



دانشگاه پیام نور

مرکز تهران

دانشکده علوم پایه

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته آمار ریاضی

عنوان پایان نامه:

برآورد پارامترهای مدل‌های رگرسیون مرتبه ای

با روشهای متفاوت

استاد راهنمای اول:

دکتر علی شادرخ

استاد راهنمای همکار:

دکتر پرویز نصیری

نگارش:

علی خاشعی و رنامخواستی

تیر ماه ۱۳۸۹

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ



دانشگاه پیام نور

مرکز تهران

دانشکده علوم پایه

پایان نامه

برای دریافت درجه کارشناسی ارشد

در رشته آمار ریاضی

عنوان پایان نامه:

برآورد پارامترهای مدل‌های رگرسیون مرتبه ای

با روشهای متفاوت

استاد راهنمای اول:

دکتر علی شادرخ

استاد راهنمای همکار:

دکتر پرویز نصیری

نگارش:

علی خاشعی ورنامخواستی

تیر ماه ۱۳۸۹

شماره:
تاریخ:
پیوست:



جمهوری اسلامی ایران
وزارت علوم، تحقیقات و فناوری
مجمع علوم پایه و کشاورزی

تصویب نامه

پایان نامه کارشناسی ارشد در رشته: آمار ریاضی

تحت عنوان:

"برآورد پارامترهای مدل‌های رگرسیون مرتبه‌ای
باروشهای متفاوت"

ساعت: ۱۰-۱۱

تاریخ دفاع: ۸۹/۰۴/۲۹

نمره پایان نامه: ۱۸,۵ درجه ارزشیابی: عالی

امضاء	مرتبه علمی	نام و نام خانوادگی	اساتید هیات داوران
		دکتر علی شادرخ	استاد راهنما
		دکتر پرویز نصیری	استاد مشاور
		دکتر احسان جمالی	استاد داور
		دکتر مسعودیاری محمدی	نماینده علمی گروه

تهران، خیابان استاد
نجات الهی، خیابان
شهید فلاح پور، پلاک ۲۷
تلفن: ۸۸۸۰۰۲۵۲
دورنگار: ۸۸۲۱۹۴۷۵
www.tpnu.ac.ir
science_agri@tpnu.ac.ir

باتشکر و قدردانی از اساتید گرامی

جناب آقای دکتر علی شادرخ

و

جناب آقای دکتر پرویز نصیری

که مرا در تهیه و تنظیم این تحقیق یاری نموده و اینجانب را مورد لطف

و عنایت خود قرار دادند.

نام خانوادگی دانشجو: خاشعی ورنامخواستی نام: علی
عنوان پایان نامه: برآورد پارامترهای مدل‌های رگرسیون مرتبه ای باروشهای متفاوت
استاد راهنما: دکتر علی شادرخ
استاد راهنمای همکار: دکتر پرویز نصیری
مقطع تحصیلی: کارشناسی ارشد رشته: آمار گرایش: آمار ریاضی دانشگاه: پیام نور مرکز تهران
دانشکده: علوم پایه تاریخ فارغ التحصیلی: تعداد صفحه:

چکیده:

در بسیاری از مطالعات علمی و پژوهشی مایل به بررسی تأثیرگذاری متغیرها بر هم و یا پیش بینی یک متغیر به وسیله گروهی از متغیرهای دیگر هستیم. در مباحث آماری این عمل اغلب با استفاده از مدل‌های رگرسیونی امکان پذیر است. اما برای استفاده از این مدل‌های رگرسیونی پیش فرضهائی نیاز است. از جمله این پیش فرضها استقلال مشاهدات و داده هاست، که باید مدنظر گرفته شود. البته ممکن است که این فرض برای همه مشاهدات برقرار نباشد. یعنی مشاهدات نسبت به هم وابسته بوده و یا ساختاری تودرتو داشته باشند. پس در این شرایط نیاز به استفاده از روشهای رگرسیونی مناسب دیگری غیر از روشهای استاندارد می باشد.

در حالت وجود همبستگی بین مشاهدات و یا داشتن ساختاری تودرتو، مدل‌های جدیدی به نام مدل‌های رگرسیون مرتبه ای و یا به عبارت دیگر رگرسیون سلسله مراتبی می تواند برای برازش داده ها و کسب نتایج قابل قبولتر، مفید باشد. در این تحقیق به معرفی مدل‌های رگرسیون مرتبه ای (سلسله مراتبی) و پارامترهای موجود در آن پرداخته ایم. سپس با استفاده از روشهای مختلف آماری از قبیل برآوردهای درستنمائی ماکسیمم، کمترین مربعات، بیز تجربی، نیم بیز و تمام بیز پارامترها را برآورد کرده و مورد بررسی و مقایسه قرار داده ایم و سعی در بهبود و اصلاح آنها نموده ایم.

در نهایت با ارائه چند نمونه عملی و کاربردی از داده های واقعی، تلاش کرده ایم که با مدلسازی رگرسیون مرتبه ای و برآورد پارامترهای آن، نتایج قابل قبولتری نسبت به مدل‌های رگرسیون معمولی نشان دهیم. نتایج کامل این تحقیق در متن پایان نامه مورد توجه قرار گرفته است.

واژه های کلیدی: رگرسیون سلسله مراتبی - داده های تودرتو- برآورد کمترین مربعات - برآورد

درستنمائی ماکسیمم- برآورد بیز تجربی- برآورد نیم بیز - برآورد تمام بیز.

فهرست مطالب

۱	فصل اول: مدل‌های رگرسیونی
۲	۱-۱ مقدمه
۳	۲-۱ رگرسیون خطی ساده
۳	۱-۲-۱ مدل رگرسیون خطی
۴	۲-۲-۱ تابع رگرسیون خطی
۵	۳-۲-۱ برآورد α و β و σ^2
۱۴	۳-۱ رگرسیون چندگانه
۱۵	۱-۳-۱ مدل با دو متغیر مستقل
۱۵	۲-۳-۱ مدل با k متغیر مستقل
۱۷	۳-۳-۱ برآورد β و σ^2
۲۵	فصل دوم: رگرسیون مرتبه ای (سلسله مراتبی)
۲۶	۱-۲ مقدمه
۲۹	۲-۲ ساختارهای تودرتوی داده ها
۳۲	۱-۲-۲ وابستگی داده ها
۳۲	۲-۲-۲ رفتار غیر چند سطحی از داده های تودرتو
۳۴	۳-۲ مدل رگرسیون دو سطحی
۳۷	۴-۲ زیر مدل‌های ساده تر
۳۷	۱-۴-۲ ANOVA یکطرفه با اثرهای تصادفی
۳۹	۲-۴-۲ روش رگرسیون میانگینها

۴۰	۳-۴-۲ ANCOVA یکطرفه با اثرهای تصادفی
۴۱	۴-۴-۲ مدل رگرسیون ضرایب تصادفی
۴۳	۵-۴-۲ روش رگرسیون عرض از مبدأها و شیبها
۴۳	۶-۴-۲ مدل با شیبهای مختلف غیر تصادفی
۴۵	۵-۲ مدل خطی چندگانه سلسله مراتبی
۴۵	۱-۵-۲ چندین X و چندین W
۴۷	۲-۵-۲ تعمیم ساختارهای خطا در سطح ۱ و سطح ۲
۴۸	۶-۲ هدف از آنالیز چند سطحی

فصل سوم: برآورد پارامترهای رگرسیون سلسله مراتبی

۴۹	۱-۳ مقدمه
۵۰	۲-۳ برآورد اثرهای ثابت
۵۲	۱-۲-۳ برآورد نقطه ای
۵۵	۲-۲-۳ برآورد فاصله ای
۵۶	۳-۲-۳ داده ها با رتبه ناقص
۵۸	۳-۳ برآورد ضرایب تصادفی سطح ۱
۵۸	۱-۳-۳ برآورد درستمنائی ماکسیمم
۶۰	۲-۳-۳ برآوردهای بیزی
۶۳	۳-۳-۳ برآورد بیز تجربی (EB)
۶۵	۴-۳-۳ برآورد نیم بیز (SB)
۶۶	۵-۳-۳ برآورد تماماً بیز (FB)
۶۶	۶-۳-۳ برآورد فاصله ای
۶۷	۷-۳-۳ قابلیت اعتماد ضرایب برآورد شده سطح ۱
۶۸	۸-۳-۳ مشاهده باقیمانده

فصل چهارم: کاربرد مدل‌های رگرسیون سلسله مراتبی

۷۰	۱-۴ مقدمه
----	-----------

۷۴	۲-۴	تحقیق در مورد رابطه موفقیت و وضعیت اجتماعی اقتصادی در مدارس
۷۴	۱-۲-۴	بررسی رابطه موفقیت و وضعیت اجتماعی اقتصادی در یک مدرسه
۷۶	۲-۲-۴	بررسی رابطه موفقیت و وضعیت اجتماعی اقتصادی در دو مدرسه
۷۷	۳-۲-۴	بررسی رابطه موفقیت و وضعیت اجتماعی اقتصادی در ۱۰ مدرسه
	۳-۴	ارتباط بین درس ریاضیات و معدل درسی دانش آموزان مدارس راهنمایی شهرستان
۸۱		لنجان
۸۵	۱-۳-۴	ANOVA یکطرفه
۸۹	۲-۳-۴	روش رگرسیون میانگینها
۹۲	۳-۳-۴	مدل ضرایب تصادفی
	۴-۳-۴	مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> و <i>N st</i>
۹۹		برای سطح مدرسه
	۵-۳-۴	مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> و
۱۱۳		<i>Space</i> برای سطح مدرسه
	۶-۳-۴	مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> ، <i>N st</i>
۱۲۲		و <i>Space</i> برای سطح مدرسه
۱۳۰	۴-۴	نتیجه گیریها

منابع

۱۳۳

پیوست

۱۳۶

۱۳۶

پیوست الف: جدول مربوط به بخش ۳-۴

فهرست جداول

۳۰	۱-۲	نمونه هائی از واحدها در سطح ۱ و سطح ۲.
۳۰	۲-۲	خلاصه ای از عبارات بیان شده برای وصف واحدها در هر یک از دو سطح.
۸۳	۱-۴	نتایج از رگرسیون نمره ریاضیات و معدل ۲۰۰ دانش آموز.
۸۵	۲-۴	آماره های توصیفی برای داده های سطح دانش آموز.
۸۵	۳-۴	آماره های توصیفی برای سطح مدرسه.
۸۶	۴-۴	برآورد ضرایب سطح ۱ در مدل ANOVA یکطرفه.
۸۷	۵-۴	نتایج از مدل ANOVA یکطرفه.
۹۰	۶-۴	نتایج از روش رگرسیون میانگینها.
۹۳	۷-۴	برآورد ضرایب سطح دانش آموز در مدل ضرایب تصادفی.
۹۶	۸-۴	نتایج از مدل ضرایب تصادفی.
۹۸	۹-۴	قابلیت اعتماد ضرایب سطح ۱ برای مدل ضرایب تصادفی.
۱۰۰	۱۰-۴	نتایج از مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> و <i>N St</i> برای سطح مدرسه.
۱۰۴	۱۱-۴	قابلیت اعتماد ضرایب سطح ۱ برای مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> و <i>N St</i> برای سطح مدرسه.
۱۰۶	۱۲-۴	برآوردهای بیز تجربی ضرایب سطح دانش آموز برای مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> و <i>N St</i> برای سطح مدرسه.
۱۰۹	۱۳-۴	قابلیت اعتماد ضرایب سطح ۱ برای مدل با پیش بین <i>MAT</i> برای سطح دانش آموز و پیش بینهای <i>Type</i> و <i>N St</i> برای سطح مدرسه. با توجه به $\tau_{..} = 0.5$ و $\tau_{11} = 0.005$.

- ۱۴-۴ برآوردهای نیم بیز ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش آموز
 ۱۰۹ و پیش بینهای $N St Type$ و برای سطح مدرسه.
- ۱۵-۴ برآوردهای تمام بیز ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش
 ۱۱۱ آموز و پیش بینهای $Type$ و $N St$ برای سطح مدرسه.
- ۱۶-۴ نتایج از مدل با پیش بین MAT برای سطح دانش آموز و پیش بینهای $Type$ و $Space$ برای
 ۱۱۴ سطح مدرسه.
- ۱۷-۴ قابلیت اعتماد ضرایب سطح ۱ برای مدل با پیش بین MAT برای سطح دانش آموز و پیش بینهای
 ۱۱۶ $Type$ و $Space$ برای سطح مدرسه.
- ۱۸-۴ برآوردهای بیز تجربی ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش
 ۱۱۷ آموز و پیش بینهای $Type$ و $Space$ برای سطح مدرسه.
- ۱۹-۴ برآوردهای نیم بیز ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش آموز
 ۱۱۹ و پیش بینهای $Type$ و $Space$ برای سطح مدرسه.
- ۲۰-۴ برآوردهای نیم بیز ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش آموز
 ۱۲۱ و پیش بینهای $Type$ و $Space$ برای سطح مدرسه.
- ۲۱-۴ نتایج از مدل با پیش بین MAT برای سطح دانش آموز و پیش بینهای $St N, Type$ و
 ۱۲۲ $Space$ برای سطح مدرسه.
- ۲۲-۴ قابلیت اعتماد ضرایب سطح ۱ برای مدل با پیش بین MAT برای سطح دانش آموز و پیش بینهای
 ۱۲۵ $St N, Type$ و $Space$ برای سطح مدرسه.
- ۲۳-۴ برآوردهای بیز تجربی ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش
 ۱۲۶ آموز و پیش بینهای $Type$ و $St N, Space$ برای سطح مدرسه.
- ۲۴-۴ برآوردهای نیم بیز ضرایب سطح دانش آموز برای مدل با پیش بین MAT برای سطح دانش
 ۱۲۸ آموز و پیش بینهای $Type$ و $St N, Space$ برای سطح مدرسه.

- الف-۱ مشخصات مربوط به نمونه انتخاب شده از مدارس (نوع مدرسه، دولتی (۰) و غیردولتی (۰) ،
 جنسیت، پسر (۰) و دختر (۱) ، تعداد دانش آموزان مشغول به تحصیل در آن و مقدار فضای
 آموزشی که مدرسه در اختیار دارد برحسب مترمربع) و دانش آموزان (نمره درس ریاضیات و
 معدل درسی) مقطع راهنمایی شهرستان لنجان در سال تحصیلی ۸۸-۸۷.
 ۱۳۶

فهرست اشکال و نمودارها

۲۹	۱-۲	ساختار داده ها برای یک مدل سلسله مراتبی دو سطحی.
۳۴	۲-۲	سازگاری سطح خرد در مقابل سطح کلان.
۷۵	۱-۴	نمودار پراکنش ارتباط بین موفقیت و <i>SES</i> در یک مدرسه فرضی.
۷۶	۲-۴	نمودار پراکنش ارتباط بین موفقیت و <i>SES</i> (متمرکز شده) در یک مدرسه فرضی.
۷۶	۳-۴	نمودار پراکنش ارتباط بین موفقیت و <i>SES</i> در دو مدرسه فرضی.
۷۸	۴-۴	نمودار میانگینها و شیبها برای ۲۰۰ مدرسه فرضی.
۸۳	۵-۴	نمودار پراکنش بین درس ریاضیات و معدل ۲۰۰ دانش آموز.
۹۲	۶-۴	نمودار ارتباط بین میانگین درس ریاضیات و میانگین معدل برای مدارس.
۹۴	۷-۴	نمودار ارتباط بین ریاضیات و معدل درسی برای مدرسه ۶.
۹۴	۸-۴	نمودار ارتباط بین ریاضیات و معدل درسی برای مدرسه ۱.
۱۰۱	۹-۴	ارتباط بین معدل و ریاضیات برای مدارس دولتی (۰) و غیردولتی (۱).
	۱۰-۴	مانده های کمترین مربعات معمولی از عرض از مبدأ و شیب برای مدارس در مدل با پیش
۱۰۷		بینهای <i>Type</i> و <i>St N</i> برای سطح مدرسه.
	۱۱-۴	مانده های بیز تجربی از عرض از مبدأ و شیب برای مدارس در مدل با پیش بینهای <i>Type</i> و
۱۰۸		<i>St N</i> برای سطح مدرسه.
	۱۲-۴	مانده های نیم بیز از عرض از مبدأ و شیب برای مدارس در مدل با پیش بینهای <i>Type</i> و
۱۱۰		<i>St N</i> برای سطح مدرسه.
	۱۳-۴	مانده های تمام بیز از عرض از مبدأ و شیب برای مدارس در مدل با پیش بینهای <i>Type</i> و
۱۱۲		<i>St N</i> برای سطح مدرسه.

- ۱۴-۴ مانده های کمترین مربعات معمولی از عرض از مبدأ و شیب برای مدارس در مدل با پیش
بینیهای *Space* و *Type* برای سطح مدرسه. ۱۱۸
- ۱۵-۴ مانده های بیز تجربی از عرض از مبدأ و شیب برای مدارس در مدل با پیش بینیهای *Type* و
Space برای سطح مدرسه. ۱۱۸
- ۱۶-۴ مانده های کمترین مربعات معمولی از عرض از مبدأ و شیب برای مدارس در مدل با پیش
بینیهای *St N*، *Type* و *Space* برای سطح مدرسه. ۱۲۷
- ۱۷-۴ مانده های بیز تجربی از عرض از مبدأ و شیب برای مدارس در مدل با پیش بینیهای *Type*
و *St N* برای سطح مدرسه. ۱۲۷

۱

مدلهای رگرسیونی

۱-۱ مقدمه

تحلیل رگرسیونی ما را قادر می سازد که بین یک متغیر مورد نظر که آن را متغیر وابسته یا پاسخ می نامیم و یک یا چند متغیر مستقل یا پیش بین رابطه ای را معین کنیم و از آن استفاده نمائیم. مثلاً، در یک مطالعه رگرسیونی، متغیر پاسخ، قد افراد برحسب سانتیمتر و متغیر مستقل، وزن افراد برحسب کیلوگرم در یک زمان معین است.

تحلیل رگرسیونی را غالباً برای پیش بینی متغیر پاسخ و یا اساساً برای بررسی وابستگی بین متغیرهای مستقل و متغیر پاسخ به کار می برند. مثلاً، مایلیم قد افراد را با استفاده از آگاهی که از وزن آنها داریم پیش بینی کنیم و یا بررسی کنیم که آیا با افزایش وزن افراد، قد آنها نیز افزایش می یابد یا نه.

عموماً متغیر پاسخ با Y و متغیر مستقل با X نشان داده می شود. همانطور که بیان شد تحلیل رگرسیونی به ما این امکان را می دهد که بین دو متغیر X و Y رابطه ای برقرار کنیم. وقتی مقدار X مشخص باشد و مقدار Y به صورتی یکتا معین شود، بین X و Y رابطه ای دقیق برقرار شده است که این رابطه، یک رابطه تابعی بین دو متغیر X و Y است. اگر براساس مقادیر X و Y نمودار پراکنشی رسم شود، بدلیل اینکه مقدار Y به صورتی یکتا از روی مقدار X تعیین شده، تمام مشاهدات بر خط رابطه بین X و Y می افتند. حال اگر مقدار X مشخص باشد و مقدار Y به صورتی یکتا معین نشود، رابطه بین X و Y دقیق نخواهد بود که این رابطه یک رابطه آماری بین دو متغیر X و Y است. برای توصیف این رابطه اگر نمودار پراکنش مشاهدات رسم شود و از محل تمرکز و تجمع نقاط خطی رسم شود، دیده می شود که این رابطه، رابطه ای کامل نیست و

مشاهدات حول این خط پراکنده اند. این دو نوع رابطه تابعی و آماری بین دو متغیر، نسبت به مشاهدات X و Y به صورت خطی یا غیر خطی نشان داده می شوند. در این فصل، از برخی مفاهیم پایه ای تحلیل رگرسیونی بحث خواهیم کرد.

۲-۱ رگرسیون خطی ساده

اگر رابطه بین دو متغیر X و Y آماری باشد، مدل رگرسیونی برای مشاهده Y_i به صورت زیر است:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (1-1)$$

مؤلفه $\alpha + \beta X_i$ که رابطه آماری را منعکس می کند، مؤلفه رگرسیونی می خوانند و مؤلفه ε_i که پراکنش تصادفی را منعکس می کند، خطا می نامند. Y_i هم متغیری تصادفی است، به این دلیل که هر مشاهده Y_i ، شامل یک مؤلفه تصادفی ε_i است.

۱-۲-۱ مدل رگرسیون خطی^۱

حال مدل رگرسیونی را که در آن رابطه آماری خطی است مورد بررسی قرار می دهیم:

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, 2, \dots, n \quad (2-1)$$

که در آن:

- Y_i ، پاسخ در i امین مورد است.
- x_i ، مقدار متغیر مستقل در i امین مورد، مقداری ثابت و معلوم فرض شده است.
- α و β ، پارامترند.

^۱) Linear regression model

• ε_i ها، متغیرهای مستقل دارای توزیع نرمال $N(0, \sigma^2)$ هستند.

مدل (۲-۱) را مدل رگرسیون خطی ساده می نامند. این مدل شامل سه پارامتر α ، β و σ^2 می باشد.

۲-۲-۱ تابع رگرسیون خطی^۱

می خواهیم مقدار امید Y_i را بدست آوریم. با توجه به شرایطی که در بالا ذکر کردیم، داریم:

$$E(Y_i|x) = E(\alpha + \beta x_i + \varepsilon_i) = E(\alpha + \beta x_i) + E(\varepsilon_i) = \alpha + \beta x_i + E(\varepsilon_i)$$

از آنجاییکه $E(\varepsilon_i) = 0$ است، در پایان با توجه به معلوم بودن x ، به نتیجه زیر می رسیم.

$$E(Y_i|x) = \alpha + \beta x_i \quad (۳-۱)$$

رابطه بین x و $E(Y_i)$ را تابع رگرسیونی می نامند. پارامتر α ، عرض از مبدأ خط رگرسیونی و پارامتر β ، هم شیب خط رگرسیونی است. حال می خواهیم واریانس Y_i را بیابیم.

$$Var(Y_i|x) = Var(\alpha + \beta x_i + \varepsilon_i)$$

از آنجاییکه $\alpha + \beta x_i$ مقداری ثابت است، داریم:

$$Var(Y_i|x) = Var(\varepsilon_i) = \sigma^2 \quad (۴-۱)$$

پس صرفنظر از مقادیر x_i ها، Y_i ها تغییرپذیری یکسانی دارند.

چون Y_i تابعی خطی از متغیر تصادفی نرمال ε_i است، می توان گفت که هر Y_i هم به صورت نرمال توزیع شده است و سرانجام، چون ε_i ها برای مشاهدات مختلف مستقل فرض شده بودند، Y_i ها نیز چنین خواهند بود. بنابراین Y_i ها متغیرهای تصادفی نرمال مستقلی هستند.

^۱) Linear regression function

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2) \quad (5-1)$$

۳-۲-۱ برآورد α و β و σ^2

معمولاً پارامترهای رگرسیونی α و β و σ^2 مجهول اند و باید از داده های نمونه برآورد شوند. بنابراین برای برآورد این پارامترها از برآوردگرهایی مانند برآوردگر کمترین مربعات^۱ و برآوردگر درستنمایی ماکسیمم^۲ استفاده خواهیم کرد.

الف) برآوردگرهای کمترین مربعات

اگر نمودار پراکنشی از مشاهدات را در نظر بگیریم و بخواهیم خط راستی را به عنوان بهترین برازش از این داده ها رسم کنیم، روش کمترین مربعات، مجموع توانهای دوم انحرافهای مشاهدات Y از خط راست را بعنوان معیار برازش در نظر می گیرد. در اینجا این مجموع را با Q نشان می دهیم. روش کمترین مربعات، خطی را بعنوان بهترین خط راست برازش داده شده در نظر می گیرد که برای آن مقدار Q یعنی مجموع توانهای دوم انحرافها کمترین مقدار ممکن را نشان دهد.

برآوردهای کمترین مربعات α و β را به ترتیب با a و b نشان می دهیم، بنابراین خط رگرسیونی برازش داده شده با برآوردهای کمترین مربعات برابر خواهد بود با:

$$\hat{Y}_i = a + bX_i$$

پس برای مقدار Q داریم:

$$Q = \sum_{i=1}^n [Y_i - (a + bX_i)]^2 \quad (6-1)$$

مطابق با اصل کمترین مربعات، مقادیر a و b را باید طوری تعیین کنیم که Q کمینه شود.

^۱) Least squares estimator

^۲) Maximum likelihood estimator

برای بدست آوردن برآوردهای a و b ، یک راه تحلیلی وجود دارد. برای ارائه ساده تر این

راه حل مجموعه‌های زیر را تعریف می‌کنیم:

$$S_X^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2 \quad (7-1)$$

$$S_Y^2 = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2 \quad (8-1)$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - n\bar{X}\bar{Y} \quad (9-1)$$

که در آن،

- \bar{X} و \bar{Y} ، میانگینهای مقادیر X_i ها و Y_i ها،
- S_X^2 و S_Y^2 ، مجموع توانهای دوم انحرافها از میانگینها،
- S_{XY} ، مجموع حاصلضربهای برداری انحرافها،

ابتدا می‌نویسیم

$$Y_i - (a + bX_i) = (Y_i - \bar{Y}) - b(X_i - \bar{X}) + (\bar{Y} - a - b\bar{X})$$

با مربع کردن دو طرف داریم:

$$(Y_i - a - bX_i)^2 = (Y_i - \bar{Y})^2 + b^2(X_i - \bar{X})^2 + (\bar{Y} - a - b\bar{X})^2$$

$$-2b(X_i - \bar{X})(Y_i - \bar{Y}) - 2b(X_i - \bar{X})(\bar{Y} - a - b\bar{X}) + 2(Y_i - \bar{Y})(\bar{Y} - a - b\bar{X})$$

هر دو طرف رابطه را بر روی مقادیر $i = 1, 2, \dots, n$ جمع بندی می‌کنیم. چون

$$\sum (X_i - \bar{X}) = 0$$

$$\sum (Y_i - \bar{Y}) = 0$$

با توجه به معادلات (7-1)، (8-1) و (9-1) بنابراین داریم:

$$Q = S_Y^2 + b^2 S_X^2 + n(\bar{Y} - a - b\bar{X})^2 - 2b S_{XY}$$

می‌توانیم جملات را اینگونه نیز مرتب کنم،

$$Q = n(\bar{Y} - a - b\bar{X})^2 + (b^2 S_X^2 - 2b S_{XY} + \frac{S_{XY}^2}{S_X^2}) + S_Y^2 - \frac{S_{XY}^2}{S_X^2}$$

$$Q = n(\bar{Y} - a - b\bar{X})^2 + (b^2 S_X^2 - 2b S_{XY} + \frac{S_{XY}^2}{S_X^2}) + S_Y^2 - \frac{S_{XY}^2}{S_X^2}$$

اگر a و b موجود در رابطه را برابر با $a = \bar{Y} - b\bar{X}$ و $b = \frac{S_{XY}}{S_X^2}$ قرار دهیم، دو جمله

اول را به کوچکترین مقدار، یعنی صفر کاهش خواهیم داد.

بنابراین برآوردکننده های کمترین مربعات برای α و β که از اینجا به بعد آنها را با $\hat{\alpha}$ و $\hat{\beta}$

نشان می دهیم، عبارت اند از:

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} \quad (10-1)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad (11-1)$$

که با توجه به تعریف S_{XY} و S_X^2 ، برای (10-1) داریم

$$\hat{\beta} = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sum(x_i - \bar{X})^2}$$

سرانجام خط رگرسیون کمترین مربعات عبارت خواهد بود از:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \quad (12-1)$$

معادله (12-1) برآوردی از تابع رگرسیونی $E(Y_i|X) = \alpha + \beta X_i$ است که آن را تابع

رگرسیونی برآورد شده می نامند.

بطور کلی در روش کمترین مربعات، برآوردکننده های $\hat{\alpha}$ و $\hat{\beta}$ را به قسمی پیدا می کنیم که

مجموع توانهای دوم انحرافها $Y_i - \hat{Y}_i$ را مینیمم سازد. برای این مینیم سازی روش دیگری نیز

وجود دارد که در اینجا به ذکر آن می پردازیم.

برای یافتن مقادیر $\hat{\alpha}$ و $\hat{\beta}$ که Q را مینیمم سازد، نسبت به $\hat{\alpha}$ و $\hat{\beta}$ مشتق می گیریم و حاصل

را برابر صفر قرار می دهیم: