Practical challenge 3 hint sheet

Model crafting

Foundations of Data Science—DJW—2018/2019

These questions are optional. They are not intended for supervision, and they are not examinable material. But it is HIGHLY RECOMMENDED that you attempt question 1.

If you want guidance, you can (i) ask your fellow students for help in the Moodle forum, (ii) come to the practical classes in the second half of term and ask demonstrators, (iii) wait until model solutions are released, after the practical classes.

Question 1. Find a 95% confidence interval for the rate of temperature increase at the Cambridge weather station, using the model

$$\mathsf{temp} \approx \alpha + \beta_1 \sin(2\pi \mathsf{t}) + \beta_2 \cos(2\pi \mathsf{t}) + \gamma \mathsf{t}. \tag{1}$$

```
url = 'https://teachingfiles.blob.core.windows.net/founds/climate.csv'
climate = pandas.read_csv(url)
df = climate.loc[climate.station=='Cambridge']
t = df.yyyy + (df.mm-1)/12
temp = (df.tmin + df.tmax)/2
```

We'll assume the standard linear regression model,

$$\mathit{temp} \sim \text{Normal}(\alpha + \beta_1 \sin(2\pi t) + \beta_2 \cos(2\pi t) + \gamma t,\, \sigma^2).$$

First, find maximum likelihood estimators $\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\gamma}$. As explained in lecture notes page 76 (also slides for lecture 11), this is equivalent to finding least squares estimates. Use code like that on page 63 of lecture notes (also slides for lecture 10).

Next, find the maximum likelihood estimator $\hat{\sigma}$. The method is described in lecture notes page 76 (also slides for lecture 11).

We want a confidence interval for γ. This is very similar to example sheet 2 question 4, and the procedure is as described in lecture notes page 39. First we have to invent the shape of the confidence interval, and a good starting point is something surrounding the maximum likelihood estimator—let's pick the confidence interval $[\hat{\gamma} - \delta_1, \hat{\gamma} + \delta_2]$ where δ_1 and δ_2 are given. The probability that this interval is correct is

$$\mathbb{P}\big(\gamma \in [\hat{\gamma} - \delta_1, \hat{\gamma} + \delta_2]\big)$$

and we can approximate this using bootstrap resampling: it is approximately

$$\mathbb{P}\big(\hat{\gamma}(x) \in [\hat{\gamma}(X^*) - \delta_1, \hat{\gamma}(X^*) + \delta_2]\big). \tag{$*$}$$

Here I have written $\hat{\gamma}(x)$ to mean $\hat{\gamma}$ computed from the actual observed dataset, and $\hat{\gamma}(X^*)$ to mean $\hat{\gamma}$ computed from a resampled version of the dataset. We've already computed $\hat{\gamma}(x)$. To create a resampled version of the dataset using parametric resampling, follow the procedure and sample code laid out in the slides for lecture 11.

By generating 10,000 resampled versions of the dataset and using Monte Carlo integration, we can estimate the confidence probability, for a given δ_1 and δ_2.

Finally, we have to tune δ_1 and δ_2 so that the confidence probability is 0.95. You could do this by brute force, repeatedly tweaking δ_1 and δ_2 until the probability comes out how you want it. There's a slicker way, which you have seen if you answered exercise sheet 2 question 4(c). Rewrite $(*)$ as

$$\mathbb{P}\Big(-\delta_1 \leq \hat{\gamma}(x) - \hat{\gamma}(X^*) \leq \delta_2\Big)$$

and choose δ_1 and δ_2 from appropriate quantiles of $(\hat{\gamma}(x) - \hat{\gamma}(X^*))$ using `np.quantile`.

Question 2. In this question, you will investigate three refinements of model (1). Answer each part separately; you don't need to combine your findings between the parts. This is a question about designing features for linear models; you don't need to find confidence intervals or conduct hypothesis tests, or devise fancy non-linear probabilistic models.

(a) The Met Office says that it's not just a question of temperature increase, but that the temperatures are getting more extreme. Fit a model to Cambridge temperatures, in which the summer high and the winter low are both changing linearly, but at different rates. What are these rates?
`https://www.bbc.co.uk/news/science-environment-46064266`

(b) Perhaps the annual cycle isn't a perfect sin curve. Fit a model to Cambridge temperatures, in which the annual cycle is an arbitrary shape, using one-hot coding as in page 66 of lecture notes. Plot a bar chart of the annual cycle, one bar per month, and superimpose the fitted sin curve from model (1).

(c) For every station separately, using temperature data from 1960 onwards, fit the model (1). Plot a scatter plot of the α coefficient as a function of latitude, with one dot per station. Plot three more scatter plots, to show how α and γ coefficients depend on latitude and altitude. Propose and fit a single model for the entire dataset (from 1960 onwards), incorporating your findings about α and γ.