
بسم الله الرحمن الرحیم

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

پایان‌نامه کارشناسی ارشد

آمار ریاضی

عنوان:

یک برآوردگر بسیار استوار برای مدل‌های رگرسیونی

نگارنده: طاهره نصراله‌زاده ممقانی

استاد راهنما: دکتر صادق رضایی

استاد مشاور: دکتر سعید رضاخواه

پاییز ۱۳۸۶

الهي!

الهی تا آموختن را آموختم، آموخته را جمله بسوختم، اندوخته را برانداختم و انداخته را بیندوختم نیست را بفروختم تا هست را بیفروختم.

الهی اگر کسی تو را به جستن یافت من تو را بگریختن یافتم، اگر کسی تو را به ذکر کردن یافت من تو را بخود فراموش کردن یافتم، اگر کسی تو را بطلب یافت من خود طلب از تو یافتم، خدایا وسیلت به تو هم تویی، اول تو بودی و آخر هم تویی.

الهی روزگاری تو را میجستم خود را می یافتم ، اکنون خود را می جویم تو را می یابم ای محب را یاد و انس را یادگار، چون حاضری این جستن به چه کار؟

الهی دانایی ده که در راه نیفتیم و بینایی ده که در چاه نیفتیم.

خواجه عبدالله انصاری

و

سپاس بی پایان نثار آنانی می کنم که در حلقه مهر خویش بینایی بندگی ، توانایی اندیشه و لذت تلاش را سرلوحه قدمهایم نمودند:

از موهبات بی دریغ خانواده، الطاف بی شائبه اساتید گرامی و مهربانی دوستان و...

فهرست مطالب

چکیده:

استخراج الگوها و مدل‌های مطلوب از مجموعه داده‌های بزرگ توجه بسیاری را در رشته‌های مختلف بخود جلب کرده است، در این خصوص استخراج اطلاعات مفید از پایگاه‌های داده[1] و داده‌کاوی[2] دو زمینه جالب توجه برای محققین در شناسایی الگوها، آمار، هوش مصنوعی و خصوصاً محاسبات در سطوح بالا ایجاد کرده است.

در این پایان نامه یک روش کارا و استوار به نام تجزیه کلاس رگرسیونی آمیخته[3] برای استخراج کلاس‌های رگرسیونی در مجموعه داده‌های بزرگ خصوصاً در شرایط آلوده ارائه می‌گردد. کلاس رگرسیونی که به عنوان یک زیر مجموعه از مجموعه داده‌هایی تعریف می‌شود که موضوع اصلی در مدل رگرسیونی است مطرح می‌گردد, آنگاه یک مجموعه از داده‌های درونی به هر کدام از این کلاس‌های رگرسیونی اختصاص می‌یابد و در نهایت مدل‌های رگرسیونی معنی‌دار در مجموعه داده‌ها تعیین می‌گردد. مجموعه داده‌های بزرگ به عنوان یک جامعه آمیخته مورد بحث قرار می‌گیرد که در آن تعداد زیاد و متناهی کلاس رگرسیونی و ساختارهای دیگر, وجود دارد. از سویی می‌دانیم که برآوردگرهای استوار کلاسیک تنها کمتر از ۵۰ درصد از داده‌های پرت را کنترل می‌کنند, اما شرایطی پیش می‌آید که در آن بیش از ۵۰ درصد از داده‌ها پرت باشد. در این پایان‌نامه همچنین یک برآوردگر بسیار استوار برای مقابله با چنین مشکلاتی تحت عنوان, برآوردگر تجزیه چگالی رگرسیونی [4] ارائه می‌گردد. این برآوردگر در مقابل کسر بالایی از داده‌های آلوده حتی بیش از ۵۰ درصد مقاوم است, این موضوع در یک مثال شبیه‌سازی شده بخوبی عمل می‌کند. بخش اعظم این پایان‌نامه بر اساس مقاله [۲۴] توسط *Leung , Ma* و *Luo* در *سال* ۲۰۰۶ می‌باشد که در فهرست مراجع نیز به آن اشاره شده‌است.

[1] Knowledge Discovery in Databases (KDD)
[2] Data Mining (DM)
[3] Regression-Class Mixture Decomposition (RCMD)
[4] Regression Density Decomposition (RDD)

کلید واژه

داده‌کاوی، ماکزیمم درستنمایی، مدلبندی آمیخته، روش تجزیه کلاس رگرسیونی آمیخته، کلاس رگرسیونی، استواری, تجزیه رگرسیونی چگالی, مدل رگرسیونی.

علائم اختصاری:

RCMD: Regression-class Mixture Decomposition
MF: Model Fitting
RDD: Regression Density Decomposition
IF: Influence Function
LMS: Least Median Squares
LTS: Least Trimmed Squares
OLS: Ordinary Least Squares
MINPRAN: Minimum Probability Random
DM: Data Mining
GAs: Genetic Algorithm Simulation
GMDD: Gaussian Mixture Density Modeling, Decompositions

فصل اول

کلیات

۱-۱ پیشگفتار

همانطور که می‌دانیم آمار علم و عمل استخراج اطلاعات مفید و الگوسازی مطلوب از داده‌های تجربی می‌باشد. داده‌ها اغلب حجیم هستند و به تنهایی قابل استفاده نمی‌باشند, بلکه دانش نهفته در داده‌ها قابل استفاده می‌باشد, بنابراین بهره‌گیری از قدرت فرآیند داده کاوی[۵] جهت شناسایی الگوها و مدلها و نیز ارتباط عناصر مختلف در پایگاه داده‌ها جهت کشف دانش نهفته در داده‌ها و نهایتاً تبدیل داده به اطلاعات, روز به روز ضروری‌تر می‌شود.

از این نقطه نظر مفهوم داده‌کاوی در آمار مطرح می‌گردد, اصطلاح داده‌کاوی مترادف با یکی از عبارت‌های استخراج دانش, برداشت اطلاعات, وارسی داده‌ها و حتی لایروبی‌کردن داده‌هاست که به عنوان پل ارتباطی میان علم آمار, علم کامپیوتر, هوش مصنوعی, الگوشناسی[۶] و فراگیری ماشینی[۷] می‌باشد. برخی, داده‌کاوی را علم استخراج اطلاعات مفید از پایگاه‌های داده[۸] تعریف می‌کنند, و برخی دیگر داده‌کاوی را در حقیقت کشف ساختارهای جالب توجه, غیر منتظره و با ارزش از داخل مجموعه وسیعی از داده‌ها معرفی می‌کنند که اساساً منطبق بر آمار و تحلیل دقیق داده‌هاست.

اصطلاح داده‌کاوی را آمارشناسان و تحلیل‌گران داده بکار می‌برند, در حالی که پژوهشگران یادگیری ماشین و هوش مصنوعی, بیشتر از KDD استفاده می‌کنند. در مقالات ارائه شده در [۲۲] و[۲۳] خصوصیات برجسته داده‌کاوی در ارتباط با داده‌های بزرگ مورد بررسی قرار گرفته است. هرچند در این زمینه فعالیت‌های زیادی صورت پذیرفته اما پیشرفت روشهای استخراج موثر برای مجموعه داده‌های بزرگ خصوصاً داده‌های آلوده[۹] همچنان یک مشکل مهم و دشوار می‌باشد.

---

[5] Data Mining
[6] Pattern recognition
[7] Machine learning
[8] Knowledge Discovery in Databases (KDD)
[9] Contaminated by noise

یکی از روشهای استخراج اطلاعات, استفاده از مدل‌های پارامتری تصادفی است که می‌تواند اطلاعات بیشتری را که در درون داده‌ها وجود دارد, بدست دهد. بنابراین ما برای بدست آوردن اطلاعات بیشتر, از مجموعه داده‌ها از مدل‌های تصادفی پارامتری استفاده می‌کنیم و از میان مدل‌های تصادفی پارامتری مدل پارامتری رگرسیون عمومی اطلاعات دقیق‌تری از داده‌های مذکور و تعبیر کمی آنها ارائه می‌دهد.

در اغلب برازشهای رگرسیونی از شیوه حداقل مربعات معمولی[10] (OLS) استفاده می‌شود, اما زمانی که خطاها دارای توزیع نرمال نباشند و یا مجموعه داده‌ها شامل داده‌های پرت باشند, روش حداقل مربعات معمولی, دیگر کارا نیست زیرا این شیوه به داده‌های پرت حساس می‌باشد و بایستی از روشهای استوار برای برآورد پارامترها استفاده نمود. رگرسیون استوار معمولاً به روشی گفته می‌شود که نه تنها وقتی خطاها دارای توزیع نرمال است و مشاهدات پرت در مدل حضور ندارند خوب عمل می‌کند بلکه نسبت به انحرافات کوچک از فرض نرمال بودن و نسبت به حضور نقاط پرت در مدل نیز حساس نمی‌باشند. رگرسیون استوار دارای تعداد زیادی برآوردگر است و تکنیک‌های آنها مکمل کمترین مربعات می‌باشد بطوریکه وقتی خطاها نرمال است و مشاهدات پرت در مدل حضور نداشته باشند, جواب‌های آنها مشابه با جواب رگرسیون حداقل مربعات است.

همانطور که می‌دانیم برآوردگرهای استوار کلاسیک تنها کمتر از ۵۰ درصد داده‌های پرت را تحت کنترل درمی‌آورند, این برآوردگرها راه حل مناسبی برای شناسایی مدل خصوصاً زمانی که داده‌های پرت بیش از نیمی از داده‌ها را تشکیل دهد ارائه نمی‌دهند. در هر صورت در عمل ممکن است شرایطی پیش آید که بیش از ۵۰ درصد داده‌ها, پرت باشد.

در هر صورت ما امیدواریم که در صورت وجود, یک مدل رگرسیونی مجرد بتواند برای یک مجموعه از داده‌های بزرگ یا پیچیده نیز بکار رود. از این رو آنالیز رگرسیون به دلایل ذیل برای مطالعه مجموعه داده‌های بزرگ خصوصاً داده های آلوده مناسب نمی‌باشد:

---

[10]Ordinary Least Squares

۱. آنالیز رگرسیون به یک مجموعه از داده‌ها بصورت کلی می‌پردازد حتی با وجود امکانات نیز روش موثری برای آنالیز کردن یک مجموعه بزرگ از داده‌ها وجود ندارد.

۲. مهم‌تر اینکه فرض استفاده از یک مدل برای تعداد زیادی داده چندان واقعی نیست و استفاده از چند مدل برای برازش به یک مجموعه از داده‌های بزرگ معقول‌تر است به عبارت دیگر یک مجموعه از داده‌ها با استفاده از یک مدل به تنهایی مدلبندی دقیق نمی‌شود.

۳. آنالیز رگرسیون کلاسیک بر پایه فرضیات دشوار می‌باشد و داده‌های واقعی بخصوص, مجموعه داده‌های بزرگ ممکن است منطبق بر این فرضیات رفتار نکند.

به منظور غلبه بر مشکلات فوق و استفاده مناسب از مدل‌های رگرسیون پارامتری, ابتدا نیاز به دانستن مفهوم مجموعه داده‌های پیچیده داریم. مهم است که بدانیم چگونه با چنین مجموعه‌ای از داده‌ها رفتار کنیم. یک روش مناسب این است که مجموعه داده‌های پیچیده را به صورت ترکیبی از جوامع ببینیم. فرض کنید یک مجموعه از داده‌ها را بصورت تعداد متناهی از مدل‌های رگرسیونی نشان دهیم. اگر هر کدام از این مدل‌ها را بصورت یک جامعه ببینیم پس یک مجموعه بزرگ از داده‌ها بصورت ترکیبی متناهی از چنین جوامعی در نظر گرفته می‌شود.

در ادبیات, مدلهای ترکیبی در واقع مدل‌بندی یک توزیع آماری با استفاده از ترکیبی از دیگر توزیع‌ها تعریف می شود که این ترکیب شامل مولفه‌ها یا کلاس‌های آن توزیع می‌باشند.

مثال ۱-۱:

نمونه ذیل یک شکل خاص از چگالی‌های نرمال آمیخته از مدل‌های رگرسیونی آمیخته می‌باشند که در آن Y بردار مشاهدات است:

$$y=\begin{cases} X^T\beta_1+e_1 & with\ probabilit\,y\ \ \lambda \\ X^T\beta_2+e_2 & with\ probabilit\,y\ \ 1-\lambda' \end{cases}$$

که در آن $e_i \sim N(0,\sigma_i^{\,2})$, $(\ i=1,2\ )$.

مدل‌های ترکیبی, درجوامع نرمال آمیخته بطور گسترده‌تر مورد مطالعه قرار گرفته و بکار برده می‌شوند. در برآورد کردن پارامترهای آمیخته, روش حداکثر درستنمایی[11] (ML) بیشترین کاربرد را دارد. هر چند استفاده از الگوریتم EM[12] نیز بسیاری از مشکلات محاسبه را کاسته است اما همچنان استفاده از این شیوه‌ها خالی از خطا نیست.

سایر روشها از قبیل روش گشتاورها و تابع مولد گشتاور[13] (MGF)هر دو مشکل شبیه سازی برآورد کردن پارامترهای آمیخته را دارند. به این ترتیب همانطور که برآورد پارامترهای آمیخته مشکل است برآورد پارامترهای مدل رگرسیونی, در یک مجموعه داده های بزرگ نیز دشوار می‌باشد.

علاوه بر کارایی یک مدل, موضوع مهم دیگر که نیاز به توجه دارد استواری است. به منظور کاربردی بودن در عمل, لازم است یک شیوه بخصوص برای مجموعه‌های بزرگ بسیار استوار باشد بدین معنا که یک روش نباید بصورت معنی‌داری تحت تاثیر انحرافات جزئی از مدل مفروض قرار گیرد و بطور چشمگیری بر اساس نوفه‌ها و داده‌های پرت رو به ضعف رود. اخیراً, تلاشهایی در این زمینه صورت پذیرفته است و در این خصوص نیاز به تحقیق بیشتری نیز هست.

۱-۲ اهداف تحقیق:

در این قسمت بصورت خلاصه اهداف این پایان‌نامه را ارائه می‌دهیم:

۱. ارائه یک روش موثر و استوار برای استخراج کلاس‌های رگرسیونی در مجموعه داده‌های بزرگ خصوصاً تحت شرایط نوفه‌ای, فرض ما بر این است که تعداد متناهی برای چنین کلاس‌های رگرسیونی در یک مجموعه داده‌های بزرگ وجود دارد. در ادامه به معرفی برآوردگر جدید بنام

[11] Maximum Likelihood
[12] Expectation Maximization
[13] Moment Generating Function

برآوردگر تجزیه کلاس رگرسیونی آمیخته [14](*RCMD*) که تنها شامل پارامترهای کلاس رگرسیونی در هر زمان t از فرآیند استخراج است, می‌پردازیم. ضمن اینکه برآوردگر فوق به درجه بالایی از استواری رسیده است.

۲. بر پایه این چارچوب به معرفی یک برآوردگر بسیار استوار برای مدل‌های رگرسیونی عمومی پرداخته که بعنوان یک نسخه جدید از برآوردگر *RCMD* برای یک مدل واحد که حالت خاصی از مدل ارائه شده قبل می‌باشد تحت شرایط نوفه‌ای ارائه گردیده است, برآوردگر ارائه شده جدید[15] تجزیه رگرسیونی تابع چگالی (*RDD*) نامیده می‌شود.

لازم به ذکر است که ما مقاله‌ای تحت عنوان"یک روش جدید استخراج کلاس‌های رگرسیونی در مجموعه داده‌های بزرگ" به اولین کنفرانس ملی داده‌کاوی در دانشگاه صنعتی امیرکبیر ارسال کردیم که این مقاله مورد پذیرش قرار گرفته است.

۱-۳ تاریخچه:

از مهمترین اهداف الگوسازی شناسایی مدل مطلوب از داده‌های نوفه‌ای با وجود تعداد زیادی داده پرت می‌باشد, توسعه روشها برای مقابله با اثرات داده‌های پرت هدف آمار استوار است. همانطور که می‌دانیم تقریباً تمامی روشهای استوار تنها در مقابل کمتر از ۵۰ درصد از داده‌های پرت می‌توانند استوار گردند, در حالی که با وجود کلاس‌های رگرسیونی در یک مجموعه از داده‌ها این روش‌ها راه حل مناسبی برای توضیح کلاس‌های رگرسیونی نخواهند داد, چرا که نسبت داده‌های پرت در مقایسه, برای یک کلاس به تنهایی ممکن است بیش از ۵۰ درصد باشد, اخیراً, بیش از چندین روش استوار توسعه

[14] Regression Class Mixture Decomposition
[15] Regression Density Decomposition

یافته است, *Stewart* در سال ۱۹۹۵ ارائه شده در [۱] یک روش تحت عنوان حداقل کردن احتمال تصادفی[16] *MINPRAN* ارائه داده که شاید اولین تکنیک به حساب می‌آید که بیش از ۵۰ درصد داده‌های پرت را تحت کنترل درآورده است. این روش فرض را براین قرار می‌دهد که داده‌های پرت دارای توزیع تصادفی نرمال در قالب دامنه پویا از گیرنده‌ها[17] ست, اما فرضیات در نظر گرفته شده برای *MINPRAN* کلیت آن را در عمل محدود می‌سازد.

پیش از آن نیز برآوردگر بسیار استواری تحت عنوان برآوردگر برازش مدل[18] (*MF*) توسط *Zhuang* وهمکارانش در سال ۱۹۹۲ در مقاله [۲] ارائه گردیده است. این برآوردگر نیاز به فرضیاتی شبیه آنچه دربالا اشاره شد, ندارد, در واقع نیازمند فرض دانستن توزیع داده‌های پرت نیست. بنابراین شیوه‌ای مناسب برای مقابله با چنین شرایطی می‌باشد. ضمن اینکه این برآوردگر بیشتر مشکلات رگرسیون عمومی را شامل نمی‌شود. در کنار *MINPRAN* و *MF* روشهای استوار دیگری نیز وجود دارند که تنها محدود به برخی مسائل مدلسازی خطی با استفاده از برخی شیوه‌های خاص هستند. به عنوان مثال شیوه‌ای که توسط *Perantonis* و همکارانش در سال ۱۹۹۸ در مقاله [۳] ارائه گردیده است, برپایه روش اصلاح شده وزنی تبدیل هاگ است. یک روش استوار برازش خطی نیز توسط *Frigui* وهمکارانش در سال ۱۹۹۸ در مقاله [۸] ارائه گردیده است, که به ارائه چندین الگوریتم پیش پردازش فازی و استوار می‌پردازد, و در واقع در خصوص یک استدلال فازی بحث می‌کند که به فیلتر کردن سیستم با حذف داده‌های آلوده پرداخته است بدون اینکه جزئیات دیگر را تخریب کند.

یک روش خوشه بندی جدید آماری برپایه تئوری گشتاوری لژاندر[19] و اصل ماکسیمم آنتروپی[20] برای برازش خطی در فضای نوفه‌ای (آلوده) توسط *Qjidaa* و*Radouane* درسال *۱۹۹۹ در مقاله* [۹] فرمولبندی شده است.

---

[16] MINimum Probability Random
[17] dynamic range of the sensor
[18] Model-fitting
[19] Legendre moment
[20] Maximum entropy principle

الگوریتم‌های پیشنهاد شده توسط *Kiryati* و *Bruckstein* در *سالهای ۱۹۹۲ و۲۰۰۰ در مقالات ارائه شده* در[۱۰] و[۱۱] نیز بر پایه تبدیل هاگ بوده که می‌توانند مسائل مدلسازی موثر در حضور داده‌های پرت در هر دو مختصات را حل کنند.

بعضی برآوردگرهای ارائه شده نیز بسیار خاص بوده و هم تنها برای مشکلات سطحی مناسب می‌باشند و یا اینکه ممکن است در عمل بکار بردن آنها سخت باشد.

اخیراً, یک برآوردگر تجزیه کلاس رگرسیونی آمیخته توسط *Leung* و همکارانش در سال ۲۰۰۱ در مقاله [۴] ارائه گردیده است, که در فصل دوم به آن خواهیم پرداخت, این برآوردگر به منظور استخراج کلاس‌های رگرسیونی متفاوت در مجموعه داده‌های بزرگ ساخته شده است. در هر مرحله از فرآیند استخراج, برآوردگر *RCMD* باید در مقابل تعداد زیادی از داده‌های پرت استوار بماند (معمولاً بیش از ۵۰ درصد). بر پایه این چارچوب در فصل سوم یک برآوردگر بسیار استوار برای مدل‌های رگرسیونی عمومی[۲۱] معرفی می‌کنیم که بعنوان یک نسخه جدید از *RCMD* در یک کلاس واحد, تحت شرایط بسیار نوفه‌ای است, برآوردگر ارائه شده[۲۲] تجزیه تابع چگالی رگرسیونی (*RDD*) نامیده می‌شود. این برآوردگر متفاوت از برآوردگر *RCMD*/ست که توسط *Leung* و همکارانش در سال ۲۰۰۱ ارائه گردیده است. در مقایسه با برآوردگرهای اشاره شده برآوردگر RDD در متودولوژی متفاوت و در عمل نیز آسان است. این برآوردگر را *Jiang-Hong Ma* و همکارانش در سال ۲۰۰۶ در مقاله [۱۲] مورد بررسی قرار داده‌اند.

۱-۴ ساختار

ساختار پایان‌نامه بصورت زیر است:

---

[21] General
[22] Regression Density Decomposition

در فصل دوم ابتدا مفهوم کلاس رگرسیونی را ارائه می‌دهیم که براساس مدل رگرسیونی تعریف شده است. بصورت کلی, یک کلاس رگرسیونی شامل اطلاعات مفیدتری می‌باشد. به عنوان مثال مدل‌های رگرسیونی ارائه شده در مثال۱-۱, دو کلاس رگرسیونی بحساب می‌آیند. بجای در نظر گرفتن کل مجموعه داده‌ها, نمونه‌گیری بصورت مشابه از کلاس‌های رگرسیونی صورت می‌پذیرد. استخراج کلاس‌های رگرسیونی در مجموعه داده‌های بزرگ خصوصاً تحت شرایط نوفه‌ای در چارچوب جدید با استفاده از الگوریتم RCMD صورت می‌پذیرد. دو الگوریتم ژنتیک و بازگشتی نیز برای برآورد پارامترها در هر کلاس رگرسیونی ارائه و تعریف می‌گردند. با در نظر گرفتن متناهی بودن کلاس‌های رگرسیونی به معرفی برآوردگر جدید *RCMD* می‌پردازیم.

در فصل سوم بر پایه چارچوب کلی ارائه شده در فصل دوم, به معرفی یک برآوردگر بسیار استوار برای مدل‌های رگرسیونی عمومی پرداخته که بعنوان یک نسخه جدید از *RCMD* در یک مدل تحت شرایط نوفه‌ای ارائه گردیده است, برآوردگر ارائه شده جدید تجزیه تابع چگالی رگرسیونی (*RDD*) نامیده می‌شود. در نهایت در قالب یک مثال شبیه سازی شده به بررسی کارایی این برآوردگر در مقایسه با برآوردگر کلاسیک LS و دو برآوردگر دیگر رگرسیون استوار بنام برآوردگرهای *LTS* و *LMS* می‌پردازیم.

فصل دوم

کلاس‌های رگرسیونی و روش تجزیه کلاس رگرسیونی آمیخته(RCMD)

۲-۱ مقدمه

در این فصل به بیان مفهوم کلاس رگرسیونی پرداخته و روش تقسیم‌بندی نمونه به تعداد متناهی از این کلاس‌های رگرسیونی را مورد بررسی قرار می‌دهیم, تاثیر داده‌های پرت را در قالب دو قضیه نشان می‌دهیم. ضمن اینکه برآوردگر تجزیه کلاس رگرسیونی آمیخته (RCMD) را به عنوان یک برآوردگر جهت استخراج پارامترهای هر کلاس رگرسیونی معرفی می‌کنیم و درانتها نیز دو الگوریتم جهت مراحل محاسباتی پارامترهای برآوردگر RCMD و استخراج کلاس‌های رگرسیونی ارائه می‌دهیم.

۲-۲ کلاسهای رگرسیونی

در این بخش مفهوم کلاس رگرسیونی[۲۳] را مطرح می‌کنیم. کلاس رگرسیونی همان مدل رگرسیونی است. به عبارتی برای یک i ثابت, کلاس رگرسیونی G_i با مدل رگرسیونی ذیل به همراه حاملان تصادفی[۲۴] تعریف می‌شود:

$$G_i : y = f_i(X, \beta_i) + e_i, \qquad i = 1, ..., m \qquad (۱-۲)$$

که در آن y متغیر پاسخ, $X \in R^P$ متغیرهای توضیحی می‌باشند که شامل حاملان یا رگرسورها هستند و یک بردار تصادفی با تابع چگالی احتمال p(.) را تشکیل می‌دهند, e_i یک متغیر تصادفی با تابع چگالی احتمال $\psi(u, \sigma_i)$ است که شامل پارامتر σ_i می‌باشد. در اینجا $f_i(.,.) : R^P \times R^{q_i} \to R$ یک تابع رگرسیونی مشخص می‌باشد و $\beta_i \in R^{q_i}$ یک ستون از پارامترهای نامعلوم می‌باشند. همانطور که دیده می‌شود بعد β_i و q_i برای G_iهای متفاوت, مختلف می‌باشد, معمولاً برای سادگی $q_i \equiv q$ می‌گیریم. پس از این فرض را بر این قرار می‌دهیم که e_i ها مطابق با توزیع نرمال ذیل است:

---

[23] Reg-class
[24] Random carriers

$$\psi(u,\sigma_i)= 1/\sigma_i \,\phi\left(\frac{u}{\sigma_i}\right) \qquad (۲-۲)$$

که در آن $\phi(.)$ تابع چگالی احتمال نرمال استاندارد است. به منظور سادگی بحث به ازای هر β_i باقیمانده‌ها را بصورت ذیل نشان می‌دهیم:

$$r_i(x,y,\beta_i)= y - f_i(x,\beta_i) \qquad (۳-۲)$$

تعریف۲-۱ :

بردار تصادفی (X,Y) به کلاس رگرسیونی G_i تعلق دارد, اگر توزیع آن مطابق با مدل رگرسیونی G_i باشد. بنابراین تحت این تعریف, بردار (X,Y) به کلاس رگرسیونی G_i تعلق دارد اگر دارای تابع چگالی احتمال ذیل باشد:

$$p_i(x,y,\theta_i)= p(x)\psi(r_i(x,y,\beta_i),\sigma_i)\ ,\theta_i=(\beta_i^T,\sigma_i^T)^T. \qquad (۴-۲)$$

به منظور تسهیل در عمل نیز تعریف ذیل معمولاً مرتبط با تعریف فوق مورد استفاده قرار می‌گیرد.

تعریف ۲-۲ :

نقطه (x,y) متعلق به کلاس رگرسیونی G_i است اگر $p_i(x,y,\theta_i)\geq b_i$ یعنی:

$$G_i \equiv G_i(\theta_i)\equiv\{(x,y): p_i(x,y,\theta_i)\geq b_i\} \qquad (۵-۲)$$

که در آن ثابت $b_i \succ 0$ بصورت ذیل بدست می‌آید:

$$P[p_i(x,y,\theta_i)\geq b_i]=a$$