Statistical Modeling and Inference – Project: Stochastic Gradient Descent

Group 3: Niti Mishra · Miquel Torrens · Bálint Ván

December 24^{th}, 2015

Exercise 1. *Find the Fisher Information matrix for logistic regression models.*

We take the definition of the Fisher information matrix as:

$$\mathcal{I}(\theta) := -\mathbb{E}[\nabla\nabla \log p(\mathbf{t}|\theta, q)].$$

Recall that logit output is Bernouilli distributed (belongs to exponential family) and uses the canonical link as a link function. In such distribution $q = 1$. Under the canonical link:

$$\begin{aligned}\mathcal{I}(\theta) &= -\mathbb{E}[\nabla\nabla \log p(\mathbf{t}|\theta, q)] \\ &= -\nabla\nabla \log p(\mathbf{t}|\theta, q).\end{aligned}$$

Thus, in such case the Fisher Information is also the Hessian operator.

We recover the exercise on GLM problem set where we generally[1] found that with the canonical link:

$$\begin{aligned}-\nabla\nabla \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) &= \sum_n q c''(\phi_n^T \mathbf{w})\phi_n \phi_n^T \\ &= \sum_n c''(\phi_n^T \mathbf{w})\phi_n \phi_n^T.\end{aligned}$$

In the second equation we plug in $q = 1$. The funciton $c''(\phi_n^T \mathbf{w})$ is the variance function, for the logit case its value is $c''(\phi_n^T \mathbf{w}) = p_n(1 - p_n)$, where p_n is the predicted value of our model for observation n. Thus:

$$\begin{aligned}-\nabla\nabla \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) &= \sum_n p_n(1 - p_n)\phi_n \phi_n^T \\ &= \mathbf{\Phi}^T \mathbf{\Gamma} \mathbf{\Phi},\end{aligned}$$

where $\mathbf{\Gamma} = \text{diag}\{p_n(1 - p_n)\}$. Hence we conclude:

$$\mathcal{I}(\theta) = \mathbf{\Phi}^T \mathbf{\Gamma} \mathbf{\Phi}.$$

[1] We assume that if the observations are weighted, such weight is normalized so that $\mathbb{E}[\gamma_n] = 1$.

Exercise 2. *Use the first 500K observations from the Higgs data set to calculate the MLE, the Fisher information matrix and, hence obtain the standard errors of the estimators when all features are present.*

The MLE parameter estimations are:

```
          int      feature1      feature2      feature3      feature4      feature5
 0.3409696367 -0.3163147631 -0.0036749536  0.0035259034 -0.3985513808 -0.0040957429
     feature6      feature7      feature8      feature9     feature10     feature11
 0.7573938467 -0.0067978234 -0.0027916526  0.0261748355  0.1773486113  0.0064906456
    feature12     feature13     feature14     feature15     feature16     feature17
-0.0078589816 -0.0604956986  0.1368764839  0.0033824428 -0.0009940343 -0.0683952332
    feature18     feature19     feature20     feature21    HLfeature1    HLfeature2
 0.1810469283 -0.0037715630 -0.0013436445 -0.0520651826 -0.1617299745  0.7064845387
   HLfeature3    HLfeature4    HLfeature5    HLfeature6    HLfeature7
 0.6468489258  0.4979919018 -1.1536753119  1.8150631015 -3.2204976071
```

The first parameter corresponds to the intercept (in R tagged as `int`).

The standard errors of the estimators are calculated by inverting the Hessian matrix and taking the squared root of the resulting diagonal. The outcome is the following:

```
        int    feature1    feature2    feature3    feature4    feature5    feature6
0.025571975 0.006152261 0.003231063 0.003102402 0.006071880 0.003058123 0.008758480
   feature7    feature8    feature9   feature10   feature11   feature12   feature13
0.003240137 0.003243072 0.004228112 0.007979249 0.003200069 0.003132400 0.003889931
  feature14   feature15   feature16   feature17   feature18   feature19   feature20
0.007424357 0.003146446 0.003079675 0.003292087 0.006792816 0.003080005 0.003040968
  feature21  HLfeature1  HLfeature2  HLfeature3  HLfeature4  HLfeature5  HLfeature6
0.002746495 0.007953374 0.016708579 0.019858700 0.010814962 0.009104218 0.026101661
 HLfeature7
0.028451938
```

For completeness we print the Hessian matrix, in three chunks:

	int	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	feature10	feature11
int	112154.10472	110862.50064	-37.012614	-87.806172	110612.89950	-102.485063	107091.92054	-125.75915	2.754321e+01	114299.08589	107868.68324	196.03932
feature1	110862.50064	143833.55998	13.600543	49.613886	103524.35477	18.253872	110345.16244	-79.86761	-7.850566e+01	112867.79764	106587.88784	206.58401
feature2	-37.01261	13.60054	113247.608993	-2.317946	-94.84685	31.968099	11.88540	31969.06666	6.687216e+01	-76.68094	-10.13000	31322.15100
feature3	-87.80617	49.61389	-2.317946	113458.555185	-139.77794	-5884.302665	-76.66452	-182.46108	-1.901072e+04	16.97715	-140.42412	92.86114
feature4	110612.89950	103524.35477	-94.846855	-139.777945	146343.09172	-166.494972	111079.30028	-178.44887	1.063936e+02	111189.80061	107242.24680	283.31390
feature5	-102.48506	18.25387	31.968099	-5884.302665	-166.49497	113737.574178	-183.80294	-178.85253	-1.753149e+04	-123.68585	-207.03452	-79.79536
feature6	107091.92054	110345.16244	11.885404	-76.664515	111079.30028	-183.802942	123220.26799	-174.15311	-4.372981e+01	109971.46434	112856.97384	131.78557
feature7	-125.75915	-79.86761	31969.066655	-182.461076	-178.44887	-178.852535	-174.15311	111896.09730	9.119807e+01	67.64050	-144.55189	28895.61732
feature8	27.54321	-78.50566	66.872161	-19010.717917	106.39360	-17531.490793	-43.72981	91.19807	1.135661e+05	178.04324	-124.13378	-161.06776
feature9	114299.08589	112867.79764	-76.680938	16.977149	111189.80061	-123.685847	109971.46434	67.64050	1.780432e+02	235091.01826	102709.78233	480.96201
feature10	107868.68324	106587.88784	-10.130004	-140.424116	107242.24680	-207.034518	112856.97384	-144.55189	-1.241338e+02	102709.78233	127666.35463	102.70176
feature11	196.03932	206.58401	31322.151002	92.861143	283.31390	-79.795358	131.78557	28895.61732	-1.610678e+02	480.96201	102.70176	113042.99474
feature12	52.10758	-33.42042	-58.326832	-10841.888525	256.03191	-7921.823713	109.84250	32.03616	-2.228918e+04	-120.55209	73.92735	-12.50922
feature13	112576.56181	110663.43237	58.578159	-60.762722	109858.23778	-148.095886	105303.81709	-113.92455	-1.620150e+02	82532.26273	116503.72570	236.29364
feature14	108992.55745	107205.98363	-105.346456	-73.001522	107885.69912	-86.948729	109373.39681	-192.91603	-5.573211e+01	103423.63497	111451.82616	236.60970
feature15	163.64875	228.55979	26929.091859	293.305870	137.04900	53.785164	184.85698	26934.09077	2.030555e+01	-96.49460	238.70569	25059.73695
feature16	11.26888	171.05309	-66.707046	-9561.247953	-98.89438	-5976.281934	87.00044	377.05272	-1.498818e+04	158.74663	36.40471	275.02236
feature17	112418.75861	111185.93109	179.392890	79.987227	111181.31418	-269.321627	105662.86740	-291.20829	-1.012949e+02	79463.68172	105103.95965	298.23039
feature18	108791.71290	106847.57763	-75.289648	-69.631289	107337.69964	-137.291916	107203.67741	-228.16171	-6.247773e+01	102777.99615	109090.46071	56.59569
feature19	-56.17402	-83.26899	21402.909262	-215.621008	-24.29308	36.452351	23.78382	22153.84077	2.080715e+01	-100.16132	59.79178	20366.73575
feature20	-231.45830	-183.91110	171.375737	-7116.364042	-219.42564	-4316.465945	-235.76964	-252.07582	-1.171791e+04	-377.20066	-206.03677	-34.44157
feature21	111726.06735	110495.46683	-72.436366	-101.147733	110834.70133	193.810526	106067.96039	-108.41738	1.862396e+02	75891.62328	105449.56181	-228.46559
HLfeature1	114226.71581	113806.31811	17.928007	31.414886	113944.83661	3.239603	114153.64000	-255.41406	-9.476665e+01	108332.28886	115421.56014	233.48356
HLfeature2	113373.06717	112373.60329	16.339407	18.325570	112465.25355	-63.601647	112265.25193	-170.19310	-7.202360e+01	113033.61393	113417.09976	189.07246
HLfeature3	117834.65722	119225.60742	-24.058664	-104.196865	118140.58929	-77.605203	112633.34436	-95.05320	-1.513544e+01	120127.54046	113355.03920	195.50917
HLfeature4	112619.71308	114488.98170	24.280454	-68.754689	118031.25074	-131.545145	112832.72709	-102.84759	8.238244e-01	120731.98341	112357.49045	245.53574
HLfeature5	107432.64771	105718.95329	9.345172	-55.025572	105438.79846	-180.081507	110716.15103	-177.83116	-7.376614e+01	124902.74656	112363.10732	192.01953
HLfeature6	113522.76984	114032.81987	15.776038	18.984590	116267.12411	-91.559974	116120.06606	-180.17738	-5.704555e+01	120464.31979	116253.06026	197.95619
HLfeature7	105088.46905	106130.88407	-6.872483	16.779070	108489.72480	-51.731875	106597.22331	-191.60732	-3.756037e+01	107585.80452	106550.40262	221.65458

	feature12	feature13	feature14	feature15	feature16	feature17	feature18	feature19	feature20	feature21	HLfeature1	HLfeature2
int	52.107579	112576.561809	108992.55745	163.648748	11.268885	112418.75861	108791.71290	-56.17402	-231.458297	111726.06735	114226.715807	113373.06717
feature1	-33.420420	110663.432369	107205.98363	228.559791	171.053088	111185.93109	106847.57763	-83.26899	-183.911104	110495.46683	113806.318114	112373.60329
feature2	-58.326832	58.578159	-105.34646	26929.091859	-66.707046	179.39289	-75.28965	21402.90926	171.375737	-72.43637	17.928007	16.33941
feature3	-10841.888525	-60.762722	-73.00152	293.305870	-9561.247953	79.98723	-69.63129	-215.62101	-7116.364042	-101.14773	31.414886	18.32557
feature4	256.031914	109858.237780	107885.69912	137.048998	-98.894380	111181.31418	107337.69964	-24.29308	-219.425636	110834.70133	113944.836613	112465.25355
feature5	-7921.823713	-148.095886	-86.94873	53.785164	-5976.281934	-269.32163	-137.29192	36.45235	-4316.465945	193.81053	3.239603	-63.60165
feature6	109.842502	105303.817086	109373.39681	184.856983	87.000437	105662.86740	107203.67741	23.78382	-235.769636	106067.96039	114153.640002	112265.25193
feature7	32.036165	-113.924552	-192.91603	26934.090773	377.052717	-291.20829	-228.16171	22153.84077	-252.075820	-108.41738	-255.414057	-170.19310
feature8	-22289.179232	-162.015011	-55.73211	20.305551	-14988.176764	-101.29488	-62.47773	20.80715	-11717.913696	186.23957	-94.766646	-72.02360
feature9	-120.552086	82532.262732	103423.63497	-96.494600	158.746626	79463.68172	102777.99615	-100.16132	-377.200660	75891.62328	108332.288863	113033.61393
feature10	73.927352	116503.725700	111451.82616	238.705692	36.404707	105103.95965	109090.46071	59.79178	-206.036766	105449.56181	115421.560141	113417.09976
feature11	-12.509224	236.293638	236.60970	25059.736946	275.022357	298.23039	56.59569	20366.73575	-34.441568	-228.46559	233.483561	189.07246
feature12	113790.266563	-33.439528	20.91294	9.254267	-10388.237200	151.14751	60.64023	56.26253	-8283.746831	392.20031	26.456230	95.00278
feature13	-33.439528	236540.988664	104350.52631	265.847216	124.956635	77581.29874	103692.46208	-96.61278	-9.254509	71571.74164	113288.290514	114726.66718
feature14	20.912939	104350.526309	130186.51516	179.937279	21.800703	121444.85311	112883.88557	-114.04749	-288.324171	106041.04946	115842.522648	114458.92468
feature15	9.254267	265.847216	179.93728	113478.902722	76.315352	-133.94079	262.17353	17105.57314	89.556527	648.34541	303.769752	198.81203
feature16	-10388.237200	124.956635	21.80070	76.315352	113355.864829	-245.91491	-37.61750	95.36303	-7527.678697	-293.57878	12.470379	15.68727
feature17	151.147511	77581.298744	121444.85311	-133.940790	-245.914910	272870.67050	105200.20762	-81.82378	-273.672421	63268.37785	119308.793132	116641.35366
feature18	60.640230	103692.462076	112883.88557	262.173525	-37.617501	105200.20762	132239.67175	-59.89012	-222.117758	123881.30258	114523.051197	113339.07128
feature19	56.262531	-96.612783	-114.04749	17105.573145	95.363026	-81.82378	-59.89012	113801.59230	-29.905603	-36.36187	-156.767477	-65.43675
feature20	-8283.746831	-9.254509	-288.32417	89.556527	-7527.678697	-273.67242	-222.11776	-29.90560	113480.384191	-115.14952	-273.722488	-271.87138
feature21	392.200306	71571.741639	106041.04946	648.345413	-293.578778	63268.37785	123881.30258	-36.36187	-115.149518	330833.30410	125066.354829	115385.54309
HLfeature1	26.456230	113288.290514	115842.52265	303.769752	12.470379	119308.79313	114523.05120	-156.76748	-273.722488	125066.35483	159021.819459	134117.45643
HLfeature2	95.002777	114726.667178	114458.92468	198.812034	15.687271	116641.35366	113339.07128	-65.43675	-271.871382	115385.54309	134117.456433	128000.95614
HLfeature3	95.351085	118311.137640	114491.67571	188.492970	1.882007	118155.15695	114237.51430	-56.63479	-241.377661	117279.78876	120153.511096	119182.25582
HLfeature4	77.327686	114774.523524	112003.80870	180.791618	-41.471399	111598.07834	110621.08423	-21.65808	-197.572925	107610.44657	117551.885820	115491.47911
HLfeature5	-2.010618	115422.675468	110945.42676	148.080030	125.814263	104122.84667	108201.09441	-67.06404	-317.127667	91547.78209	110032.968587	111491.01463
HLfeature6	115.354608	115974.542949	115157.16912	191.389370	18.342353	113059.27239	113134.95668	-90.92572	-258.466780	109845.34742	125776.571755	122390.49608
HLfeature7	83.414042	106273.459729	106125.59281	171.240615	2.904418	105799.04174	104648.66448	-74.62884	-225.760778	104608.69074	115966.844750	112579.68820

	HLfeature3	HLfeature4	HLfeature5	HLfeature6	HLfeature7
int	117834.657221	1.126197e+05	107432.647715	113522.76984	105088.469049
feature1	119225.607416	1.144890e+05	105718.953289	114032.81987	106130.884071
feature2	-24.058664	2.428045e+01	9.345172	15.77604	-6.872483
feature3	-104.196865	-6.875469e+01	-55.025572	18.98459	16.779070
feature4	118140.589294	1.180313e+05	105438.798460	116267.12411	108489.724799
feature5	-77.605203	-1.315451e+02	-180.081507	-91.55997	-51.731875
feature6	112633.344365	1.128327e+05	110716.151028	116120.06606	106597.223312
feature7	-95.053199	-1.028476e+02	-177.831160	-180.17738	-191.607320
feature8	-15.135444	8.238244e-01	-73.766139	-57.04555	-37.560373
feature9	120127.540460	1.207320e+05	124902.746563	120464.31979	107585.804523
feature10	113355.039199	1.123575e+05	112363.107320	116253.06026	106550.402623
feature11	195.509174	2.455357e+02	192.019530	197.95619	221.654580
feature12	95.351085	7.732769e+01	-2.010618	115.35461	83.414042
feature13	118311.137640	1.147745e+05	115422.675468	115974.54295	106273.459729
feature14	114491.675714	1.120038e+05	110945.426761	115157.16912	106125.592812
feature15	188.492970	1.807916e+02	148.080030	191.38937	171.240615
feature16	1.882007	-4.147140e+01	125.814263	18.34235	2.904418
feature17	118155.156952	1.115981e+05	104122.846667	113059.27239	105799.041738
feature18	114237.514296	1.106211e+05	108201.094409	113134.95668	104648.664479
feature19	-56.634794	-2.165808e+01	-67.064043	-90.92572	-74.628842
feature20	-241.377661	-1.975729e+02	-317.127667	-258.46678	-225.760778
feature21	117279.788764	1.076104e+05	91547.782090	109845.34742	104608.690745
HLfeature1	120153.511096	1.175519e+05	110032.968587	125776.57175	115966.844750
HLfeature2	119182.255820	1.154915e+05	111491.014634	122390.49608	112579.688199
HLfeature3	126772.456029	1.192047e+05	112892.932042	119513.45617	110696.444392
HLfeature4	119204.658689	1.298048e+05	113991.692117	122456.47532	112425.995379
HLfeature5	112892.932042	1.139917e+05	128138.022401	118282.83664	106400.697672
HLfeature6	119513.456173	1.224565e+05	118282.836642	127279.89062	115657.419603
HLfeature7	110696.444392	1.124260e+05	106400.697672	115657.41960	106993.939326

Exercise 3. *Describe in approximately one page, the methodology of stochastic gradient descent and its default implementation in the sgd R package.*

Extremely high time complexity requirement for estimations of big datasets sparked interest in algorithms that utilize only the gradient computations such as Stochastic Gradient Descent (SGD). SGD is a modification of the Robbins-Monro procedure for recursive estimation that requires linear time complexity in N and sublinear in p, which is much better than traditional estimation algorithms such as the Fisher scoring. SGD is defined through the following iteration:

$$\theta_n^{\text{SGD}} = \theta_{n-1}^{\text{SGD}} + \gamma_n C_n \nabla \log f(\mathbf{y}_n | \mathbf{x}_n, \theta_{n-1}^{\text{SGD}}),$$

where, the learning rate γ_n is defined such that $n\gamma_n \to \gamma > 0$ as $n \to \infty$. The sequence C_n is a sequence of positive-definite matrices, such that $C_n \to C$. It is used to better condition the iteration. This method, known as explicit SGD, is efficient, because γ_n is just a scalar sequence, C_n is numerically tractable and the log-likelihood is evaluated only in the observation n and not the entire dataset. It is also statistically correct, as it can be shown that it converges to a point where $\mathbb{E}(\nabla \log f(\mathbf{y}_n | \mathbf{x}_n, \theta) = 0$. In exponential familiy models this point is a unique optimum, so it converges to the true parameter value (it is unbiased).

We should mention, however, that it is hard to find a learning rate γ that converges fast and does not cause numerical divergence, and even for well-behaved values of γ convergence and stability are not guaranteed. To solve this, the improvement made is known as the Averaged Implicit Stochastic Gradient Descent (AI-SGD) method. This is the default implementation of the `sgd` R package. The AI-SGD procedure is:

$$\theta_n^{\text{im}} = \theta_{n-1}^{\text{im}} + \gamma_n C_n \nabla \log f(\mathbf{y}_n | \mathbf{x}_n, \theta_n^{\text{im}}),$$

with $\bar{\theta}_n = (1/n) \sum_{i=1}^{n} \theta_i^{\text{im}}$. Implicit update is the first key component of AI-SGD. Note that θ_n^{im} is present on both sides. One can show that this implicit update is a *shrinked* version of the explicit update. This makes it robust to misspecifications of the learning parameter. The second key component is the averaging, which guarantees optimal statistical efficiency under relatively relaxed conditions.

A key assumption here is that the direction of the gradient of the likelihood does not depend on the θ. This implies that the implicit update can be performed once a scalar value is found which will scale the gradient appropriately. Hence, the gradient for the implicit iterate θ is a scaled version of the gradient of the previous iterate.

It is also possible to regularize the implicit SGD by adding a elastic net penalty to the log-likelihood. Thus, AI-SGD is effectively a recursive estimation method that is statistically optimal, numerically stable and applicable to big datasets. This analysis leads to an algorithm which implements the most general update of implicit SGD.

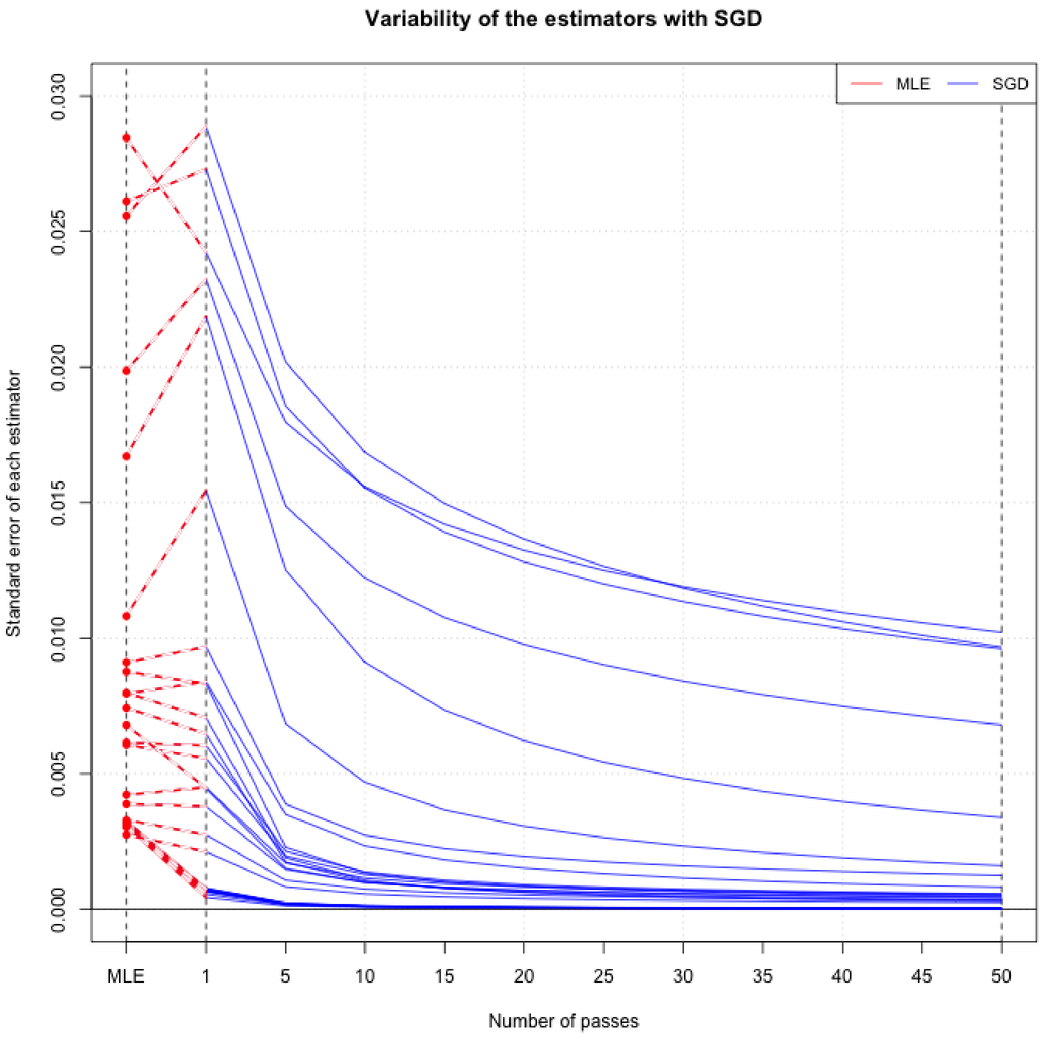
Generally, the algorithm works as follows. To minimize the objective function using SGD, we choose initial vector of parameters θ and the learning rate γ. The default choice for the parameters in R is $\theta = 0$, and the default learning rate is one-dimensional, of the form:

$$\gamma_n = \gamma_0 (1 + a\gamma_0 n)^c,$$

where $\gamma_0, a, c \in \mathbb{R}$ are fixed constants set so that they lead to statistical efficiency. Then, we perform the described iterations on independent random permutations of the dataset with a specified number of passes through the data (in R by default, 3) until we obtain an approximate minimum of our objective function. This is proxied by a stopping rule, namely when improvement of the likelihood is small enough (in R this is `1e-05`). At that point, parameters are stable and the result can be drawn.

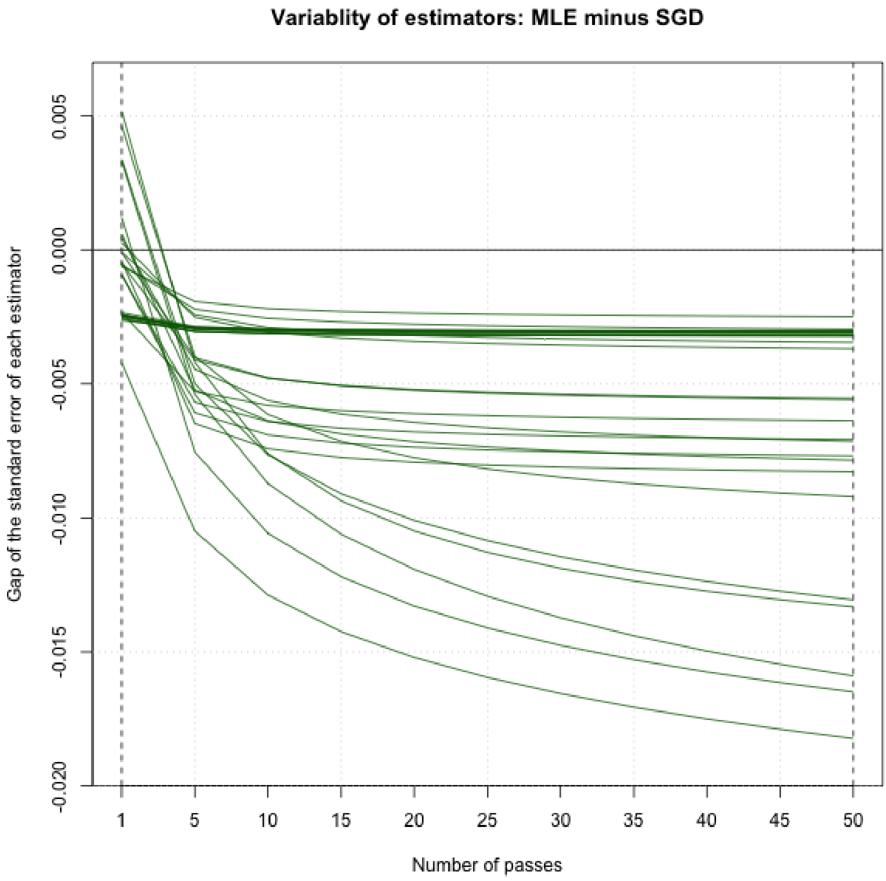
Exercise 4. *Fit the same logistic regression model using stachastic gradient descent. You should do this for each of 1, 5, 10, ..., 50 passes through the data, starting from the MLE. For each number of passes, repeat the estimation for 50 independent random permutations of the data. As an outcome, you should produce an appropriate figure that illustrates the variability of the estimators due to permutation as a function of the number of passes and compares it to the varibility of MLE.*

We produce the following plot:



In the first vertical dashed line on the left, we see as red dots the standard error of each of the 29 estimated coefficients under MLE (anonymized). The red dashed lines point the dots towards their respective standard error using SGD. The second vertical dashed line represents these same standard errors evaluated performing SGD starting from the MLE, with 1 pass through the data using 50 independent random permutations of the data. From there we see how this variability evolves as we change the number of passes, up to 50. We observe that with SGD more passes through the data help the estandard error of the estimators decrease.

To better see these differences, as an extra we also plot the difference in standard errors for MLE and SGD. Positive values mean that those from MLE are greater, and viceversa. For a small enough number of passes there is no dominant method, but as the number grows SGD exhibits smaller variability for all estimators.



Variablity of estimators: MLE minus SGD
Gap of the standard error of each estimator
0.005
0.000
-0.005
-0.010
-0.015
-0.020
1
5
10
15
20
25
30
35
40
45
50
Number of passes