**Predicting Particle Pollution Levels Based on Car Population**

**Noam Benkler**

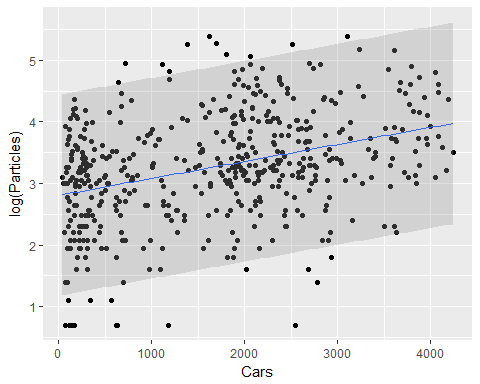
**Introduction**

Particulate matter (pm), or particle pollution, is a complex mixture of minuscule particles, including acids, organic chemicals, metals, and soil or dust particles. Many studies have linked particle exposure to health problems. Given potential negative health impacts of PM, scientists often study the relationship between various human activities and PM concentrations in the air. In this case study we examine the relationship between the number of PM particles and the number of cars, and whether the traffic be used to predict PM levels.

**Data**

Our data set consists of PM levels (in number of particles) and number of Cars that passed through an intersection. No data points stood out as outliers that needed to be removed before analysis (Table 1).

*Table 1. Summary statistics for Particulate Matter data.*  
========================================  
Statistic N Mean St. Dev. Min Max   
----------------------------------------  
Particles 500 37.9 35.0 2 220   
Cars 500 1,683.2 1,154.5 45 4,239  
logPm 500 3.3 0.9 0.7 5.4   
----------------------------------------

Exploratory graphs of the number of Cars and PM levels showed strong skewness and outliers, so analysis was performed on the log-transformed PM variable. (Figure 1)

*Fig 1. Scatterplot of Particle Pollution levels (log scale) vs. Number of Cars in an Intersection. The fitted regression model with 95% prediction bands hat been added.*

**Results**

We consider the simple linear regression model

μ[log(Particles) | Cars] = b0 + b1 log(Particles)

where Particles and Cars denote measured PM levels and number of Cars, respectively. The estimates of the model parameters are given in Table 2.

*Table 2. Model Coefficients*

Linear Model Coefficients (standard error in parenthesis)  
===============================================  
 Dependent variable:   
 ---------------------------  
 log(Particles)   
-----------------------------------------------  
Cars 0.0003\*\*\*   
 (0.00003)   
   
Constant 2.805\*\*\*   
 (0.066)   
   
-----------------------------------------------  
Observations 500   
R2 0.130   
Adjusted R2 0.128   
Residual Std. Error 0.828 (df = 498)   
F Statistic 74.137\*\*\* (df = 1; 498)   
===============================================  
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

We find that a 1 Car increase is associated with a multiplicative change in median PM levels of 1.00. This means our data shows a nearly 1/1 relationship between the number of Cars on the road and an increase in the log of PM levels. This model explains 13% of the variability in the log of PM levels.

For a day with 2000 cars on the road, we predict a log (PM level) of 3.36 (95% prediction interval of 1.73 to 4.99).

**Discussion**

Our regression model meets all the assumptions necessary for accurate linear regression. The residuals plot and qq-plot created from our data, showed the data conformed to a normal distribution, with independence, and constant error variance. Despite conforming to these assumptions and having p-values below 0.01 on our model coefficients, there were some problems that arise from our model. First, there were several points outside of the prediction interval, which make it difficult to conclude accurate predictions 100% of the time. Second, though the prediction interval itself covers most of the data, our linear regression only explains 13% of the variability in in the log of PM levels, meaning, while the interval may be accurate, the predicted value given by the linear model may be off. However, with some stipulation, our simple linear model can be used to predict a range of PM levels from number of cars.

Stats Case Study #1

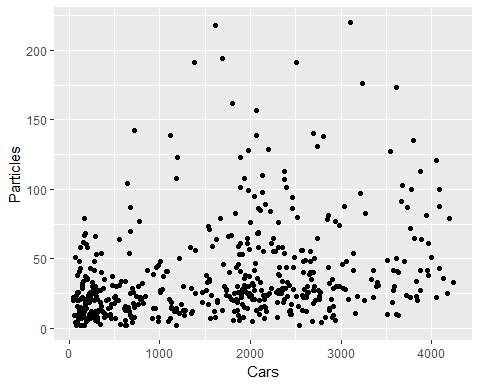
Noam Benkler

# scatterplot of cars vs. pm levels and residuals for log grapj

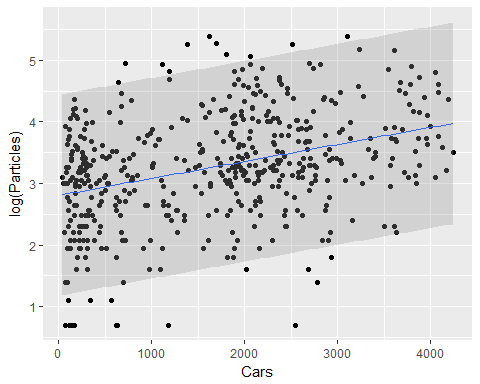
summary(pm)

## Particles Cars   
## Min. : 2.00 Min. : 45.0   
## 1st Qu.: 15.00 1st Qu.: 523.5   
## Median : 27.00 Median :1851.5   
## Mean : 37.88 Mean :1683.2   
## 3rd Qu.: 48.00 3rd Qu.:2509.5   
## Max. :220.00 Max. :4239.0

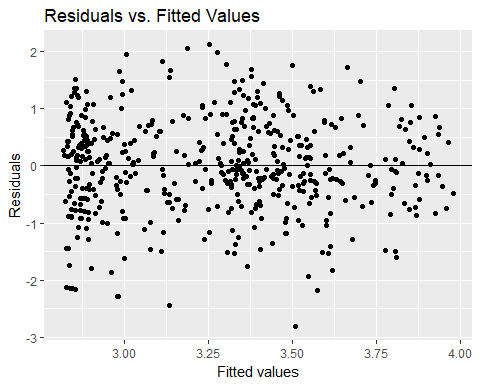
gf\_point(Particles~Cars, data = pm)



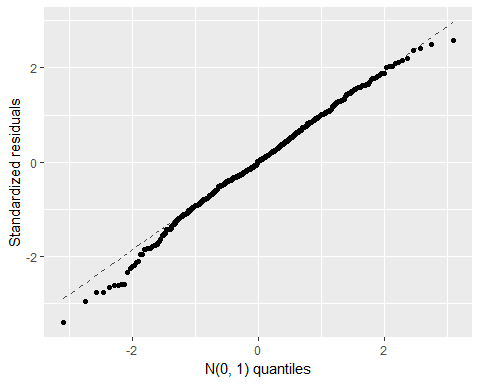
#try to log the dep var. to make graph more linear  
gf\_point(log(Particles)~Cars, data = pm) %>% gf\_lm(log(Particles)~Cars, data = pm, interval = "prediction")



logPm.lm <- lm(log(Particles)~Cars, data = pm)  
resid.logPm <-augment(logPm.lm)  
gf\_point(.resid~.fitted, data=resid.logPm) %>% gf\_hline(yintercept = 0, col = "blue", lty = 2) %>% gf\_labs(x = "Fitted values", y = "Residuals", title = "Residuals vs. Fitted Values")



gf\_qq(~.std.resid, data = resid.logPm) %>% gf\_qqline() %>% gf\_labs(x = "N(0, 1) quantiles", y = "Standardized residuals")



# Tables

summary(logPm.lm)

##   
## Call:  
## lm(formula = log(Particles) ~ Cars, data = pm)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.81497 -0.49851 0.01011 0.55772 2.13160   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.805e+00 6.553e-02 42.81 <2e-16 \*\*\*  
## Cars 2.765e-04 3.211e-05 8.61 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8282 on 498 degrees of freedom  
## Multiple R-squared: 0.1296, Adjusted R-squared: 0.1278   
## F-statistic: 74.14 on 1 and 498 DF, p-value: < 2.2e-16

pm <- pm %>% mutate(logPm = log(Particles))  
stargazer(pm,type = "text", title = "Summary of PM dataset", digits = 1)

##   
## Summary of PM dataset  
## ========================================  
## Statistic N Mean St. Dev. Min Max   
## ----------------------------------------  
## Particles 500 37.9 35.0 2 220   
## Cars 500 1,683.2 1,154.5 45 4,239  
## logPm 500 3.3 0.9 0.7 5.4   
## ----------------------------------------

stargazer(logPm.lm, type = "text", title = "Linear Model Coefficients (standard error in parenthisis)", flip = TRUE)

##   
## Linear Model Coefficients (standard error in parenthisis)  
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log(Particles)   
## -----------------------------------------------  
## Cars 0.0003\*\*\*   
## (0.00003)   
##   
## Constant 2.805\*\*\*   
## (0.066)   
##   
## -----------------------------------------------  
## Observations 500   
## R2 0.130   
## Adjusted R2 0.128   
## Residual Std. Error 0.828 (df = 498)   
## F Statistic 74.137\*\*\* (df = 1; 498)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

newdata = data.frame(Cars = 2000)  
predict(logPm.lm, newdata, interval="predict")

## fit lwr upr  
## 1 3.358522 1.729596 4.987448