of the (8 hour) Shift", ylabel = "Average No of
Moves at the end of the Shift")
```
~~~

Below scatters plot from above Julia code is clearly indicating a curvy linear relationship of (exponential decreasing) over a period.



Also, when you fit a any curvy linear model of GLM nature you will get good results:

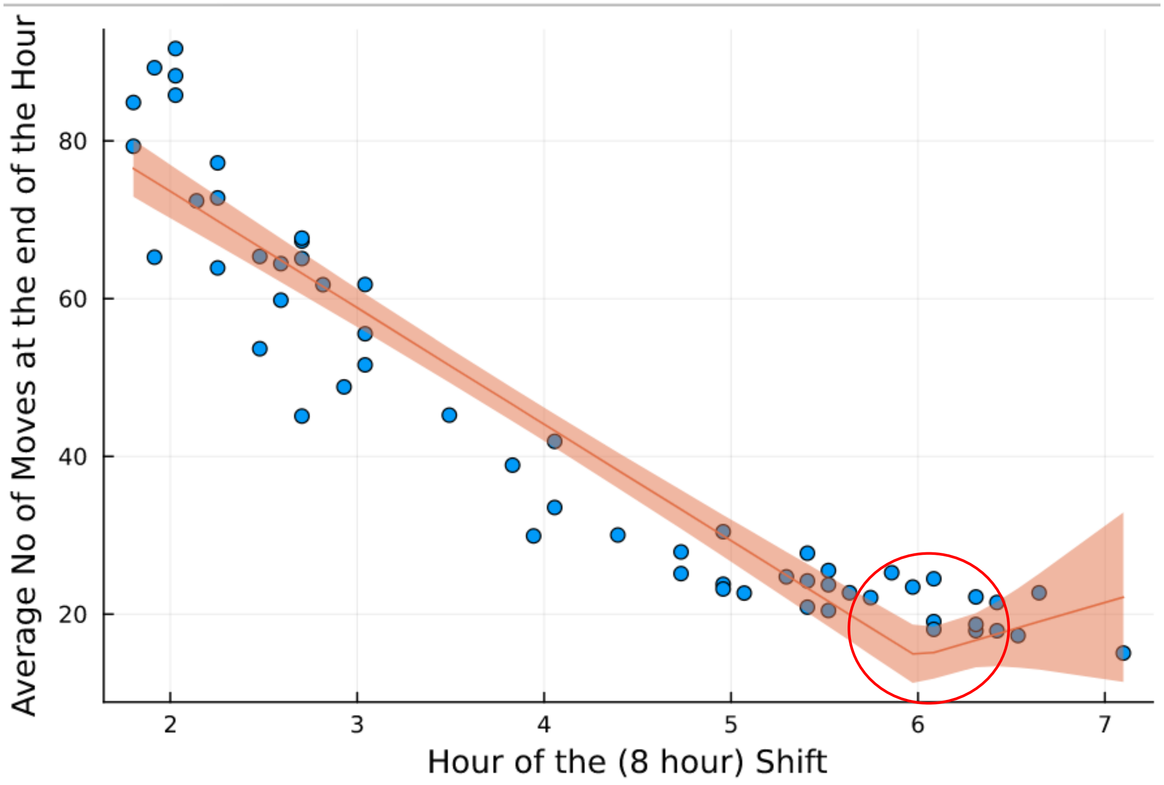
“reg_lm = lm(@formula(avg_crane_moves ~ hour_of_the_shift), cranedata); coeftable(reg_lm):”

```
──────────────────────────────────────────────────────────────────────
                  Coef.  Std. Error       t  Pr(>|t|)  Lower 95%  Upper 95%
──────────────────────────────────────────────────────────────────────
(Intercept)        98.7678   2.96497    33.31    <1e-37    92.8283   104.707
hour_of_the_shift  -13.3366    0.658338  -20.26    <1e-26   -14.6554   -12.0178
──────────────────────────────────────────────────────────────────────
```

However, when you go deep, we see different relationship.

Yes, when we went deep into why curvy linear estimates were not working better; ***<u>what we observed not only different distributions across the shifts of the crane and operators across the terminal of the port, but a different latent state of QC that can be inferred by dividing into a timely segments of the operations. In simple words, mean, variance, trend change after some time interval.</u>***

In order to identify the same, study used piecewise regression as a first/quick attempt and **below graph depicts the potential shift (highlighted in red circle) after sixth hour.** Thus, **study identified that multiple slopes exist at different length of interval** and these need to be considered instead of blindly assuming single linear curvy relationship when identifying crane performances and those inputs for simulations.


Average No of Moves at the end of the Hour
Hour of the (8 hour) Shift