Foundations of Statistical and Machine Learning for Actuaries -

Classical Regression Modeling

Edward (Jed) Frees, University of Wisconsin - Madison
Andrés Villegas Ramirez, University of New South Wales

July 2025

Schedule

Day and Time	Presenter	Topics	Notebooks for Participant Activity
Monday Morning	Jed	Welcome and Foundations Hello to Google Colab	Auto Liability Claims
	Jed	Classical Regression Modeling	Medical Expenditures (MEPS)
Monday Afternoon	Andrés	Regularization, Resampling, Cross-Validation	Seattle House Sales
	Andrés	Classification	Victoria road crash data
Tuesday Morning	Andrés	Trees, Boosting, Bagging	
	Jed	Big Data, Dimension Reduction and Non-Supervised Learning	Big Data, Dimension Reduction, and Non-Supervised Learning
Tuesday Afternoon	Jed	Neural Networks	Seattle House Prices Claim Counts
	Jed	Graphic Data Neural Networks	MNIST Digits Data
Tuesday 4 pm	Fei	Fei Huang Thoughts on Ethics	
Wednesday Morning	Jed	Recurrent Neural Networks, Text Data	Insurer Stock Returns
	Jed	Artificial Intelligence, Natural Language Processing, and ChatGPT	
Wednesday After Lunch	Dani	Dani Bauer Insights	
Wednesday Afternoon	Andrés	Applications and Wrap-Up	

Monday Morning IB - Classical Regression Modeling

- This module reviews:
  - linear regression,
  - logistic regression, and
  - generalized linear models.
- During lecture, participants may follow
  - Chapter 2 of Loss Data Analytics
  - Chapters 11 and 13 of Frees Regression Book in Spanish
- During lab, participants may follow the notebook Medical Expenditures (MEPS)

Data Analytic Concepts

Underpinning the elements of data analytics are:

- **Data Driven**. Conclusions and decisions made through a data analytic process depend heavily on data inputs.
  - In comparison, econometricians have long recognized the difference between a data-driven model and a structural model.
- **EDA** - exploratory data analysis - and **CDA** - confirmatory data analysis.
  - The purpose of EDA is to reveal aspects or patterns in the data without reference to any particular model.
  - CDA techniques use data to substantiate, or confirm, aspects or patterns in a model.

Statistical Inference: Hypothesis Testing, Estimation and Prediction

- Medical statisticians test the efficacy of a new drug and econometricians estimate parameters of an economic relationship.
- In insurance, predictions of yet to be realized random outcomes are critical for financial risk management (e.g., pricing) of existing risks in future periods.

Comparison of Exploratory Data Analysis and Confirmatory Data Analysis

	EDA	**CDA**
Data	Observational data	Experimental data
Goal	Pattern recognition, formulate hypotheses	Hypothesis testing, estimation, prediction
Techniques	Descriptive statistics, visualization, clustering	Traditional statistical tools of inference, significance, and confidence

Data Modeling

- With a “probability” or “likelihood” based model, our main goal is to understand the target (Y) distribution, typically in terms of the explanatory variables (X).
- Classical data models are particularly useful for:
  - the goal of explanation
  - understanding the uncertainty of our predictions
  - interpretability.
- Let us review three important cases:
  - the normal distribution
  - the Bernoulli (0-1) distribution
  - the exponential family of distributions (for GLM models)

The Normal Distribution

Linear Regression Model Assumptions

Observable Data Representation

$E\left[y_i\right] = \mu_i$	regression mean
$\quad = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$	
$\{x_{i1}, \ldots, x_{ik}\}$	non-stochastic explanatory variables
$\beta_0, \beta_1, \ldots, \beta_k$	unknown regression parameters
$\mathrm{Var}[y_i] = \sigma^2$	regression variance
$\{y_1, \ldots, y_n\}$	independent random variables
$\{y_1, \ldots, y_n\}$	normally distributed

Interpretability is key. For example,

$$\beta_j = \frac{\partial \mathrm{E}[y]}{\partial x_j}$$

we can think about β_j as the expected change in y per unit change in x_j, holding other explanatory variables fixed.

Model Fitting

It is customary to fit a regression model using the **method of maximum likelihood**.

- The joint probability density (mass) function is viewed as a function of the realized data, with the parameters held fixed.
- In contrast, the likelihood is viewed as a function of the parameters, with the data held fixed.
- The method of maximum likelihood means finding the values of β that maximize the likelihood.

Model Fitting 2

- In the benchmark (standard), observations are independent and so the joint density is a product of marginal densities.
  - Determining arguments that maximize a function yield the same results achieved when maximizing the log of the function.
  - Method of maximum likelihood, find the values of the parameters $\boldsymbol{\theta}$ that maximize the log-likelihood

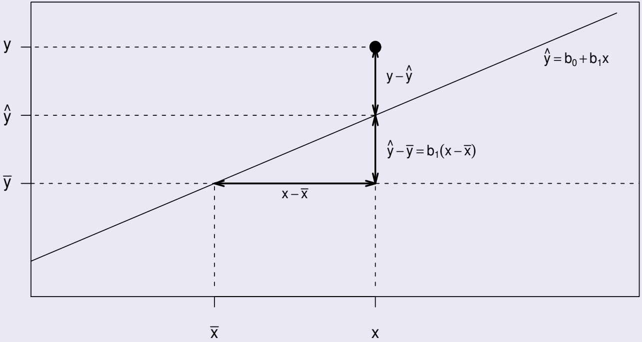
$$L(\boldsymbol{\theta}) = L(\beta, \sigma^2) = \sum_{i=1}^{n} \left\{ \log f(y_i; \mathbf{x}_i' \beta, \sigma^2) \right\}.$$

Here, f is a normal distribution. Let us call the values that maximize this $\boldsymbol{\theta}_{MLE}$.

How well does the model fit?

With the estimated regression coefficients, say β_{MLE}, one can compute the fitted values $\hat{y}_i = \mathbf{x}_i'\beta_{MLE}$.

How close are the fitted values to the observed values y?



$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{Total\ \ SS} = \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{Error\ \ SS} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{Regression\ \ SS}$$

Define R-square (*coefficient of determination*):

$$R^2 = \frac{Regression\ SS}{Total\ SS}$$

to be the proportion of variability explained by the model.

Model Adequacy and Goodness of Fit

Naturally, there are many other measures for how well a fitted model fits the training data, including

- t-statistics of individual coefficients
- a version of R^2, R_a^2, that is adjusted for model complexity
  - Information criteria = measure of fit plus penalty for model complexity, e.g.
  - $AIC = -2 \times \text{log-likelihood} + 2 \times \text{number of parameters}$
  - smaller is better
- Residual analysis

How reliable are the estimated coefficients?

Inference – Standard errors

An estimator of the asymptotic variance of $\boldsymbol{\theta}$ may be calculated taking partial derivatives of the log-likelihood.

$$\mathbf{I}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} L(\boldsymbol{\theta}).$$

It is known as the *information matrix*.

The square root of the jth diagonal element of this matrix evaluated at $\boldsymbol{\theta}_{MLE}$ yields the standard error for $\theta_{j,MLE}$.

Example: Medical Expenditures

- Data from the Medical Expenditure Panel Survey (MEPS), conducted by the U.S. Agency of Health Research and Quality (AHRQ).
  - A probability survey that provides nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian population.
  - Collects detailed information on individuals of each medical care episode by type of services including
    - physician office visits,
    - hospital emergency room visits,
    - hospital outpatient visits,
    - hospital inpatient stays,
    - all other medical provider visits, and
    - use of prescribed medicines.

MEPS

- For MEPS, inpatient admissions include persons who were admitted to a hospital and stayed overnight.
- In contrast, outpatient events include hospital outpatient department visits, office-based provider visits and emergency room visits excluding dental services.
  - Hospital stays with the same date of admission and discharge, known as *zero-night stays*, were included in outpatient counts and expenditures.
  - Payments associated with emergency room visits that immediately preceded an inpatient stay were included in the inpatient expenditures.
  - Prescribed medicines that can be linked to hospital admissions were included in inpatient expenditures, not in outpatient utilization.

MEPS

- This detailed information allows one to develop models of health care utilization to predict future expenditures.
- We consider MEPS data from the first panel of 2003 and take a random sample of $n = 2,000$ individuals between ages 18 and 65.

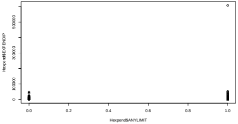
Let's try fitting a regression model!

A regression model for fitting inpatient expenditures (EXPENDIP) produces poor results. For example, The $R^2 < 2\%$.

	Coefficient	Standard Error	t-Statistic
(Intercept)	1349.05	2258.51	0.60
AGE	-38.97	26.15	-1.49
GENDER	-411.57	640.50	-0.64
factor(RACE)BLACK	213.34	1775.94	0.12
factor(RACE)NATIV	-44.84	3373.79	-0.01
factor(RACE)OTHER	-228.92	2896.10	-0.08
factor(RACE)WHITE	384.70	1572.54	0.24
factor(REGION)NORTHEAST	-1850.44	1099.75	-1.68
factor(REGION)SOUTH	-1723.80	893.71	-1.93
factor(REGION)WEST	-2138.81	946.93	-2.26
factor(EDUC)HIGHSCH	109.87	805.44	0.14
factor(EDUC)LHIGHSC	1516.30	970.07	1.56
factor(PHSTAT)FAIR	-263.28	1294.16	-0.20
factor(PHSTAT)GOOD	1180.45	876.49	1.35
factor(PHSTAT)POOR	4311.43	1971.09	2.19

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
AGE	1	8727179.28	8727179.28	0.04	0.83
GENDER	1	34854272.62	34854272.62	0.18	0.67
factor(RACE)	4	79472284.83	19868071.21	0.10	0.98
factor(REGION)	3	1131118494.70	377039498.23	1.91	0.13
factor(EDUC)	2	709807132.76	354903566.38	1.80	0.17
factor(PHSTAT)	4	2082459376.03	520614844.01	2.64	0.03
MNHPOOR	1	14255471.25	14255471.25	0.07	0.79
ANYLIMIT	1	1725581242.09	1725581242.09	8.76	0.00
factor(INCOME)	4	431499864.93	107874966.23	0.55	0.70
insure	1	441684091.58	441684091.58	2.24	0.13
Residuals	1977	389273542465.67	196901134.28	NA	NA

The statistical significance of ANYLIMIT is due to a single observation.



Back to the Beginnings

- As is common with actuarial data sets,such as the MEPS data, most subjects have zero expenditures.
- When expenditures do occur, they tend to be long-tail.
- Although regression methods work well with non-normal data (in part due to central limit theorems), they do not give reasonable results for data that contains massive amounts of zeroes mixed with long-tailed out outcomes.
- So, the approach is to think instead about different components of the response variable (in this case health care expenditures).

Table 11.4. **Positive Expenditures by Explanatory Variable**

Category	**Variable**	**Description**	**Percent of data**	**Percent Positive Expend**
Demography	*AGE*	Age in years between 18 to 65 (mean: 39.0)		
	GENDER	1 if female	52.7	10.7
		0 if male	47.3	4.7
Ethnicity	*ASIAN*	1 if Asian	4.3	4.7
	BLACK	1 if Black	14.8	10.5
	NATIVE	1 if Native	1.1	13.6
	WHITE	Reference leve*l*	79.9	7.5
Region	*NORTHEAST*	1 if Northeast	14.3	10.1
	MIDWEST	1 if Midwest	19.7	8.7
	SOUTH	1 if South	38.2	8.4
	WEST	Reference level	27.9	5.4
Education	COLLEGE	1 if college or higher degree	27.2	6.8
	HIGHSCHOOL	1 if high school degree	43.3	7.9
	Reference level is lower than high school degree		29.5	8.8
Self-rated	*POOR*	1 if poor	3.8	36.0
physical	*FAIR*	1 if fair	9.9	8.1
health	*GOOD*	1 if good	29.9	8.2
	VGOOD	1 if very good	31.1	6.3
	Reference level is excellent health		25.4	5.1
Self-rated	*MNHPOOR*	1 if poor or fair	7.5	16.8
mental health		0 if good to excellent mental health	92.6	7.1
Any activity	*ANYLIMIT*	1 if any functional/activity limitation	22.3	14.6
limitation		0 if otherwise	77.7	5.9
Income	*HINCOME*	1 if high income	31.6	5.4
compared to	*MINCOME*	1 if middle income	29.9	7.0
poverty line	*LINCOME*	1 if low income	15.8	8.3

The Bernoulli Distribution

- y_i has a Bernoulli distribution, resulting in a 0 or 1 outcome.
  - The probability that the response equals 1 by $\pi_i = \Pr(y_i = 1)$.
  - The mean response is
    $\mathrm{E}\ y_i = 0 \times \mathrm{Pr}(y_i = 0) + 1 \times \mathrm{Pr}(y_i = 1) = \pi_i$.
  - Thus, the variance is related to the mean through the expression $\mathrm{Var}\ y_i = \pi_i(1 - \pi_i)$.

Logistic and probit regression models

The approach is to use a **known** nonlinear function of the explanatory variables

- The linear combination of explanatory variables, $\mathbf{x}_i'\beta$, is sometimes known as the *systematic component.*
- We consider a function of explanatory variables, $\pi_i = \pi(\mathbf{x}_i'\beta) = \Pr(y_i = 1|\mathbf{x}_i)$.
- We focus on two special cases of the function $\pi(.)$:
  - $\pi(z) = \frac{1}{1+\exp(-z)} = \frac{e^z}{1+e^z}$, the logit case, and
  - $\pi(z) = \Phi(z)$, the probit case.
  - $\Phi(.)$ is the standard normal distribution function.
- Note that $\pi(z) = z$ yields the linear probability model.
- The inverse of the function, π^{-1}, is linear in the explanatory variables, that is, $\pi^{-1}(\pi_i) = \mathbf{x}_i'\beta$.
- The logit and probit are really close.

Model Fitting - Logistic Regression

- Define $\pi_i = \pi(\mathbf{x}_i'\beta)$, the probability of a one for the ith subject.
- The log-likelihood of a single observation is

$$\begin{cases} \ln \pi(\mathbf{x}_i'\beta) & \text{if } y_i = 1 \\ \ln\left(1 - \pi(\mathbf{x}_i'\beta)\right) & \text{if } y_i = 0 \end{cases}$$
$$= y_i \ln \pi(\mathbf{x}_i'\beta) + (1 - y_i) \ln\left(1 - \pi(\mathbf{x}_i'\beta)\right).$$

- Assuming independence, the log-likelihood of the data set is

$$L(\beta) = \sum_{i=1}^{n} \left\{ y_i \ln \pi(\mathbf{x}_i'\beta) + (1 - y_i) \ln\left(1 - \pi(\mathbf{x}_i'\beta)\right) \right\}.$$

Maximum Likelihood - Logistic Regression

- The customary method of finding the maximum is taking partial derivatives with respect to the parameters of interest and finding roots of the these equations.
- In this case, taking partial derivatives with respect to β yields the *score equations*

$$\frac{\partial}{\partial\beta}L(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \pi(\mathbf{x}_i'\beta)\right) \frac{\pi'(\mathbf{x}_i'\beta)}{\pi(\mathbf{x}_i'\beta)(1-\pi(\mathbf{x}_i'\beta))} = \mathbf{0}.$$

- The solution of these equations, say $\mathbf{b}_{MLE}$, is the maximum likelihood estimator.

- To illustrate, for the logit case, the score equations reduce to

$$\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \pi(\mathbf{x}_i' \boldsymbol{\beta}) \right) = \mathbf{0}.$$

  where $\pi(z) = 1/(1 + \exp(-z))$.
- When the model contains an intercept term, we can write the first row of this expression as $\sum_{i=1}^{n} \left(y_i - \pi(\mathbf{x}_i' \mathbf{b}_{MLE}) \right) = 0$, so the sum of observed values equals the sum of fitted values.

Paul Embrechts and Mario V. Wüthrich emphasize the "balance property" as key...

MEPS

Table 11.5. **Comparison of Binary Regression Models**

	Logistic		Logistic		Probit	
	Full Model		Reduced Model		Reduced Model	
	Parameter		Parameter		Parameter	
Effect	Estimate	*t*-ratioo	Estimate	*t*-ratio	Estimate	*t*-ratio
Intercept	−4.239	−8.982	−4.278	−10.094	−2.281	−11.432
AGE	−0.001	−0.180				
GENDER	0.733	3.812	0.732	3.806	0.395	4.178
ASIAN	−0.219	−0.411	−0.219	−0.412	−0.108	−0.427
BLACK	−0.001	−0.003	0.004	0.019	0.009	0.073
NATIVE	0.610	0.926	0.612	0.930	0.285	0.780
NORTHEAST	0.609	2.112	0.604	2.098	0.281	1.950
MIDWEST	0.524	1.904	0.517	1.883	0.237	1.754
SOUTH	0.339	1.376	0.328	1.342	0.130	1.085
COLLEGE	0.068	0.255	0.070	0.263	0.049	0.362
HIGHSCHOOL	0.004	0.017	0.009	0.041	0.003	0.030
POOR	1.712	4.385	1.652	4.575	0.939	4.805
FAIR	0.136	0.375	0.109	0.306	0.079	0.450
GOOD	0.376	1.429	0.368	1.405	0.182	1.412
VGOOD	0.178	0.667	0.174	0.655	0.094	0.728
MNHPOOR	−0.113	−0.369				
ANYLIMIT	0.564	2.680	0.545	2.704	0.311	3.022
HINCOME	−0.921	−3.101	−0.919	−3.162	−0.470	−3.224
MINCOME	−0.609	−2.315	−0.604	−2.317	−0.314	−2.345
LINCOME	−0.411	−1.453	−0.408	−1.449	−0.241	−1.633
NPOOR	−0.201	−0.528	−0.204	−0.534	−0.146	−0.721
INSURE	1.234	4.047	1.227	4.031	0.579	4.147
$Log - Likelihood$	−488.69		−488.78		−486.98	
AIC	1, 021.38		1, 017.56		1, 013.96	

Linear Exponential Family of Distributions

GLM Ingredients

- This extension of the linear model is so widely used that it is known as **the** *generalized linear model*, or as the acronym GLM.
- GLM generalizes the linear model in three ways:
- 1. Mean as a function of linear predictors
    - Call the linear combination of explanatory variables the *systematic component*, denoted as $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$
    - The **link** function relates the mean to the systematic component

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} = \mathrm{g}\left(\mu_i\right).$$

    - g(.) a smooth, invertible function. The inverse $\mu_i = \mathrm{g}^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$, is the mean function.
    - Some examples we have seen:
        - $\mathbf{x}_i'\boldsymbol{\beta} = \mu_i$, for (normal) linear regression,
        - $\mathbf{x}_i'\boldsymbol{\beta} = \exp(\mu_i)/(1+\exp(\mu_i))$, for logistic regression and
        - $\mathbf{x}_i'\boldsymbol{\beta} = \ln(\mu_i)$, for Poisson regression.

GLM Ingredients II

- 2. The GLM extends linear modeling through the use of the **linear exponential family** of distribution
  - *Not* the exponential distribution - it is a generalization.
  - This family includes the normal, Bernoulli and Poisson distributions as special cases.
- 3. GLM modeling is robust to the choice of distributions.
  - The linear model sampling assumptions focused on:
    - the form of the mean function (assumption F1),
    - non-stochastic or exogenous explanatory variables (F2),
    - constant variance (F3) and
    - independence among observations (F4).
  - GLM models maintain assumptions F2 and F4
  - GLM models extend F1 through the link function.
  - To extend F3, the variance depends on the choice of distributions

Linear Exponential Family of Distributions

- *Definition.* The distribution of the *linear exponential family* is

$$\mathrm{f}(y;\theta,\phi)=\exp\left(\frac{y\theta-b(\theta)}{\phi}+S\left(y,\phi\right)\right).$$

- y is a dependent variable and θ is the parameter of interest.
- ϕ is a scale parameter, often assumed known.
- $b(\theta)$ depends only on the parameter θ, not the dependent variable.
- $S(y,\phi)$ is a function of y and the scale parameter, not the parameter θ.
- Example: Normal distribution - use $\theta=\mu$ and $\phi=\sigma^2$,

$$\begin{aligned}\mathrm{f}(y;\mu,\sigma^2) &= \tfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\tfrac{(y-\mu)^2}{2\sigma^2}\right)\\ &= \exp\left(\tfrac{(y\mu-\mu^2/2)}{\sigma^2}-\tfrac{y^2}{2\sigma^2}-\tfrac{1}{2}\ln\left(2\pi\sigma^2\right)\right).\end{aligned}$$

* Also $b(\theta)=\theta^2/2$ and $S\left(y,\phi\right)=-y^2/(2\phi)-\ln\left(2\pi\sigma^2\right)/2$.

MEPS Fit

Table 13.5. **Median Expenditures by Explanatory Variable Based on a Sample of $n = 157$ with Positive Expenditures**

Category	Variable	Description	Percent of data	Median Expend
	COUNTIP	Number of expenditures (median: 1.0)		
Demography	*AGE*	Age in years between 18 to 65 (median: 41.0)		
	GENDER	1 if female	72.0	5,546
		0 if male	28.0	7,313
Ethnicity	*ASIAN*	1 if Asian	2.6	4,003
	BLACK	1 if Black	19.8	6,100
	NATIVE	1 if Native	1.9	2,310
	WHITE	*Reference level*	75.6	5,695
Region	*NORTHEAST*	1 if Northeast	18.5	5,833
	MIDWEST	1 if Midwest	21.7	7,999
	SOUTH	1 if South	40.8	5,595
	WEST	*Reference level*	19.1	4,297
Education	*COLLEGE*	1 if college or higher degree	23.6	5,611
	HIGHSCHOOL	1 if high school degree	43.3	5,907
		Reference level is lower than high school degree	33.1	5,338
Self-rated physical health	*POOR*	1 if poor	17.2	10,447
	FAIR	1 if fair	10.2	5,228
	GOOD	1 if good	31.2	5,032
	VGOOD	1 if very good	24.8	5,546
		Reference level is excellent health	16.6	5,277
Self-rated mental health	*MPOOR*	1 if poor or fair	15.9	6,583
		0 if good to excellent mental health	84.1	5,599

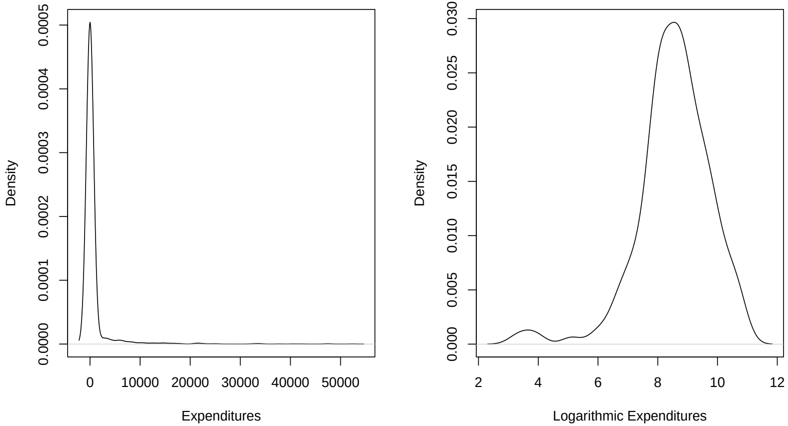

Density
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0
10000
20000
30000
40000
50000
Expenditures
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
2
4
6
8
10
12
Logarithmic Expenditures

Table 13.6. **Comparison of Gamma and Inverse Gaussian Regression Models**

	Gamma		Gamma		Inverse	Gaussian
	Full	Model	Reduced	Model	Reduced	Model
Effect	*Parameter Estimate*	*t*-value	*Parameter Estimate*	*t*-value	*Parameter Estimate*	*t*-value
Intercept	6.891	13.080	7.859	17.951	6.544	3.024
COUNTIP	0.681	6.155	0.672	5.965	1.263	0.989
AGE	0.021	3.024	0.015	2.439	0.018	0.727
GENDER	−0.228	−1.263	−0.118	−0.648	0.363	0.482
ASIAN	−0.506	−1.029				
BLACK	−0.331	−1.656	−0.258	−1.287	−0.321	−0.577
NATIVE	−1.220	−2.217				
NORTHEAST	−0.372	−1.548	−0.214	−0.890	0.109	0.165
MIDWEST	0.255	1.062	0.448	1.888	0.399	0.654
SOUTH	0.010	0.047	0.108	0.516	0.164	0.319
COLLEGE	−0.413	−1.723	−0.469	−2.108	−0.367	−0.606
HIGHSCHOOL	−0.155	−0.827	−0.210	−1.138	−0.039	−0.078
POOR	−0.003	−0.010	0.167	0.706	0.167	0.258
FAIR	−0.194	−0.641				
GOOD	0.041	0.183				
VGOOD	0.000	0.000				
MNHPOOR	−0.396	−1.634	−0.314	−1.337	−0.378	−0.642
ANYLIMIT	0.010	0.053	0.052	0.266	0.218	0.287
MINCOME	0.114	0.522				
LINCOME	0.536	2.148				
NPOOR	0.453	1.243				
POORNEG	−0.078	−0.308	−0.406	−2.129	−0.356	−0.595
INSURE	0.794	3.068				
Scale	1.409	9.779	1.280	9.854	0.026	17.720
Log − Likelihood	−1, 558.67		−1, 567.93		−1, 669.02	
AIC	3, 163.34		3, 163.86		3, 366.04	

Session IB - Classical Regression Modeling Summary

In this module, we:

- reviewed some basic data analytic concepts such as summarized in Chapter 2 of Loss Data Analytics
- reviewed some fundamental regression models such as summarized in Chapters 11 and 13 of Frees Regression Book in Spanish
  - These models include linear regression, logistic regression, and generalized linear models.
- During this workshop, we build and extend these fundamental models using statistical and machine learning techniques
  - (*Hint of thing to come* - as we have seen, regression models employ extensively categorial variables. We intend to suggest that these can be done more efficiently using machine learning **embedding** concepts).
- During lab, participants may follow the notebook Medical Expenditures (MEPS)

Resources For Future Studies

- Regression Modeling with Actuarial and Financial Applications, Frees, 2010
  - Frees Regression Book in Spanish
- James et al. (2023), An Introduction to Statistical Learning with Applications in Python