Broadway data analysis

Julianna Szabo

12/23/2020

Executive summary

Research question

Is there a correlation between the occupancy percentage of a show and the revenue per attendant?
I will be looking for causality.

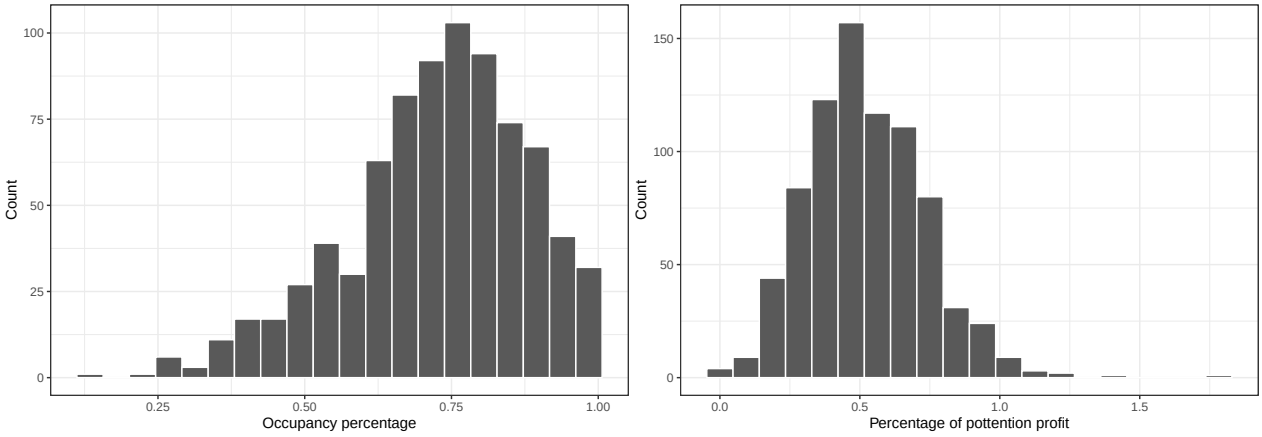
Data

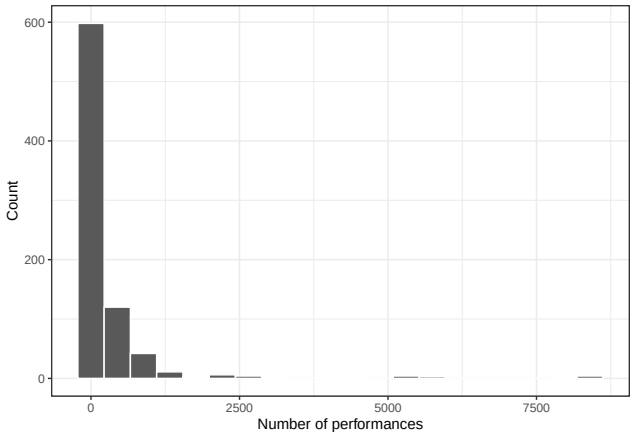
The data is very complete and representative. I have removed some missing values during the cleaning process but it was a very small percentage. Further some measures were lost by switching from a time series to a cross sectional data set. However, I aggregated on the show name, which lets me keel the most amount of detail. Most of the variables are quantitative so that means they measure what they decribe. I will use using Revenue / Attendant, where Revenue is measured as the gross revenue of the show, and attendants which are measured as total number of people who attended the show.

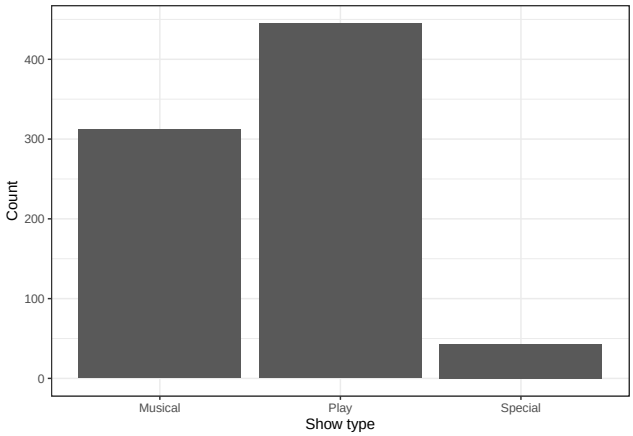
My x variable will be Occupancy percentage (capacity_filled) My y variable will be Revenue / Attendant which I will calculate based on revenue and attendant

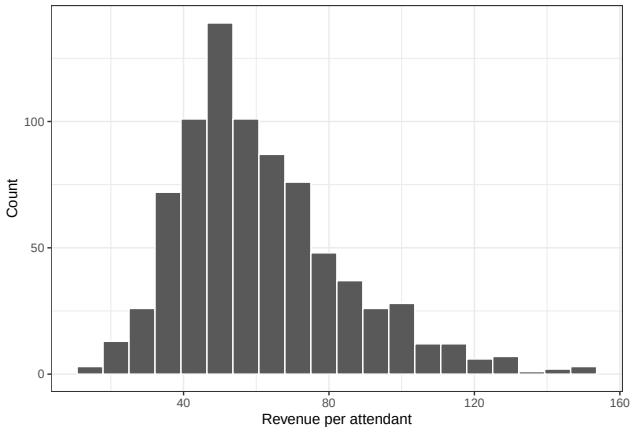
There may be some measurement error in y, which is classic and doesn't affect the slope. There may be some measuement error in x which could also be classic, which does affect the slope.

Summary of variables









variable	n	mean	median	min	max	sd
Occupancy percentage	800	0.7250182	0.7423252	0.1500000	1.000000	0.1542145
Percentage of possible profit	800	0.5247919	0.5039474	0.0136364	1.796364	0.2107851
Number of performances	800	345.4887500	101.0000000	0.0000000	8400.000000	944.3725797
Revenue per Attendant	800	61.2937476	56.5722427	12.6415157	148.397589	22.9658889

Looks like they are distributed somehwat normall, but y has a long right tail, while x has more of a left tail. Also looking at x, there are a few outliers, since a percentage should not be larger than 1. Therefore I will remove these from the set.

Ln transformations

Appendix Level- log makes the most sense

Regression Models

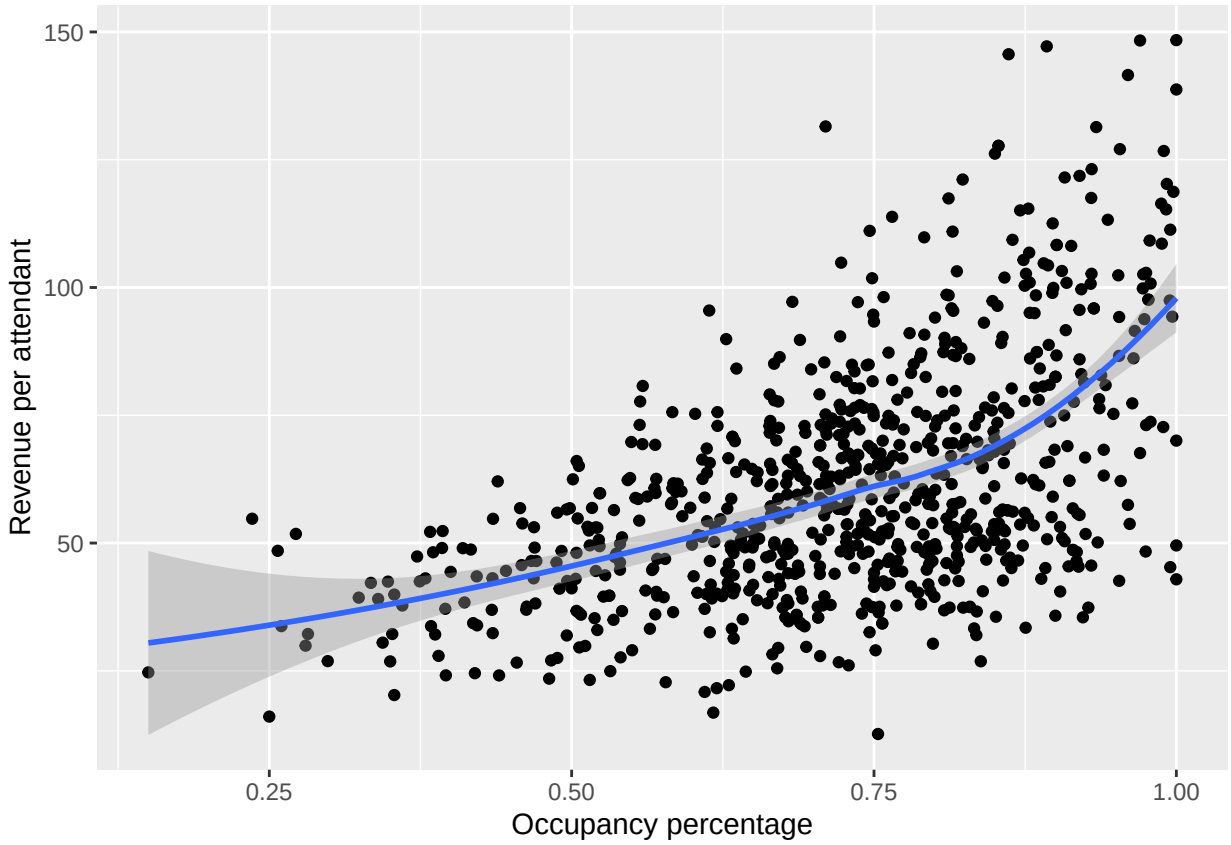
Decided on ..... model

Appendix

Ln transformation

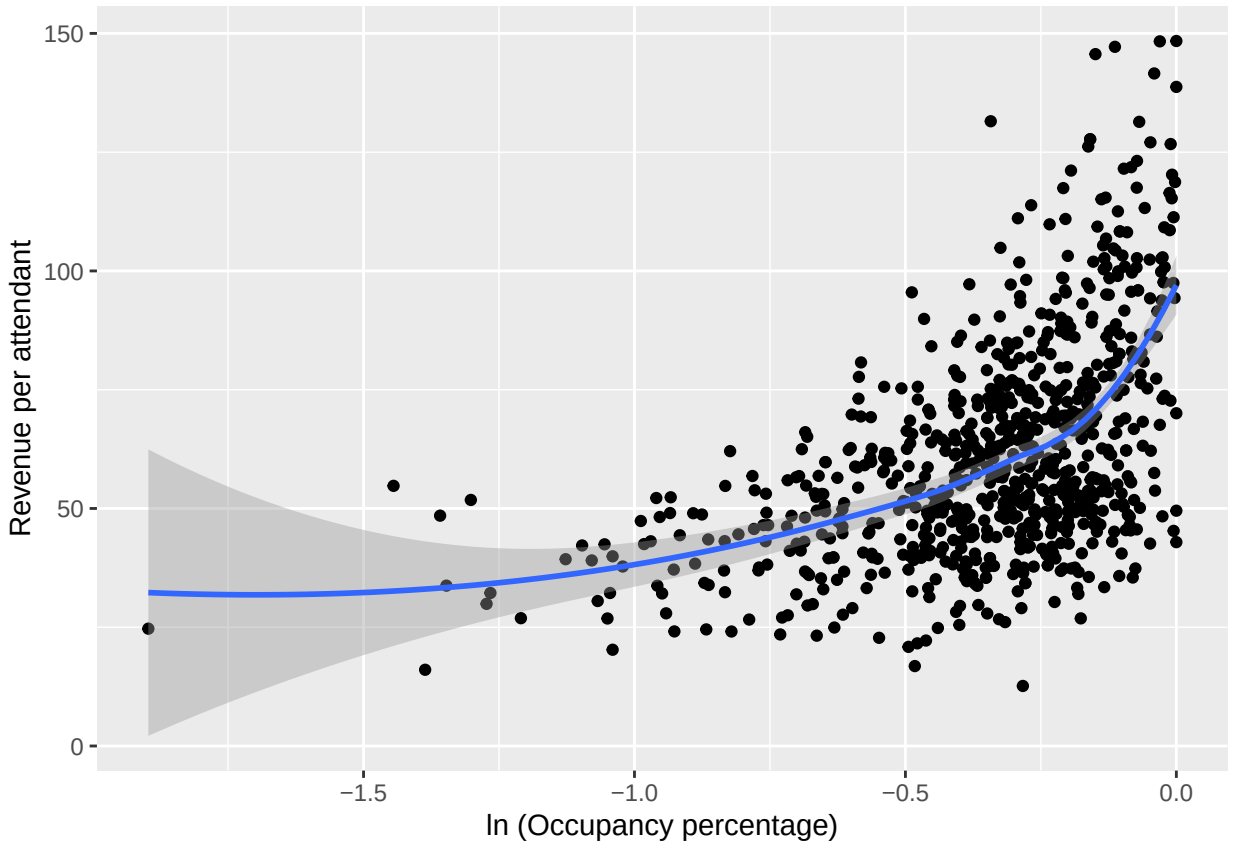
Level - level regression

```
## `geom_smooth()` using formula 'y ~ x'
```



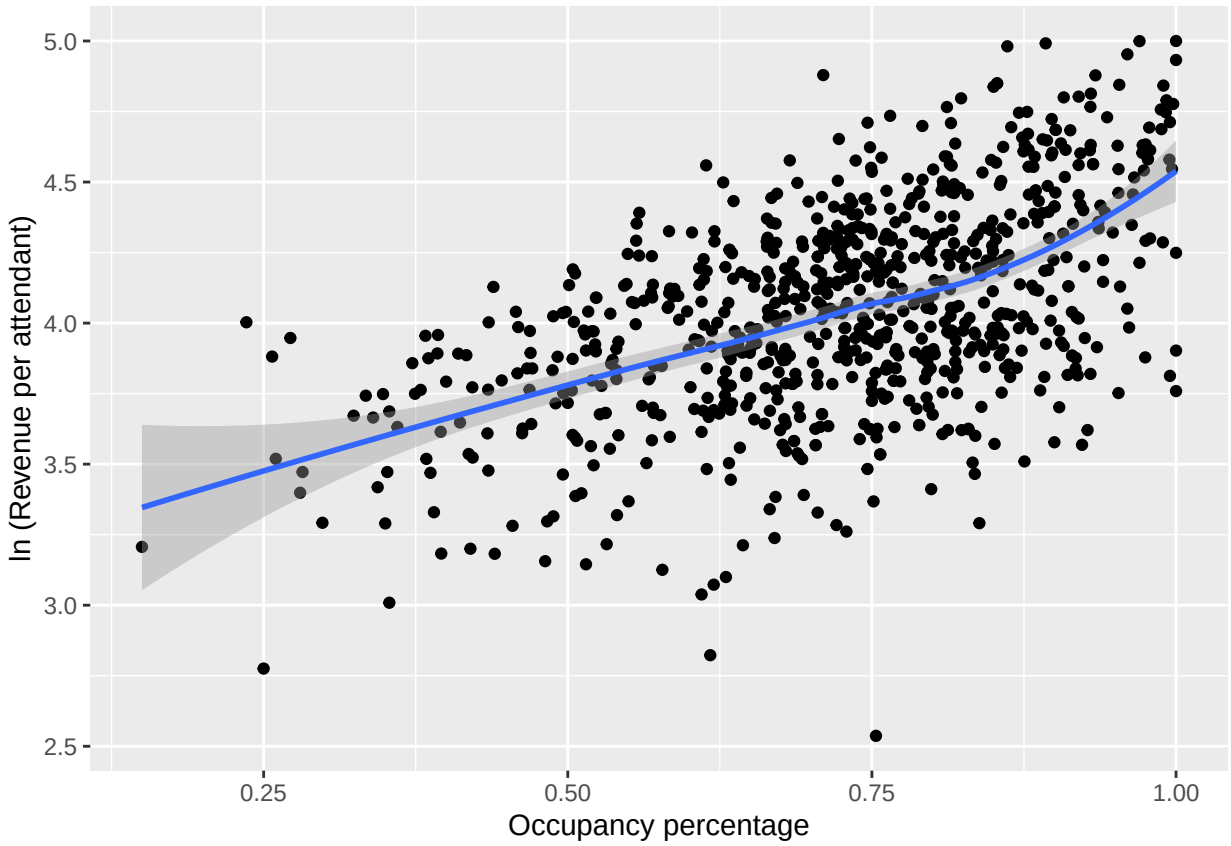
Log - level regression

```
## `geom_smooth()` using formula 'y ~ x'
```



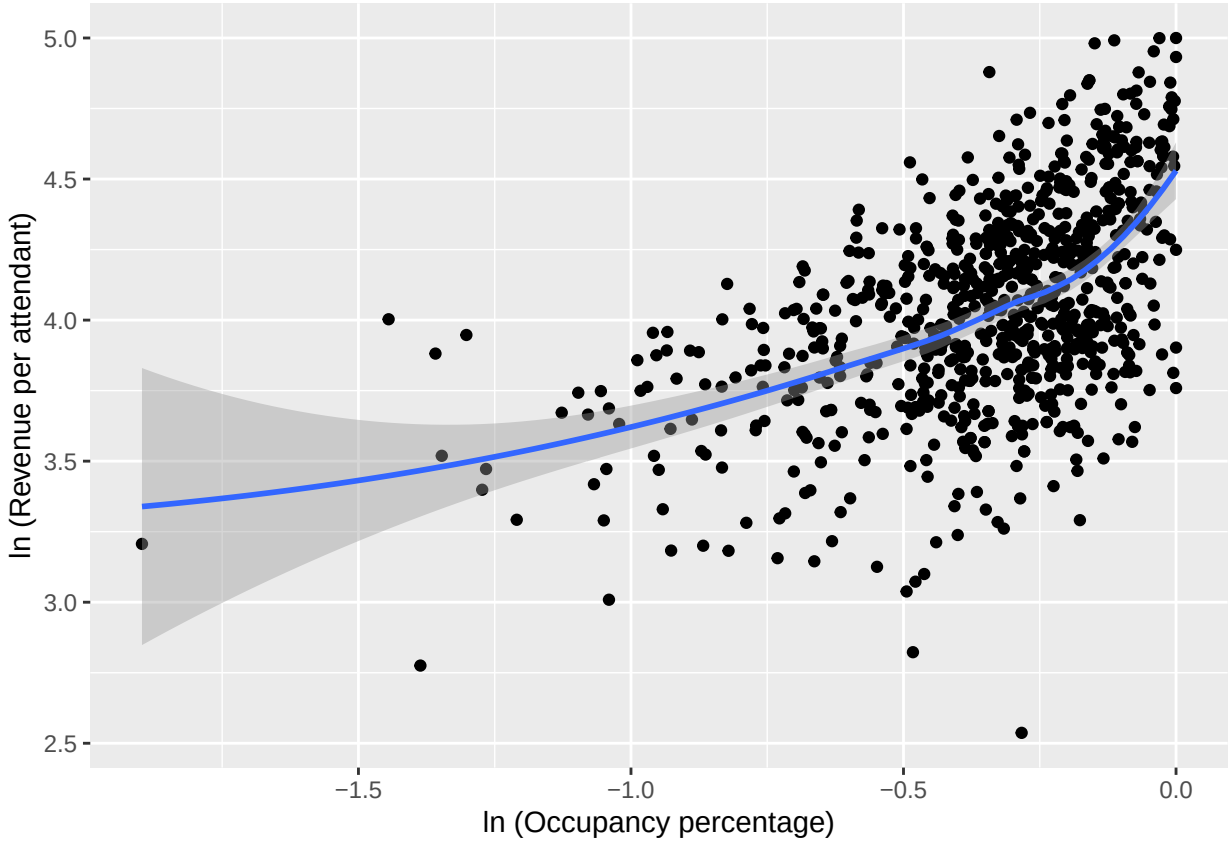
Level - log regression

```
## `geom_smooth()` using formula 'y ~ x'
```



Log - log regression

```
## `geom_smooth()` using formula 'y ~ x'
```

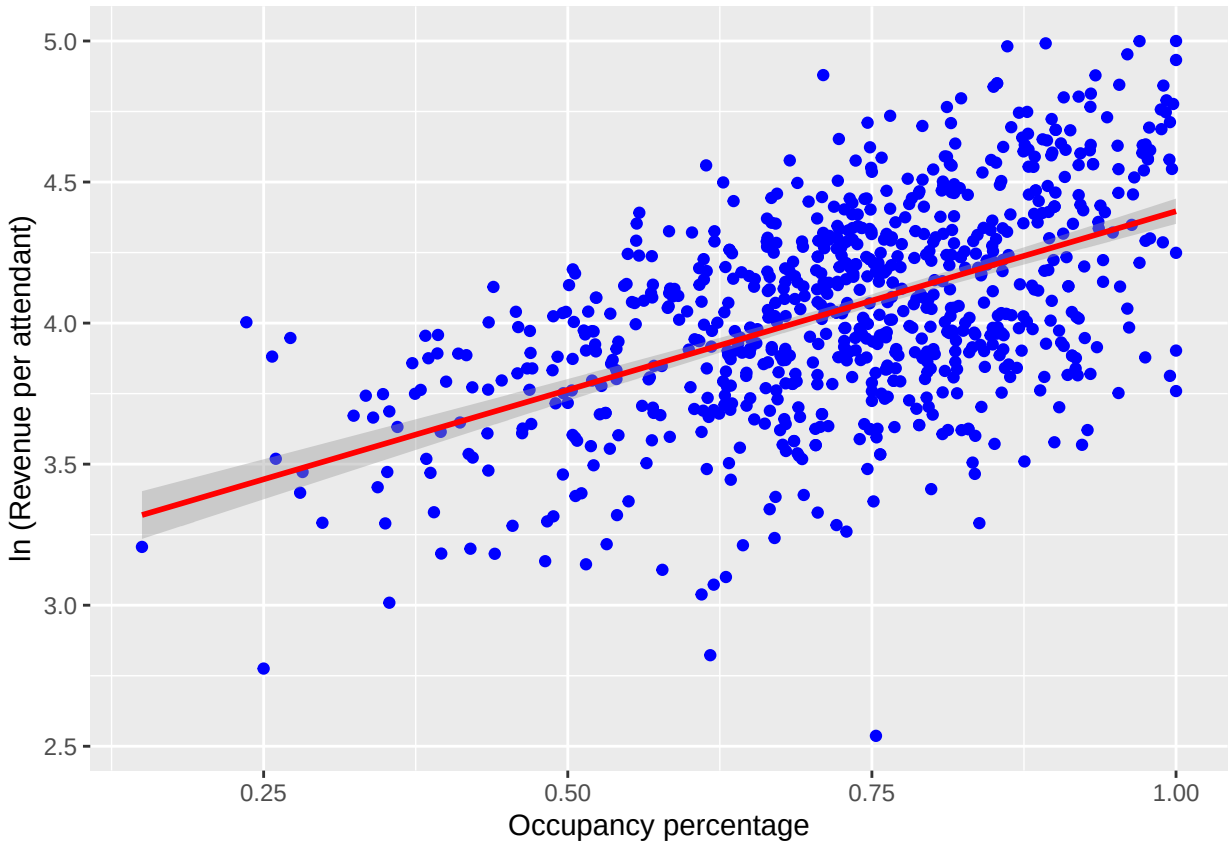


Regression modes

Regression 1 - Simple linear regression

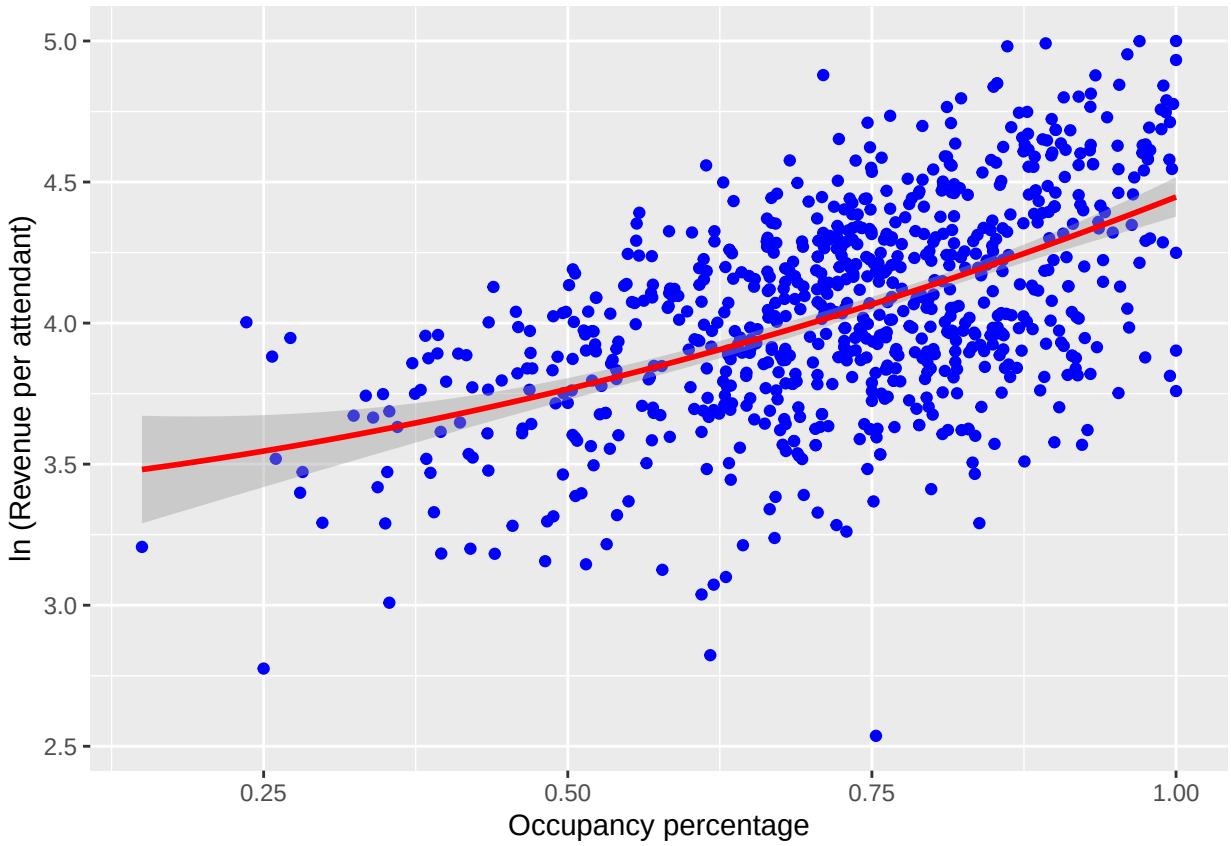
```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled, data = df,
##     se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)        3.130    0.05222   59.93 1.204e-297    3.027    3.232 798
## capacity_filled    1.267    0.07224   17.54  1.769e-58    1.125    1.409 798
##
## Multiple R-squared:  0.2774 ,    Adjusted R-squared:  0.2765
## F-statistic: 307.5 on 1 and 798 DF,  p-value: < 2.2e-16

## `geom_smooth()` using formula 'y ~ x'
```



Regression 2 - Quadratic (linear) regression

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + capacity_filled_sq,
##     data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                    Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)          3.4069     0.1663 20.4852 2.982e-75   3.0804    3.733 797
## capacity_filled      0.3969     0.5041  0.7873 4.314e-01  -0.5927    1.387 797
## capacity_filled_sq   0.6433     0.3721  1.7289 8.422e-02  -0.0871    1.374 797
##
## Multiple R-squared:  0.2805 ,    Adjusted R-squared:  0.2787
## F-statistic: 153.8 on 2 and 797 DF,  p-value: < 2.2e-16
```



Regressipn 3 - Piecewise linear spline regression

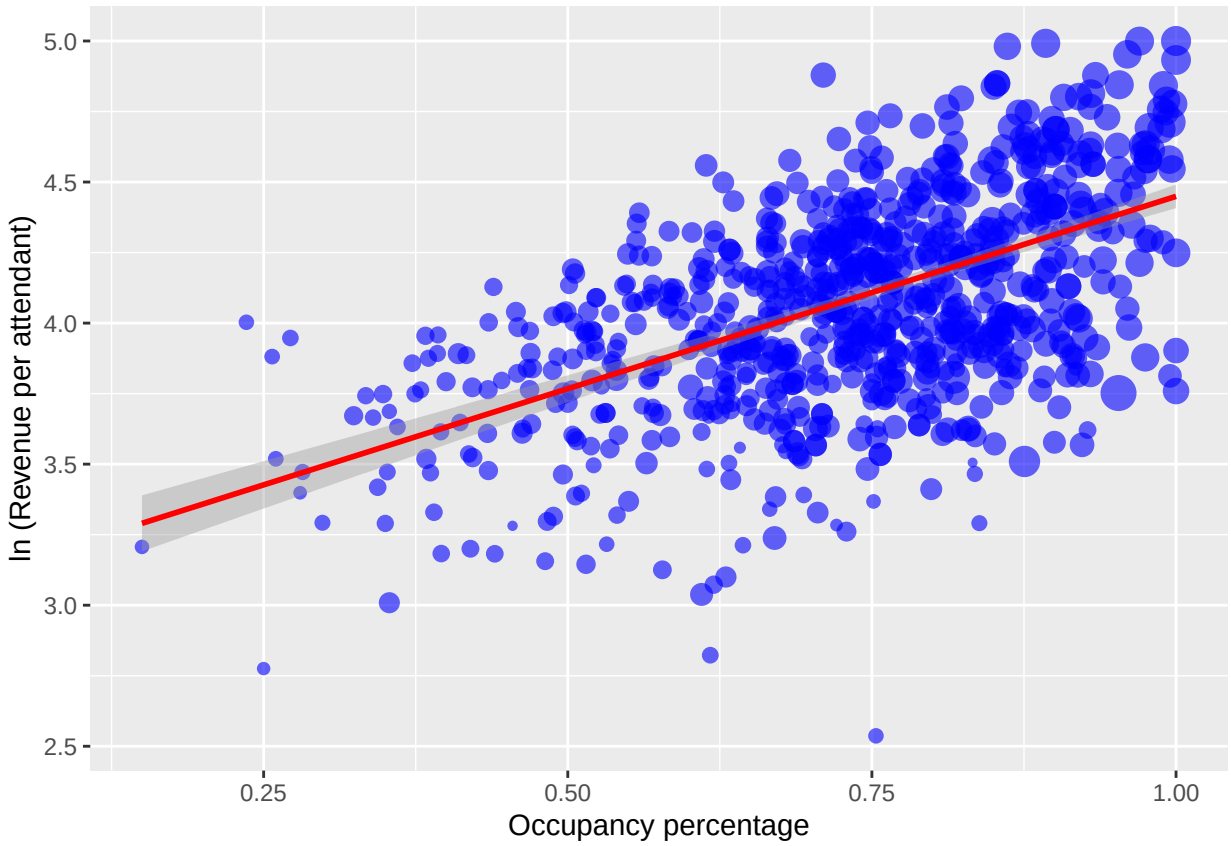
Using 0.5 as a cutof point

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ lspline(capacity_filled,
##     cutoff), data = df)
##
## Standard error type:  HC2
##
## Coefficients:
##                                  Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                         3.237    0.17047  18.989 1.293e-66
## lspline(capacity_filled, cutoff)1   1.031    0.36241   2.845 4.561e-03
## lspline(capacity_filled, cutoff)2   1.304    0.09117  14.302 1.799e-41
##                                  CI Lower CI Upper  DF
## (Intercept)                        2.9023    3.572 797
## lspline(capacity_filled, cutoff)1  0.3195    1.742 797
## lspline(capacity_filled, cutoff)2  1.1250    1.483 797
##
## Multiple R-squared:  0.2779 ,    Adjusted R-squared:  0.2761
## F-statistic: 153.1 on 2 and 797 DF,  p-value: < 2.2e-16
```

Regression 4 - Weighted linear regression, where weights = percentage of total revenue

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled, data = df,
##     weights = percentage_of_poss_profit)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)        3.086    0.06155   50.13 8.467e-249    2.965    3.206 798
## capacity_filled    1.364    0.08552   15.94  6.827e-50    1.196    1.531 798
##
## Multiple R-squared:   0.27 , Adjusted R-squared:  0.2691
## F-statistic: 254.2 on 1 and 798 DF,  p-value: < 2.2e-16

## `geom_smooth()` using formula 'y ~ x'
```



Regression 5 - Weighted linear regression, where weights = number of performances

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled, data = df,
##     weights = num_of_performances)
##
## Weighted, Standard error type:  HC2
##
## Coefficients:
##                 Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)        2.887    0.09215   31.33 3.863e-141    2.707    3.068 798
## capacity_filled    1.638    0.12320   13.29  1.467e-36    1.396    1.879 798
##
## Multiple R-squared:  0.3523 ,    Adjusted R-squared:  0.3515
## F-statistic: 176.7 on 1 and 798 DF,  p-value: < 2.2e-16

## `geom_smooth()` using formula 'y ~ x'
```

Model Comparison

```
## The table was written to the file '/Users/Terez/OneDrive - Central European University/Data_Analysis_
```

Additional models

Check if it becomes better if one of the weights are included as variables

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + percentage_of_poss_profit,
##     data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                            Estimate Std. Error t value   Pr(>|t|) CI Lower
## (Intercept)                  3.2975    0.05251  62.795 8.278e-311   3.1944
## capacity_filled              0.4541    0.13818   3.286  1.059e-03   0.1829
## percentage_of_poss_profit    0.8029    0.13674   5.872  6.327e-09   0.5345
##                            CI Upper  DF
## (Intercept)                  3.4005 797
## capacity_filled              0.7254 797
## percentage_of_poss_profit    1.0713 797
```

```
##
## Multiple R-squared:  0.3715 ,    Adjusted R-squared:  0.3699
## F-statistic:   170 on 2 and 797 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + num_of_performances,
##     data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                      Estimate Std. Error t value   Pr(>|t|)  CI Lower  CI Upper
## (Intercept)         3.146e+00  5.315e-02   59.19 5.963e-294 3.042e+00 3.250e+00
## capacity_filled     1.232e+00  7.513e-02   16.40  2.758e-52 1.084e+00 1.379e+00
## num_of_performances 2.524e-05  8.169e-06    3.09  2.071e-03 9.208e-06 4.128e-05
##                      DF
## (Intercept)         797
## capacity_filled     797
## num_of_performances 797
##
## Multiple R-squared:  0.2814 ,    Adjusted R-squared:  0.2796
## F-statistic: 179.3 on 2 and 797 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + percentage_of_poss_profit +
##     num_of_performances, data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
## Coefficients:
##                             Estimate Std. Error t value   Pr(>|t|)  CI Lower
## (Intercept)                3.306e+00  5.268e-02  62.755 2.016e-310 3.202e+00
## capacity_filled            4.410e-01  1.360e-01   3.244  1.230e-03 1.741e-01
## percentage_of_poss_profit 7.949e-01  1.372e-01   5.795  9.853e-09 5.256e-01
## num_of_performances        1.538e-05  6.579e-06   2.337  1.968e-02 2.462e-06
##                             CI Upper  DF
## (Intercept)                3.409e+00 796
## capacity_filled            7.079e-01 796
## percentage_of_poss_profit 1.064e+00 796
## num_of_performances        2.829e-05 796
##
## Multiple R-squared:  0.3729 ,    Adjusted R-squared:  0.3706
## F-statistic:   156 on 3 and 796 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm_robust(formula = ln_revenue_per_att ~ capacity_filled + percentage_of_poss_profit +
##     num_of_performances + as.factor(show_type), data = df, se_type = "HC2")
##
## Standard error type:  HC2
##
```

```
## Coefficients:
##                               Estimate Std. Error t value   Pr(>|t|)   CI Lower
## (Intercept)                  3.397e+00  5.389e-02 63.0466 2.449e-311  3.292e+00
## capacity_filled              4.036e-01  1.339e-01  3.0136  2.663e-03  1.407e-01
## percentage_of_poss_profit    8.113e-01  1.368e-01  5.9324  4.455e-09  5.429e-01
## num_of_performances         -9.732e-07  6.310e-06 -0.1542  8.775e-01 -1.336e-05
## as.factor(show_type)Play    -1.143e-01  2.083e-02 -5.4885  5.456e-08 -1.552e-01
## as.factor(show_type)Special -7.156e-02  6.557e-02 -1.0914  2.754e-01 -2.003e-01
##                               CI Upper  DF
## (Intercept)                  3.503e+00 794
## capacity_filled              6.664e-01 794
## percentage_of_poss_profit    1.080e+00 794
## num_of_performances          1.141e-05 794
## as.factor(show_type)Play    -7.343e-02 794
## as.factor(show_type)Special  5.715e-02 794
##
## Multiple R-squared:  0.3928 ,    Adjusted R-squared:  0.3889
## F-statistic: 109.8 on 5 and 794 DF,  p-value: < 2.2e-16
```

Explore again

```
## The table was written to the file '/Users/Terez/OneDrive - Central European University/Data_Analysis
```