Exercise Sheet 6
Bayesian Linear Regression

Solutions. You do not need to submit your exercise solutions to us. Of course you are free to ask questions during tutorial hours! Exercise solutions will be published on Moodle at the same time with the exercise sheet. Use solutions responsibly, otherwise you risk not being able to solve the exam to a satisfactory level.

Exercise 1 Modeling Credit Data with Simple Linear Regression

In this exercise, you will perform the basic steps of Bayesian linear regression. To this end, you will work with the credit dataset that you can load with Pandas from `credit_data.csv`.

a) You would like to predict the balance (current credit card debt in USD) of a person using only the limit set by the bank (in USD) as a predictor. Write down the structure of a linear regression model that achieves this using the mathematical notation from the lecture (priors, likelihood, etc.). Thereby you do not have to specify numeric values for the prior distribution parameters (but you may think about their sizes if you want).

b) Fit a linear regression model with bambi, letting it automatically select default priors for you. Extract the model that it created both through `.backend.model` (mathematical notation) and through `.graph()` (graphical model). Does the model align with your own model in a)? If no, what are the differences?

c) Return a summary of the posterior regression coefficient distributions using `pm.summary()` and plot them visually using `pm.plot_posterior()` (for both using 95% credible intervals). Write down the mean regression model as a formula. How can the coefficients be interpreted in terms of the effect of the associated predictors on credit card balance?

d) Create a plot with data and regression line using `bmb.interpret.plot_predictions()`. Use a 95% HDI for the regression line uncertainty. Using the plot, give an interpretation why the recovered intercept is negative. Also, visually judge aleatoric and epistemic uncertainty.

Hint: Whether `bmb.interpret.plot_predictions()` plots an HDI of the posterior regression line distribution or of the predictive distribution can be controlled with the `pps` parameter.

e) How good is your model? Compute the Bayesian predictive versions of RMSE, MAE and R^2 score. On this occasion, make sure that you understand the Bayesian interpretation of R^2 proposed in the lecture.

f) You try to predict the balance for a customer with a limit of 3000 USD. Compute the predictive distribution and summarize and visualize it (use a 90% HDI). Does the recovered distribution make sense to you? Give also a summary of its aleatoric and epistemic uncertainty.

Hint: For epistemic and aleatoric uncertainty use `.predict()` on your model and pass your data point to get the predictive distribution conditioned on it. The epistemic uncertainty σ_e^2 is the variance of the predicted means μ and the predictive uncertainty σ_p^2 the variance of the predictions for `Balance`. You may recover the aleatoric uncertainty using $\sigma_p^2 = \sigma_e^2 + \sigma_a^2$.

Exercise 2 Alternative Likelihoods

You are a bit frustrated by the negative balances predicted by your model in Exercise 1. You have the brilliant idea to change the likelihood of your model to produce only positive values. Instead of $y|\beta_0, \beta_1, \sigma \sim N(\beta_0 + \beta_1 x, \sigma^2)$, you choose the likelihood

$$y \sim \text{TruncatedNormal}(\mu = \beta_0 + \beta_1 x, \sigma^2, \text{lower} = 0, \text{upper} = 2000),$$

since the maximum observed balance in your data is 2000.

a) Since the TruncatedNormal likelihood is not available in Bambi, implement your model in PyMC (slight modification of code used in lecture) and run the posterior simulation.

b) Compute the mean values for β_0 and β_1 and overplot a scatterplot of the observed data with the mean regression line. Interpret your results. Would you use this model in favour of the model in Exercise 1?

Exercise 3 Multiple Linear Regression on Insurance Data

In this exercise, we are going to look at a sample insurance dataset that you can find in `insurance.csv`. The goal is to use the set of predictors provided in the data to predict the yearly charges in USD for clients to get an estimate of the overall reserve budget.

a) Load the dataset with Pandas and get an overview over the nature of the different predictors that are provided. For a first analysis, we will use the predictors `age`, `bmi` and `smoker` for an estimation of the total charges. To this end, perform an exploratory data analysis and plot `age` and `bmi` against `charges` in two separate plots. Since `smoker` is a categorical variable, use it to color your observations (e.g. in seaborn with the `hue` parameter).

b) Fit a multiple Bayesian linear regression model with Bambi that predicts `charges` from `age`, `bmi` and `smoker`. Summarize the 95% HDIs for all involved parameters and give an interpretation for the role of each parameter in the prediction. Are all parameters significant in the Bayesian sense? Compute Bayesian RMSE and R^2 to quantify the performance of the model.

c) Fit a model with all the available predictors. Do all predictors significantly contribute to the predictions of the model? Compute Bayesian RMSE and R^2 to quantify the performance of the model. Which model would you rather use, the previous model including only `age`, `bmi` and `smoker` as predictors or the more complex current one?

d) Could your more complex model get better with more data? Compute epistemic and aleatoric variance of the predictive distribution on your dataset using `.predict()` and setting `data = insurance_data`.

Hint: The rest is like in Exercise 1.

e) It is often said that there is an aggravating effect when an individual is both overweight and a smoker, with even bigger impacts than just the sum of the two on personal health. You want to verify this on the given dataset by adding an interaction term between `bmi` and `smoker`. Fit a model and report and interpret your results with `pm.summary()`. Did your prediction performance in terms of R^2 increase?

f) Go back to your initial plots in a). Is evidence for the significant interaction between overweight and smoking visible in one of the plots?

Exercise 4 Multiple Linear Regression on Advertisement Data

In this exercise you will perform multiple linear regression on the advertisement dataset introduced by the famous book *Introduction to Statistical Learning*. It contains data of an advertisement campaign with the budget used for TV, radio and newspaper ads in 1000 USD against the number of units that were sold (in thousands).

a) Load the dataset with Pandas and get an overview over the different predictors that are provided. Create some exploratory plots to guess which variables will probably be related to the sales variable.

b) Fit a multiple linear regression model including all the available predictors. Which contribute significantly to the model? What is your interpretation of the coefficients?

c) Compute predictive RMSE and R^2 score. Is this a good model? Give an interpretation in particular of RMSE.

d) Refit your model without `newspaper`. Do RMSE and R^2 change significantly?

e) There might be synergy or redundancy between TV and radio advertisement budget: either viewers who get the ad through both channels are even more inclined to buy the product (synergy) or it is enough that they hear it through one channel and the other one is an unnecessary repetition (redundancy), or something in-between of course.

   Introduce an interaction between `TV` and `radio` and interpret your results. Is the interaction term significant? Can we observe a synergy or a redundancy? How does the predictive performance change?

f) You might have noticed in the exploratory analysis performed in part a) that you are rather in a heteroscedastic than a homescedastic setting, where variance is not constant anymore but depends on the value of the predictor.

   To alleviate this, model the *logarithm* of `sales` instead of just `sales`. Show using scatter plots that this is a reasonable approach and compare the performance of your resulting model (including interactions) with your model from d).

Exercise 5 Robust Linear Regression

In this exercise, you will model body fat from BMI (data in `bodyfat.csv`) as done in the lecture, however including the real outliers in the dataset.

a) Load the dataset with Pandas, create a scatter plot of `BMI` against `BodyFat` and convince your self of the presence of outliers that might influence the coefficients of linear models.

b) Fit a simple linear regression model with normal likelihood with Bambi and assess its predictive performance in terms of RMSE and R^2. In addition, plot the regression model with `bmb.interpret.plot_predictions()`.

c) As an alternative, fit a robust model that uses a Student's t distribution as likelihood instead of a normal distribution (see exercises in week 5 for insights into Student's t distribution). You may do this with Bambi by choosing `family="t"` instead of `family="gaussian"`. Do the non-robust and the robust model differ significantly? Compare RMSE and R^2 and plot both models in the same plot with `bmb.interpret.plot_predictions()`. Which model would you use at the end?

d) As an addition model selection tool, you consider computing ELPDs for both models and to compare them (see week 5). Compute the log-likelihood for both model traces with `pm.compute_log_likelihood()` (in the context of the underlying PyMC model that can be accessed through `.backend.model`) and then use `pm.compare()` and `pm.plot_compare()`. Which do you trust more in this case, RMSE and R^2 or ELPD?

e) Just for fun: Fit a **robust** multiple linear regression model with all variables (except `Density` that is used to compute `BodyFat`) and check whether it is an improvement on the models in b) and c) in terms of RMSE and R^2.

Check the posterior distribution of ν. Would you at the end rather use a normal or a Student's t likelihood?

Exercise 6 Leaking Satellite Fuel Tank

You work for the European Space Agency in the maintenance team for a satellite. Next to scientific data, the satellite sends back a host of maintenance data, among other things the fuel level of its thrusters that are needed to keep it on the designed orbit. Since sending satellite data down to Earth is expensive (antenna network distributed over the Earth) and a large part of the bandwidth is reserved for scientific data and redundancy, you get only one fuel level measurement per day.

Three days ago you have discovered that the fuel tank level has decreased by a larger amount than usual. You suspect that the satellite has been hit by space debris that has produced a whole in the fuel tank. The goal of this exercise is to compute the remaining time in the satellite mission, before the thrusters fail by lack of fuel and the satellite will leave its carefully calibrated orbit.

a) Load the dataset from `fuel_data.csv`. It contains the time t in days (starting from zero), the tank level in milligrams and the suspected state of the fuel tank (nominal / leaking). Plot the tank level against the time t and verify that the point in time where the tank level decreases more quickly is visible (at $t = 19$).

b) First you fit a model to the nominal data points with the yet intact fuel tank to guess the regular rate r of fuel consumption in mg. To this end, fit a linear regression model to $t \in [0, 19]$ and guess a 95% HDI of r.

   Because there is prior information available from your engineering team, you decide to choose your own priors instead of letting Bambi choose weak default priors:

   - Bambi chooses for intercept not the value at $t = 0$, but the value in the **center of the data distribution**. In this case, the center falls between 9 and 10, so you choose as intercept $y_0 = \frac{1}{2}(\text{tank\_level}[9] + \text{tank\_level}[10])$. The engineers in your project team say that there is an uncertainty of $\sigma = 0.2$ mg in the tank level.
   - For fuel rate, your engineering team says that the thrusters consume at around 40μg per day, possibly also something between 20μg and 60μg. Use PreliZ to devise an appropriate normal prior. **Hint:** Make sure that the rate is negative (decrease).

   Pass your priors to bambi and fit a model to the first 20 data points ($t \leq 19$). Visualize your model predictions with `bmb.interpret.plot_predictions()` with `pps=True` and estimate a 95% HDI for the regular fuel consumption rate r.

c) Now fit a model to the data points where you suspect the leak ($t \geq 19$). This time let Bambi choose default priors for you. Visualize your model predictions again with `bmb.interpret.plot_predictions()` with `pps=True` and estimate a 95% HDI for the leaking fuel consumption rate r^*.

d) Mission control asks you how long they can expect the satellite mission to continue, i.e. how many days the mission can last until the fuel tank is empty. Use the 95% HDI on r^* to give an estimate.