Homework 6 Report

Chenning Xu

1 Linear regression

Table 1: Simple linear regression results

	Number of Employees	95% CI
management score	6777.187***	[3314.151,10240.224]
Constant	-13993.993***	[-23755.571,-4232.415]
Observations	393	
$R2$	0.060	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

I have chosen Singapore for my analysis. The results of the linear regression of the number of employees on the management quality score show a slope coefficient of 6777. This indicates that for every one-point increase in a firm's management quality score, the expected number of employees is associated with an increase of 6777, holding other factors constant.

The 95% confidence interval for the slope coefficient ranges from 3314 to 10240, meaning we can be 95% confident that the true slope lies within this range. We are 95% confident that for every one-point increase in a firm's management quality score, the expected number of employees is associated with an increase of from 3314 to 10200, holding other factors constant. Since this interval does not include zero, it suggests that the relationship between management quality and the number of employees is statistically significant at the 5% significant value.

The intercept is of no meaning since firms would not have a negative number of employees, regardless of their management quality scores.
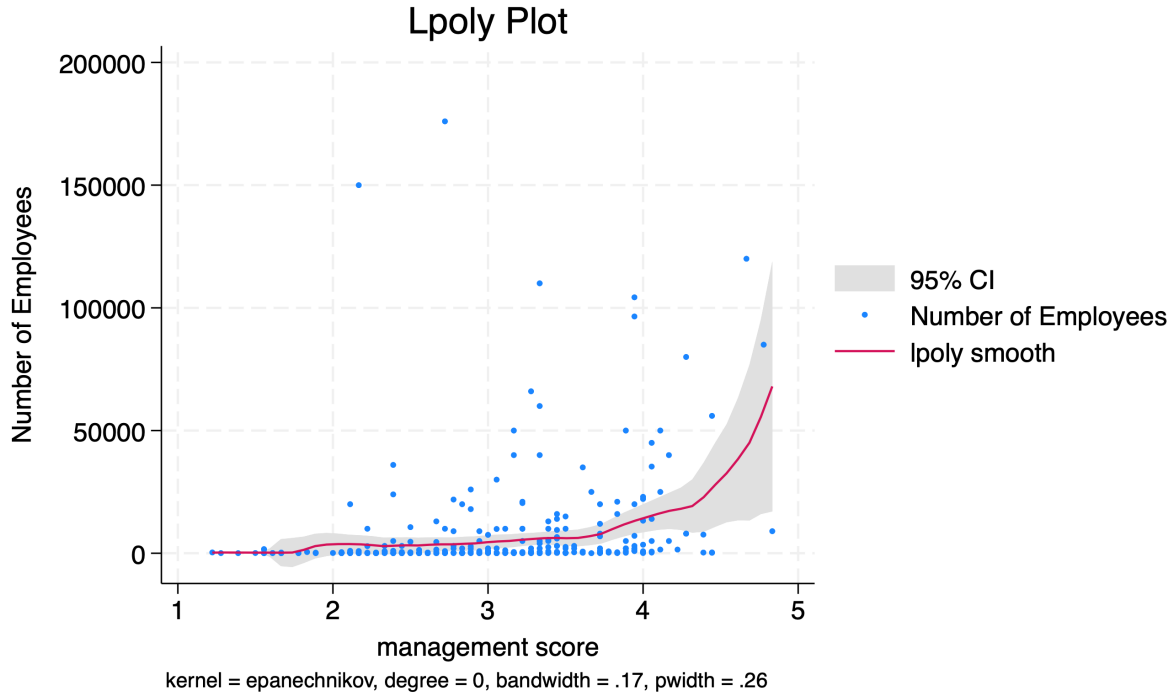
2 Non-linear regression

Figure 1: Lpoly plot

The lpoly graph indicates a non-linear relationship between management score and the number of employees. It shows that the slope of the fitted line steepens as the management score increases. However, the widening confidence intervals (CIs) and the lack of observations for firms with higher management scores suggest that this steepening may be driven by influential observations rather than a robust underlying trend.

To address this, I designed a log-level regression with splines at management scores of 2 and 4. The rationale for the log-level setup is that firm size tends to grow exponentially—most firms are very small, while a few are exceptionally large. Taking the natural log of firm size reduces the influence of these outlier firms. As for the splines, I observed in Figure 1 that most firms cluster around management scores between 2 and 4. For firms with scores from 1 to 2, they are almost uniformly small, while for those with scores from 4 to 5, the positive relationship between employee size and management score appears to be more pronounced.

Table 2: Log-level regression with splines results

	l_emp	95% CI
management: (.,2)	-0.025	[-1.987,1.937]
management: (2,4)	1.435***	[1.095,1.774]
management: (4,.)	3.003***	[0.777,5.230]

2

Table 2: Log-level regression with splines results

	l_emp	95% CI
Constant	4.873**	[1.100,8.646]
Observations	393	
R2	0.227	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results show that the 2 for this new model is significantly higher than that of the simple linear model, indicating greater explanatory power. Supporting my observations, there is no significant relationship between the log of the number of employees and the management score when the score is between 1 and 2. However, for management scores between 2 and 4, and from 4 to 5, the number of employees is positively associated with the management score. Specifically, for firms with management scores between 2 and 4, a one-unit increase in the management score is associated with a 144% larger employee size. For firms with management scores between 4 and 5, a one-unit increase in the management score is associated with a 300% larger employee size.
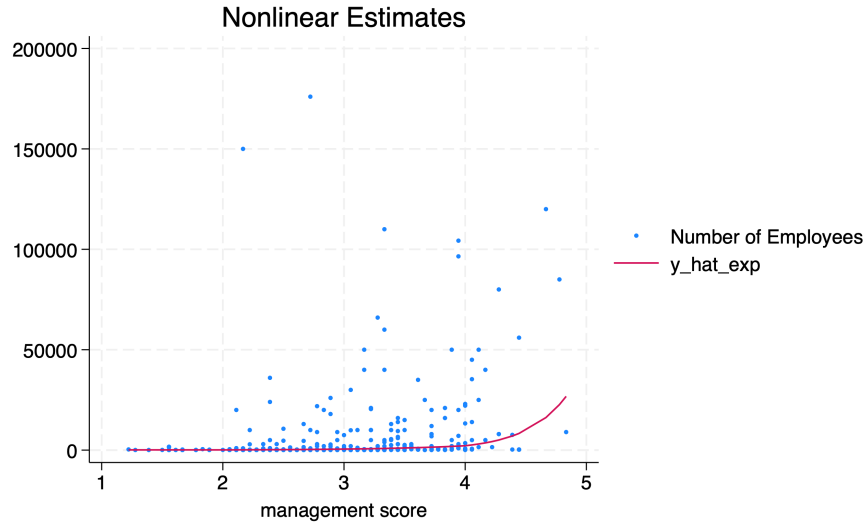
This plot shows the estimates made by the new model, which seem smaller than the ones made by the simple linear model.

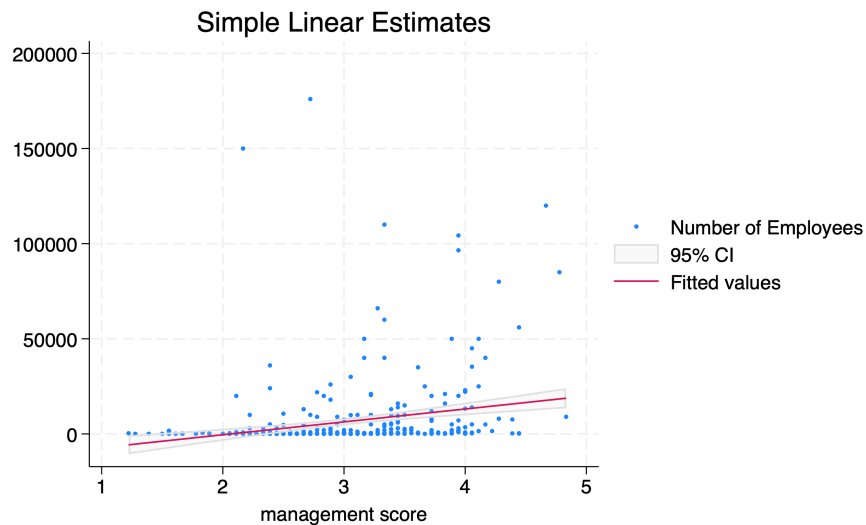Figure 2: Nonlinear estimates plot

3 Comparison

Figure 3: Simple linear regression estimates

I believe I can reject the hypothesis that the linear approximation is adequate for the population of firms represented by the data. First, the new model has an R^2 that is more than twice as high as the simple linear model, indicating greater explanatory power. Second, the linear model produces many negative estimates for employee size when the management score is below 2, as shown in Figure 3, which is unrealistic. The usage of splines in the new model avoided that issue. Additionally, the new model is less susceptible to the influence of outlier observations and captures the non-linear relations, making it more robust.