Comparison of econometric methods to examine factors influencing inpatient length of stay in adults with insulin-dependent diabetes

Lisa Irvine & Ed Wilson, University of East Anglia on behalf of DIPSat Study group

Introduction and background

It is estimated that up to 10% of inpatient populations have diabetes (1). Although diabetic patients may be admitted for any condition, and diabetes may not be the primary diagnosis, the management of diabetes in inpatients can complicate admissions, and may contribute to excess length of stay in this population. The cause of this prolonged length of stay (hereafter LOS) is unknown, but factors such as duration of diabetes, insulin use, loss of control over insulin injections, poor blood glucose control, and variations in standards of care may influence the inpatient experience, and for this analysis, the budgetary impact in terms of length of stay.

The DIPSat study set out to investigate the relationship between inpatient treatment satisfaction and various aspects of diabetes management. The recently developed Diabetes Treatment Satisfaction Questionnaire for Inpatients (DTSQ-IP) is a 19-item questionnaire to evaluate inpatient satisfaction with diabetes related aspects of their care. The study also presented the opportunity to examine any relationship between patient length of stay and inpatient care, amongst other demographic variables, which is the focus of this paper.

Understanding the determinants of hospital length of stay is important to fully characterise demand for inpatient beds, and the impact this demand has on running hospitals in general. Inpatient bed days are a substantial part of quantitative measures in estimating hospital performance, and often the single greatest factor contributing to overall hospital costs (2). Each additional bed day for one particular patient represents an opportunity cost to the hospital, unable to allow another patient to be admitted, leading to increases in waiting lists and reduction in overall efficiency.

With ever increasing demand on hospital resourcing, forecasting inpatient bed usage may allow policy makers to better predict the relationship between how diabetes is managed and increasing efficiency in care delivery. Many different models have been used when determining the association between patient characteristics and hospital length of stay (3-6). However, no regression model has been uniformly recommended with which to analyse such data, nor is it known which method is best.

Our research objectives in this paper are twofold:

1. To what extent is length of stay explained by the characteristics of patients admitted?
2. Which choice of econometric model is most appropriate for our dataset?

Method

Contextual setting and Data

A self-complete questionnaire was designed and distributed among insulin-dependent diabetic inpatients in 61 UK Hospitals. Working toward an overall sample size of 1400 responses, diabetes inpatient nurses (DISNs) in each hospital were provided with 160 hard-copy questionnaires to distribute between April 2008 and January 2010. All inpatients fulfilling the following inclusion criteria were invited to take part: aged 18 or older, with Type 1; Type2; or newly-diagnosed diabetes, who were *managed with insulin during that hospital admission*. We collected data on age, gender, ethnicity, first language, duration of diabetes, duration of insulin use, frequency of insulin use, insulin regimen used, type of ward (medical, surgical or orthopaedic), administering of insulin (by self or staff), blood glucose monitoring (by self or by staff), and self-reported length of stay. In addition, patients were asked to complete the Diabetes Treatment Satisfaction Questionnaire for Inpatients (DTSQ-IP).

Length of stay and reasons for admission were crosschecked retrospectively from hospital databases. Data cleaning and imputation was performed in SPSS – unanswered responses were assumed not to have taken place or not relevant (e.g. relating only to surgical admissions), unless the entire page of responses was left missing. Questionnaires with more than one page missing were excluded. Stata 10 (STATA Corporation, 2007) was used for all subsequent analysis (7).

Description of the modelling methods

We hypothesised different econometric models would result in different conclusions about the impact of patient, hospital, and treatment characteristics on the length of stay, and that different models would have varying abilities to correctly predict inpatient length for stay for insulin dependent diabetics

We compared

- Linear regression;
- Linear regression with log-transformed length of stay;
- Generalized linear models with the following distributions: Poisson, Gamma, Negative binomial; and
- Semi-parametric survival models

Length of stay data tends to be positively skewed and subject to outliers. Hence, the distribution of the error term also tends to be positively skewed, as opposed to normal (see Appendix 6a).

We first look at the most basic regression framework: ordinary least squares (OLS). An advantage of using OLS regression is that the model is additive, with the regression coefficients interpretable as the increase in LOS for a one-unit increase in a given predictor variable. The assumptions of ordinary least squares (OLS) regression include the residual errors being normally distributed, equal variance at all levels of the independent variables (homoskedasticity), and uncorrelated residual errors with the independent variables. The assumption of uniform variance is unrealistic in length of stay data.

Linear regression on log transformed

A linear regression with logarithm of length of stay as the dependent variable may be used to model demand for inpatient beds. As the distribution of costs tend to be approximately normalised by taking its logarithm (see Appendix 6b), it is more likely that the distributional assumptions of linear regression are satisfied, allowing valid inferences about the statistical significance of the regressors to be drawn from the model. A drawback of this method is that the model is interpreted on a multiplicative scale. This may be more difficult to interpret and may lead to potential confusion about the outcome. Additionally, estimated predictions must subsequently be retransformed, using techniques such as Duan's smearing,(8) which may lead to efficiency losses (9, 10).

GLM

There are several options available to relax the assumption of normal distribution. Generalized linear models have been explored in econometric literature for analysing continuous outcomes that are subject to skewness (such as cost data) (11). Unlike the classical regression model, this places probability mass at nonnegative integer values only. Generalized linear models incorporate both a *distribution from an exponential family*, and *link function*, which provides the relationship between the linear predictor and the mean of the distribution function. With the identity link function, covariates act additively on the mean, hence interpretations of coefficients are unchanged. With the log link function, covariates act multiplicatively on the mean, therefore interpretation of each coefficient is less straightforward.

GLM distributional families are:

- Gaussian: constant variance ($\lambda = 0$) (Equivalent to standard OLS)
- Poisson: variance proportional to the mean ($\lambda = 1$)
- Gamma: variance proportional to the square of the mean ($\lambda = 2$)
- Inverse Gaussian: variance proportional to the cube of the mean ($\lambda = 3$)

Generalized linear models are commonly used when dealing with count data, such as bed days, health care costs, or occurrence of events, as only positive values are estimated, regardless of how many values are close to or at zero. An assumption inherent in Poisson modelling is equi-dispersion: that the mean is equal to the variance(7). The negative binomial model is a variant of the Poisson model that does not make the assumption that the mean is equal to the variance, and as a result, it can provide a better fit if there is over-dispersion of the data. Finally, both Gamma and inverse Gaussian offer means of accounting for heteroskedasticity in length of stay. By modelling variance as proportional to the square

(Gamma) or cube (Inverse Gaussian) of the mean LOS distribution, these can reflect the shape of raw length of stay data more closely than Gaussian distribution. However, both can suffer substantial efficiency losses, as they places less weight on extremely long length of stays, which contribute most severely to hospital costs (5).

Survival analysis

Survival analysis concerns analysing the time to the occurrence of an event, and naturally lends itself to the study of length of stay (the 'event' namely discharge from hospital) (12). In survival analysis, the main focus is not only the duration of events (length of stay in our case) but also the *conditional probability that the event will end in the next period,* given that it lasted at least till the period t. The duration of stay is formalised in terms of the *hazard rate*. Hazard rate can vary from zero (no risk at all of discharge/failure at that point) to infinity (certainty of discharge/failure at that instance). A disadvantage of the use of survival analysis is that the regression coefficients are interpreted as the logarithm of relative *hazard ratios*: the 'hazard' of reaching final length of stay is a concept that is difficult to translate into resource utilisation.

Distribution for time of an event is almost certainly non-symmetric. Parametric survival models are unrestricted and allow for non-normal data patterns. As with standard regression models, there are a range of parametric distributions which can be used which vary in how the baseline hazard is assumed. With the true distribution of length of stay not known, it is advisable to test a number of models as opposed to selecting one method ex-ante (13).

Three commonly employed parametric models were fitted to the data using the following alternative distributions:

- Exponential (Poisson model): constant hazard rate, hence suitable to model LOS when the probability of discharge in the next short time interval does not depend of LOS
- Weibull: generalisation of the exponential model: suitable for modelling data with monotone hazard rates that either increase or decrease exponentially with time. Coefficient higher than 1 means an hazard increases with survival time and thus, lower expected duration
- Generalised gamma: this function is flexible, allowing a large number of possible shapes, encompassing Weibull, exponential, and log-normal as special cases. It is commonly used for evaluating and selecting an appropriate model for data. Here, a negative coefficient is associated with a shorter expected time to discharge

Hazard ratios tell us by how much groups are different: e.g. a hazard ratio of 0.2 means that group 1 has a 80% smaller hazard than the reference category. A hazard ratio of 1.4 means that group 1 has a 40% higher hazard than the reference category. Interpretation of hazard ratios is equal to that of odds ratios.

Analysis plan

Each of the regression and survival models were run independently. In the first instance, all potential predictors of length of stay were included. Coefficients from each model were then assessed, and where model variables failed to show statistically significant ($p<0.05$) association with length of stay in any of the models, this variable was subsequently excluded. The models were then re-run with the reduced number of predictors. Goodness of fit, as reported here, was based on the *later* set of covariates (see Appendix 2).

Discriminating between models

The ability to predict length of stay is assessed using Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC); Root Mean Squared Error (RMSE), and residual plots.

Information criteria are log-likelihood criteria with degrees of freedom adjustment. The AIC is a test between models, which not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. The Bayesian information criterion (BIC) further discourages over-fitting. In both cases, the model with the smallest information criteria is preferred.

Root mean squared error

RMSE is a statistical measure of the magnitude of a varying quantity. It is measured in the same units as the data (here, number of days), rather than in squared units, and so is more sensitive than other measures to any occasional large error, as the squaring process can give disproportionate weight to very large errors. To generate the Root Mean Squared Error (RMSE), the dataset must be split to estimation and validation sets, allowing validation of predictive models. The model derived using the first subsample is used to obtain fitted values for the data in the second subsample.

Graphical presentation

Finally, we compare scatter plots and simple graphs plotting estimated values against the DIPSAT dataset for various regression and duration models. For GLM models, probability plots are used to compare the length of stay distribution with hypothesised distributions (Poisson, gamma, inverse Gaussian). For survival models, we calculate the Cox-Snell residual(14). If the model fits well then a plot of these residuals against the cumulative hazard, with the Cox-Snell residual as the time variable should have slope 1. Plots for each of the parametric models are presented in Appendix 7, together with a 45^{o} line for reference.

Results

We received questionnaire returns from 1421 inpatients in 61 hospitals. Seven of these responses came from participants who had already completed the questionnaire at a previous admission. Excluding

respondents with irretrievable length of stay data (n=199) or missing data exceeding acceptable limits (defined as more than one page of questionnaire left blank) (n=17), we had a sample size of 1210.

Length of stay

The median length of stay was 7 nights. The mean was 12.19 (SD 16.8), due to a small number (n=12) of inpatient stays greater than 100 days.

Participant characteristics

The mean (SD) age of respondents and duration of diabetes were 59.3 (17.4) and 14.9 (12.8) years, respectively. We used age and age^2 to take account of potential nonlinear relationships between age and length of stay.

Reason for admission

Large and significant differences in mean and median length of stay were found between medical, surgical, and orthopaedic surgery patients. Reasons for admission were retrieved retrospectively by hospital staff using hospital admissions databases. In some hospitals full details of reason for admission were not available, whilst a proportion of patients did not consent to having their medical records checked, thus admissions data is available for 45.87% (N = 555) of participants. Where at least 15 participants were found with the same reason for admission, dummy variables were created and added to the dataset. This included the following broad diagnoses: diabetic complications (including foot ulcers, cellulitis); diabetic control (diabetic ketoacidosis; hypoglycaemia; hyperglycaemia); Stroke; Cardiovascular disease; COPD; Maternity; Joint surgery; Amputations; Myocardial infarction; and Cancer care. Additionally, a new dummy “Unknown reason admission” was created.

Diabetes management factors

Approximately 73% of respondents reported using insulin regularly before that admission, whilst for the remaining 27% this stay was their first experience with insulin management. Median LOS ranges were relatively stable depending on who measured blood sugars, however a large increase (1 week) in LOS was observed between self-injecting insulin, compared with insulin administered by nurses. Almost 20% of respondents reported having had a severe hypoglycaemic attack whilst in hospital.

Model estimation and selection

Akaike's and Bayesian information criterion were applied to all models. In the GLM format, lowest values (best fitting) AIC and BIC values were found using Gamma distribution. Inverse Gaussian model gave the second best performance, and performed very marginally better when applying the Root Mean Square Error. Therefore either of these models could be selected. For survival models, the Gamma distribution had the lowest AIC and BIC values and is therefore judged model of best fit.

Graphically, both Gamma and Inverse Gaussian distributions appear superior to the Poisson distribution, as their residuals plot closely along the 45^{o} line. In terms of deviance residuals, the Gamma model performs best, denoted by the most widely dispersed plots of the deviance scatter plots.

For survival data, Cox snell plots were fitted for Poisson, Gamma, and Inverse Gaussian distributions. Again, Gamma and Weibull's distributions were superior to Poisson models. Weibull's survival analysis appears, through this qualitative assessment, to have best fit. Note that this is in contrast to assessment of AIC and BIC values, where highest values in Gamma PH denoted best fit to our dataset.

Econometric results

Appendix 4 reports results (coefficients and standard errors, with p-values) obtained from all regression models. Regression coefficients on sex, marital status, years diagnosed, frequency of injections and type of insulin medication do not follow any clear patterns. Additionally, the study findings suggest no compelling robust evidence that length of stay significantly correlates with inpatient satisfaction. These parameters were subsequently dropped from analysis.

Parameters associated with a highly significant ($p<0.001$) decrease in LOS in all models were participants self-administering own insulin; or their admission being due to diabetes management itself. Conversely, longer length of stay was associated with undergoing surgery; having a severe hypoglycaemic episode during the stay, or having insulin administered by a nurse or through intravenous drip.

Consistency, in terms of statistically significant associations, was high between our models of choice. This is unsurprising as the Gamma distribution was used for both GLM regression and survival modelling. In the Weibull proportional hazards model, being given insulin for the first time during that admission suggested a statistically significant ($p<0.05$) decrease in length of stay. However, this association was not found to be significant in any other model. Similarly, older patients were associated with longer LOS stay in the Inverse Gaussian GLM model, but the significance of this association did not hold in other analyses. Note, contrary to common logic, the age coefficient is not statistically significant in all models. This may owe to the inclusion of other covariates closely related with age, particularly reasons for admission.

Where reasons for admission have been reported, longer hospital spells were strongly associated with stroke, and diabetic foot problems, including amputation. Shorter spells were associated with hypo- and hyperglycaemic episodes (including ketoacidosis), and also maternity admissions. Although the linear regression model was poorly specified, its easily interpretable coefficients are noteworthy. All other things remaining constant, here an admission due to stroke leads to an additional 34.5 bed days ($p<0.001$, constant =16.9), whereas an admission to stabilise diabetes is associated with a reduction in bed days of 9.9 ($p<0.001$). Clearly, the medical reason for hospital admissions will be the greatest single factor in determining length of stay.

Discussion / implications

Six characteristics were found to be significantly associated with length of stay in all models. Self-administering ones own insulin was associated with shorter duration of stay, whereas having insulin administered by a nurse or through intravenous drip was associated with longer stays. Admissions due to diabetes management were comparatively short, whilst patients undergoing surgery had longer stays. An increase or decrease in such variables is accompanied by a significant increase or decrease in inpatient bed days. Models disagreed upon the statistical significance of remaining characteristics and their relationship to LOS. Note also, in all models, overall predictive ability was low.

The Gamma distribution GLM and survival analysis performed best in model estimation, namely the (lowest) value of Akaike's and Bayesian information criterions, Root Mean Square Error, (highest) value of R^2 statistic, and best fit of residual plots (Appendix 3). When assessing the adequacy of models to describe the distribution of length of stay, our conclusions are in agreement with Dodd and colleagues (2006) (4), who modelled the costs incurred in the treatment of inflammatory bowel disease. Elsewhere, inverse Gaussian has been recommended (15), with special reference to psychiatric inpatients, whilst other authors have approached length of stay data from an exponential model (16), also in psychiatric wards. The choice of model will depend on specific data to be analysed, however by applying the necessary estimation tests (AIC, RMSE, residual graphics) the most suitable model can be found with relative ease.

Merits

This analysis provides some insight into why length of stay may vary from one patient to another. As would be expected, the patient's diagnostic reason for admission is the most significant explanatory variable, but within hospital factors over and above the diagnosis, including how insulin is managed, are significant. This information is not collected routinely (e.g. within Hospital Episode Statistics or NHS Reference Costs), therefore the use of survey data is warranted.

Limitations

Knowledge of the factors affecting length of stay can be useful to assist policy making where the goal is to reduce length of stay. However, this should be adopted with caution as the model *does not predict changes in health outcomes,* and is therefore an incomplete analysis. A full decision analysis would involve an economic evaluation of a policy to reduce length of stay. This would draw on not only these results, but also outcomes data.

Unexplained factors

With regard to the goodness of fit of the regression function, the R^2 value was 0.30 in the log(LOS) linear regression, and 0.16 in for linear LOS. This somewhat poor fit may be explained by a number of limitations in our dataset. After controlling for patient, hospital, and diabetes treatment characteristics a

substantial amount of unexplained variation in length of stay remains at departmental level. Xiao (1997) (17) cites the availability of nursing homes; organisation of discharge; availability of interpreters; level of family support; turnaround time for lab/x-ray results; and socioeconomic status of patients all as potential factors in determining length of stay.

Transfers to other hospitals & deaths

A second limitation with this dataset concerns selection bias: we cannot link several episodes, including transfers to different hospitals or destination at discharge. More complex cases are more likely to be transferred to another hospital, thus shifting risks and costs. At the most extreme, a large proportion of diabetic inpatients may have died during the stay, which of course must remain unaccounted for. Patients who die in hospital are most likely to cost the most & cannot be included. On the other hand, those with shorter length of stay may be discharged before any nurse has had a chance to give them the questionnaire.

Modelling very long inpatient stays

It is common practice to eliminate outliers in statistical analyses as firstly, they may be errors, and secondly, their inclusion inflates the variance. If estimates are less precise, the risk of Type II error is increased. However, the validity of this depends on the purpose to which the results are to be put: assuming the outliers are not errors, the patients with the longest length of stay have the biggest impact on the estimate of the mean. Some authors advocate the use of median as measure of central tendency and quantile regression to minimise the impact of outliers(18). However, the most useful statistic for budgetary purposes is the mean, as the total cost is simply the mean multiplied by n. Excluding outliers will bias the mean downwards leading to incorrect inferences on changes in the mean.

The longest length of stay recorded in our dataset is 324 days (stroke rehabilitation). In total twelve patients experienced admissions over 100 days, and 94 were admitted for over 30 days (1% and 7.8% of total sample, respectively). These Influential observations appear to contain valid data and do not lead us to change the data or adjust the model. Sensitivity analyses (not reported here due to space constraints) was conducted to show how inpatient stays over 100 days and 30 days affected our model. Whilst these limits changed the shape of LOS distribution, the overall conclusions in terms of statistically significant parameters were unaffected.

Conclusions

As several competing regression models exist to examine the relationship between patient characteristics and length of stay, it is worthwhile to look at several methodologies before choosing the most appropriate. For our dataset, GLM and survival models using the Gamma distribution family appears to be most suitable. Our study sheds some light on the factors affecting length of stay for inpatients with diabetes.

Points for discussion

- How to decide between survival analysis and standard regression techniques? Are there specific tests? Or is this decision based on qualitative/intuitive assessment of data at hand?
- Debate regarding outliers: Potential merits of median and quantile regression?
- Unclear relationship between inpatient satisfaction and length of stay
- Next steps: More narrow analysis of patients admitted solely due to diabetes related problems, and how small changes to in-hospital insulin management may affect length of stay

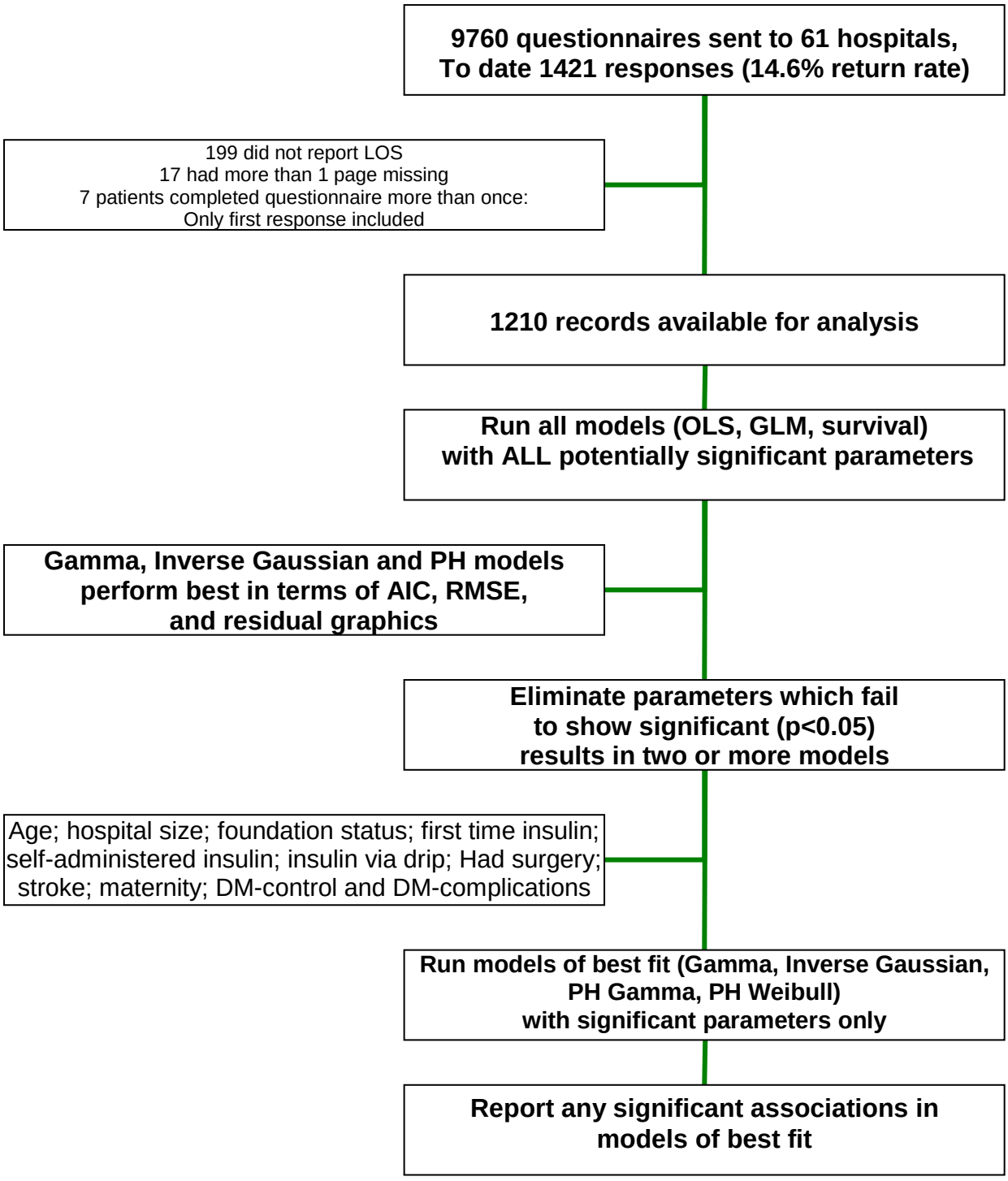
Acknowledgements

We would like to thank members of the DIPSat team: Prof Mike Sampson, Prof Clare Bradley, Dr Claire Rutter, Mrs Christine Jones, Mrs Esther Walden, Dr Ketan Dhatariya, Mrs June James, and nurses involved in all our participating hospitals. We would also like to that all the respondent who took part in the survey. This study is funded by a project grant from DiabetesUK.

Appendix 1: Basic descriptive statistics: Key variables

Total sample size, n=1210				
VARIABLE	**mean**	**Skewness**	**kurtosis**	**Sd**
Length of Stay	12.669	7.323	89.393	20.516
PATIENT CHARACTERISTICS				
AGE	59.313	-0.598	2.677	17.358
MALE SEX	0.557	-0.228	1.052	0.497
SINGLE	0.164	1.810	4.277	0.371
MARRIED	0.591	-0.370	1.137	0.492
REGULAR INSULIN MANAGEMENT				
NORMALLY INSULIN TREATED	0.750	-1.152	2.327	0.433
YEARS DIAGNOSED WITH DIABETES	14.930	1.098	4.230	12.761
YEARS ON INSULIN	9.250	1.751	5.949	12.085
# INSULIN INJECTIONS PER DAY	2.111	0.442	3.026	1.763
THIS STAY 1ST INSULIN USE	0.270	1.035	2.071	0.444
INPATIENT INSULIN MANAGEMENT				
SELF ADMINISTERED INSULIN ONLY	0.426	0.297	1.088	0.495
NURSE ADMINISTERED INSULIN ONLY	0.207	1.443	3.082	0.406
SELF MEASURED BLOOD SUGARS ONLY	0.089	2.881	9.302	0.285
NURSE MEASURED BLOOD SUGARS ONLY	0.612	-0.458	1.210	0.488
INSULIN THROUGH INTRAVENOUS DRIP	0.482	0.073	1.005	0.500
MEAN SATISFACTION SCORE (7=MAX, 0=MIN)	4.542	-0.991	3.858	1.284
SPECIFIED REASONS FOR ADMISSION				
UNKNOWN REASON	0.541	-0.166	1.028	0.498
HAD SURGERY DURING STAY	0.304	0.852	1.725	0.460
ADMITTED DUE TO HIGH BLOOD SUGARS	0.275	1.007	2.013	0.447
ADMITTED DUE TO LOW BLOOD SUGARS	0.056	3.854	15.854	0.230
CARDIOVASCULAR	0.117	2.378	6.654	0.322
DIABETIC KETOACIDOSIS	0.040	4.662	22.736	0.197
FOOT ULCER	0.035	5.084	26.845	0.183
CELLULITUS	0.018	7.212	53.019	0.134
STROKE	0.014	8.258	69.191	0.118
AMPUTATION	0.008	10.863	119.008	0.091

Appendix 2: Framework for analysis



Appendix 3: Prediction Error of various models

Model	AIC	BIC	R-squared	RMSE
Linear Modelling				
OLS Linear regression	9727.922	9893.054	0.1559	19.768
OLS with log transformed Length of Stay	2870.623	3050.428	0.3014	6.564
Generalised Linear Modelling			**(Pseudo R^2)**	
Poisson GLM	6633.363	6774.985	0.0651	24.596
Negative binomial GLM	7460.053	7630.188	0.0636	
Gamma GLM	**3614.143**	**3755.765**	0.0816	24.378
Inverse Gaussian	4288.407	4430.028	0.0834	24.354
Survival Analysis				
Exponential PH	3112.744	3277.876		
Weibull PH	3081.736	3251.871		
Gamma PH	**2870.117**	**3045.256**		

Appendix 4: Association between all treatment characteristics and length of stay using OLS, GLM and survival models

(Reporting coefficient and standard error for each parameter)

* p<0.05, ** p<0.01, ***p<0.001

	Additive models		Multiplicative models				Survival Models		
	OLS	**Log (OLS)**	**Neg Binom**	**Poisson**	**Gamma (log)**	**Inverse Gaussian (log)**	**Weibull**	**Gamma**	**Exponential**
Age	-0.215	0.004*	0.002	0	-.002	0.003	1.01	0.005*	1.01
	0.2	0	0	0	0	0	0	0.1	0
Age^2	0.001	0	0	0	0	0	0.998	0	0.999
	0	0	0	0	0	0	0	0	0
Sex	0.277	-0.048	-0.036	0.023	-0.05	-0.133*	0.053	-0.044	0.05
	-1.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Single	-0.108	0.064	0.002	0.002	0	0.03	0.011	0.071	0
	-1.8	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Satisfaction score	0.525	0.016	0.035	0.042	0.035	0.036	0.976	0.010	0.979
	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Hospital Size	-0.006	-0.000*	-0.000*	0	0	0	0	-0.000*	0
	0	0	0	0	0	0	0	0	0
Foundation status	3.259*	0.065	0.169**	0.245*	0.155	0.077	-0.194**	0.056	-0.155*
	-1.3	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
urban location	-4.919	-0.278*	-0.320**	-0.368	-0.309*	-0.262	0.363**	-0.280*	0.309*
	-2.6	-0.1	-0.1	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1
district Hospital	-5.503*	-0.324**	-0.357**	-0.413*	-0.345*	-0.269	0.408**	-0.332**	0.345*
	-2.8	-0.1	-0.1	-0.2	-0.2	-0.1	-0.1	-0.1	-0.1
[1]Surgical Ward	4.707*	0.062	0.252**	0.314**	0.244*	0.206	-0.308**	0.041	-0.244*
	-1.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
[1]Orthopaedic Ward	0.939	-0.037	0.112	0.108	0.101	-0.007	-0.138	-0.046	-0.101
	-2.7	-0.1	-0.1	-0.2	-0.2	-0.2	-0.1	-0.1	-0.1
Years Diagnosed	-0.061	-0.001	-0.003	-0.005	-0.003	0	0.003	-0.001	0.003
	-0.1	0	0	0	0	0	0	0	0
Regular Insulin	6.008*	0.005	0.175	0.423*	0.151	-0.045	-0.191	-0.01	-0.151
	-2.3	-0.1	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1
# dailys insulin injections	-0.764	-0.015	-0.029	-0.054	-0.028	-0.012	0.033	-0.013	0.028
	-0.5	0	0	0	0	0	0	0	0
First Time Insulin	4.515*	0.151	0.204*	0.315*	0.186	0.061	-0.211*	0.147	-0.186
	-1.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
[2]Self Insulin	**-3.442***	**-0.216*****	**-0.243*****	**-0.262***	**-0.239****	**-0.219*****	**0.272*****	**-0.212*****	**0.239*****
	-1.4	**-0.1**	**-0.1**	**-0.1**	**-0.1**	**-0.1**	**-0.1**	**-0.1**	**-0.1**
[2]Self	-0.23	0.089	-0.026	-0.01	-0.038	-0.102	0.067	0.102	0.038

Bloods	-2.2	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
	1.659	0.151*	0.135*	0.145	0.135	0.161*	-0.155*	0.157**	-0.135
Drip	-1.4	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Blood sugars too low	3.428	0.236*	0.132	0.241*	0.115	0.027	-0.106	0.248*	-0.115
	-2.7	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Blood sugar too high	-0.62	0.095	-0.035	-0.048	-0.035	-0.038	0.063	0.109	0.035
	-1.9	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Had surgery	2.741	0.223**	0.220**	0.182	0.229*	0.274**	-0.262**	0.220**	-0.229*
	-1.7	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Severe Hypo during stay	3.885**	0.321***	0.284***	0.280**	0.279***	0.220**	-0.303**	0.323***	-0.279***
	-1.4	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
Cardiovascular admission	-0.328	-0.089	-0.136	-0.021	-0.152	-0.239*	0.184	-0.078	0.152
	-2.1	-0.1	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.1
Stroke admission	34.490***	0.816***	1.103***	1.234**	1.108**	1.247*	-1.302***	0.797***	-1.108***
	-5.5	-0.2	-0.2	-0.4	-0.4	-0.5	-0.3	-0.2	-0.3
Amputation	20.318**	0.680*	0.793**	0.796**	0.803*	1.025*	-0.938*	0.666*	-0.803*
	-7.1	-0.3	-0.3	-0.3	-0.3	-0.4	-0.4	-0.3	-0.4
MATERNITY	-6.516	**-0.730****	**-0.805****	**-0.757***	**-0.821****	**-0.851*****	**0.925****	**-0.700****	**0.821****
	-5.5	-0.2	-0.3	-0.4	-0.3	-0.2	-0.3	-0.2	-0.3
Admission type unknown	-0.095	-0.089	-0.079	-0.011	-0.092	-0.169*	0.102	-0.086	0.092
	-1.5	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
DM Foot complications	15.884	0.748*	0.399	0.418	0.389	0.198	-0.387	0.793*	-0.389
	-8.3	-0.4	-0.4	-0.3	-0.3	-0.4	-0.4	-0.4	-0.4
DM control admission	-9.853***	-1.089***	-1.233***	-1.296***	-1.219***	-1.235***	1.403***	-1.073***	1.219***
	-2.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
constant	**16.917****	**2.384*****	**2.882*****	**2.808*****	**2.899*****	**3.002*****	**-3.366*****	**2.315*****	**-2.899*****
	-5.4	**-0.2**	**-0.2**	**-0.3**	**-0.3**	**-0.3**	**-0.3**	**-0.2**	**-0.3**

(Reporting coefficient and standard error for each parameter)

* p<0.05, ** p<0.01, ***p<0.001

[1] Omitted variable medical ward

[2] Omitted variables nurse/other measured blood sugars and administered insulin

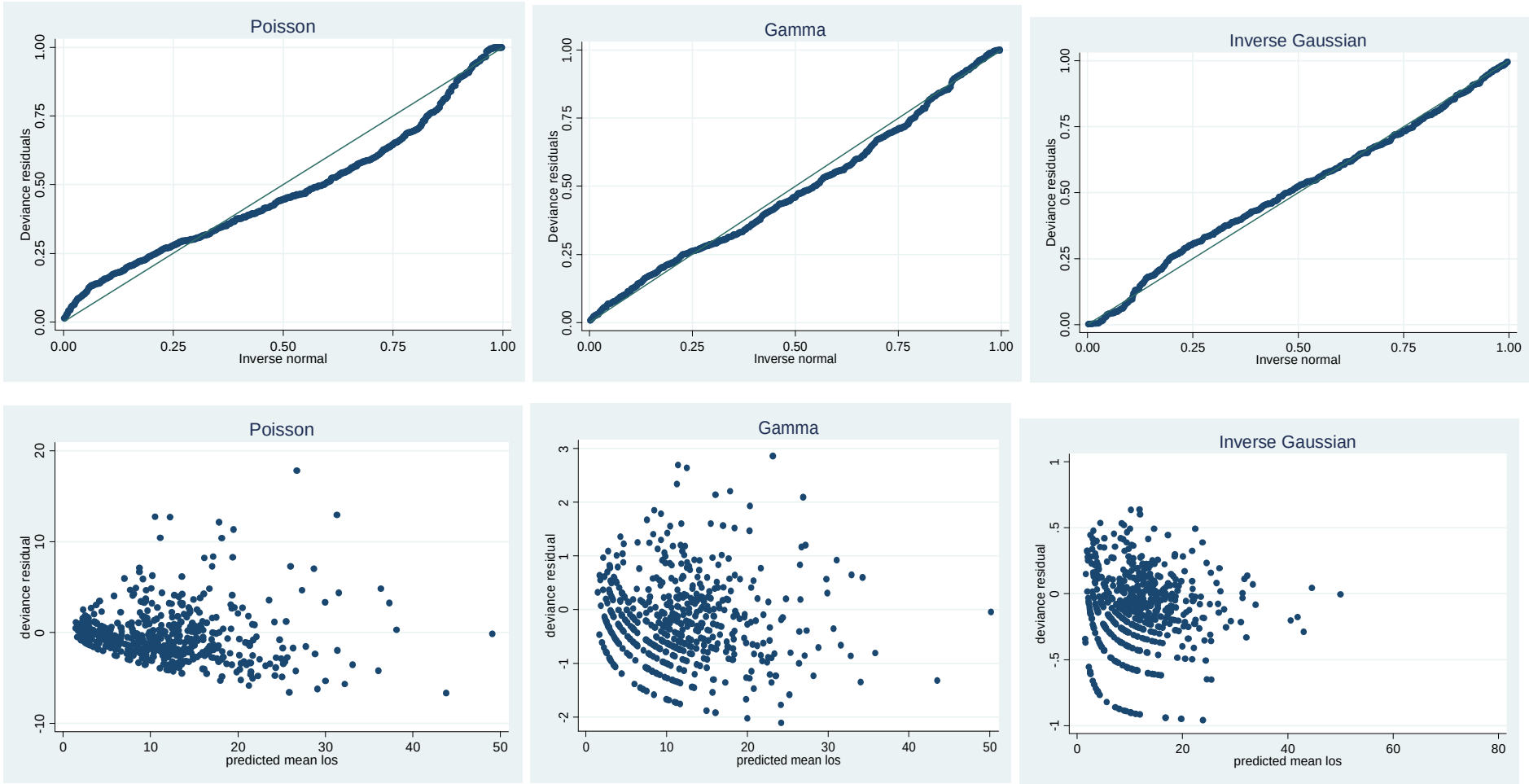
Appendix 5: Re-run analysis eliminating non-significant variables from model

	GLM Gamma	GLM Inverse Gaussian Identity	PH Weibull	PH Gamma	OLS
	b/se	b/se	b/se	b/se	b/se
Age	0.002	0.022*	1.000	0.025	0.002
se	0.0	0.0	0.1	0.0	0.0
Hospital Size	0.000	0.001	1.000	0.000	-0.005
se	0.0	0.0	0.0	0.0	0.0
Foundation status	0.158*	-0.115	0.825**	0.040	3.031*
se	-0.1	-0.3	0.1	0.1	-1.3
First time Insulin	0.133	0.307	0.811*	0.068	2.054
se	-0.1	-0.4	0.1	0.1	-1.4
Self administered insulin	-0.241***	-0.769*	1.315***	-0.252	-3.115*
se	-0.1	-0.3	0.1	0.2	-1.2
Insulin through Drip	0.150*	0.900*	0.856**	-0.080	1.582
se	-0.1	-0.4	0.1	0.2	-1.2
Had surgery	0.320***	3.009*	0.776**	0.273***	4.541***
se	-0.1	-1.2	0.1	0.1	-1.3
Stroke admission	1.183***	33.314	0.272***	1.085***	32.103***
se	-0.3	-27.4	0.1	0.3	-5.0
Maternity admission	-0.911**	-0.843	2.51***	-0.321	-7.878
se	-0.3	-0.7	0.7	0.3	-5.1
DM control admission	-1.282***	-8.758***	4.061***	-0.130	-10.333***
se	-0.1	-0.7	0.5	0.1	-1.6
DM Foot complications	0.420**	5.649	0.669	0.917**	6.902*
se	-0.2	-4.3	0.3	0.4	-2.7
constant	2.444***	10.405***	-3.366	1.164753***	14.071***
se	-0.2	-1.2	-0.2	0.8	-3.4
AIC	8077.8	9389.4	3429.744	3590.3	10486.5
	**** Best fit (AIC)**		**** Best fit (AIC)**		

(Reporting coefficient and standard error for each parameter)

* p<0.05, ** p<0.01, ***p<0.001

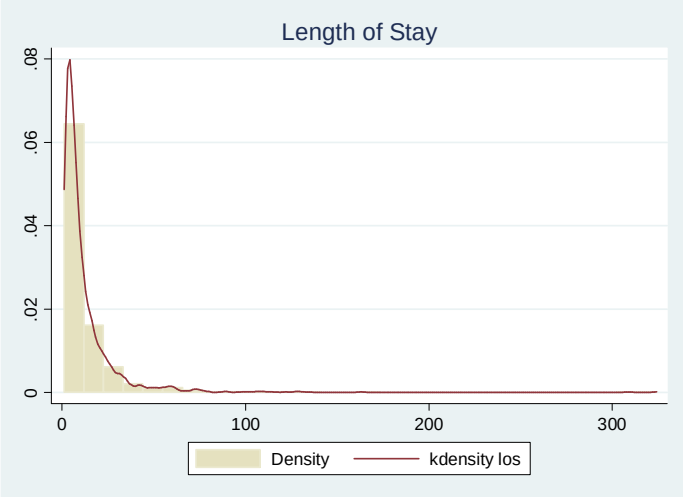
Appendix 5: Residual Graphics



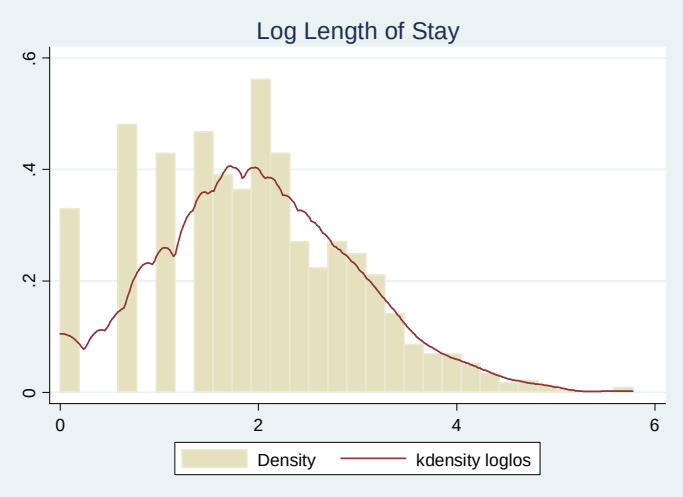
Graphically, both Gamma and Inverse Gaussian distributions appear superior to the Poisson distribution, as their residuals plot closely along the 45^{o} line. In terms of deviance residuals, the Gamma model performs best, denoted by the most widely dispersed plots of the deviance scatterplots.

Appendix 6: Distribution of length of stay, and Log(length of stay) in DIPSat dataset

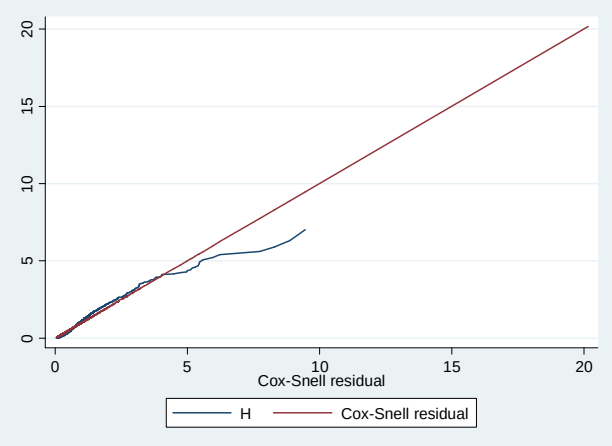
6(a) Distribution of Length of Stay in DIPSat participants



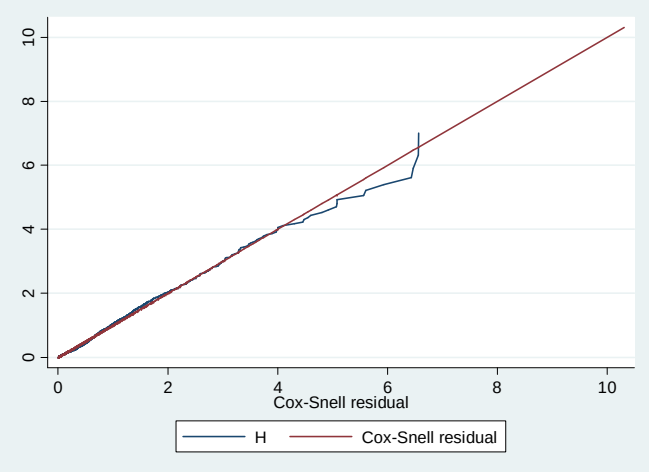
6(b) Distribution of Log(Length of Stay) in DIPSat participants



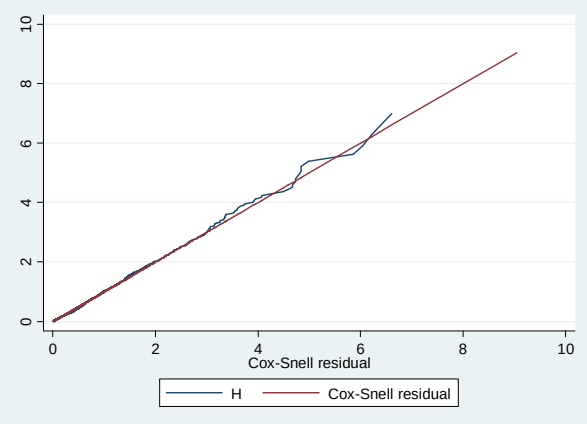
Appendix 7: Cox-Snell residual plots to assess the fit of Survival models



Snell Exponential



Snell Gamma



Snell Weibull's

References

1. Sampson MJ, Crowle T, Dhatariya K, Dozio N, Greenwood RH, Heyburn PJ, et al. Trends in bed occupancy for inpatients with diabetes before and after the introduction of a diabetes inpatient specialist nurse service. Diabet Med. 2006 Sep;23(9):1008-15.
2. Zemencuk JK, Hofer TP, Hayward RA, Moseley RH, Saint S. What effect does physician "profiling" have on inpatient physician satisfaction and hospital length of stay? BMC Health Serv Res. 2006;6:45.
3. Austin PC, Ghali WA, Tu JV. A comparison of several regression models for analysing cost of CABG surgery. Stat Med. 2003 Sep 15;22(17):2799-815.
4. Dodd S, Bassi A, Bodger K, Williamson P. A comparison of multivariable regression models to analyse cost data. J Eval Clin Pract. 2006 Feb;12(1):76-86.
5. Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ. 2005 May;24(3):465-88.
6. Moran JL, Solomon PJ, Peisach AR, Martin J. New models for old questions: generalized linear models for cost prediction. J Eval Clin Pract. 2007 Jun;13(3):381-9.
7. Cameron AC, Trivedi PK. Microeconometrics using Stata. College Station, Texas: Stata Press; 2009.
8. Duan N. Smearing Estimate: A Nonparametric Retransformation Method. Journal of the American Statistical Association. [American Statistical Association]. 1983 Sept 1983;78(383):605-10.
9. Glick HA, Polsky D. Analytic approaches for the evaluation of costs. Hepatology. 1999 Jun;29(6 Suppl):18S-22S.
10. Manning WG. The logged dependent variable, heteroscedasticity, and the retransformation problem. J Health Econ. 1998 Jun;17(3):283-95.
11. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001 Jul;20(4):461-94.
12. Cleves MA, Gould WW, Gutierrez RG. An introduction to survival analysis using STATA. revised ed. College Station, Tex.: Stata Press; 2004.
13. Harrell FE. Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.
14. Cox DR, Snell EJ. A general definition of residuals.
15. Whitmore GA. The inverse Gaussian distribution as a model of hospital stay. Health Serv Res. 1975 Fall;10(3):297-302.
16. Priest RG, Fineberg N, Merson S, Kurian T. Length of stay of acute psychiatric inpatients: an exponential model. Acta Psychiatr Scand. 1995 Oct;92(4):315-7.
17. Xiao J, Douglas D, Lee AH, Vemuri SR. A Delphi evaluation of the factors influencing length of stay in Australian hospitals. Int J Health Plann Manage. 1997 Jul-Sep;12(3):207-18.
18. Lee AH, Fung WK, Fu B. Analyzing hospital length of stay: mean or median regression? Med Care. 2003 May;41(5):681-6.