Exercises for Lecture 8: Binary regression

EBA3500 Data analysis with programming

Jonas Moss

1 Maximum likelihood

Recall the definition of the maximum likelihood estimator. If $X_1, X_2, \dots, X_n \sim (\theta)$ are iid for some unknown θ, the maximum likelihood estimator is

$$\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{n} \log f(X_i; \theta),$$

where θ can be either a scalar or a vector.

1.1 Bernoulli distribution

Let $X_1, ..., X_n$ be independent $0-1$ variables with success probability p, i.e., Bernoulli variables. Use differentiation to show that the maximum likelihood estimator of p is $\hat{p} = \overline{X}$, the mean of $X_1, ..., X_n$. (*Hint*: Find the expression for the likelihood in the lecture slides or online. Google "differentiation maximization" or something if you have forgotten how to optimize using differentiation.)

1.2 Normal distribution

1.2.1 Rederive the maximum likelihood estimator of μ

I showed in class that $\hat{\mu}_{ML} = \overline{x}$, the empirical mean, when the observations x are sampled from a normal distribution $N(\mu, \sigma)$. Try to show this yourself!

1.2.2 Maximum likelihood estimator of σ

Show that the maximum likelihood estimator of σ is $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$. (*Hint:* You must use the expression for the maximum likelihood of μ that you have already derived to derive the maximum likelihood estimator of σ^2. Then use the invariance principle.)

1.3 The exponential distribution

1.3.1 Maximum likelihood estimator of λ

Let $X_1, X_2, \ldots, X_n$ be iid from an exponential distribution with density $f(x;\lambda) = \lambda \exp(-\lambda x)$. Calculate the maximum likelihood estimator of λ. What is the maximum likelihood estimator of $\sin(\lambda)$?

1.3.2 Asymptotics of $\hat{\lambda}$

Make a Python function that samples n exponentials with parameter $\lambda = 2$ and calculates its maximum likelihood estimator. Make a histogram out of $N = 10,000$ samples from this function when $n = 100$. What does the histogram look like?

1.3.3 Rescaling the histogram

Rescale the histogram, i.e., use $\sqrt{n}(\hat{\lambda}_{ML} - \lambda)$, and display it for $n = 100, 1000, 10000$. What do you see?

1.3.4 The asymptotic variance (i)

Use the function in (b) to estimate the *asymptotic variance* $n \operatorname{Var}(\hat{\theta}_{ML})$ as n varies. What do you see?

1.3.5 The asymptotic variance (ii)

1. Calculate $\frac{\partial^2}{\partial \lambda^2} \log f(X;\lambda)$, i.e., the second derivative of the log-density of X with respect to λ. (*Hint*: Take the logarith first, then differentiate with respect to λ twice!)
2. Calculate the expectation of the expression you just found, $I(\lambda) = E\left[\frac{\partial^2}{\partial \lambda^2} \log f(X;\lambda)\right]$
3. What is $1/I(\lambda)$ when evaluated in $\lambda = 2$? Do you recognize it?

1.4 The uniform distribution

The uniform distribution on $[0, b]$ has density

$$f(x;b) = \begin{cases} \frac{1}{b} & 0 \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_1, X_2, \ldots X_n \sim f(x;b)$.

1. Verify that $f(x; b)$ is a density.
2. Calculate the expectation of $X \sim f(x, b)$. Can you imagine a reasonable estimator of b?
3. Try to show that the maximum likelihood estimator of b is $\max_{i=1}^{n} x_i$. (*Hint:* Don't use differentiation!)
4. Compare the estimator you derived in (2) to the maximum likelihood estimator using the same techniques as in the exercise on the exponential distribution. But you might have to multiply with n instead of $\sqrt{n}$ for the histograms to stabilize! Which estimator do you prefer, and why?

N.B. This maximum likelihood estimator is *irregular* since it does not follow a normal distribution, as you can observe from the histogram.

2 Logistic regression

2.1 Practical regression

This page works with logistic regression on a particular dataset.

2.1.1 Download data

Load the data set at `https://userpage.fu-berlin.de/soga/200/2010_data_sets/hurricanes.xlsx` into a data frame `hurricanes`. Make a `sns.pairplot` and look at the correlation matrix. For more information about the data, look at the link. Some of the correlations are extremely high. Why? (*Hint*: Look at the lecture notes. Be sure to remove columns that aren't numeric using `drop`. You can use `df.corr()` to calculate the correlation matrix.)

2.1.2 Correlation plot

The correlation matrix is hard to read. Modify the code from here or here to make it readable. (*Hint:* Start out with `sns.heatmap(dataframe.corr())`.)

2.1.3 Fitting a logistic regression

Make a new column in `hurricanes` called `Type.new`. A value in this column equals 0 if `Type == 0` and 1 otherwise. Use `sns.lmplot` to plot a logistic regression `"Type_new ~ FirstLat"`. (*Hint:* First make a column `c` that is 1 if `Type == 0` and 0 otherwise. Then modify it to be 0 if `Type == 0` and 1 otherwise using `1 - c`, or `(1 - 1 * (hurricanes["Type"] == 0))`.)

2.1.4 Finding other predictors

Using the `Type.new` variable as a response, find other reasonable predictors and plot them. (*Hint:* Can you use `hurricanes.corr()` for this?)

2.2 Link functions

We don't have to use the logistic or probit function, as `statsmodels` support many more. Consult the documentation for `statsmodels` and try out the Cauchy link function on the data set of the previous exercise. For instance, the probit link would look like:

```
mod_probit = smf.glm(formula="Type_new ~ FirstLat", data=hurricanes, family=sm.families.Bi
```

The documentation for the link functions is here. It may be hard to get things to run, but persevere!

Now plot the predicted values in the same plot, as we did in the lectures with probit and logit, with different colors for each. Is there i noticable difference in the curves?