Assignment 2

VERSION: 2018-10-25 · 23:38:23

Instructions:

- Due Tuesday, October 30 at 11:59pm.
- Assignments must be submitted online as a single PDF file. PDFs can be created in multiple ways:
  1. By scanning handwritten assignments (Illegible scans will not be graded.)
  2. By converting proprietary formats such as DOCX to PDF via an online tool such as this one.
  3. By integrating text and **R** code + plots directly with **R** Markdown.
- Online submission will be via LEARN or Crowdmark. Further instructions for submission will be given shortly.
- You may work on this assignment in groups and/or use any reference material you find online. However, you must include the names of all collaborators and list all external sources, otherwise this is considered plagiarism.
- Each student must turn in their own assignment. Exact replicas will obtain a grade of zero.
- Include all **R** code. Proper programming habits and techniques (labelled figure axes, efficient coding, etc.) are expected for full marks. Uncommented code will not be graded.

The file **airfare.csv** contains data on $n = 1000$ commerical flights in the US in 2002. The variables in the dataset are:

- `dep_city`: City of flight departure.
- `arr_city`: City of flight arrival.
- `fare`: Average fare price (USD).
- `dist`: Distance of flight (miles).
- `pass`: Average weekly number of passengers per flight.
- `lead_aline`: Name of market leading airline for that flight.
- `lead_share`: Market share of leading airline (×100%).
- `lead_fare`: Average fare price for leading airline (USD).
- `low_aline`: Name of lowest price airline for that flight.
- `low_share`: Market share of lowest price airline (×100%).
- `low_fare`: Average fare price for lowest price airline.

Q1.

(a) Load the dataset into a variable called `air` and produce pair plots for each of the continuous variables in the dataset (all variables except `dep_city`, `arr_city`, `lead_aline`, `low_aline`). Why is there a distinct diagonal line in the scatterplot of `lead_share` vs `low_share`?

(b) Use **R** to fit and display the summary of the linear regression model

$$E[\texttt{fare} \,|\, \texttt{dist}, \texttt{pass}, \texttt{lead\_fare}] = \beta_0 + \beta_1 \texttt{dist} + \beta_2 \texttt{pass} + \beta_3 \texttt{lead\_fare} + \beta_4 \texttt{lead\_fare}^2.$$

Based on this calculation, is there significant evidence of a nonlinear effect of `lead_fare` on average `fare`, in the presence of `dist` and `pass`? Justify your answer.

(c) Based on the model fit in Q1(b), estimate the difference in expected fare between Flight 1 and Flight 2, where Flight 1 has a leading airline fare of 160USD, Flight 2 has a leading airline fare of 120USD, and Flight 1 has 50 more weekly passengers than Flight 2, and both flights are of the same distance. Use the **R** function `predict` to obtain full marks.

Q2.

(a) Create a categorical variable (i.e., a `factor` variable in **R**) called `lead_rate`, indicating for each of the $n = 1000$ flights whether the average fare per mile of the corresponding *leading* airline[1] is low (less than .14$/mile) medium (.14$ to .21$/mile) or high (more than .21$/mile). You can check your calculations by running the following code:

```
# first few values of lead_rate
head(lead_rate)
```

```
[1] high med  high high high low
Levels: low med high
```

```
# standard deviation of fare price by lead_rate group
tapply(air$fare, lead_rate, sd) # see ?tapply for details
```

```
     low      med     high
41.62865 59.76481 59.98990
```

Hint: See the `cut` function in **R**.

(b) Consider a linear regression model of $E[\texttt{fare} \,|\, \texttt{pass}, \texttt{lead\_rate}]$ for which there is a linear relationship between `pass` and expected `fare` with different intercept and different slope for each level of `lead_rate`.

i. In order to fit this model with multiple linear regression, we must be able to write

$$E[\texttt{fare} \,|\, \texttt{pass} = s, \texttt{lead\_rate} = k] = x'\beta,$$

where $s > 0$ and $k \in \{\texttt{L}, \texttt{M}, \texttt{H}\}$. If x is a (mathematical) covariate vector for which the first element is 1, how are the rest of the elements of x determined as a function of s and k? In other words, write down the covariate vector x corresponding to any given s and k.

ii. Use **R** to fit and display the summary of the linear model defined above.

(c) Based on the model fit in Q2(b), use an F-test to calculate a p-value against the null hypothesis that there is no interaction between `pass` and `lead_rate`. Calculate the p-value in two ways: (i) using sum-of-square residuals from full and reduced models and (ii) a built-in **R** function.

Q3.

(a) Use **R** to fit and display the summary of the log-additive model

$$\log(\texttt{fare}_i) = \beta_0 + \beta_1 \log(\texttt{dist}_i) + \beta_2 \log(\texttt{pass}_i) + \eta_i, \qquad \eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

[1] I.e., not the avg fare/mile for the flight itself.

(b) Consider a multiplicative model of the form

$$\texttt{fare}_i = \gamma_0 \texttt{dist}_i^{\gamma_1} \texttt{pass}_i^{\gamma_2} \cdot \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} \log\text{-}\mathcal{N}(-\tfrac{1}{2}\sigma^2, \sigma^2),$$

such that the ε_i are iid log-Normal error terms with $E[\varepsilon_i] = 1$. Calculate the MLE of $\gamma = (\gamma_0, \gamma_1, \gamma_2)$.

Hint: Take logs on both sides of the multiplicative error model, and let $\eta_i = \log \varepsilon_i + \frac{1}{2}\sigma^2$. Moreover, take for granted the following results:

1. If you have a function g such that $\gamma = g(\beta, \sigma)$, then $\hat{\gamma}_{\text{ML}} = g(\hat{\beta}_{\text{ML}}, \hat{\sigma}_{\text{ML}})$.
2. The MLE of σ is given by $\hat{\sigma}_{\text{ML}} = \sqrt{e'e/n}$. That is, it divides the sum-of-square residuals by n instead of $n - p$.