Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 8 (due: 11/18/16)

Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

Ai. Using the summary command, the model I have found is:
log(price) = 13.27 – 1.226*(cityLong Beach) – 0.3118*(citySanta Monica) –
0.6161*(cityWestwood) +0.1744*(bed) – 0.02825*(bath) – 0.0001731*(sqft)

To interpret the intercept, we first transform the equation so that we are not interpreting log(price).

$e^{log(price)} = e^{13.27}$ → price = $579,545.82.

To interpret this, we say that the mean price of empty lots (homes with 0 square footage and thus no bedrooms or bathrooms) is $579,545.82.

What about the city variable? 1/2

Aii. Since Beverly Hills is not in the dataset (R automatically makes the decision alphabetically), the slopes of the city will be in relation to Beverly Hills homes. So, to interpret the slope of cityWestwood, we say that among homes with the same number of bedrooms, bathrooms, and square footage, the price of a home in Westwood, on average, differs by a factor of $e^{-0.6161} = 0.54$ to a home in Beverly Hills.

In transforming the city variables, we obtain the average price of a home with the same number of bedrooms, bathrooms and square footage in the city "…" differing by a factor of "…" to a home in Beverly Hills:

Long Beach: $e^{-1.226} = 0.29$
Santa Monica: $e^{-0.3118} = 0.73$
Westwood: $e^{-0.6161} = 0.54$

Since all of the values calculated above are below a factor of 1, these cities, in comparison to Beverly Hills, cost less on average. Thus, the most expensive city, on average, is Beverly Hills, while the least expensive city, on average, is Long Beach, since its price differs by the smallest factor to Beverly Hills in comparison to the other cities.

Aiii. To interpret the slope, we first transform the equation so that we are not interpreting log(price).

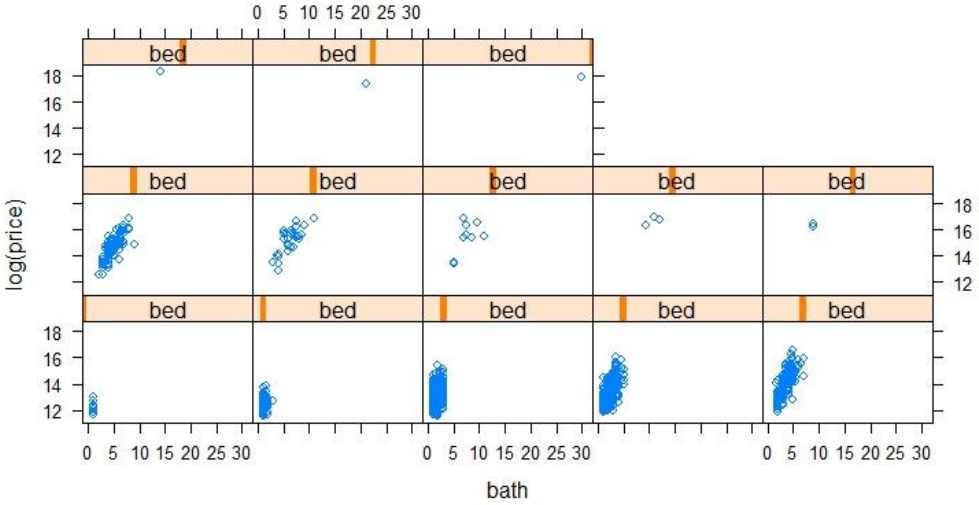
$e^{log(price)} = e^{0.1744}$ → 1.19

To interpret this, we say that among houses within the same city and with the same number of bedrooms and square footage, the mean price of a home increases by a factor of 1.19 per additional bathroom. So, yes, by perspective of value by cost, an additional bathroom does increase monetary value to a house.

Aiv. A high p-value for bathroom means that given that all of the other variables in the model are significant, the bathroom variable is not significant in explaining the total variation in the price of a home.

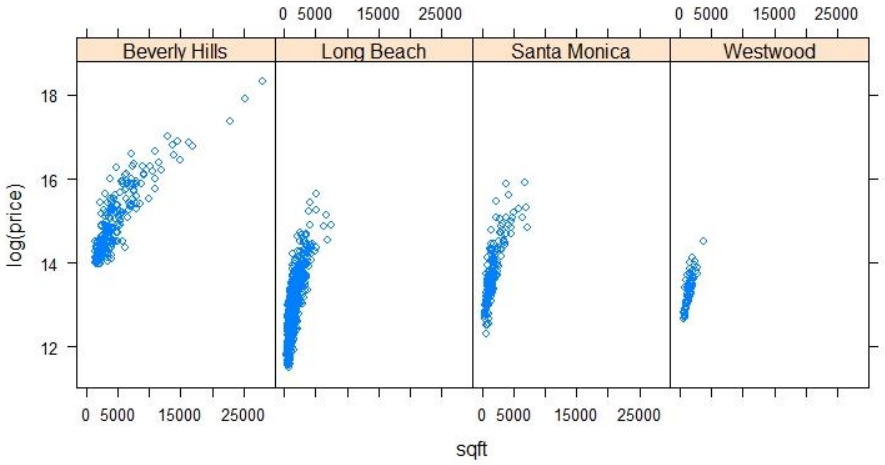
Av. Now that bedrooms is not in the model, bathroom is significant because bedrooms and bathrooms intrinsically have an interacting co-linear relationship in which there is an association between bedrooms and bathrooms. Homes with more bedrooms have more bathrooms as well. Thus, by having bedroom in the model, having bathroom will not be significant in the model since the important information it provides in relation to price of a home is already provided by the first variable.

Avi.



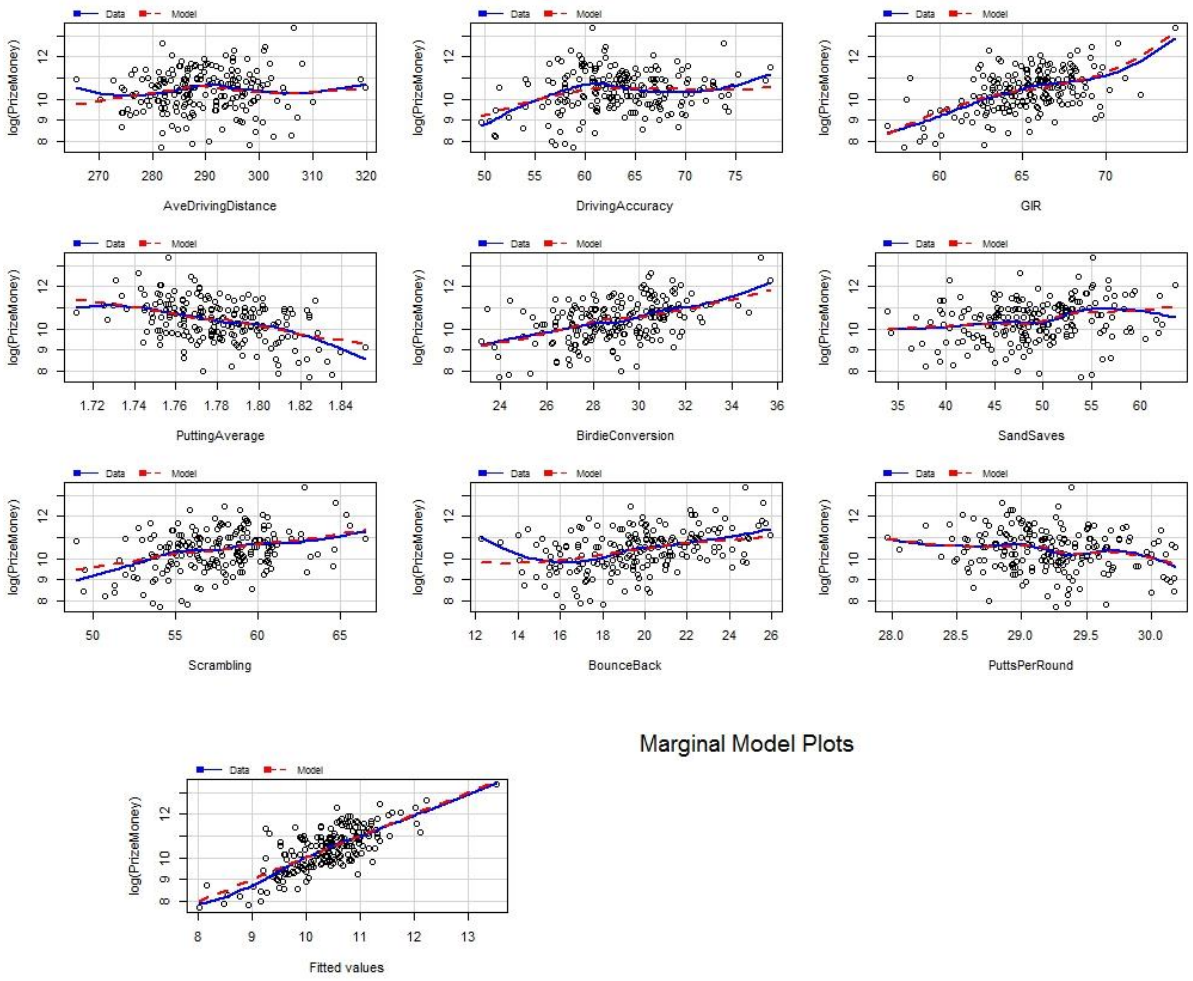
As we can see from the xy-plot, the slope between bath and log(price) at different intervals of beds does not change, suggesting a co-linearity association via interaction of the two variables. In other words, bathroom isn't explaining anything significant that bedroom hasn't done so already.

Avii.



Unfortunately, as seen by the above xy-plot, the relationship between square footage and log(price) varies slightly between different cities since in each graph there is a slight different slope between these two variables. By inspection, the assumption made seems to be in violation, but not by much, so perhaps that is okay. Perhaps if there were a way to quantitatively assess whether the difference in slope means is significant would provide more useful information on whether the initial assumption was violated.

Ba.

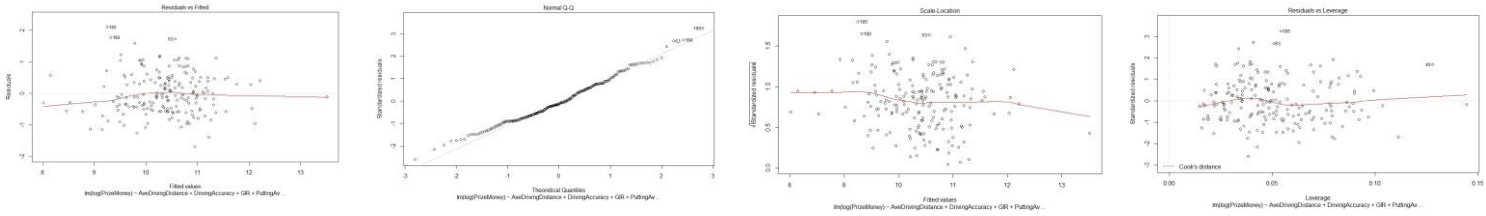


As seen by the above MMP plots after the log transformation, all of the nonparametric estimates as solid blue line curves are almost indistinguishable from the smooth fitted values shown as dashed red lines. Thus, the log transformation on the response variable is a valid model for the data.

Bb. Using the summary function, the full regression model is log(PrizeMoney) = 0.10950 – 0.01317*AveDrivingDistance – 0.01495*DrivingAccuracy + 0.21633*GIR + 2.42834*PuttingAverage + 0.17922*BirdieConversion + 0.01330*SandSaves + 0.04415*Scrambling + 0.01363*BounceBack – 0.41254*PuttsPerRound

You should justify the assumptions for this model or at least give some justification for why this is the best model and no transformations are needed. 1/2

Bc. Using the MMP plots from Ba, we see that from these scatterplots of the data that the log transform works very well on the data since the nonparametric estimates are almost indistinguishable from the smooth fitted values.



From left to right, we have the residual plot, the normal QQ plot, the scale residual plot, and the residuals-leverage plot. The residual plot has a fairly flat trendline indicative of constant variance in the errors as well as a linear relationship since there is no apparent trend in the residual plot. The normal QQ plot is a good plot since the majority of the data points lie on the straight line, indicative of normality of the errors. The scale-location is fairly flat except for a slight dip toward the end, suggesting some sense of constant variance in the errors. Thus, in conclusion, our model in which the response variable is log transformed is valid for the data since it meets our assumptions.

Bd. As seen in the last graph from part Bc, there are no influential points from the last graph since there are no data points outside any contour lines. There are, however, 3 high leverage points, points 40, 63, and 185, in which only point 40 is a good high leverage point since it is the only point contained within the standardized residual interval of [-2, 2]. Thus, the two points that should be further investigated are points 63 and 185.

Be. To remove all variables with insignificant t-values would be foolish because the very backbone of the partial F test is testing significance for a particular variable given that all of the other variables are significant in the model. If a variable is discarded, then the new p-values will have a different meaning than the original model's interpretation, and then you will end up comparing apples to oranges. Basically, even though a variables may be insignificant, it does not imply that the variable won't affect the response variable, especially if the variable actually does have principled reasons for being in the model, which is definitely appropriate in our case, as seen from the MMPs and the close association between all of the variables' nonparametric estimates to the smooth fitted values.