Understanding Reenlistment in the Union Army during the American Civil War

Mario Trottini and John Robertson

Statistics Department, Carnegie Mellon University, Pittsburgh (USA)
History Department, Carnegie Mellon University, Pittsburgh (USA)

Abstract: *This paper focuses on the analysis of the factors that influenced the reelinstment decision of the volunteers who fought in the Union Army during the American Civil War. We examine 15 of the companies which served in four infantry regiment recruited in Pennsylvania. Individual soldiers' covariates as well as company covariates and regiment covariates are considered. A two-stage hierarchical logistic regression model with a company's random effect is used. Based on our analysis the occupation of the soldier and the date in which he started his service are significant predictors of the soldier reenlistment decision. In particular we found that non-stakeholders are more likely to reenlist than stakeholders, and soldiers that started their service later are more likely to reenlist than soldiers that started their service earlier in 1861. The evidence in our data set suggests that other variables, such as the size of the company or the number of battles that the company went through or the area in which the company was recruited, did not have any effect on the decision of reenlistment. Some of these results confirm results from previous studies, as for example the non relevance of the number and type of battles for the decision of reenlistment. The modest sample size and the presence of potential confounding variables not included in the study such as the marital status of the soldier or the economic condition of his family, suggest caution in interpreting these results.*

1 Introduction and motivation

The creation of the armies that fought the American Civil War was both a monumental and absolutely unprecedent task. In the entire course of the war there were nearly 3 million soldiers involved. The regular army, though enlarged for the war, could not supply such a number of men. Both the Union and the Confederacy therefore were forced to relay on "volunteers". The volunteers were recruited territorially. People joined companies made up from their own town and villages, and then were put into units by counties and states. In the initial call for volunteers in 1861, Abraham Lincoln called for one-half million men to enlist for three years. The war however lasted longer than expected, and at the beginning of 1864, almost three years after the first battle, there was still uncertainty about the duration and the resolution of the conflict. The unexpected duration of the war was a problem for both sides. As 1864 began almost half of the Union and Confederate regiments' enlistment terms were ending. However, while the Confederacy retroactively changed all of its enlistments from a fixed period to the duration of the war, the Union could not do the same. A previous decision of the Supreme Court established that the Government could not force soldiers to continue their service after the expiration of their enlistment terms (see Geary (1991a)). The expiration of the enlistments of these three-years men brought a manpower crisis in 1864. The federal government used aggressive measures to induce soldiers to reenlist. To get the soldiers to reenlist, the government appealed to the ideal of the "Union cause" and offered a Federal bonus of 400 dollars, a thirty-day furlough, the title of "Veteran Volunteer", and a sleeve chevron. In addition, if three quarters of the men of a regiment reenlist, it would go home as a unit on the furlough, and retain its regimental number, flag, and organization.

1.1 Goal of the analysis

This paper focuses on the analysis of the factors that influenced the reelinstment decision of the volunteers who fought in 15 companies of 4 regiments in the Union Army. In particular we consider individual soldiers' covariates such as rank, age and occupation of the soldier; company covariates such as the area in which the company was recruited, the company's regiment, the percentage of soldiers in the company that died from disease or were killed in action, and the size of the company at the beginning of the war in 1861 and just before the reenlistment campaign in 1864; and regiment covariates as, for example, the number and the type of battles that the regiment fought in the first three years of the war between 1861 and January 1864. These variables do not represent an exhaustive list of all possible factors that might be related with the decision for reenlistment. For example the marital status of the soldier, the economic conditions of his family, the quality of the captain in his company, are obvious factors that might have affected the decision of reenlistment. However we do not have information on all these variables. Most of the variables that are considered in existing theories of reenlistment are contained in our analysis, in this respect the set of the covariate that we work with seems to be adequate for the study. However we have to keep in mind that there might be confounding variables in our analysis that we cannot adjust for.

1.2 Motivation

The analysis of the factors that influenced the reenlistment decision of the soldiers who fought in the Union Army during the American civil war is of great historical interest. The patriotic volunteers of 1861 play a very special role in American history. They are considered the Union's best soldiers and the most socially representative army in United States history (see for example Kohn (1981), Karsten (1984)). Volunteers filled the ranks with white collar and skilled workers and farmers. This profile leads many historians to find that reenlistment of these gentlemen warriors of 1861 was crucial to the Union's ultimate victory. However even if there is general agreement on this initial construct, the characteristics of those who reenlisted in 1864 and those who went home at the end of their three-year term of enlistment have undergone very little exploration. Previous studies of the social background of the people who joined the Union army explore soldiers' ethnicity, class, prewar employment, life in the army, and postwar life (see Kohn (1981), Karsten (1984), Paludan (1988a), Rorabaugh (1986), Vinovskis (1990) or Thomas (1994)). Although excellent, these studies have many limitations as they usually use a small sample size or are limited to recruits from only one or two small towns and most of all they fail to examine reenlisting as rigorously as they did enlistment. Our analysis represents the first serious attempt to study the decision of reenlistment. The sample size that we use is huge compared to previous studies but still small compared to the size of the Union Army (the whole Union Army in 1861 consisted of 500 regiments). However our analysis is useful at least in two respects: it offers a methodology that can be used in richer data sets of the same type available in the future; it gives insight in understanding which factors affected the decision of reenlistment.

The paper is organized as follows. Sections 2 and 3 contain a description of the data set that we used and a first exploratory analysis. A more formal analysis is presented in section 4 where we study the relationship between the decision of reenlistment and the covariates of interest using a two-stage hierarchical logistic regression model with a company's random effect. Section 5 contains a diagnostic analysis for the model in section 4. We study the overall fit of the model as well as its predictive performance. Final remarks and conclusions are in sections 6.

2 The data

This study examines 15 of the companies, 5 from Pittsburgh and 10 from rural Western Pennsylvania, which served in four infantry regiments (the 38^{th}, 46^{th}, 61^{st}, 62^{nd} Pennsylvania Volunteers) recruited in Pennsylvania to fight for the Union Army in the American Civil War. Each regiment consists of 10 companies and each company consists approximately of 100 soldiers. The total number of soldiers involved in the study is 1355. The four Pennsylvania regiments selected are excellent sources of data for this type of study for several reasons:

- They are the only ones of the 68 three-years regiments formed in Pennsylvania in 1861 that included companies raised in rural Western Pennsylvania and companies raised in Pittsburgh. Using the 15 companies from these two areas makes an urban versus rural comparison possible;

- They all have a similar service record. All four regiments were involved in at least five of six major battles fought in the East during the three years of their original enlistment term and they lost approximately 15 percent of their initial complement to death in battle or to disease and had just under half of their original enlistment still under armies at reenlistment time;

- For 15 companies in these four regiments all the variables of interest in our study are available.

The data were collected by John Robertson, a Ph.D. graduate student in the History Department at Carnegie Mellon University. The personal information for soldiers is found in the muster out rolls in Samuel Bates' *History of the Pennsylvania Volunteers 1861-1865* (Bates,1869a). Using the corrections of Bates from Richard Sauers' *Advance the Colors* (Sauers 1987,1994), Robertson was able to compile a reasonably accurate roster of the 4000 soldiers who served in the 4 selected regiments. For each soldier the roster lists name, initial rank, date of enrollment, term of enlistment, final rank and condition under which active service ended (i.e if he reenlisted or musters out). He then completed the data set linking the information in Bates's muster out rolls with the information on age and occupations of the soldiers contained in the enlistment rolls that the companies filed in Harrisburg in the summer of 1861. These records are paperwork that the company officers were supposed to file to register their company with the state. Individuals in the two sources are considered a match if the enlistment roll lists them under the same company, name, initial rank, and date of enlistment as Bates. Only fifteen companies out of the forty of the four regiments of interest had successful link for a total of 1355 soldiers. We have no reason to believe that these fifteen companies are anyhow different from those for which we do not have complete information. In other words we do not expect the matching procedure to introduce any bias in the analysis and we can think of the 15 companies as the equivalent of a random sample of the companies in the four regiments under study. During the first three years of war almost half of the 1355 soldiers who enlisted in these 15 companies "left" their companies. Some of them died from disease, some were killed in action, some deserted and others found legal *solutions*. At the beginning of 1864 only 616 soldiers out of the initial 1355 were still active in service in their companies. These 616 observations represent our data set.

2.1 Problems and limitations

The main difficulty in our study comes from the nested nature of the problem. We have soldiers nested in companies that are nested in regiments. These relationships induce a dependency on the data. The reenlistment decision of a soldier might depend some what on the other soldiers in his company and on what other companies in his regiment do. When we measure the effect of the different factors of interest on the reenlistment decision we have to adjust for this dependency in the data. As we already said, our analysis is based on 15 companies which served in four regiments recruited in Pennsylvania. The sample size of the data set is huge if compared to previous studies but small compared to the size of the Union Army. As a consequence the results of this study cannot be extended to the whole Union army, not even to the Pennsylvania Regiments of the Union army. Finally, there are possible confounding variables that are not included in our analysis.

3 Exploratory analysis

The variables involved in this study are:

- REENLISTMENT: binary response variables which takes value 1 if the soldier reenlists and 0 if he musters out (i.e., if he does not reenlist);

- AGE: age of the soldier at the time in which he enlisted;

- RANK: Initial rank of the soldier, i.e., rank that was given to soldiers when they enlisted in 1861. It is a categorical variable with categories: private, corporal, sergeant, 1^{st}sergeant, 2^{st}lieutenant, 1^{st}lieutenant, musician;

- OCCUPATION: occupation of the soldier at the time in which he enlisted. It is a categorical variable with 6 categories: white collar worker, farmer, skilled worker, laborer, no occupation listed, other. Note that occupation data was self-reported and absence of a listed occupation could mean the soldier chooses not to state his or that the officer neglected to write it down.

- REGIMENT: categorical variable with four categories, the four regiments in our data set. It takes values 1, 2, 3, or 4 if the regiment is the 38^{th}, 46^{th}, 61^{st}, or 62^{nd} Pennsylvania Volunteers respectively;

- COMPANY: categorical variable with 15 categories, the 15 companies in our study, takes values 1 to 15 as described in table A1 in the appendix;

- SIZE.1861: discrete variable, for each company represents the size of the company at the beginning of the war in 1861;

- SIZE.1864: discrete variable, for each company represents the size of the company just before the reenlistment campaign at the beginning 1864;

- %KILLED: for each company expresses the percentage of soldiers in the company killed in action during the first three years of war, between 1861 and the beginning of 1864;

- %DIED: for each company it represents the percentage of soldiers who died from disease or accident in the company during the first three years of war, between 1861 and the beginning of 1864;

- RURAL: binary variable which takes value 1 if the company was recruited in a rural area and 0 if the company was recruited in a urban area;

- MONTHS.LEFT: discrete variable. Soldiers start their service at different time, but they all enlist for three years so, even if a soldier decides to muster out in January 1864, he has to offer his service until the three years period is completed. For each company MONTHS.LEFT represents the number of months of service left after January 1864 for a soldier that decides to musters out;

- N.BATTLES: discrete variable, for each company represents the number of battles that the company fought between the beginning of its service and January 1864;
- B.BATTLES: discrete variable, for each company represents the number of "bad" battles that the company went through; these include either battles that the army lost, battles in which the company suffered many losses, or battles in which the company experienced a traumatic event (the classifications of the battles as "bad" or "good" was done by John Robertson according with the criteria described above);

Figure 1 shows the barplot for OCCUPATION at the beginning of the war in 1861 and just before the reenlistment campaign in 1864.

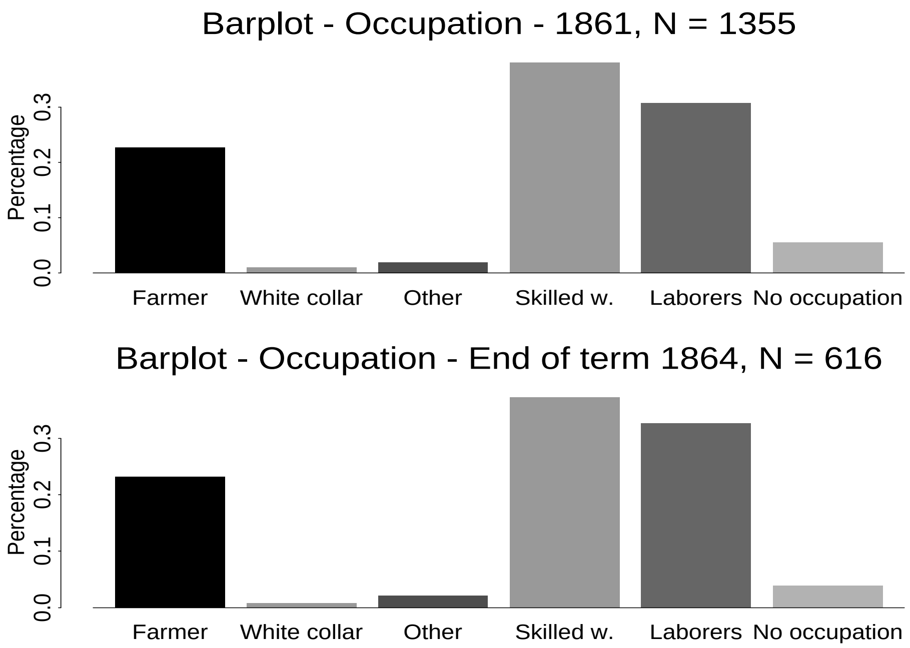


Figure 1

Almost half (616) of the 1355 soldiers that enlisted in 1861 were still active in service in their companies during the reenlistment campaign. It seems that the fortunes of the war were fair in eliminate soldiers. The distribution of occupation at the beginning of the war and just before reenlistment look almost identical. The soldiers that left their company in the three year period between the beginning of the war and the reenlistment campaign do not seem to be different in terms of occupation from those who stayed.

Actually in our analysis we will use a dichotomization of the variable OCCUPATION based on the nature of work listed for each soldier at the time of enlistment. The two social classes model is drawn from Katz, Doucet and Stern (1982) (see also Blumin (1989) and Lee (1999)). In particular we code white collar workers, skilled workers, farmers and others as stakeholder and we code laborers and the men listing no occupation as non-stakeholder. We define a new dummy variable, NON.STAKEHOLDER which takes value 1 for non-stakeholder and 0 for stakeholder. The reason for this dichotomization is that the job market opportunities and the salary for the two groups were very different at the time of the reenlistment campaign. There were great opportunities for stakeholders and very few for non-stakeholders (see Muller (1989), Gates (1965), David (1975), Rasmussen (1962), and Bates (1876)). In addition, while, for example, the salary of a skilled worker ranged between 300 and 500 dollars per year, for a laborer it was about 200 dollars per year (see Berryman (1977) and Paludan (1988b)). The 400 dollars Federal bonus for reenlisting therefore might look extremely appealing for a non-stakeholder but not very attractive for a stakeholder. Based on this consideration we expected non-stakeholders to be more likely to reenlist than stakeholders and the barplots of OCCUPATION before and after the reenlistment campaign in figure 2 seem to confirm this impression. In the group of those who reenlisted we observe a higher percentage of laborers (they constitute the main part of non-stakeholder) and a smaller percentage of skilled workers, white collars and farmers (i.e stakeholder).

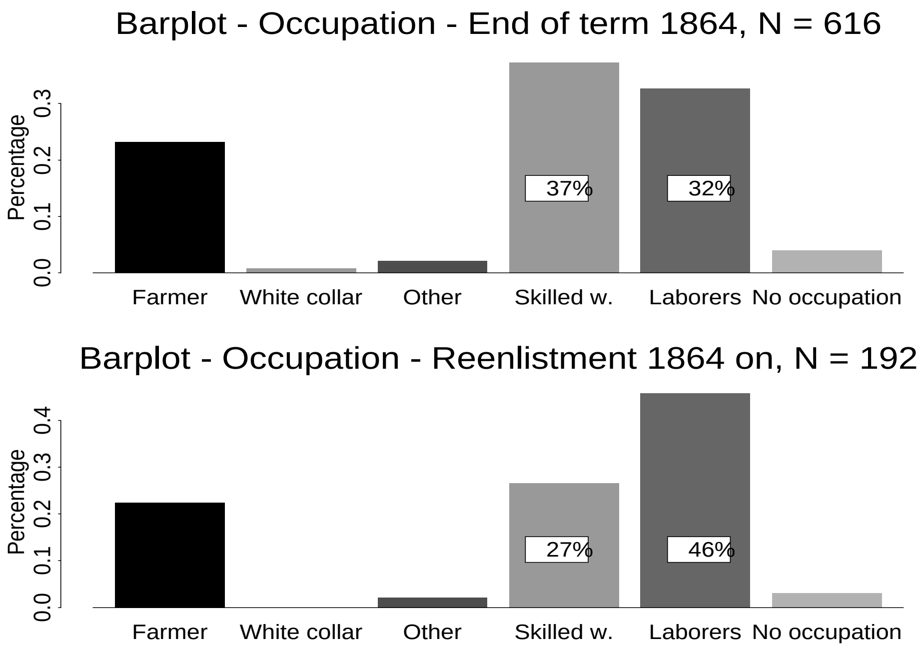


Figure 2

Note that of the 616 soldiers that had the opportunity to reenlist roughly one third of them (192) reenlisted.

The scatterplot in figure 3 shows percentage of reenlistment for each of the 15 companies in our study, grouped by regiment.

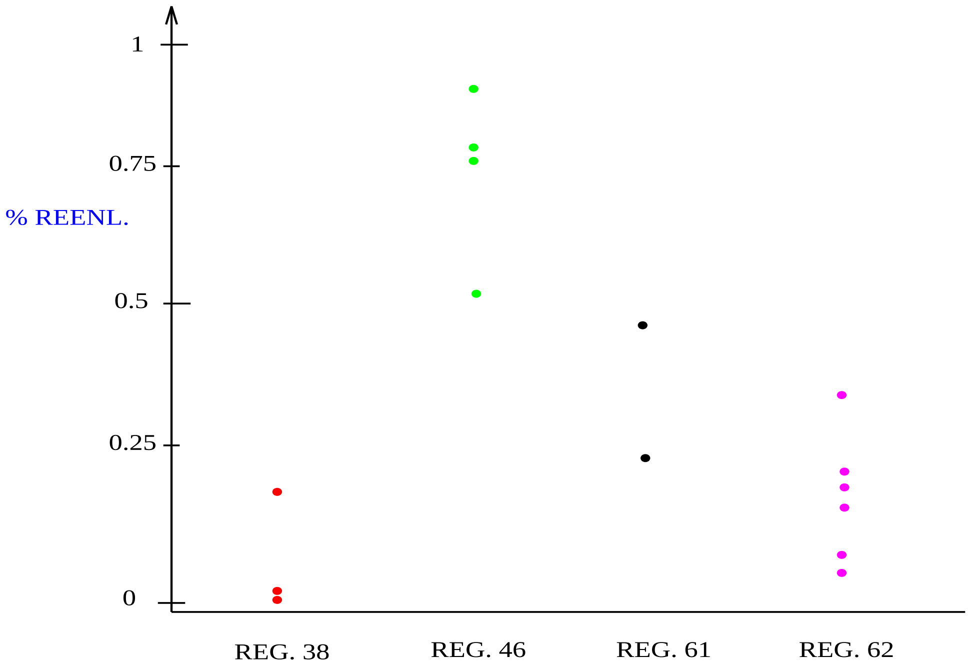


Figure 3: Companies reenlistment

From the scatterplot it is clear that companies in different regiments exhibit quite different percentages of reenlistment. In particular the percentage of reenlistment is very low in the the 38^{th} regiment and very high in the 46^{th} regiment. The analysis of the regiments' variables: (i) months of service left, and (ii) number and type of battles, might suggest an explanation of the differences in percentage of reenlistment observed in the different regiments.

As shown in figure 4 the different regiments started their service at different times. The 38^{th} for example started its service in July 1861 while the 46^{th} started its service in November 1861. However they all enlisted for three years. In January 1864, at the moment of reenlistment soldiers in different regiments had very different perspectives. The soldiers in the 46^{th} regiment have to stay in the army 10 more months, even if they decide not to reenlist, while the soldiers in the 38^{th} regiment have to stay only 6 more months. Soldiers in the 46^{th} regiment therefore have to choose between: (i) not reenlist, lose the Federal bonus of 400 dollars and stay 10 more months and (ii) reenlist, get the Federal bonus, go home for one month and stay for three years or less if the war ends before. It is not a surprise that most of them decided to reenlist especially if we take into account that at the moment of reenlistment there was a general expectation that the war would have not last more then one other year.

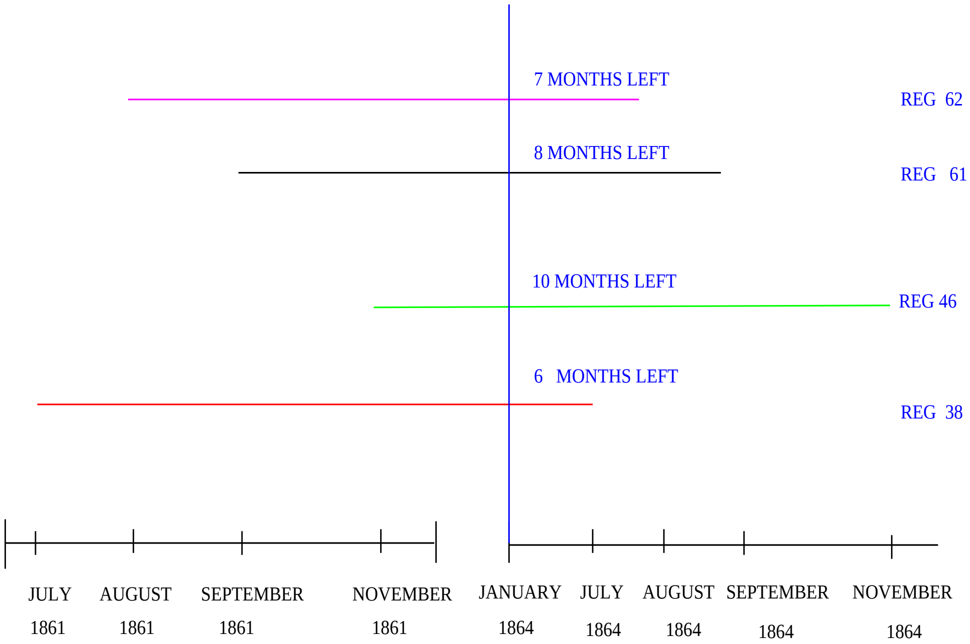


Figure 4: Number of months left after January 1864

In figure 5 a bar plot with bars proportional to the months of service left for each regiment after January 1864 if soldiers muster out is placed on top of the scatterplot in figure 3 which describes the percentage of reenlistment for each of the 15 companies in our study grouped by regiment. The number in the circles indicate the number of months of service left.

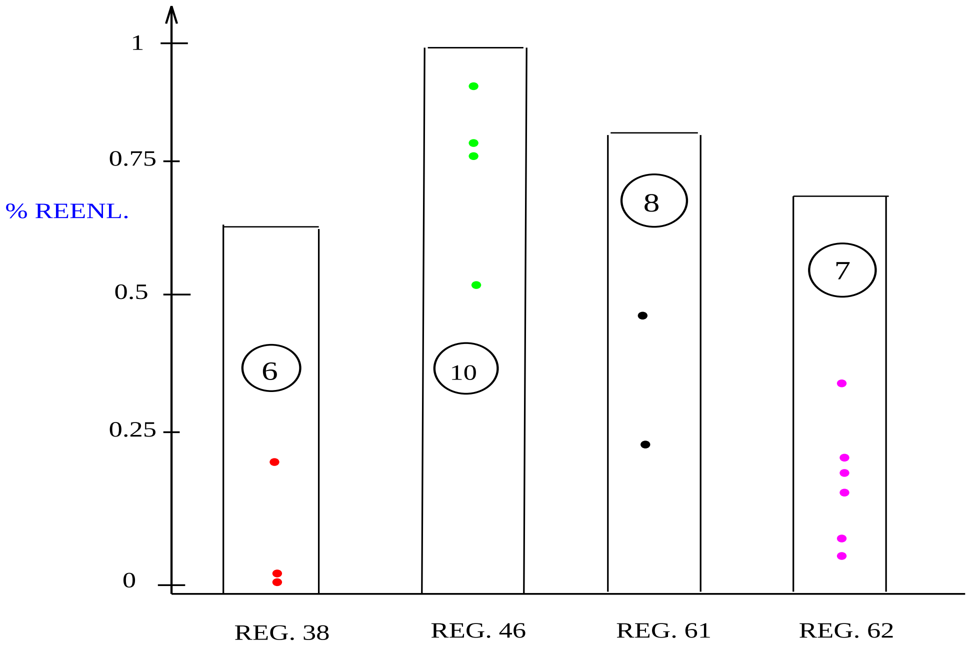


Figure 5: Companies reenlistment and number of months left

The barplot exhibits the same trend observed in the scatterplot. The time at which the regiment starts its service seems to be a very good predictor of the regiment's behavior in terms of reenlistment.

Another possible explanation for the differences in percentage of reenlistment observed in the different regiments could be the number and the type of battles that the regiment went through. Table 1 summarizes this information (see Dyer (1959), Bates (1869b)). For each combination battle-regiment an X indicates that the regiment participated to that battle. In particular "bad" battles are indicated in red and "good" battles in black.

		REG 38	REG 46	REG 61	REG 62
MAY 62	Battle of Fair Oaks				X
JUN-62	Seven Days Battle	X		X	X
JUL-62	Bat. of Malvern Hill				X
AUG-62	2nd Bull Run Battle	X			X
SEP-62	Battle of Antietam	X	X	X	X
DEC- 62	Battle of Fredericksburg	X		X	X
MAY- 63	B. of Chancellorsville		X		X
JUL - 63	Battle of Gettysburg	X	X	X	X

Table 1: Number and type of battles

As shown in table 1, the two regiments that exhibit the smallest percentage of reenlistment (the 38^{th} and the 62^{nd}) are those that were involved in the largest number of battles and also in the largest number of "bad" battles. On the other hand the 46^{th} regiment, which is the one with the highest percentage of reenlistment, was involved in just three battles. This suggests that the number and the type of battles that the regiments went through might be an important factor in the reenlistment decision.

None of the other variables seem to show any interesting change before or after the reenlistment campaign. Figure 6, for example, shows a multiple boxplot of the variable AGE at the beginning of the war, just before the reenlistment campaign and after the reenlistment campaign.The distribution of AGE seems to be very stable throughout the war.

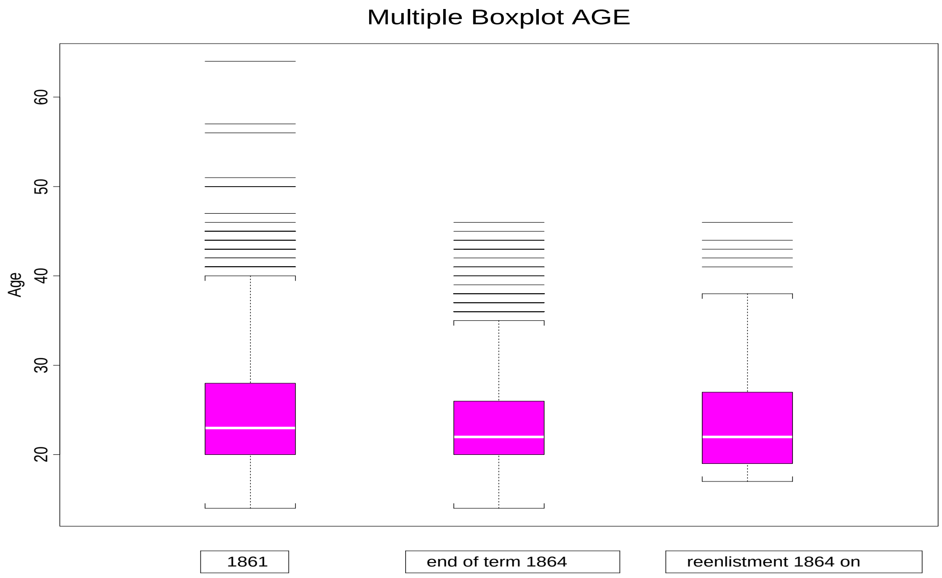


Figure 6

Overall the explanatory analysis suggests that the occupation of the soldier, the time at which his company started its service and the number and type of battles in which the the company was involved might be good predictors of the soldier's behavior in terms of reenlistment. In particular, "ceteris paribus", we expect non stakeholders to be more likely to reenlist than stakeholders and in the same way soldiers that served in companies that started their service late in 1861 to be more likely to reenlist than soldier that started their service early in 1861. Along the same line we expect soldiers that served in companies that were involved in few battles being more likely to reenlist than soldiers that served in companies involved in many battles.

Note that we could make other tables similar to table 1 refining the informations contained in table 1. Instead of distinguishing between "good" and "bad" battles we could use a finer classification which would also include "very traumatic" battles and "lightly engaged" battles. For example during the Battle of Fredericksburg the 62^{nd} regiment spent a day and two nights under heavy fire without having the opportunity to retreat (see Bates (1869c)) . That battle for the 62^{nd} regiment represented a very traumatic experience that cannot be compared with those of the other three regiments. The problem of what is the best classification to be used is not addressed in this work.

4 Two-stage hierarchical logistic regression model

In this section we analyze the relationship between the reenlistment decision and the covariates of interest. The hierarchical structure of the data (soldiers within companies within regiments) makes hierarchical models a natural choice for our problem. In particular we consider a two-stage hierarchical logistic regression model with a company's random effect. Let N_j be the size of company j at the moment of the reenlistment campaign in 1864 and let y_{ij} be the decision of reenlistment for the i^{th} soldier in the j^{th} company; $y_{ij} = 1$ if the soldier reenlisted, $y_{ij} = 0$ if the soldier musters out, $j = 1, \ldots, 15$, $i = 1, \ldots, N_j$. Denoting with p_{ij} the probability of reenlistment of the soldier i^{th} in the j^{th} company, we have:

$$y_{ij}|p_{ij} \sim Bernoulli(p_{ij}) \quad , j = 1, \ldots, 15, \quad i = 1, \ldots, N_j.$$

Stage 1 We model the logit of the probability of reenlistment of each soldier as a linear function of the soldier's covariate, AGE, RANK and NON.STAKEHOLDER which we treat as fixed effect, and a random effect COMPANY, which allows soldiers in different companies to have different probabilities of reenlistment,

$$\begin{aligned} logit(p_{ij}) = \alpha_1 \cdot Rank1_{ij} + \ldots \alpha_6 \cdot Rank6_{ij} + \alpha_7 \cdot Age_{ij} + \\ + \alpha_8 \cdot NonStakeholder_{ij} + Company_j \end{aligned}$$

Stage 2 For the random effect COMPANY we assume a normal distribution whose mean is a linear function of some of the company's covariate: the recruiting area of the company (RURAL), the months of service left after January 1864 (MONTHS.LEFT), the number of battles (N.BATTLES), the percentage of soldiers that were killed in action (%KILLED), the percentage of soldiers in the company that died from disease (%DIED) and the size of the company at the beginning of the war and just before reenlistment (SIZE.1861 and SIZE.1864 respectively),

$$Company_j \sim N(\mu_j, \tau^2),$$

$$\begin{aligned} \mu_j = \mu + \beta_1 RURAL_j + \beta_2 MONTHS.LEFT_j + \beta_3 N.BATTLES_j + \\ + \beta_4 \%KILLED_j + \beta_5 \%DIED_j + \beta_6 SIZE.1861_j + \beta_7 SIZE.1864_j. \end{aligned}$$

The specification of the model is completed by assigning prior distributions for the parameters in the model. In particular we assume:

$$\tau \sim Inv.Gamma(0.2, 0.2), \;\; \alpha_k \sim N(0, 5), \beta_h \sim N(0, 5), \;\; \mu \sim N(0, 5) \;\; h = 1, \ldots, 8; k = 1, \ldots, 7$$

The choice of the company's variables to include in the second stage of the hierarchical model was made by taking into account their correlations. Table 2 summarizes some of these correlations,

	MONTHS.LEFT	N.BATTLES	B.BATTLES	%KILLED	%DIED
MONTHS.LEFT	1.00	-0.67	-0.84	-0.39	0.68
N.BATTLES	-0.67	1.00	0.97	0.01	-0.26
B.BATTLES	-0.84	0.97	1.00	0.14	-0.43
%KILLED	-0.39	0.01	0.14	1.00	-0.29
%DIED	0.68	-0.26	-0.43	-0.29	1.00

Table 2: Correlation matrix

As we expect, the months of service left is negatively correlated with the numbers of "bad" battles, with the number of battles and with the percentage of soldiers in the company killed before January 1864. The earlier a soldier started his service in 1861 the more opportunities he had to be involved in battles and therefore more opportunities to be killed. However while months of service left is highly correlated with the number of "bad" battles (correlation $= -0.84$), it is only moderately correlated with the number of battles (correlation $= -0.67$) and very weakly correlated with the percentage of soldier killed in the company (correlation $= -0.39$). Also, as we expect, the number of battles and the number of "bad" battles are highly correlated (correlation $= 0.97$). What was unexpected, was the moderate positive correlation between months of service left and the percentage of soldiers that died in the company before January 1864 (correlation $= 0.68$). We would have expected here a negative correlation similar to the months of service left and percentage of soldiers killed in the company (the earlier a soldier started his service in 1861 the more opportunities he had to die from disease or have his health ruined some other way). From the variables listed in table 2 we decided to exclude from the analysis at the second stage of the hierarchical model only the number of "bad" battles that appears to correlate with months of service left and with the number of battles. We decided to keep in the analysis all the others variables. Note also that the correlation between the size of the company at the beginning of the war and just before reenlistment (not reported in table 2) is positive, as we expected, but weak (correlation 0.53). We will return to the correlation structure of the company's covariate later when we attempt to interpret the coefficients of the hierarchical model.

The choice of the priors and the parameterization used are briefly discussed in the appendix.

The posterior distributions for the parameters of interest were obtained using the statistical software BUGS 0.5 (Spiegelhalter D, Thomas A., Best N., Gilks W. 1996). We used 100000 simulations and a "burn in" of 10000. Diagnostic analysis showed convergence for all the parameters in the model.

4.1 Results

Figure 7 shows the 95% credibility intervals (CI) for the parameters in the model. The red lines indicate the 95% credibility intervals that do not contain zero, the corresponding numeric values are in parenthesis.

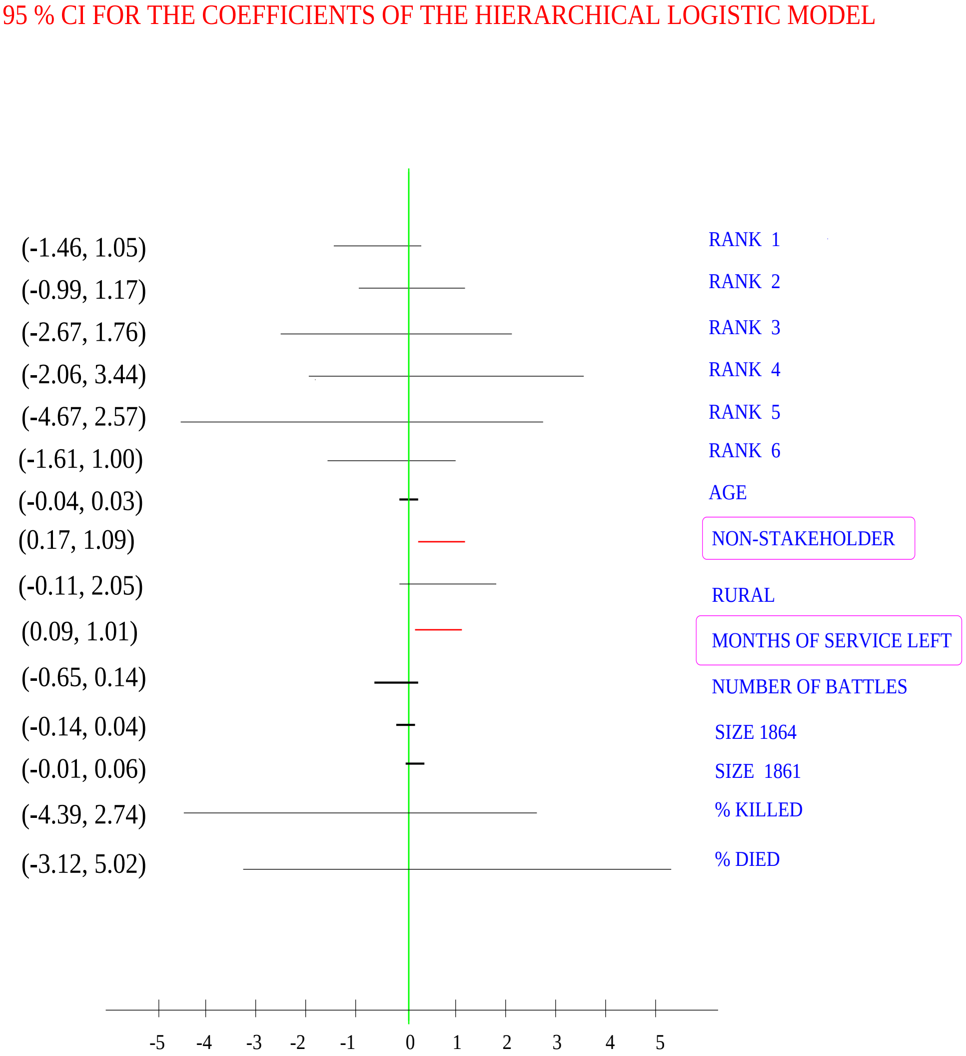


Figure 7

As shown in figure 7 there are two variables whose 95% confidence interval do not contain zero: the non-stakeholder variable and the months of service left variable. As we expect in both cases the lower bound for the 95% confidence interval for the coefficient is positive, meaning that non-stakeholders are more likely to reenlist than stakeholders, and soldiers that started their service later are more likely to reenlist than soldiers that started their service earlier in 1861.

In order to interpret the coefficient for NON.STAKEHOLDER it might be useful to translate the 95% CI for NON.STAKEHOLDER into a statement about the probability of reenlistment that we expect for a non-stakeholder given the probability of reenlistment for a stakeholder with a similar profile. Let P_1 and P_2 be the probability of reenlistment for non-stakeholders and stakeholders respectively. The coefficient α_8 for NON.STAKEHOLDER has an interpretation in terms of odds ratio. The odds of reenlistment for non-stakeholders is $exp(\alpha_8)$ times the odds of reenlistment for stakeholders if all the other covariates stay the same, i.e.

$$\frac{P_1}{1-P_1} = exp(\alpha_8) \cdot \frac{P_2}{1-P_2}. \tag{1}$$

The previous statement in terms of odds ratio can be translated in an equivalent statement about how P_1 varies as a function of P_2 as follows:

$$P_1 = \frac{exp(\alpha_8) \cdot \frac{P_2}{1-P_2}}{1 + exp(\alpha_8) \cdot \frac{P_2}{1-P_2}} = g(P_2).$$

In figure 8 we plot the function $g(P_2)$ using as values for α_8 the lower bound (0.17) and the upper bound (1.1) of the 95% CI for NON.STAKEHOLDER

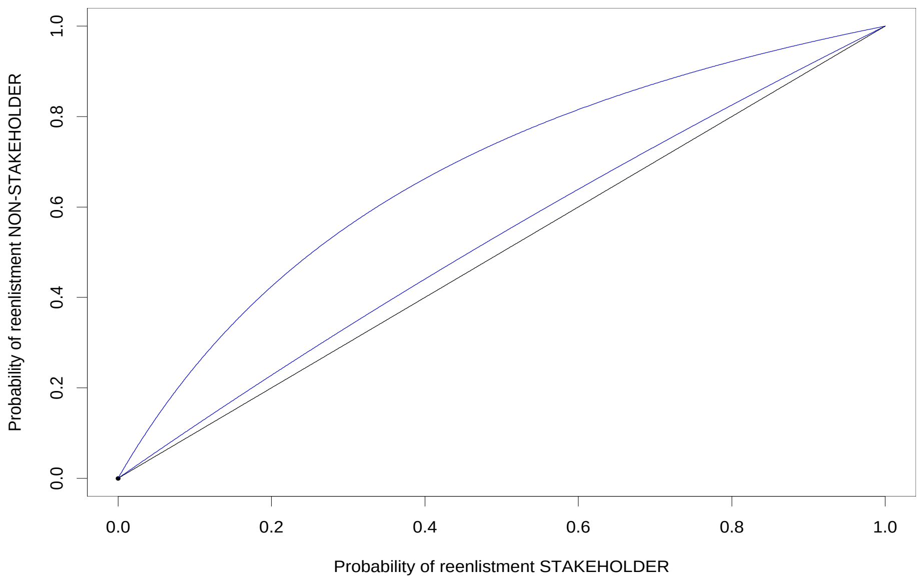


Figure 8: Interpretation coefficient for NON-STAKEHOLDER

On the x-axis we have the probability of reenlistment for a stakeholder, on the y-axis we have the probability of reenlistment for a non-stakeholder of similar profile. The 45 degrees line (in black) corresponds to the case in which there is no difference between the two groups ($\alpha_8 = 0$). As shown in figure 8 the effect of NON.STAKHOLDER is not very strong, the function $g(P_2)$ corresponding to the lower bound $\alpha_8 = 0.17$ is very close to the 45 degrees line of no effect at all. It seems however that prewar employment has an effect on reenlistment.

Figure 9 shows a similar graph for the other "significant" variable, the months of service left after January 1864.

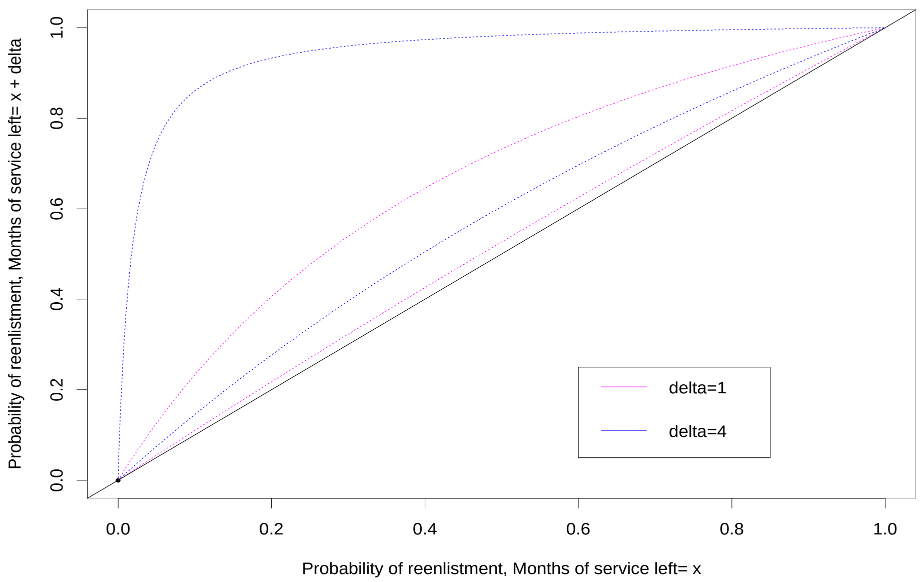


Figure 9: Interpretation of the coefficient for MONTHS OF SERVICE LEFT

On the x-axis we have the probability of reenlistment for a soldier that has x months of service left after January 1864 if he decides to musters out, on the y-axis instead we have the probability of reenlistment for a soldier that has $(x + delta)$ months of service left after January 1864 if he decides to musters out. In particular we consider two cases, $delta = 1$ (magenta lines) and $delta = 4$ (blue lines). The 45 degrees line (in black) represents the curve that we would see if the months of service left was independent of reenlistment. Again even if the lower bound for the two cases ($delta = 1$ and $delta = 4$) is close to the 45 degrees line, the effect of the variable months of service left is clear. In interpreting the coefficient for months of service left after January 1864 is important to remember that the variable MONTHS.LEFT is moderately correlated with other variables like percentage of soldiers died in the company (correlation = 0.68) and the number of battles (correlation $= -0.67$). This moderate correlation prevents us from interpreting the magnitude of the coefficients for these correlated variables.

It is also interesting to note that the number of battles is not significant. This result is consistent with previous studies where it was found that the number of battles or the type of battles that soldiers fought did not play a crucial role in the decision of reenlistment (see for example Linderman (1987), Hess (1997), McPherson (1997) Mitchell (1993), Frank (1991) and Geary(1991b)).

5 Diagnostic analysis

In order to validate the interpretation of the model in this section we perform a diagnostic analysis. In particular we consider the overall fit of the model (subsection 5.1) as well as its predictive performance (section 5.2).

5.1 Overall fit

Let y be the vector of the observed response, θ the vector of parameters in our model and let J be the number of covariates patterns, i.e. the number of different sets of values for the covariates in the model. Given $\Theta = \theta$ we have,

$$Y|\Theta = \theta \sim \prod_{h=1}^{J} Binomial(logit^{-1}(\theta \cdot X_h, m_h)$$

where X_h is the h^{th} covariate pattern and m_h is its frequency. Given $\Theta = \theta$ the expected response is therefore, $E(Y|\Theta = \theta)$. For a fixed θ, consider a partition of the data set in 10 subgroups based on the decile of the vector $E(Y|\Theta = \theta)$. The first group will contain the 10% of the soldiers with the smallest expected probability of reenlistment given $\Theta = \theta$ the last group instead will contain the 10% with the highest probability of reenlistment given $\Theta = \theta$. Denoting with n_j and $\tilde{n}_j(\theta)$ respectively the observed and the expected number of soldiers who reenlisted in group j when $\Theta = \theta$, we can define a measure of discrepancy between what we observe and our model in terms of the following test quantity:

$$T(y, \theta) = \sum_{i}^{10} \frac{(n_j - \tilde{n}_j(\theta))^2}{\tilde{n}_j(\theta)}.$$

In order to assess the overall fit of the model we evaluated the Bayesian p-value for our test quantity T. Following A. Gelman, J.B. Carlin, H.S Stern and D.B. Rubin (*Bayesian data analysis* 1984, pag. 168-169) we define y^{rep} as the replicated data "that could have been observed, or, to think predictively, as the data we would see tomorrow if the experiment that produced y today were replicated with the same model and the same value of θ that produced the observed data". The Bayesian p-value for T is defined as the probability that the replicated data could be more extreme than the observed data in terms of the test quantity T:

$$Bayes\ p - value = Pr(T(y^{rep}, \theta) > T(y, \theta)), \tag{2}$$

where the probability is taken with respect the joint posterior distribution of θ and y^{rep}:

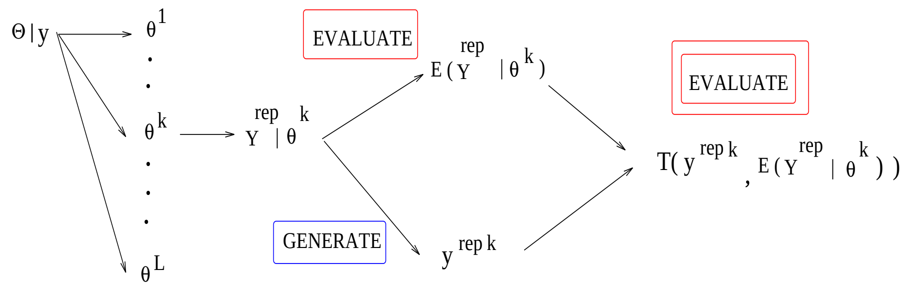
$$Bayes\, p - value = \int\int I_{T(y^{rep},\theta) \geq T(y,\theta)} \cdot p(\theta|y) \cdot p(y^{rep}|\theta) d\theta dy^{rep}. \tag{3}$$

Since y^{rep} is a realization of our model when $\Theta = \theta$ then we have that $T(y^{rep}, \theta)$ is the discrepancy that we expect to observe if the data comes from our model. If the observed data, y, do not come from our model we expect $T(y, \theta)$ to be bigger than $T(y^{rep}, \theta)$ and therefore the probability in (2) to be small. Therefore "small" values of the Bayes p-value indicate a bad fit of the model. A reasonable range for the p-value is between 0.05 and 0.95. We obtain the Bayes p-value for T, using simulation as described in figure 10. We first draw L values of θ from the posterior distribution of $\Theta|y$. For each simulated θ we draw one y^{rep} from the predictive distribution $Y|\theta$. In this way we obtain L draws from the joint posterior distribution of θ and y^{rep}. The estimated Bayes p-value is therefore the proportion of these L simulations for which $T(y^{rep}, \theta) \geq T(y, \theta)$.

(I) $$T(y, E(Y\mid\theta)) = \sum_{h=1}^{10} \frac{(n_h - \tilde{n}_h)^2}{\tilde{n}}$$

MEASURE OF DISCREPANCY

(II)



(III) $$\frac{\#\,(T(y^{rep\,k}, E(Y^{rep}\mid\theta^k)) > T(y, E(Y\mid\theta^k)))}{L}$$

Figure 10: Overall fit

Using L=1000 simulations we obtain a Bayes p-value for T of 0.33 which is in the range of a reasonable fit of the model. The overall fit of the model seems to be good.

5.2 Predictive performance of the model

In order to check the predictive performance of the model we split the original data set in two subsets, a training data set, A, and a test data set, B. As before, let θ be the parameter vector for our model. The idea is to use the training data set A to obtain posterior distribution for θ and then use this posterior distribution to get a predictive distribution for the test data set B. The assessment of the predictive performance of the model is then based on the comparison between the predictive distribution and the observed response Y_B in the test data set. Denote with Y_A and X_A respectively the response vector and the covariate matrix in the training data set. Similarly let Y_B, X_B be the response vector and the covariate matrix for the test data set. We consider a partition of the test data set B in 8 subgroups, $B_1, \ldots, B_8$, according to the occupation of the soldiers (stakeholder or non-stakeholder), the area in which their company was recruited (urban or rural) and the size of the company at the time of the reenlistment campaign in 1864 (size<42 or size>43, where 43 is the median of the size of the company in 1864). Let $p_1, \ldots, p_8$ be the observed percentage of reenlistment in the 8 groups. We now draw L values of θ from the posterior distribution of $\Theta|Y_A$. For each simulated θ we draw N response vectors $\tilde{y}_1(\theta), \ldots \tilde{y}_N(\theta)$ from the predictive distribution of $Y|\theta, X_B$ and for each $\tilde{y}$ we obtain a prediction $\tilde{p}_1, \ldots, \tilde{p}_8$ for the probability of reenlistment in the 8 groups $B_1, \ldots, B_8$. In this way for each group B_j, $j = 1, \ldots, 8$ we have $L \cdot N$ predictions for the probability of reenlistment in that group and we can compare the histogram of the predicted probability of reenlistment in group j with the observed percentage of reenlistment in the group, p_j. Figure 11 shows the histogram of the predicted probability of reenlistment in each group together with the observed value indicated by a vertical red line (L=100 and N=1000 are used). In order to take into account the different sizes of the eight groups in the test data, in figure 11 we superimpose the posterior distribution of p_j (curve in red), obtained from the test data using a uniform prior on p_j, to the histogram of the predicted probability of reenlistment in group j, $j = 1, \ldots, 8$.

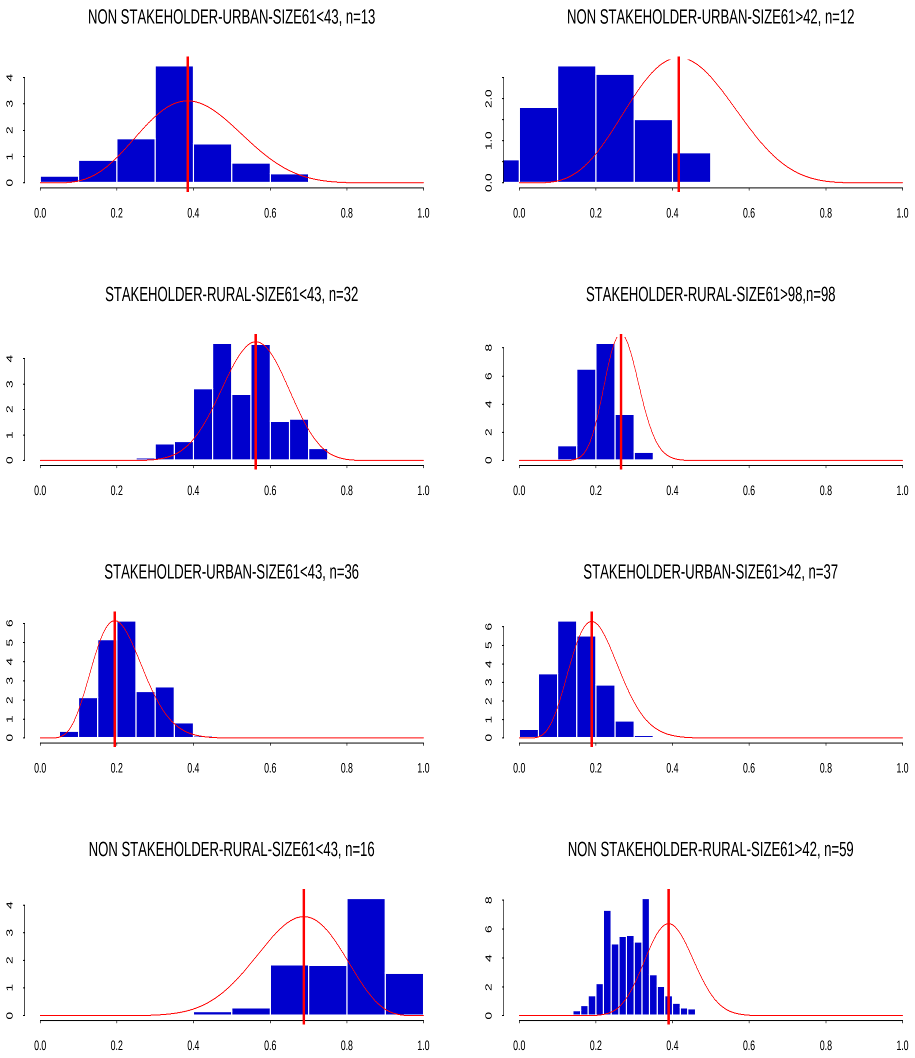


Figure 11: Histograms of the predictive probability of reemlistment in different groups

As shown in figure 11 the predictive performance of the model is decent but not great. In particular the model seems to fairly estimate the variability in the estimate of the probability of reenlistment in each group, but for some groups it overestimates the probability of reenlistment (see $NON-STAKEHOLDER-RURAL-SIZE61 < 43$) while for others it underestimates the probability of reenlistment (like in the $NON-STAKEHOLDER-URBAN-SIZE61 > 42$ or in the $NON-STAKEHOLDER-RYRAL-SIZE61 > 42$ groups). This result is not very surprising. As we said at the very beginning, there are variables that we could not include in our model and that are very likely to play an important role in the decision of reenlistment such as the marital status of the soldiers, or the economic conditions of his family, or the qualities of his captain. Including these variables in the model would probably improve both the general fit and the prediction performance of the model. However it is worthwhile to notice that the location of the histograms of the predicted probability of reenlistment in each group are not too far from the observed probability of reenlistment in each group, the maximum bias is less than 0.2 and in most of the cases it is less than 0.1.

6 CONCLUSIONS

In our analysis we studied the reenlistment decision in 15 companies which served in four infantry regiments recruited in Pennsylvania to fight for the Union Army in the American Civil War. Given the small sample size of the data set that we used, the results of this study cannot be extended to the whole Union army, not even to the Pennsylvania Regiments of the Union army. However a richer data set should be available in the near future and the methodology developed in this paper can be used to analyze the new data. Although with some limitations, as for example the presence of possible confounding variables, the results of our analysis give insight into understanding the reenlistment decision of the volunteers of 1861. We found that the occupation of the soldiers and the time in which the soldier started his service are important factors for the reenlistment decision. In particular we found that, "ceteris paribus", non stakeholders were more likely to reenlist than stakeholders and similarly soldiers that served in companies that started their service late in 1861 were more likely to reenlist than soldier that started their service early in 1861. These results have a natural interpretation, considering the changes that the job market went through in the first three years of the war, and the general expectation about the duration of the war at the moment of reenlistment.

The evidence in our data set suggests that other variables, such as the size of the company or the number of battles that the company went through or the area in which the company was recruited, did not have any effect on the decision of reenlistment. Some of these results confirm results from previous studies, as for example the non relevance of the number and type of battles for the decision of reenlistment. It would be interesting to include in the analysis other variables such as the marital status of the soldiers and the economic conditions of his family, these are likely to be important confounding variables whose inclusion in the model might improve both the overall fit and the predictive performance of the model as well as give a more reliable estimate of the coefficients of the other variables.

As a final remark we note that in the model that we use the variable RANK denotes the initial rank of the soldier. We ran the same analysis using the final rank of the soldiers (that in addition to the 6 categories listed for the initial rank contains also the category "captain") but the results did not change.

7 APPENDIX

Companies used in the study. Table A1 reports the 15 companies used in this study.

Regiment's name	Company's name	COMPANY code	Regiment's name	Company's name	COMPANY code
38^{th}	A	1	61^{st}	H	9
38^{th}	F	2	62^{nd}	C	10
38^{th}	I	3	62^{nd}	D	11
46^{th}	A	4	62^{nd}	E	12
46^{th}	B	5	62^{nd}	I	13
46^{th}	G	6	62^{nd}	L	14
46^{th}	H	7	62^{nd}	M	15
61^{st}	A	8	62^{nd}		

Table A1

Choice of the priors For the parameters of the hierarchical model in section 4 we tried to use non-informative priors. The choice of the priors was based on heuristic considerations about the meaning of the coefficients of the logistic model using graphs similar to those showed in figures 6, 7 and 8. We compared the results from our two-stage hierarchical model with those of a fixed effect logistic model and they were very close. This gives us enough confidence that the choice of our priors as non informative priors was reasonable.

Parameterization The variable RANK has been coded using 6 dummy variables, RANK1,..., RANK6 as described in table A_1

	RANK1	RANK2	RANK3	RANK4	RANK5	RANK6
private	0	0	0	0	0	0
corporal	1	0	0	0	0	0
sargeant	0	1	0	0	0	0
1^{st}sargeant	0	0	1	0	0	0
2^{st}lieutenant	0	0	0	1	0	0
1^{st}lieutenant	0	0	0	0	1	0
musician	0	0	1	0	0	1

Table A_2

References

Bates S. P., (1869a). *History of Pennsylvania Volunteers 1861-5* Vol. 1-2, Harrisburg: B. Singerly, State Printer, 1869.

Bates S. P., (1869b). *History of Pennsylvania Volunteers 1861-5*, Vol. 1, pag. 784-812, Vol. 2, pag. 407-521, Harrisburg: B. Singerly, State Printer, 1869.

Bates S. P., (1869c). *History of Pennsylvania Volunteers 1861-5*, Vol. ?, pag 455-456, Harrisburg: B. Singerly, State Printer, 1869.

Bates S. P., (1876). *Martial Deeds of Pennsylvnia*, Philadelphia, 1876, pag. 133

Berryman S. E., (1977). *Who Serves ? The Persistent Myth of the Underclass Army.*, Boulder, 1977, pag. 26

Blumin S., 1989. *The Emergence of the Middle Class: Social experience in the American City 1760-1900*, New York, 1989

David P. A., 1975. *Technical Choice Innovation and Economic Growth: essays on American and British Experience in the Nineteenth Century*, London, 1975, pag. 195-232

Dyer F., 1959. *Compendium of the War of the Rebellion*, Vol. 3, *regimental Histories*, New York, 1959, pag. 1581, 1589, 1594-1596

Frank J. A., 1991. "Profile of a Citizen Army: Shiloh's Soldiers", *Amed Forces and Society*, XVIII, pag. 97-110.

Gates P., 1965. *Agriculture and the Civil War*, New York 1965, pag. 229-235

Geary, J. W., 1991a. *We need men: The Union Draft in the Civil War*, DeKalb, 1991, pag. 6

Geary, J. W., 1991b. *We need men: The Union Draft in the Civil War*, DeKalb, 1991, pag. 112-113

Gelman A., Carlin J. B., Stern, H. S. and Rubin D. B., 1984 *Bayesian data analysis*

Hess H. J., 1997. *The Union Soldier in Battle: Enduring the Ordeal of Combat*, Modern War Studies, edited by Theodore A. Wilson, Lawerence, 1997, pag. 89, 101, 150.

Karsten, P. 1984. "The "New" American Military History: A Map of the Territory, Explored and Unexplored", *American Quarterly*, XXXVI, pag 389-414

Katz M., Doucet M. J., and Stern M. J., 1982 *The Social Organization of Early Industrial Capitalism*, Cambridge, MA, 1982

Kohn, R. H. 1981. "The Social History of the American Soldier: A review and Prospectus for Research", *The American Historical Review*, LXXXVI, pag. 553-567

Lee C., 1999. "Selective Assignment of Military Positions in the Union Army: Implications for the Impact of the Civil War", *Social Science History*, XXIII, 69-97

Linderman G. R., 1987. *Embattled Courage: The Experinece of Combat in the American Civil War*, New York 1987, pag. 261-264.

McPherson J. M., 1997. *For Cause and Comrades: Why Men Fougth in the Civil War*, New York, 1997

Mitchell R., 1987. *The Vacant Chair: The Northern Soldier Leaves Home*, New York, 1993, 158-159

Muller E. K., 1989 "Metropolis and Region: a Framework for Inquiry into Western Pennsylvania", *City at the Point: Essays on the Social History of Pittsburgh*, ed. Samual P. Hays, Pittsburgh, 1989, pag. 188-190

Paludan P. S., 1988a "A People's Contest", *The Union and Civil War 1861-1865'*, pag. 178

Paludan P. S., 1988b "A People's Contest", *The Union and Civil War 1861-1865'*, pag. 182

Rasmussen W. D., 1962. "The impact of Thecnological Change on American Agriculture,1862-1962", *Journal of Economic History*, XX, pag. 578-591

Rorabaugh W. J., 1986. "Who Fought for the North in the Civil War ? Concord Massachussets, Enlistments", *Journal of American History*, LXXIII, pag. 695-701

Sauers, R. A. (1987). *Advance The Colors! Pennsylvania Civil War Battle Flags* Commonwealth of Pennsylvania: The Capital Preservation Committee 1987

Sauers, R. A. (1994). "Samule P. Bates Pennsylvania Historian", *History of Pennsylvania Volunteers 1861-5*, Index Vol. 1, ed. Janet Hewett, Wilmington, North Carolina 1994, X-XVI.

Spiegelhalter D., Thomas A., Best N., Gilks (1996). *BUGS 0.5 Bayesian inference using Gibbs Sampling.*

Thomas, T. A., 1994. "For Union, not for Glory: Memory and the Civil War Volunteers of Lancaster Massachusetts",*Civil War History a Journal of the Middle Period*, XL, pag. 25-47

Vinovskis, M. A., 1990 "Have Social Historians Lost the Civil War ?", essay in Vinoskis, Maris A., ed. *Toward a Social History of the American Civil War Exploratory Essays*, Cambridge, 1990, pag 16-17