TA2 MidTerm Review and Project Guidance

Contents

Review of Frequentist Concepts that Carryover into Bayesian Analysis

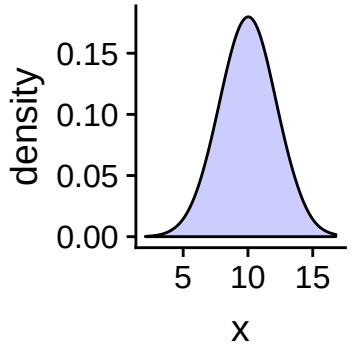
As we transition into Bayesian Analysis, let's review a few concepts that will be important to fully understand the material ahead. Particularly, estimation of parameters and confidence intervals.

Parameters in Distributions

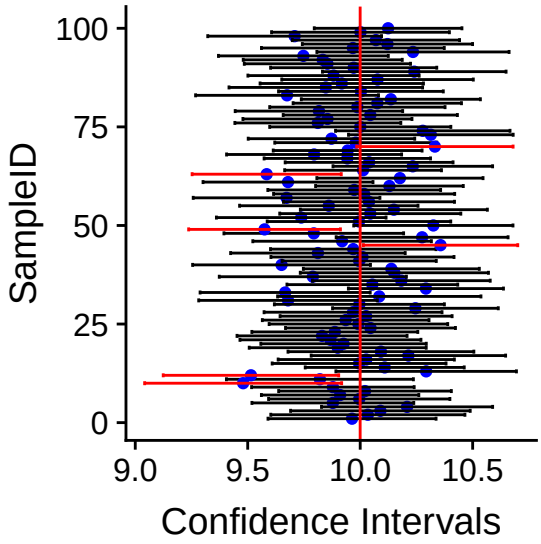
Recall from DA1, that we estimated model parameters *(keep in mind, a distribution is a model)*, and we computed confidence intervals for those parameters. For example, when we estimated the mean of a population based on a sampling process, and we computed the standard error as:

$$\frac{\sigma}{\sqrt{n}}$$

and to find a 95% **level** of that estimate, we used *qnorm* to calculate a z-score, and multiplied that by the standard error *(i.e., the standard deviation of the sampling distribution)* to give us the **confidence interval**. For example, given a population of 10,000 with a mean of 10 and a standard deviation of 2:

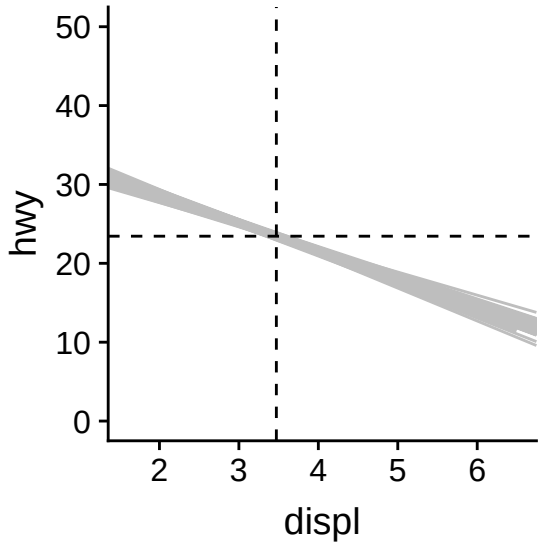


From that population, we can draw 100 random samples of 100 ea, and compute the confidence **interval** around our estimate of the mean parameter, using an alpha confidence **level** of 95%. And not surprisingly, we'll get ~95 samples that include the real *(and unknown)* mean.



Parameters in Simple Linear Models

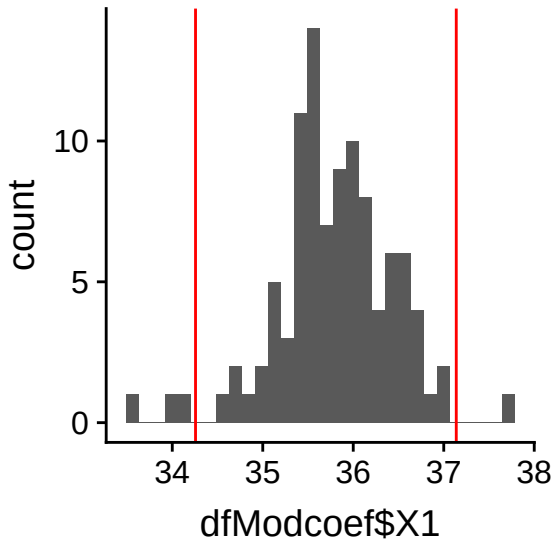
Of course, the mean of a distribution is just one parameter we might be interested in. We can apply the same confidence intervals *(with some modification of our formula)* to other models and parameters: for example, a regression intercept. Again, let's draw 100 random samples of 100 ea. *(from mpg data)* and create linear models for each sample:



And if we plot these 100 models, we'll see a fairly tight grouping of parameters *(intercept and slope)*. And we can compute the confidence interval for the intercept as:

$$\pm 2 * \sqrt{\frac{SSE}{df}}$$

and not surprisingly, ~95% of the estimated sample intercepts fall within that interval:



proving out this 95% CI may not happen every time we draw **samples**, but it will if we do it **frequently** enough :) *(and assuming the data are normally distributed - a BIG assumption)*

Parameters in Genearlized Linear Models

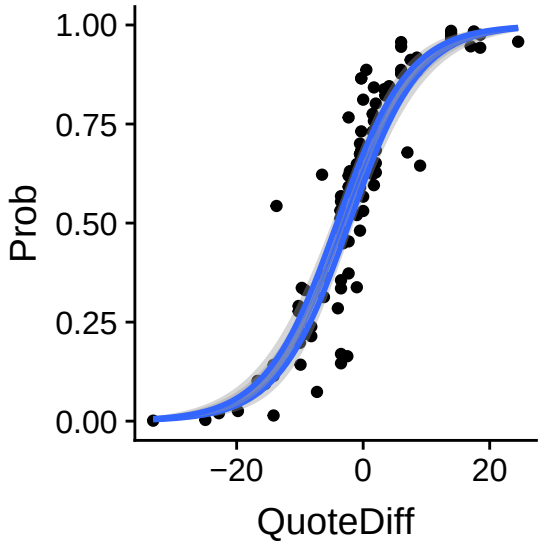
With Generalized Linear Models, we still review parameter confidence intervals, but now there are some now complicating issues:

1. The logit link function adds a little complexity: the dependent variable is transformed by the logit function, and so is SE.
2. GLM uses likelihood to determine parameters, and many glm link functions will transform the variables to non-linear space
3. Recall from DA1 that SE is determined differently with multivariate models.

Fortunately, glm provides SE values, so you can use something like:

- test$lcl <- test$Prob - testPred$se.fit
- test$ucl <- test$Prob + testPred$se.fit

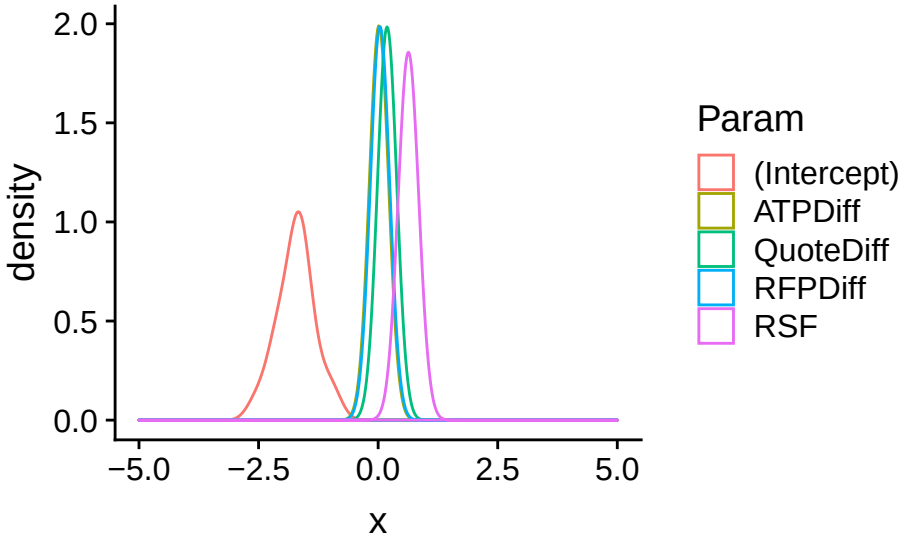
To build you confidence intervals, as shown below:



GLM provides a function to compute parameter confidence intervals:

- confint(glm.fit) # this uses profile likelihood to compute CIs
- confint.default(glm.fit) # this uses likelihood to compute Wald CIs - we'll use Wald here *(more traditional symmetric method)*

Using these CIs, I **simulated** the parameter distributions shown below *(see R file for method)*:



These intervals give us a metric for quantifying our *confidence* in the model :).

Transitioning into Bayesian Analysis

First, there are "philosophical"" differences between frequentist and Bayesian approaches:

- For **frequentists**, the parameters are fixed but unknown and the data are random. Again, review how we estimated the parameters *(the mean of the distribution, the intercpept of the linear model, all of the parameters of the generalized linear model - all of these are fixed, we're just trying to figure out what*

they are by taking samples of the population). Frequentist processes fit well when the population size is unknown *(e.g., a survey of likely buyers, a trail for new drug)* and we have no **experience** with the data. Think of it this way: **if a frequentist model is not working, the solution is to redesign the sampling process**.

- For **Bayesians**, the parameters are random and the data are fixed. Think back to our study of *Maximum Likelihood* estimation, when we held the data fixed and varied the parameters, measuring the likelihood of each parameter value *(choosing the most likely value).* We could do that because we held the data fixed *(or else it would have been a total cluster :))* Bayesian modeling produces analyses like the glm parameter analysis above *(which might be the reason we're doing this project :).* Think of it this way: **If a Bayesian model is not working, the solution is to redesign the modeling process**.

Bayesian modeling is usually a better approach for **business transaction** environments. There are many reasons, including:

1. **In transaction environments, the data are usually fixed** *(and the models are complex).* So, the challenge is to fit a model that supports inference *(parametric).* Sampling is neither difficult nor expensive.
2. **Inference** *(i.e., projection)* **is integral to Bayesian modeling** because inference is based on parameters, and **parameters** *(rather than data)* are the **outcome of Bayesian modeling** *(e.g., we sample parameters, not the data).* We can then use these parameters to generate data *(simulation)* for inference *(projection)*
3. **Transaction environments are hierarchical** *(or multi-level)*, and the effects of each level influences the other levels. Bayesian modeling adapts naturally to multi-level data and gives us a way to quantify and vary the effects within *and* between all the levels. This is so essential to the structure and characteristics of transaction environments that many believe that single level *(fully pooled)* models are fundamentally inadequate *(frequentist models for multi-level, mixed effects scenarios exist, but they are limited and difficult in transacation envionments)*
4. **Bayesian models have fine control over Generalization**. Generalization can be tuned by parameter, dimension and level. Overfitting is a big problem in modeling transaction environments, and this level control over generalization is often necessary for relevancy.
5. In Business environments, we usually have some experience with the data, and that **experience is quantified and integrated into Bayesian models**. Often times, experience is more informative than the likelihood data *(especially in dynamic transacation environments)*, and these beliefs can often encapsulate characteristics where models become intractable *(this ventures into the domain of "professional judgement", but with added data and parameters to validate beliefs and support infererence)*
6. From an Operational perspective, transaction environments are high volume and dynamic, and models can become irrelevant quickly. Bayesian models can be architected so that parameters can be constantly passed between analysis / training models and operational models *(BIG deal in fintech, and becoming a big deal everywhere else)*

Hopefully, these points motivate you to begin Section II. But before we do, let's finish the GLM project:

Project Outline

You have been assigned a project to build and compare two-class models in AML and R *(using GLM with logistic regression).* Your report should be in R Markdown *(RMD)*, with a minimum of the following sections:

1. Description of your modeling process in AML with analysis of results. Model selection is your choice. Model development / training should include resampling and tuning *(where appropriate - your judgment).* For 2 class models, an overall accuracy of 80% is a good target, but cases where a minority classes or other factors present challenges will be considered. The RMD file should include a narrative with

references to the model in AML *(be sure and save your model to the class workspace)*. Also, any relevant analysis plots to support your narrative should be in the RMD.

2. GLM model with GLM predictions and **validation using a Logistic Regression equation** *(there should be no difference in the predictions - see exercise V2)*. Your Analysis should summarize the modeling process with findings including Comparison/Contrasting to the AML model.
3. Analysis of GLM parameters with Confidence Intervals. Just use the same methodology as presented in this document and the updated file *(DA_LogReg_Exercise_V2)*.
4. Analysis of other factors and processes *(e.g., SMOTE, cross-sampling, variable information analysis / dimension reduction)* and how these processes influenced the final model. This should include a discussion of your decisions and outcomes during the modeling process.

Presentation. PPTX decks are optional - you can just walk through the paper if you like. Presentations will take place in class 2/28.