Assignment A1

The assignment follows in the next page.

Exercise 1 (=6 points): here we consider the `bikesharing` dataset found on the Canvas course page. The following assumes that you have already applied the data-formatting via the chron package, as well as defined categorical covariates, as suggested in the previous section. However, you won't make use of several of these variables.

(i) (0 points) Let's start with something basic: produce a figure reporting the boxplots of the number of shared bikes for the different seasons. Also report a figure showing the boxplots of the number of shared bikes at the several hours. Comment on the main findings from these two figures.

(ii) (2 points) Plot the cnt variable (y axis) vs humidity (x axis). You notice there is a lot of variability. Let's simplify things a bit and create a subset of the full dataset: from the `bikesharing` data, create a dataset named e.g. `data_norush` pertaining only "no-rush" data, that is a dataset that *excludes* all observations between hours $[7am, 9am]$ and between [17.00,18.00] and keeps the rest. Use this to create a subset of `data_norush` named `data_norush_spring` that only considers no-rush data pertaining spring (to check that you got this right: `data_norush_spring` should have 3478 rows and 12 columns).

For `data_norush_spring` we wish to apply linear regression to study the number of bikshares as response variable and the humidity level as covariate: however, even if `data_norush_spring` is less noisy than the full data, it seems not appropriate to fit the model directly. Instead, first apply a Box-Cox transformation to obtain a new response variable. *[Tip: only inside `boxcox()` add a 1 to `cnt` to avoid the error you will get from occasionally having 0 counts, and afterwards (when not using `boxcox()`) no need to keep adding a 1 to cnt].* Then fit a linear regression model on such transformed variable with humidity as covariate. Report:

(a) the results from the Box-Cox procedure (the optimal λ and the log-likelihood plot) and hence clearly report the transformed response;

(b) a plot of the transformed response vs humidity;

(c) the plot of the model residuals versus humidity.

Comment the plots (b)-(c) and explain whether you think these look satisfying enough to support the use of a linear regression model.

(iii) (1 points) In (ii) you have fitted a linear model on the transformed responses: denote with $y(\hat{\lambda}) = \hat{\beta}_0 + \hat{\beta}_1 hum + \varepsilon$ the model for the transformed response (we are still talking of `data_norush_spring`). After having applied some simple algebra on the latter model, you can obtain a corresponding model for the untransformed response y, then use this to simulate (untransformed) responses y when humidity is 40. Use $B = 2,000$ of such simulations to produce a 95% prediction interval for the original untransformed y. You are required to interpret the interval.

*[**Note**: just because it will be easier to grade, when simulating data, please place `set.seed(321)` before the `for` loop. So we get the same results.]*

(iv) (1.5 points) We still refer to `data_norush_spring`. We have explored response transformation, and now we still use the Box-Cox transformed response, but explore covariate transformations by selecting a suitable power γ so that the covariate is hum^γ, assuming $\gamma \in [0.5, 2]$. So we allow γ to take real values in the interval, not exclusively integer ones, so we are not doing polynomial regression. You should use I(`hum^gamma`) to specify the covariate within `lm()`. In practice consider values of γ in the sequence of sixteen equispaced values [0.5, 0.6,...,1.9, 2.0]. Then, for each value of γ compute the corresponding $\sqrt{pMSE}$ using training data having size `n=floor(0.8*N)` (this is a function rounding to the closest integer from below) where N is the number of rows of `data_norush_spring`. Plot the values of $\sqrt{pMSE}$ against each γ and discuss which value of γ would you suggest to pick, exclusively by looking at the plot.
*[**Note**: just because it will be easier to grade, please place `set.seed(321)` just before sampling training and testing data. So we get the same results.]*

(v) (0 points) Repeat what you did in (iv) independently for 10 times *[use two nested `for` loops and place `set.seed(321)` before the outermost `for` loop]* and report the corresponding 10 plots of $\sqrt{pMSE}$ vs γ. Do you consistently find the same best value of γ or do you rather identify an interval of possible values? Discuss.

(vi) (1.5 points) Here we have a question similar to the one found in the previous notes for the recap of linear regression. As usual we refer to `data_norush_spring`: assume to have obtained the fitted model $E(y(\hat{\lambda})) = \hat{\beta}_0 + \hat{\beta}_1 hum$, therefore here we take $\gamma = 1$ but we have a transformed response. We wish to verify empirically the theoretical result that the true value of a parameter is included into confidence intervals with some pre-specified probability. Denote with $(\hat{\beta}_0, \hat{\beta}_1)$ the least squares estimates obtained by fitting the model above to data. Let's pretend that $(\hat{\beta}_0, \hat{\beta}_1)$ are the *true* parameter values (β_0^*, β_1^*) that really generated data and therefore we write $(\beta_0^*, \beta_1^*) \equiv (\hat{\beta}_0, \hat{\beta}_1)$. Write an R code using `for` loops to produce 2000 sets of parameter estimates and confidence intervals according to the following reasoning:

1. Plug the (β_0^*, β_1^*) in the following linear model and use it to produce simulated observations $\mathrm{y}(\hat{\lambda})_i^{new} = \beta_0^* + \beta_1^* \mathrm{hum}_i + \epsilon_i^{new}$, where the hum_i are the same values as in the given dataset, and where you have simulated the n values of the $\epsilon_i^{new} \sim N(0, s^2)$ by using the same s as obtained when fitting the real data. So now we have a simulated dataset $\mathcal{D}_1 = (\mathrm{hum}_i, \mathrm{y}(\hat{\lambda})_i^{new})_{i=1,\dots,n}$. Fit $\mathcal{D}_1$ and denote the parameter estimates with $(\hat{\beta}_0^{(1)}, \hat{\beta}_1^{(1)})$.
2. Repeat the procedure: use again the original (β_0^*, β_1^*) to produce new simulated observations $y(\hat{\lambda})_i^{new} = \beta_0^* + \beta_1^* \mathrm{hum}_i + \epsilon_i^{new}$, where the ϵ_i^{new} **are generated anew via `rnorm`** (they are not the same ϵ_i^{new} as in the previous step). Call the new dataset $\mathcal{D}_2 = (\mathrm{hum}_i, \mathrm{y}(\hat{\lambda})_i^{new})_{i=1,\dots,n}$. Fit $\mathcal{D}_2$ via linear regression and denote the parameter estimates $(\hat{\beta}_0^{(2)}, \hat{\beta}_1^{(2)})$.

Repeat the two steps above until you have obtained 2000 sets of estimates $(\hat{\beta}_0^{(j)}, \hat{\beta}_1^{(j)})$, $j = 1, \dots, 2000$ *[place `set.seed(321)` before the `for` loop]*. Construct confidence intervals from each set of estimates using $1-\alpha = 0.80$, so in the end you have obtained 2000 confidence intervals for β_0^* and 2000 intervals for β_1^*. At this point, you can finally compute the proportion of intervals that include the original value β_0^*. Then do the same for β_1^*. Show that both proportions are very close to $1 - \alpha$, as expected from the theory.

Variables:

Here follow the variables definition. You won't need to use most of them.

- timestamp: day of the year and hour
- "cnt" - the number of new bike shares in the considered timestamp
- "t1" - real temperature in Celsius
- "t2" - perceived temperature in Celsius
- "hum" - humidity in percentage
- "wind_speed" - wind speed in km/h
- "weather_code" (see below for a description)
- "is_holiday" - 1 means holiday / 0 for non holiday
- "is_weekend" - 1 if the day is weekend
- "season" - meteorological seasons: 0-spring ; 1-summer; 2-fall; 3-winter.

"weather_code": 1 = Clear ; 2 = scattered clouds / few clouds; 3 = Broken clouds; 4 = Cloudy; 7 = Rain; 10 = rain with thunderstorm; 26 = snowfall; 94 = Freezing Fog.