Predicting Hospital Charges for Inpatient Care in the State of California

The United States health care system has been on a transformative journey over the last few years. The two dominant agents of change are 1) the shift of payment towards value-based care, from volume-based care and relatedly, 2) the shifting of cost from insurers to patients.

The first dynamic, the shifting of towards a value-based payment model, means that hospitals and other care settings, have a distinct motivation to be more efficient in the way they deliver care. The second dynamic, the shifting of more care costs to patients, has given rise to consumerism in health care. Individuals who are now paying more for care have proven to be more interested to shop around and consider cost as a factor before receiving care.

Even with these market dynamics at play, the true cost of health care and what a patient can expect to pay for care is extremely complicated. To attempt to provide a view into what drives hospital charges and examine any differences between hospitals and payer categories, analysis will be run on the 2014 California Inpatient Database, which includes patient discharge data for 3.8 million hospital stays in 2014.

The response variable here is **total charges** and the explanatory variables are:

1. Length of Stay
2. Payer Category (a categorical variable with 9 levels)

Problem Statement

Develop a model that predicts total charges in an acute care setting with consideration for length of stay and the nine categorical levels of payer category.

Constraints and Limitations

It's well known that what hospitals list as charges and what a patient or insurer ends up paying are not the same. Many private insurers have a set percentage of the total charges that they pay as a standard and public insurance pays a set amount, regardless of the actual cost. It should also be said that this is a retrospective, observational study, so no cause and effect relationships between the explanatory and response variables can be determined.

Data Set Description

The response variable is the total charges per patient episode of care. The data cover 3.8 million patient encounters in all of 2014 and the focus is on acute care, or hospital care settings. Also included is a categorical explanatory variables with 9 levels, specifically:

- Payer Category (pay_cat) – this variable details who is the expected source of payment for the encounter. The 8 levels of this variable include:
    - Medicare (01)
    - Medi-Cal (02)
    - Private Coverage (03)
    - Workers' Compensation (04)
    - County Indigent Programs (05)
    - Other Government (06)
    - Other Indigent (07)
    - Self Pay (08)
    - Other Payer (09)

- Length of Stay (los) – fairly self-explanatory, this variable counts the number of days a patient is in the acute care setting.

Exploratory Data Analysis

The descriptive statistics from Figure 1 show that the response variable show fairly consistent standard deviations, with the exception of those who are “Other Government (06)” and “Self-pay (08)”. There appears to be an arbitrary minimum ($1,000) and maximum ($10,000,000) for the Total Charge variable, so the question can be raised if the response variable is truly continuous.

The MEANS Procedure

Payer Category	N Obs	Variable	Label	N	Mean	Maximum	Minimum	Range	Std Dev
01	980897	los	Length of Stay	980897	5.1707	1111.0	0	1111.0	6.9516
		charge	Total Charges	980897	89771.3	10000000	1000.0	9999000	131370
02	1044580	los	Length of Stay	1044580	4.0593	5452.0	0	5452.0	10.7916
		charge	Total Charges	1044580	52175.9	10000000	1000.0	9999000	130064
03	813746	los	Length of Stay	813746	3.6070	630.0	0	630.0	6.4175
		charge	Total Charges	813746	61675.9	10000000	1000.0	9999000	142282
04	16928	los	Length of Stay	16928	3.9221	314.0	0	314.0	7.0805
		charge	Total Charges	16928	111223	7502000	1000.0	7501000	151387
05	7852	los	Length of Stay	7852	5.1445	317.0	0	317.0	11.5290
		charge	Total Charges	7852	75798.9	6134000	2000.0	6132000	165410
06	63916	los	Length of Stay	63916	5.6745	810.0	0	810.0	12.6043
		charge	Total Charges	63916	91069.6	10000000	1000.0	9999000	245627
07	2633	los	Length of Stay	2633	4.1952	139.0	0	139.0	7.2929
		charge	Total Charges	2633	73743.3	2061000	1000.0	2060000	118340
08	96098	los	Length of Stay	96098	3.0230	223.0	0	223.0	4.5907
		charge	Total Charges	96098	43258.9	9378000	1000.0	9377000	89870.4
09	19449	los	Length of Stay	19449	4.1738	387.0	0	387.0	7.3555
		charge	Total Charges	19449	90857.3	7080000	1000.0	7079000	166165

Figure 1 – Means by Payer Categories and Length of Stay

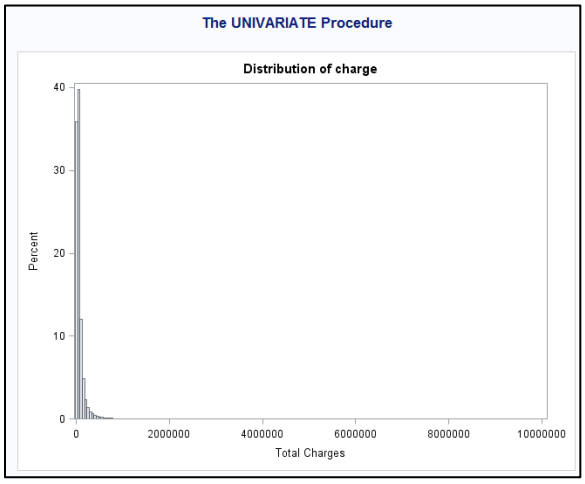

Figure 2 – Distribution of Charges

Figure 2 shows the distribution of charges for the 3.8 million patient visits, with some definite right skewness to be seen. Because the data is unbalanced, the least squares mean function is appropriate instead of arithmetic means. Based on Figure 3 and Figure 4 below, there are statistically significant differences between charges and payer categories at almost each level of the categorical variable, except for those boxed in red (7/1, 7/3 and 76).

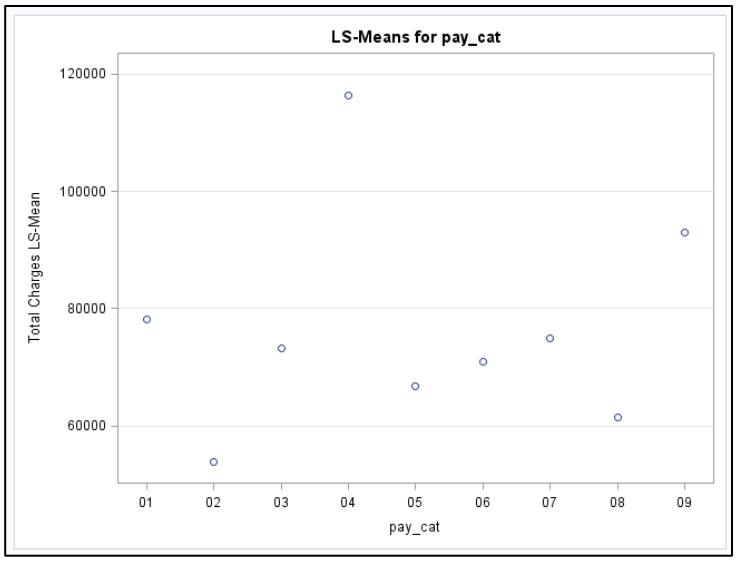

Figure 3 – Least Squares Means of Payer Category and Charges

Least Squares Means for Effect pay_cat
t for H0: LSMean(i)=LSMean(j) / Pr > |t|
Dependent Variable: charge

i/j	1	2	3	4	5	6	7	8	9
1		173.2021 <.0001	33.17285 <.0001	-49.4108 <.0001	10.08235 <.0001	17.91865 <.0001	1.629764 1.0000	48.05139 <.0001	-20.5242 <.0001
2	-173.202 <.0001		-131.071 <.0001	-80.959 <.0001	-11.4447 <.0001	-41.7185 <.0001	-10.9005 <.0001	-21.8918 <.0001	-54.3234 <.0001
3	-33.1728 <.0001	131.0712 <.0001		-55.7646 <.0001	5.670163 <.0001	5.690806 <.0001	-0.93573 1.0000	33.45993 <.0001	-27.3832 <.0001
4	49.41084 <.0001	80.95901 <.0001	55.76461 <.0001		36.40217 <.0001	52.76816 <.0001	19.82466 <.0001	65.73754 <.0001	22.3337 <.0001
5	-10.0824 <.0001	11.44466 <.0001	-5.67016 <.0001	-36.4022 <.0001		-3.41452 0.0230	-3.67124 0.0087	4.543478 0.0002	-19.6476 <.0001
6	-17.9186 <.0001	41.71852 <.0001	-5.69081 <.0001	-52.7682 <.0001	3.414517 0.0230		-2.10045 1.0000	18.18557 <.0001	-27.0997 <.0001
7	-1.62976 1.0000	10.90055 <.0001	0.935728 1.0000	-19.8247 <.0001	3.671244 0.0087	2.100445 1.0000		6.894942 <.0001	-8.69001 <.0001
8	-48.0514 <.0001	21.89176 <.0001	-33.4599 <.0001	-65.7375 <.0001	-4.54348 0.0002	-18.1856 <.0001	-6.89494 <.0001		-40.0298 <.0001
9	20.52418 <.0001	54.32344 <.0001	27.38324 <.0001	-22.3337 <.0001	19.64757 <.0001	27.09972 <.0001	8.690011 <.0001	40.02982 <.0001	

Figure 4 – Least Squared Mean Comparisons

Model Selection

Before splitting the payer category variable to understand if variance in total charges exists between the different types of payers, Figure 5 shows that both length of stay and the payer category seem to contribute to the model. The r-squared value is low at .41, but splitting out the impact of the categorical variables should have a positive impact going forward. The equation for the model is:

$$Y_{Charges} = \beta_0 + \beta_{los} + \beta_{pay\_cat}$$

The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	2.3988989E16	2.6654432E15	240306	<.0001
Error	3.05E6	3.3786874E16	11091886619		
Corrected Total	3.05E6	5.7775862E16			

R-Square	Coeff Var	Root MSE	charge Mean
0.415208	154.8576	105318.0	68009.58

Source	DF	Type I SS	Mean Square	F Value	Pr > F
los	1	2.3575605E16	2.3575605E16	2125482	<.0001
pay_cat	8	4.133835E14	5.1672937E13	4658.62	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
los	1	2.3094751E16	2.3094751E16	2082130	<.0001
pay_cat	8	4.133835E14	5.1672937E13	4658.62	<.0001

Figure 5 – Model Selection

Going an additional step forward to break out the payer categorical variable into its nine levels, the resulting analysis is in Figure 6. The corresponding equations to calculate charges based on length of stay and payer category are found to the right of the image, based on the results found in the table.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	18997.63891	B	822.100882	23.11	<.0001
los	17216.68602	B	97.208957	177.11	<.0001
pay_cat 01	2100.00421	B	831.621637	2.53	0.0116
pay_cat 02	5327.24369	B	828.682836	6.43	<.0001
pay_cat 03	-17371.04002	B	831.822438	-20.88	<.0001
pay_cat 04	37879.15106	B	1201.437640	31.53	<.0001
pay_cat 05	1902.58437	B	1481.320326	1.28	0.1990
pay_cat 06	-11326.10037	B	928.946897	-12.19	<.0001
pay_cat 07	2991.43889	B	2387.922652	1.25	0.2103
pay_cat 08	-18774.44853	B	907.845236	-20.68	<.0001
pay_cat 09	0.00000	B	.	.	.
los*pay_cat 01	-3935.39366	B	98.281942	-40.04	<.0001
los*pay_cat 02	-10355.72744	B	97.628451	-106.07	<.0001
los*pay_cat 03	-568.82622	B	98.723169	-5.76	<.0001
los*pay_cat 04	-3360.30292	B	145.485972	-23.10	<.0001
los*pay_cat 05	-6545.45327	B	137.759087	-47.51	<.0001
los*pay_cat 06	-2519.71740	B	102.121435	-24.67	<.0001
los*pay_cat 07	-4880.20601	B	283.683891	-17.20	<.0001
los*pay_cat 08	-2980.52718	B	119.830137	-24.87	<.0001
los*pay_cat 09	0.00000	B	.	.	.

Figure 6 – Unpacking Categorical Response Variables

For those on Medicare (payer category 1):

$$Y_{Charges} = \$18{,}997 + \$17{,}216(length\ of\ stay)$$

For those on Medi-Cal (payer category 2):

$$Y_{Charges} = \$18{,}997 + \$2{,}100 + (\$17{,}216 - \$3{,}935)(length\ of\ stay)$$

For those on Private Coverage (payer category 3):

$$Y_{Charges} = \$18{,}997 + \$5{,}327 + (\$17{,}216 - \$10{,}355)(length\ of\ stay)$$

For those on Workers Comp (payer category 4):

$$Y_{Charges} = \$18{,}997 - \$17{,}371 + (\$17{,}216 - \$568)(length\ of\ stay)$$

For those on County Indigent Programs (payer category 5):

$$Y_{Charges} = \$18{,}997 + \$37{,}879 + (\$17{,}216 - \$3{,}360)(length\ of\ stay)$$

For those on Other Government (payer category 6):

$$Y_{Charges} = \$18{,}997 + 1{,}902 + (\$17{,}216 - \$6{,}545)(length\ of\ stay)$$

For those on Other Indigent (payer category 7):

$$Y_{Charges} = \$18{,}997 - 11{,}326 + (\$17{,}216 - \$2{,}519)(length\ of\ stay)$$

For those on Self Pay (payer category 8):

$$Y_{Charges} = \$18{,}997 + \$2{,}991 + (\$17{,}216 - \$4{,}880)(length\ of\ stay)$$

For those on Other Payer (payer category 9):

$$Y_{Charges} = \$18{,}997 - \$18{,}774 + (\$17{,}216 - \$2{,}980)(length\ of\ stay)$$

To understand the practical significance of these numbers, it helps to plug in a few variables to calculate the estimated charges. If a 10-day hospital day was assumed, the total charges, per payer would average out to:

Payer	Charges for 10-day Stay
County Indigent	$195,454
Medicare	$191,157
Workers Comp	$168,106
Private	$157,134
Other Indig	$154,641
Medi-Cal	$153,907
Self Pay	$145,348
Other Payer	$142,583
Other Gov Payer	$127,609

Conclusion

Without a doubt, the equation hospitals use to calculate patient charges is complex and with an r-squared value of only .41, it considers more than just payer and length of stay. Interesting to me is the variation in the charges based on payer. Could it be that patient health is different among these groups, causing differences in the medical intervention required? Certainly, more research is needed.

Appendix

```
proc glm data=workingset;
model charge = los | DumMediCal| DumPriv| DumWrkrComp |DumIndig |DumOthrGv |DumOthrInd |DumSlfPy |DumOthrPyr/solution;
run;
```

```
proc glm data=workingset;
class pay_cat;
model charge = los pay_cat;
run;
```

```
proc sgscatter data=workingset;
matrix charge los;
run;
```

```
data workingset;
set "C:\Users\trogers\Desktop\Desktop Items\School Stuff\_StatsII\_Project\finalcut";
DumMediCal=(pay_cat='02');
DumPriv=(pay_cat='03');
DumWrkrComp=(pay_cat='04');
DumIndig=(pay_cat='05');
DumOthrGv=(pay_cat='06');
DumOthrInd=(pay_cat='07');
DumSlfPy=(pay_cat='08');
DumOthrPyr=(pay_cat='09');
run;
```

```
proc print data=workingset(obs=20);
run;
```

```
proc glm data="C:\Users\trogers\Desktop\Desktop Items\School Stuff\_StatsII\_Project\finalcut" PLOTS=(DIAGNOSTICS RESIDUALS);
class pay_cat;
model charge= pay_cat los pay_cat*los;
lsmeans pay_cat / pdiff tdiff adjust=bon;
estimate 'Category and LOS' pay_cat -1 1 0;
estimate 'Category and Interaction' pay_cat -1 0 1;
run;
```

```
proc glm data="C:\Users\trogers\Desktop\Desktop Items\School Stuff\_StatsII\_Project\finalcut";
model charge=pay_cat;
run;
```