IDS 702 - Data Analysis Assignment 1

Aarushi Verma

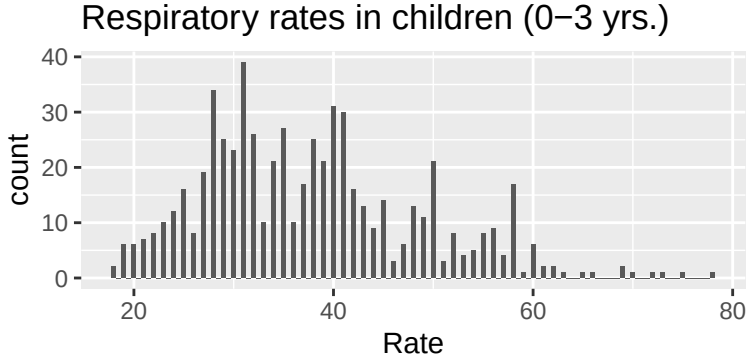
2021/09/04

Question 1 - Respiratory Rates for Children

1. Do exploratory analysis on the data and include a useful plot that a physician could use to assess a "normal" range of respiratory rates for children of any age between 0 and 3

Based on EDA our data has 618 rows and 2 columns. The 2 columns are Age (in months) and Respiratory Rates. The range of age variable is 0.10 months ot 36 months where as the range is for rate is 18 to 78. Based on the summary, we can see that the mean and median value are not very far away from each other implying that the variables are not extremely skewed and more normally distributed.

For a physician to assess the normal respiratory rate for children between the age 0 and 3 we can plot the data on a histogram.



Based on the histogram, we can see that majority of the children have rates between 29 and 45, with most children with a respiratory rate of 31. The physician with subject matter expertise can use the histogram to infer and assess or "normal range" of respiratory rates for children between the age of 0 and 3 with a more informed perspective.

2. Write down a regression model for predicting respiratory rates from age. Make sure to use the right mathematical notation.

Our regression model will be:

$Rate \sim \beta_0 + \beta_1(Age)$

Our response variable is Rate and our explanatory variable is Age. Based on the model: β_0 is the intercept term. The intercept term gives us the average value of our response variable when the explanatory variables are 0. β_1 is the coefficient for the age variable. For a unit change in age variable we can infer the change in our response variable based on the value of β_1.

Intuitively, interpreting the intercept i.e. when Age is 0 is not sensible. In order to interpret it more meaningfully, we use mean centering and create a new regression model:

$Rate \sim \beta_0 + \beta_1(Age - \bar{Age})$

3. Fit the model to the data and interpret your results.

Based on the two models mentioned int he previous question we can interpret the following:

Model 1: $Rate \sim 47.05 + (-0.70)Age$

Based on the result of this model we can interpret that, for age at 0 months, the average respiratory rate for children is 47.05. Age is a significant predictor with a p value of <0.05. For an increase in age by a month, the respiratory rate for a child on average falls by 0.70 units.The adjusted $R2$ value is 0.48 which means that this model explains 48% of the variation in respiratory rates.

Model 2: $Rate \sim 37.74 + (-.0.70)(Age - \bar{Age})$

Based on the result of this model, we can interpret that the mean respiratory rate for a child with sample Age (13.39 months) is 37.74 units.

4. Include a table showing the output from the regression model including the estimated intercept, slope, residual standard error, and proportion of variation explained by the model.

Table 1: Linear Regression on Respiratory Rate vs. Age

	Dependent variable:	
	Rate	
	(1)	(2)
Age	-0.70^{***}	
	(0.03)	
age_cent		-0.70^{***}
		(0.03)
Constant	47.05^{***}	37.74^{***}
	(0.50)	(0.32)
Observations	618	618
R^2	0.48	0.48
Adjusted R^2	0.48	0.48
Residual Std. Error (df = 616)	7.84	7.84
F Statistic (df = 1; 616)	560.92^{***}	560.92^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

5. Is there enough evidence that the model assumptions are reasonable for this data? You should consider transformations (think log transformations, etc) if you think there's a violation of normality and/or linearity.

To assess the model we check if any of the assumptions are violated. In order to check the linearity assumption we plot the residuals of the model against age. The points are randomly distributed and there is no visible pattern, therefore we can conclude that the linearity assumption is not violated.

To check the independence and equal variance assumptions we plot the Residuals against the fitted values. The points seem randomly distributed with no discernible pattern and the spread of variables seems constant above and below the line. There does seem to be some points on the x axis that may violate the equal variance of errors assumptions however, they are only few and we can go ahead and say that neither of the assumptions are violated.

To check for normality, we plot the Q-Q plot. For our model we can observe that majority of the points lie on the 45 degree line. There may be outliers present in the data, however there are not too many to say that the normality assumption is violated

6. Demonstrate the usefulness of the model by providing 95% prediction intervals for the rate for three individual children: a 1 month old, an 18 months old, and a 29 months old.

95% prediction intervals for the rate for three individual children:

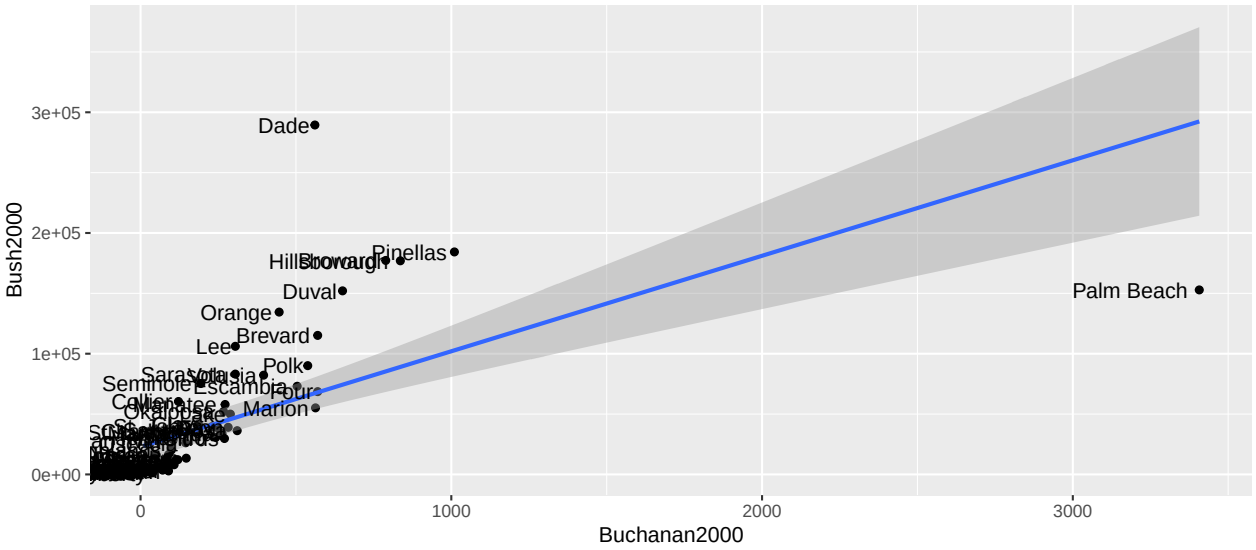
1 month old: (21.627365, 52.45370) Keeping everything constant, based on our model we can predict that the respiratory rates for children of age 1 month will be between 21.63 and 52.45 units with 95% confidence.

18 months old: (9.765408, 40.66140) Keeping everything constant, based on our model we can predict that the respiratory rates for children of age 1 month will be between 9.77 and 40.66 units with 95% confidence.

29 months old: (2.056974, 33.06414) Keeping everything constant, based on our model we can predict that the respiratory rates for children of age 1 month will be between 2.06 and 33.06 units with 95% confidence.

Question 2 - The Dramatic U.S. Presidential Election of 2000

1. Make a scatterplot of the variables Buchanan2000 and Bush2000. What evidence is there in the scatterplot that Buchanan received more votes than expected in Palm Beach County?



Based on the scatter plot, we cab observe that the point for Palm Beach county does not follow the trend when looking at the votes received by Buchanan in other counties. It seems that the Palm Beach datapoint has high leverage and we should investigate whether it carries a large influence on the other points as well.

2. Fit a linear regression model to the data to predict Buchanan votes from Bush votes, without using Palm Beach County results. You should consider transformations for both variables if you think there's a violation of normality and/or linearity.

Since Palm Beach County was an outlier, we remove the data point and fit a linear regression model of the form:

$Buchanan2000 \sim \beta_0 + \beta_1(Bush2000)$

While checking for the assumptions by examining the residual plots, we transform both the variables using a log transformation.

3. Include the output from the final regression model that you used, as well as evidence that the model fits the assumptions reasonably well.

[1] "X" "County" "Buchanan2000" "Bush2000"

Table 2: Linear Regression on Election Results

	Dependent variable:
	log(Buchanan2000)
log(Bush2000)	0.73^{***}
	(0.04)
Constant	-2.34^{***}
	(0.35)
Observations	66
R^2	0.87
Adjusted R^2	0.86
Residual Std. Error	0.42 (df = 64)
F Statistic	413.02^{***} (df = 1; 64)
Note:	*p<0.1; **p<0.05; ***p<0.01

The Table above shows the output of the regression model.

Linearity: There are some points clustered around the origin on the X axis, however the points seem randomly distributed. We can conclude that the linearity assumption is not violated

Independence and Equal Variance: Based on the plots, we can see that our data has outliers. However majority of the points are randomly distrbuted without any obvious relationship. The points are also equally distributed above and below the y=0 line. Therefore, we can conclude that the independence and equal varaince assumptions are not violated

Normality: From the QQ plot we can see that most points lie on the 45 degree line. There are some points at the beginning and end of the line that seem to scatter away however based on visual cues we can cocnlude that the normality assumption holds.

4. Obtain a 95% prediction interval for the number of Buchanan votes in Palm Beach from this result, assuming the relationship is the same in this county as in the others. If it is assumed that Buchanan's actual count contains a number of votes intended for Gore, what can be said about the likely size of this number from the prediction interval?

95% prediction interval for Buchanan: (250.8001, 1399.164)

Keeping everything constant, based on our model we can predict that Buchanan should get between 250.80 and 1399.16 votes in Palm Beach county with 95% confidence. As per the data Buchanan received 3407 votes in Palm Beach county. Based on the assumption that voters selected Buchanan instead of Gore by mistake, we can estimate that Gore should have received votes between 2007.84 and 3156.2 additionally from Palm Beach County.

Question 3 - Airbnb listings for Seattle, WA

1. Analyze the data using host_is_superhost,host_identity_verified, room_type, accommodates, bathrooms and bedrooms as predictors. You should start by doing EDA, then model fitting, and model assessment. You should consider transformations if needed.

Based on the EDA we can infer the following: Our data gas 305 rows and 8 columns. Our response variable is price and the other variables are the predictors. Variables like host_is_superhost, host_identify_verified and room_type are categorical variables. The price of airbnbs ranges from 32 to 1650

We also use boxplots to infer the following: 1. There is a positive linear relationship between price and no. of bedrooms as well as price and no. of bathrooms. 2. The median prices are higher when "host_is_superhost" is False. We also have some outliers. 3. The median prices are higher when "host_identity_verified" is False 4. Based on room types, median prices are highest for shared rooms and lowest for private rooms. There are multiple outliers present for entire home/ apartments.

Regression Model: In the following table, we use log transformation on the dependent variable, 'accommodates' and 'bathroom' variable and run the equation:

$log(price) \sim \beta_0 + \beta_1(host_is_superhost) + \beta_2(host_identity_verified) + \beta_3(room_type) + \beta_4(log(accommodates)) + \beta_5(log(bathrooms)) + \beta_6(bedrooms)$

We check the linearity assumption for the variables 'bedrooms', 'accommodates' and 'price'. We can observe that the points are randomly distributed and no trends are visible. These three variables can only be whole numbers i.e. they are discrete, the plots look discrete as well. Therefore we can conclude that linearity assumption holds for these 3 predictors. We don't check the linearity assumption for the rest of the variables as they are factor variables.

Checking for the independence and equal variance assumption, we can see that although there are outliers present in our data, the points are randomly distributed and are equally spread out. Therefore, we can conclude that both assumptions hold.

For the normality assumption, we see that most points lie on the 45 degree line and therefore conclude that the normality assumption holds.

2. Include the output from the final regression model that you used, as well as evidence that the model fits the assumptions reasonably well. Your regression output should includes a table with coefficients and SEs, and p-values or confidence intervals.

Table 3: Results

	Dependent variable:
	log(price)
host_is_superhostTrue	−0.002
	(0.04)
host_identity_verifiedTrue	−0.06
	(0.04)
room_typePrivate room	−0.33***
	(0.07)
room_typeShared room	0.47*
	(0.26)
log(accommodates)	0.33***
	(0.06)
log(bathrooms)	0.40***
	(0.08)
bedrooms	0.11***
	(0.03)
Constant	4.38***
	(0.08)
Observations	305
R^2	0.69
Adjusted R^2	0.68
Residual Std. Error	0.36 (df = 297)
F Statistic	93.22*** (df = 7; 297)
Note:	*p<0.1; **p<0.05; ***p<0.01

3.Interpret the results of your fitted model in the context of the data.

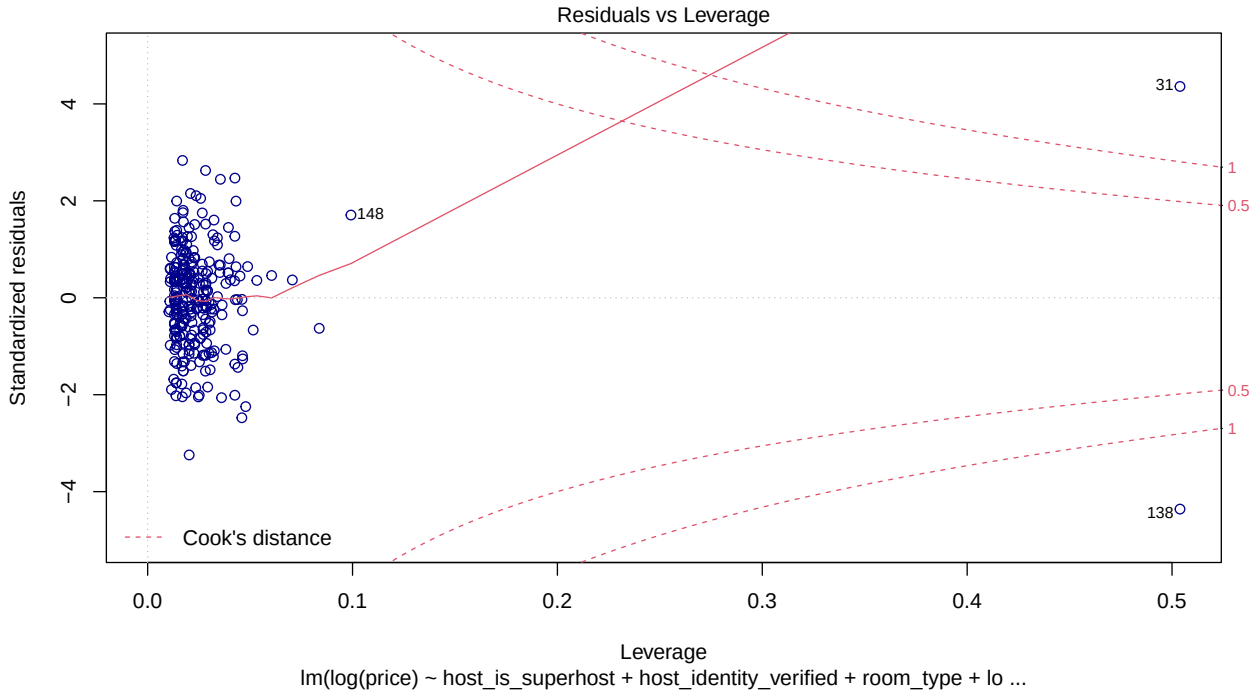
Since we have taken the log transformation, we exponentiate the coefficients of the model and subtract 1 from them. Other variables remaining constant, on average with 95% confidence:

1. On comparison with entire homes/apartments, the price of a private room is 28.09% lesser in price. This variable is significant as p value is less than 0.05.
2. On comparison with entire homes/apartments, the price of a private room is 59.47% lesser in price. This variable is not significant.
3. For a % increase in the number of people the listing can accommodate there is an increase in average price of the property by 38.82%. This variable is significant as p value is less than 0.05.
4. For a % increase in the number of bathrooms in the listing there is an increase in average price of the property by 49.32%. This variable is significant as p value is less than 0.05.
5. For a unit increase in the number of bedrooms in the listing there is an increase in average price of the property by 11.71%. This variable is significant as p value is less than 0.05.
6. On comparison with not superhosts, the price of a property with superhost is 0.23% lesser in price. This variable is not significant.
7. On comparison with unverified hosts, the price of a property with a verified host is 6.18% lesser in price. This variable is not significant.

Overall, the model explains 67.99% of the variation in price of the listing.

4.Are there any (potential) outliers, leverage points or influential points? Provide evidence to support your response. Also, if there are influential points and/or outliers, exclude the points, fit your model without them, and report the changes in your overall conclusions.

Based on the leverage plots and Cook's distance in the following plots, we can observe that: There are 2 points (Point 31 and 38) which are above the value of 0.5 on the leverage plot. These points also have a standardized residuals which are greater than 2. In term's of Cook's distance, the points lie beyond the distance = 1 point. Therefore, we can conclude that these 2 points are outliers, leverafe points and influential points.



On removing the 2 points which we indentified as outliers, we can note that the value of $R2$ improved to 0.69 or 69%. The variable accommodates relationship to our responsible variable price is not significant. The assumptions of linearity, independence, constant variance and normal distribution of residuals still hold. The model is a better fit for the data.

5.Overall, are there any potential limitations of this analysis? If yes, what are two potential limitations?

One limitation of this analysis, is the size of the data. As mentioned in the question, this data is a small subset of the available data. We cannot be certain whether this data representats the population well. Further, the larger our sample size is the better our model can predict values.

Another possible limitation in our analysis could be the presence of multicollinearity or high correlation between our predictors. We are not looking at the interaction effect of any of the variables in our model and perhaps that effect could explain more 69% of the variation in our model.