Model Regresi Bayesian Zero-Inflated Bernoulli = Bayesian Zero-Inflated Bernoulli Regression Model

Ivana Ratanaputri, author

Abstrak

Data biner merupakan tipe data yang memiliki tepat dua kemungkinan nilai, seperti sukses dan gagal atau ya dan tidak, yang lebih lanjut direpresentasikan dalam respon 0 dan 1. Data biner kerap dijumpai dalam kehidupan sehari-hari. Namun, tidak jarang pula ditemukan data biner yang mengalami zero-inflation. Fenomena zero-inflation ini merujuk kepada data dengan dua sumber nilai nol yang berbeda, yang dikenal dengan istilah structural zeros dan sampling zeros. Oleh karena itu, dikembangkanlah suatu model alternatif, yakni model regresi Zero-Inflated Bernoulli untuk memodelkan data biner yang mengalami zero-inflation. Dalam inferensi statistika, terdapat dua jenis pendekatan yang umumnya digunakan, yaitu pendekatan frekuentis dan pendekatan Bayesian. Pada tugas akhir ini, dikonstruksi suatu model regresi Zero-Inflated Bernoulli menggunakan pendekatan Bayesian. Pendekatan Bayesian digunakan karena dianggap lebih unggul dibandingkan pendekatan frekuentis. Dalam data yang mengalami zero inflation, pendekatan frekuentis tidak mampu membedakan structural zeros dan sampling zeros. Hasil konstruksi model yang terbentuk diberi nama model regresi Bayesian Zero-Inflated Bernoulli. Salah satu hal penting dalam pendekatan Bayesian adalah mendapatkan distribusi posterior. Namun, sering kali nilai parameter dari distribusi posterior sulit ditemukan secara analitik karena distribusi posteriornya memiliki formula terbuka. Oleh karena itu, dalam tugas akhir ini estimasi parameter sekaligus pembangunan sampel posterior dicari melalui teknik komputasional dengan algoritma No-U-Turn Sampler (NUTS). Selanjutnya, model regresi Bayesian Zero-Inflated Bernoulli diimplementasikan untuk masalah klasifikasi pada data sickness presenteeism. Dalam tugas akhir ini, dibangun dua buah model regresi Bayesian Zero-Inflated Bernoulli, yakni model tanpa kovariat dan model dengan kovariat. Dari model tanpa kovariat, diperoleh estimasi parameter distribusi variabel respon adalah p1 = 0.38 dan p2 = 0.75. Lebih lanjut, hasil estimasi probabilitas yang diperoleh mendekati nilai empirisnya. Pada model dengan kovariat, digunakan dua kovariat untuk dua bagian yang berbeda, yakni evaluasi kondisi kesehatan (gh) pada seluruh sampel dan kovariat frekuensi merasakan perasaan takut tergantikan apabila tidak masuk kerja (remplz) pada sampel at-risk, hasil estimasi parameter regresi akan menghasilkan persamaan regresi yang dapat digunakan memberikan prediksi klasifikasi variabel respon kondisi pekerja yang masuk kerja pada saat sedang sakit. Diperoleh, berturut-turut tingkat akurasi dari model dengan kovariat gh dan kovariat remplz adalah sebesar 72.44% dan 69.58%, tingkat sensitivitas sebesar 14.65 % dan 100.00%, serta tingkat specificity sebesar 94.35% dan 0.00%.
.....Binary data is type of data that have exact two outcomes, for instance, success and failure or yes and no, that usually represent in 0 and 1. Binary data can be easily find on daily basis. However, there is binary data that experienced with zero-inflation. Zero-inflation phenomenon is caused by two different sources of zeros, which is called structural zeros and sampling zeros. Therefore, Zero-Inflated Bernoulli regression model is constructed for modeling binary data that experienced zero-inflation. There are two statistical inferences that is commonly used, that is frequentist and Bayesian inference. This thesis constructed Zero-Inflated Bernoulli regression model with Bayesian inference. Bayesian inference is selected because it is more superior than

frequentist inference on modeling binary data with two different source of zeros. Frequentist inference unable to distinguish the difference between structural zeros and sampling zeros. Constructed model is called Bayesian Zero-Inflated Bernoulli regression model. In Bayesian inference, it is important to get the predicted posterior distribution. However, in some cases, the analytic estimation of the posterior distribution is difficult to calculate because it has open formula. Therefore, posterior estimator is searched using computational techniques name No-U-Turn Sampler algorithm (NUTS). Furthermore, this regression model is implemented on classification problem sickness presenteeism data. In this thesis, we constructed two models, that is model without covariates dan model with covariates. From model without covariates, the parameter from response variable distribution can be estimated and we got p1 = 0.38 dan p2 = 0.75. This results is closed to the empirical value. Then, from model with covariates, two covariates is considered on implementation for different parts, i.e. general state of health (gh) covariate for all sample and feeling for being replaced (remplz) covariate for at-risk sample. From the estimated regression parameters, the regression equation is able give classification predictions for attend work while sick as response variable (sp recod). The results are the model give 72.44% and 69.58% accuracy rate.