

Linear Regression Modeling using the SAS® System: A Real-World Example Comparing the Factors Influencing Charity Care and Bad Debt in Wisconsin Hospitals

Lynnette A. Gerstner, Office of Health Care Information, State of Wisconsin

Introduction

This paper outlines an example of a real-world multiple variable linear regression analysis that was done using the SAS® system. Unlike the fictitious data and analyses used in most manuals and tutorials, it presents the real step-by-step process used in a research project. It is one example of what was done using SAS®, what output was produced, and how the output was used. It is intended to stimulate discussion and expand use of the SAS® system, including SAS/STAT®. It is not intended to be a comprehensive instruction in linear regression analysis or the use of PROC REG. In any research project there are a number of things to consider (e.g., hypothesis, intended audience, available input data, applicability of statistics, intended use of the results) in deciding what approach is the most appropriate.

Research Objectives

Investigate differences in factors influencing the proportion of charity care, bad debt, and total uncompensated health care in Wisconsin GMS hospitals. Provide recommendations for public policy on uncompensated health care.

Study Design

Multivariate least squares regression analysis of system, structure, utilization, and finance variables to predict factors influencing the proportion of charity care, bad debt, and total uncompensated health care provided by a GMS hospital.

The regression procedure used was ordinary least squares. This method results in a line that minimizes the sum of square vertical distances from the observed data points to the regression line that is the best straight (linear) line between the dependent variable and the independent variables; any other line has a larger sum (Schroeder, Sjoquist, and Stephan 1990). Because predicting the best fitting line is a goal of regression, the ordinary least squares procedure is used frequently. The least squares technique requires no distributional assumptions and will fit a line even when the actual theoretical regression curve may not be a straight line (Hays 1981).

Data Definition and Source

The financial, utilization, and staffing data in this analysis are drawn from two surveys conducted annually by the Wisconsin Office of Health Care Information (OHCI): the Hospital Fiscal Survey and the Annual Survey of

Hospitals. Data were also used from OHCI's inpatient discharge file to create the Herfindahl-Hirschman Index. The HCFA Medicare case mix index was obtained from the Federal Register.

In this analysis, three dependent variables were used to test for differences in the factors influencing them:

CHAR_ALL—Charity Care—Health care a hospital provides to a patient who is determined by the hospital to be financially unable to pay all or a portion of the charges.

BADD_ALL—Bad Debt—Claims arising from rendering patient care services that the hospital determines are uncollectable.

UNCO_ALL—Total Uncompensated Health Care—Charity care plus bad debt.

The three dependent variables were each calculated as a percent of total gross patient revenue (e.g., the deduction for charity care divided by total gross patient revenue).

The nine binomial (coded YES=1, NO=0) independent variables used were:

GOVT_OWN—Government Owned Hospital—Hospital controlled by agencies or departments of state, county, or city governments.

RELIG—Owned by Religious Organization—Hospital owned by religious organization (e.g., Catholic hospitals).

NOT_PROF—Nongovernment, Not-for-Profit Hospital—Hospital controlled by nongovernment, not-for-profit organization, such as religious organizations, community hospitals, cooperative hospitals, and hospitals operated by fraternal societies.

SYSTEM—Part of a Health Care System—Hospital that is part of a corporate entity that may own and/or manage health provider hospitals or health related subsidiaries, as well as non-health related hospitals, including freestanding hospitals and/or subsidiary corporations.

NETWORK—Part of a Health Care Network—Hospital that is part of a group of hospitals, clinics, physicians, other health care providers, insurers, and/or community agencies that work together to coordinate and deliver a broad spectrum of services to their communities.

URBAN—Urban Hospital—Hospital located in federally designated metropolitan statistical area.

HILLBURT—Outstanding Hill-Burton Obligation—Hospital with outstanding obligations under the Hill-Burton program, a federal program that, between 1946 and 1974, provided funds to assist in the construction or renovation of public or not-for-profit hospitals. In

return for the funding, hospitals agreed to provide a reasonable amount of care without charge, or at reduced rates, to patients unable to afford health care.

ER_DEPT—Emergency Department—Hospital that has an emergency department, a facility for the provision of unscheduled outpatient services to patients whose conditions are considered to require immediate care. It must be staffed 24 hours a day.

TRAUMA—Self-designated Trauma Center—Hospital that is self-designated to provide emergency and specialized intensive care to critically ill and injured patients.

The 12 continuous independent variables used were:

ADJ_DIS—Adjusted Discharges (Inpatient + Outpatient)—A calculation that adjusts the number of discharges (hospitalizations) to reflect the impact of both inpatient and outpatient volume.

CASE_MIX—Medicare Case Mix Index—A measure of the complexity of the Medicare cases treated by a hospital relative to the complexity of the national average of all Medicare hospital cases, using Diagnosis Related Groups (DRGs) as a measure of relative complexity of treatment.

IR_BEDS—Ratio of Interns and Residents to Beds—The number of interns and residents employed by the hospital divided by the total number of beds set-up and staffed on the last day of the fiscal year.

PCTMDC—Percentage of Medicare Revenue—Medicare gross patient revenue divided by total gross patient revenue.

PCTMAS—Percentage of Medical Assistance Revenue—Medical Assistance (Medicaid) gross patient revenue divided by total gross patient revenue.

PCTHMO—Percentage of HMO Revenue—HMO gross patient revenue divided by total gross patient revenue.

CAPCOMP—Capital Component Percentage of Expenses—Measures the proportion of operating expenses that are capital components. High capital percentages may reflect recent renovation or construction projects, or a greater investment in high technology equipment. Low percentages may reflect either an older physical plant or a tendency not to use debt financing for major projects. It is calculated as: $(\text{Depreciation} + \text{interest expense} + \text{amortization of financing expenses} + \text{rents and leases} + \text{capital component of insurance premiums}) \div \text{total operating expenses}$.

PROFMGN—Total Hospital Profit Margin—Indicates how much the hospital generated (kept) from all sources. A negative number means the hospital operated at a loss. It is calculated as: $\text{Net income} \div (\text{Total revenue} + \text{nonoperating gains} - \text{nonoperating losses})$.

CURATIO—Current Ratio—One of the most widely used measures of liquidity. The higher the ratio value, the better the hospital's ability to meet its current liabilities. A stable current ratio of 1.5 to 2.0 is the

hospital industry norm and is viewed as appropriate by credit rating agencies. It is calculated as: $\text{Current assets} \div \text{Current liabilities}$.

LTD_EQ—Long-term Debt to Equity Ratio—Measures the relative importance of long-term debt in the hospital's permanent capital structure (long-term debt and fund balances). Hospitals with high ratios have relied more heavily on debt than equity and may have difficulty in obtaining future debt financing for major projects. It is calculated as: $\text{Long-term debt} \div \text{Unrestricted fund balance}$.

TOT_TURN—Total Asset Turnover—Measures the relationship between revenue and assets. A high value for this relationship implies that the hospital's total investment is being used efficiently and that a large number of services are being provided to the community from a limited resource base. It is calculated as: $\text{Total revenue} \div \text{Total assets}$.

HHI—Herfindahl-Hirschman Index—Measures the market concentration taking into account the number of competitors and their relative share of the market. It is calculated as the weighted mean of the sum of the squares of the market shares of the competing hospitals in each of the ZIP codes in a hospital's market area. The weight is the proportion of the hospital's discharges from each ZIP code in the market. Values range from 0 (most competitive) to 1 (a monopoly).

Modeling Using the SAS® System

Preliminary Data Examination and Testing for Validity, Multicollinearity, and Outliers

PROC CORR

The inclusion of a large number of independent variables can result in multicollinearity, a high degree of correlation among several variables. PROC CORR was used to identify variables which were highly correlated, indicating a need for further analysis to determine if they were measuring the same effect.

The output also included the minimum, maximum, mean, and standard deviation. For continuous variables, an examination of these statistics can reveal possible data outliers.

Program code:

```
PROC CORR DATA=statlib.char_reg;
  VAR urban govt_own not_prof
      relig system network
      hillburt er_dept trauma
      ir_beds pctmdc pctmas pcthmo
      profmgn ltd_eq curatio
      tot_turn adj_dis case_mix
      hhi;
RUN;
```

Selected parts of the PROC CORR output:

Correlation Analysis								
24 'VAR' Variables: CHAR_ALL BADD_ALL UNCO_ALL URBAN GOVT_OWN NOT_PROF RELIG SYSTEM NETWORK HILLBURT ER_DEPT TRAUMA IR_BEDS PCTMDC PCTMAS PCTEMO PROFMGN CAPCOMP LTD_EQ CURATIO TOT_TURN ADJ_DIS CASE_MIX HHI								
Simple Statistics								
Variable	N	Mean	Std Dev	Minimum	Maximum			
CHAR_ALL	123	0.009000	0.006921	0	0.0373			
BADD_ALL	123	0.0182	0.0135	0.001862	0.1242			
UNCO_ALL	123	0.0272	0.0157	0.006854	0.1519			
URBAN	123	0.4959	0.5020	0	1.0000			
GOVT_OWN	123	0.0407	0.1983	0	1.0000			
NOT_PROF	123	0.9512	0.2163	0	1.0000			
RELIG	123	0.2764	0.4491	0	1.0000			
SYSTEM	123	0.4634	0.5007	0	1.0000			
NETWORK	123	0.2276	0.4210	0	1.0000			
HILLBURT	123	0.1382	0.3465	0	1.0000			
ER_DEPT	123	0.9756	0.1549	0	1.0000			
TRAUMA	123	0.4390	0.4983	0	1.0000			
IR_BEDS	123	0.0139	0.0800	0	0.7225			
PCTMDC	123	0.4660	0.0689	0.2515	0.6360			
PCTMAS	123	0.0749	0.0554	0.0121	0.4586			
PCTEMO	123	0.1134	0.1031	0	0.4158			
CAPCOMP	123	0.0904	0.0231	0.0402	0.1410			
PROFMGN	123	0.0551	0.1353	-1.2849	0.2101			
LTD_EQ	123	6.1011	59.8469	-27.7287	661.8276			
CURATIO	123	2.6806	2.5836	0.4464	21.3079			
TOT_TURN	123	1.0912	0.4821	0.4594	4.5289			
ADJ_DIS	123	7027	6880	621.7173	35975			
CASE_MIX	123	1.2262	0.2282	0.8716	2.0353			
HHI	123	0.3332	0.1683	0.0163	0.9021			
Pearson Correlation Coefficients / Prob > R under Ho: Rho=0 / N = 123								
CHAR_ALL	CHAR_ALL	BADD_ALL	UNCO_ALL	URBAN	GOVT_OWN	NOT_PROF	RELIG	SYSTEM
	1.00000	0.07803	0.50928	0.07677	0.06942	-0.01437	0.06778	0.08031
	0.0	0.3909	0.0001	0.3987	0.4455	0.8747	0.4563	0.3772
BADD_ALL	0.07803	1.00000	0.89772	-0.02802	0.29631	-0.28038	-0.09806	-0.18989
	0.3909	0.0	0.0001	0.7583	0.0009	0.0017	0.2806	0.0354
UNCO_ALL	0.50928	0.89772	1.00000	0.00974	0.28646	-0.24839	-0.05469	-0.12843
	0.0001	0.0001	0.0	0.9149	0.0013	0.0056	0.5480	0.1569
URBAN	0.07677	-0.02802	0.00974	1.00000	-0.03950	0.07365	0.07774	0.28473
	0.3987	0.7583	0.9149	0.0	0.6645	0.4182	0.3927	0.0014
GOVT_OWN	0.06942	0.29631	0.28646	-0.03950	1.00000	-0.90899	-0.12723	-0.19130
	0.4455	0.0009	0.0013	0.6645	0.0	0.0001	0.1608	0.0340
NOT_PROF	-0.01437	-0.28038	-0.24839	0.07365	-0.90899	1.00000	0.13997	0.21045
	0.8747	0.0017	0.0056	0.4182	0.0001	0.0	0.1226	0.0195
RELIG	0.06778	-0.09806	-0.05469	0.07774	-0.12723	0.13997	1.00000	0.44636
	0.4563	0.2806	0.5480	0.3927	0.1608	0.1226	0.0	0.0001
SYSTEM	0.08031	-0.18989	-0.12843	0.28473	-0.19130	0.21045	0.44636	1.00000
	0.3772	0.0354	0.1569	0.0014	0.0340	0.0195	0.0001	0.0
...								

The output from PROC CORR includes the correlation coefficient and the associated *p* value (significance probability for testing the null hypothesis that the corresponding population correlation is zero). In this analysis, any pair of variables with a correlation coefficient greater than 0.195 combined with a $p < 0.05$ was examined. In this analysis CAPCOMP and LTD_EQ were highly correlated, and it was determined they were measuring the same effect. CAPCOMP was dropped from the model.

For the continuous variables, an analysis of the basic descriptive statistics (i.e., min, max, mean, std) was done. The minimum and maximum values were checked for logical values. The mean plus or minus two standard deviations was examined to identify other outliers.

Output from PROC CORR can also be used to identify any missing values. If there are any, the number of pairs with missing values will be given below each of the *p* values. However, there were no missing values in this analysis.

PROC PLOT

Each of the continuous independent variables was plotted against each of the three dependent variables. An examination of the output can reveal if there is an obvious linear relationship between the variables. In many cases the results are not easy to interpret. The general purpose of a linear regression analysis is to determine if there is a relationship, and if so, its significance.

The PROC PLOT output was used to identify data outliers. Unusual outlier points on the PROC PLOT output were investigated to see if they should be dropped from the analysis.

Program code:

```
PROC PLOT DATA=statlib.char_reg;
  PLOT char_all * (ir_beds pctmdc
    pctmas pcthmo profmgn
    ltd_eq curatio tot_turn
    adj_dis case_mix);
  PLOT badd_all * (ir_beds pctmdc
    pctmas pcthmo profmgn
    ltd_eq curatio tot_turn
    adj_dis case_mix);
  PLOT unco_all * (ir_beds pctmdc
    pctmas pcthmo profmgn
    ltd_eq curatio tot_turn
    adj_dis case_mix);
RUN;
```

Examination of the PROC PLOT in this analysis revealed two facilities that did have distinctly different characteristics and performance than that of the other GMS facilities in the model. These two facilities were then excluded from the analysis.

PROC FREQ

For the binomial variables, the output from PROC FREQ was examined to identify variables that may cause problems due to small frequencies in one cell.

```
PROC FREQ DATA=statlib.char_reg;
  TABLE govt_own relig not_prof
    system network urban
    hillburt er_dept trauma;
RUN;
```

Analysis of the variable NOT_PROF showed a frequency of YES=117 and NO=5. Because of the small number of NO responses, it was noted that this may result in unstable estimates for this variable. However, this variable was maintained in the model because of policy interest in not-for-profit status.

Running the Initial Model

PROC REG

Once the data have been examined for validity, multicollinearity, and outliers, the initial regression model can be run using PROC REG. In this analysis, each of the dependent variables was modeled using all the independent variables. The BACKWARD selection option was used to eliminate independent variables from the models. During each step, the partial coefficients are examined, and the variable with the largest *p* value (contributes the least to the fit of the model) is dropped. This step-by-step process continues until all remaining variables are significant at the 0.10 level.

Program code:

```
PROC REG DATA=statlib.char_reg;
  MODEL char_all = urban govt_own
    not_prof relig system network
    hillburt er_dept trauma ir_beds
    pctmdc pctmas pcthmo profmgn
    ltd_eq curatio tot_turn adj_dis
    case_mix hhi
    /SELECTION=BACKWARD;
  MODEL badd_all = urban govt_own
    not_prof relig system network
    hillburt er_dept trauma ir_beds
    pctmdc pctmas pcthmo profmgn
    ltd_eq curatio tot_turn adj_dis
    case_mix hhi
    /SELECTION=BACKWARD;
  MODEL unco_all = urban govt_own
    not_prof relig system network
    hillburt er_dept trauma ir_beds
    pctmdc pctmas pcthmo profmgn
    ltd_eq curatio tot_turn adj_dis
    case_mix hhi
    /SELECTION=BACKWARD;
RUN;
```

Selected parts of the PROC REG Output:

Model: MODEL1

Dependent Variable: CHAR_ALL Charity Care/Total Gross

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	20	0.00253	0.00013	3.899	0.0001
Error	102	0.00331	0.00003		
C Total	122	0.00584			

Root MSE	0.00570	R-square	0.4332
Dep Mean	0.00900	Adj R-sq	0.3221
C.V.	63.32022		

...



Backward Elimination Procedure for Dependent Variable CHAR_ALL

Step 0 All Variables Entered R-square = 0.43324511 C(p) = 21.00000000

...



Step 1 Variable TOT_TURN Removed R-square = 0.43317443 C(p) = 19.01272026

...



Step11 Variable PCTMDC Removed R-square = 0.39503765 C(p) = 5.87627295

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	9	0.00230883	0.00025654	8.20	0.0001
Error	113	0.00353575	0.00003129		
Total	122	0.00584458			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-0.01738657	0.00562780	0.00029864	9.54	0.0025
URBAN	-0.00230306	0.00123614	0.00010861	3.47	0.0650
NOT_PROF	0.00560748	0.00275186	0.00012992	4.15	0.0439
HILLBURT	0.00694177	0.00147720	0.00069097	22.08	0.0001
ER_DEPT	0.00642789	0.00340927	0.00011123	3.55	0.0619
IR_BEDS	0.01397855	0.00770071	0.00010310	3.30	0.0721
PCTMAS	0.03465864	0.01022686	0.00035937	11.49	0.0010
PROFMGN	-0.01423937	0.00440525	0.00032692	10.45	0.0016
LTD_EQ	0.00002115	0.00000854	0.00019177	6.13	0.0148
CASE_MIX	0.01046396	0.00288472	0.00041171	13.16	0.0004

Bounds on condition number: 1.689357, 106.3333

All variables left in the model are significant at the 0.1000 level.

...



The original model for the proportion of charity care had an r-squared (coefficient of multiple determination) value of 0.4332. This means that the independent variables in the model explained 43% of the variation in the proportion of charity care. Because there are multiple independent variables, the value of the adjusted r-squared value was also noted as 0.3221. The adjusted r-squared accounts for the number of variables in the model and adjusts for the cumulative effect that may be caused by adding variables that really do not add to the fit of the model. Maximization of r-squared is not the purpose of regression analysis, and should never be used as the sole determinate of the fit of the model (Schroeder, Sjoquist, and Stephan 1990).

Selecting the Final Model

The final three PROC REG models were run using only those independent variables that were determined to be significant for each of the three dependent variables. The STB option was used to include the standardized regression coefficients in the output. The standardized coefficients measure the change in the dependent variable (measured in standard deviations) that results from one-standard-deviation change in the independent variables.

In this analysis, PROC GLM was not used since it is not normally used for ordinary multiple regression models (Freund and Littell 1991). PROC REG is more efficient and provides more desirable diagnostic and output options.

Program code:

```
PROC REG DATA=statlib.char_reg;
  MODEL char_all = urban not_prof
    hillburt er_dept ir_beds pctmas
    profmgn ltd_eq case_mix /STB;

  MODEL badd_all = system er_dept
    trauma ir_beds pctmdc pcthmo
    profmgn /STB;

  MODEL unco_all = er_dept trauma
    pctmdc pctmas pcthmo profmgn
    /STB;
RUN;
```

Selected parts of the PROC REG Output:

Model: MODEL1						
Dependent Variable: CHAR_ALL Charity Care/Total Gross						
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F	
Model	9	0.00231	0.00026	8.199	0.0001	
Error	113	0.00354	0.00003			
C Total	122	0.00584				
Root MSE		0.00559	R-square	0.3950		
Dep Mean		0.00900	Adj R-sq	0.3469		
C.V.		62.15411				
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T	Standardized Estimate
INTERCEP	1	-0.017387	0.00562780	-3.089	0.0025	0.00000000
URBAN	1	-0.002303	0.00123614	-1.863	0.0650	-0.16704610
NOT_PROF	1	0.005607	0.00275186	2.038	0.0439	0.17522939
HILLBURT	1	0.006942	0.00147720	4.699	0.0001	0.34755055
ER_DEPT	1	0.006428	0.00340927	1.885	0.0619	0.14384344
IR_BEDS	1	0.013979	0.00770071	1.815	0.0721	0.16154228
PCTMAS	1	0.034659	0.01022686	3.389	0.0010	0.27734417
PROFMGN	1	-0.014239	0.00440525	-3.232	0.0016	-0.27828559
LTD_EQ	1	0.000021146	0.00000854	2.476	0.0148	0.18284265
CASE MIX	1	0.010464	0.00288472	3.627	0.0004	0.34496704



Analysis of Significance

The final model for the proportion of charity care had an r-squared value of 0.3950 and an adjusted r-squared value of 0.3459. This final model had a better overall fit than the original model and accounted for 39% of the variation in the dependent variable.

The standardized regression coefficients (standardized estimate) were analyzed along with their corresponding t-values (T for H0). An absolute t-value greater than 1.98 was considered significant. This suggests that the variable is a significant contributor (i.e., has an important influence on the dependent variable).

Conclusion

The result of this analysis supported the hypothesis that the factors influencing charity care and the factors influencing bad debt were different. They should be analyzed separately when evaluating their status and the effect of public policy.

The most important step in any successful regression analysis isn't writing the most elegant program code, but having quality data. Careful definition and examination of the dependent and independent variables is critical. Data should be examined to identify missing values, invalid data, outliers, multicollinearity, and other possible data problems that could invalidate the results.

References:

- Cody, Ronald P. and Jeffrey K. Smith. *Applied Statistics and the SAS® Programming Language*. 1991. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.
- Freund, Ph.D., Rudolf J. and Ramon C. Littell, Ph.D. *SAS® System for Regression*. 1991. 2nd ed. Cary, NC: SAS Institute Inc.
- Hays, William L. *Statistics*. 1981. 3rd ed. New York: Holt, Reinhart, and Winston.
- Schroeder, Larry D., David L. Sjoquist, and Paula E. Stephan. *Understanding Regression Analysis: An Introductory Guide*. 1990. Newbury Park, CA: Sage Publications.

SAS® and SAS/STAT® are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Author:

Lynnette A. Gerstner, Senior Research Analyst
Office of Health Care Information, State of Wisconsin
1 West Wilson Street, Room 372
P.O. Box 7984
Madison, WI 53707-7984
phone: (608) 267-0245
fax: (608) 264-9881
e-mail: lgerstne@mail.state.wi.us

Acknowledgments:

The author acknowledges the hard work and contribution of her colleagues Barbara A. Rudolph, Ph.D. and E. Brian Hutchinson, the co-investigators and co-authors of the original analysis. Dr. Rudolph directed the statistical theory and application, and Mr. Hutchinson provided research background and editorial guidance.

