به نام خدا

دانشگاه صنعتی اصفهان

دانشکده علوم ریاضی

بررسی نرخ شادی جهانی سال ۲۰۱۹

پروژه کارشناسی ارشد یادگیری آماری گرایش علوم داده

پریسا شفائی مهر

استاد

دکتر ساره گلی

تیر۱۴۰۱

فهرست مطالب

- تحلیل رگرسیونی

مدل خطی ساده

مدل خطی چند گانه

- نتیجه گیری و چکیده
- پیوست ها

جداول و نمودار ها

کدها

مقدمه

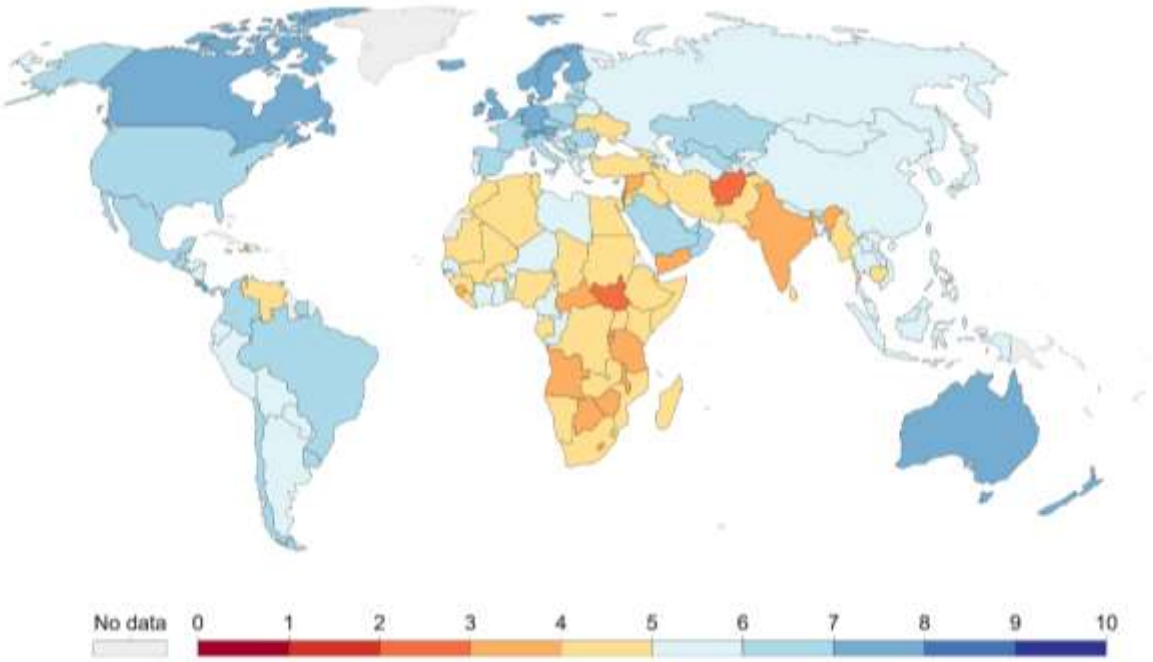
هدف از انجام پروژه و اهمیت آن

این پروژه با هدف بررسی سطح شادی انسان ها در سراسر جهان و بررسی فاکتور های تاثیر گذار مختلف بر روی این نرخ انجام شده است.

تحلیل های این پروژه مبتنی بر داده های منتشر شده سال ۲۰۱۹ میلادی است.

گزارش جهانی شادی یک بررسی مهم از وضعیت شادی جهانی است. اولین گزارش در سال ۲۰۱۲، گزارش دوم در سال ۲۰۱۳، سومین گزارش در سال ۲۰۱۵ و چهارمین گزارش در به روز رسانی ۲۰۱۶ منتشر شد. شادی جهانی ۲۰۱۷، که ۱۵۵ کشور را بر اساس سطح شادی آنها رتبه بندی می کند، در سازمان ملل متحد در رویدادی به مناسبت روز جهانی شادی در ۲۰ مارس منتشر شد. این گزارش همچنان به رسمیت شناخته شدن جهانی ادامه می‌دهد زیرا دولت‌ها، سازمان‌ها و جامعه مدنی به طور فزاینده‌ای از شاخص‌های شادی برای اطلاع‌رسانی در تصمیم‌گیری‌های سیاست‌گذاری خود استفاده می‌کنند. کارشناسان برجسته در سراسر زمینه ها – اقتصاد، روانشناسی، تجزیه و تحلیل نظرسنجی، آمار ملی، بهداشت، سیاست عمومی و موارد دیگر – توضیح می دهند که چگونه اندازه گیری های شاخص های مختلف می تواند به طور موثر برای ارزیابی پیشرفت کشورها استفاده شود. این گزارش وضعیت شادی را در جهان امروز بررسی می کند و نشان می دهد که چگونه علم داده میتواند بااستفاده از ابزارهای مختلفی که در دست دارد، عوامل مختلف تاثیر گذار بر شادی را بررسی کند.

نقشه پراکندگی مربوط به میزان نرخ شادی در سطح جهان در سال ۲۰۱۹



بررسی مجموعه داده

دلیل انتخاب این مجموعه داده آن است که هدف من در پروژه عوامل تاثیر گذار برروی نرخ شادی در شرایط عادی زندگی بوده. ازپایان سال ۲۰۱۹ به بعد با گسترش ویروس کرونا و شروع پاندمی عواملی برروی نرخ شادی انسان ها تاثیر گذار بودند که متفاوت با شرایط عادی زندگی مردم در کشورهای مختلف است.

در این مجموعه کشور فنلاند شادترین کشور و سودان جنوبی غمگین ترین کشور ارزیابی شده است.

امتیازات و رتبه بندی شادی استفاده شده در این پروژه از داده های نظرسنجی جهانی گالوپ استفاده می کند. نمرات بر اساس پاسخ به سوال اصلی ارزیابی زندگی مطرح شده در نظرسنجی است. این سوال که به عنوان نردبان کانتریل شناخته می شود، از پاسخ

دهندگان می خواهد که به نردبانی فکر کنند که بهترین زندگی ممکن برای آنها ۱۰ باشد و بدترین زندگی ممکن ۰ باشد و به زندگی فعلی خود در آن مقیاس رتبه بندی کنند.
ستون‌های بعد از امتیاز شادی میزانی را تخمین می‌زنند که هر یک از شش عامل – تولید اقتصادی، حمایت اجتماعی، امید به زندگی، آزادی، عدم وجود فساد و سخاوت – در بالا بردن ارزیابی‌های زندگی در هر کشور نقش دارند.

مجموعه داده شامل ۹ ستون و ۱۵۶ مشاهده است که مربوط است به ۷ ویژگی بررسی شده مربوط به ۱۵۶ کشور در کل جهان که هدف بررسی و تحلیل ۶ مورد از متغیر های بررسی شده برای روی نرخ شادی در هر کشور هستند.

در طول پروژه از ابزار های R و Rapidminer استفاد شده است که البته قسمت اعظم کار با زبان برنامه نویسی rانجام گردیده.

(تمامی کدهای استفاده شده در پیوست گزارش قابل مشاهده هستند)

در جدول زیر اطلاعات مربوط به مجموعه داده را که بااستفاده از کد های پیوست شده در نرم افزار rخروجی گرفته ام مشاهده میکنید.

کد شماره ۱

متغیر	توضیح	نوع	کمترین مقدار	بیشترین مقدار	میانگین	میانه
Overall.rank	رتبه کشور بر اساس امتیاز شادی.	Int	۱	۱۵۶		
Country.or.region	نام کشور و یا منطقه	Chr				
Score	معیاری که با پرسیدن این سوال از افراد نمونه اندازه گیری شد: "شادمانی خود را چگونه ارزیابی می کنید.	Num	۲.۸۵۳	۷.۷۶۹	۵.۴۰۷	۵.۳۸۰
GDP.per.capita	میزان تولید ناخالص داخلی	Num	۰.۰۰۰۰	۱.۶۸۴	۰.۹۰۵۱	۰.۹۶۰۰
Social.support	میزان حمایت دولت	Num	۰.۰۰۰	۱.۶۲۴	۱.۲۰۹	۱.۲۷۲
Healthy.life.expectancy	میزان امید به زندگی	Num	۰.۰۰۰۰	۱.۱۴۱۰	۰.۷۲۵۲	۰.۷۸۹۰
Freedom.to.make.life.choices	میزان حق انتخاب و آزادی مردم	Num	۰.۰۰۰۰	۰.۶۳۱۰	۰.۳۹۲۶	۰.۴۱۷۰
Generosity	حمایت افراد جامعه از یکدیگر	Num	۰.۰۰۰۰	۰.۵۶۶۰	۰.۱۸۴۸	۰.۱۷۷۵
Perceptions.of.corruption	نرخ فساد	Num	۰.۰۰۰۰	۰.۴۵۳۰	۰.۱۱۰۶	۰.۰۸۵۵

آماده سازی داده

آماده سازی داده

کد شماره ۱

درباره پیش پردازش داده ها بطور مفصل در سمینار ارائه شده گفته شد.

پیش‌پردازش نقشی اساسی در روند پردازش داده‌ها و نتایج حاصل از آن‌ها ایفا می‌کند.

ابتدا داده ها ازمنظر داده های از دست رفته و داده های تکراری بررسی شدند

این مجموعه داده فاقد داده از دست رفته و تکراری است.

هم چنین مجموعه داده را از منظر وجود داده پرت بررسی کردیم ، که در بخش مصور سازی داده و نمودار ها ، نمودار جعبه ای این مجموعه داده را مشاهده خواهید کرد. .

توزیع متغیر ها

بررسی نرمال بودن متغیر هدف (score)

کد شماره ۲

فرض نرمال بودن نه تنها برای تعیین ماهیت توزیع مرتبط است، بلکه به انجام آزمون های آماری مختلف مانند رگرسیون نیز کمک می کند. برخی از آزمون‌ها مانند t-test، ANOVA یک طرفه و دو طرفه برای تجزیه و تحلیل نیاز به یک نمونه توزیع شده نرمال دارند. بدون اینکه مجموعه داده به طور معمول توزیع شود، نتایج به دست آمده از تجزیه و تحلیل ماهیت ضعیفی خواهند داشت. بنابراین، آزمون نرمال بودن متغیر قبل از هر آزمون استنباطی دیگر ضروری است.

به طور کلی دو دسته وجود دارد که بر اساس آنها آزمون نرمال بودن می تواند انجام شود، یعنی گرافیکی و آماری.

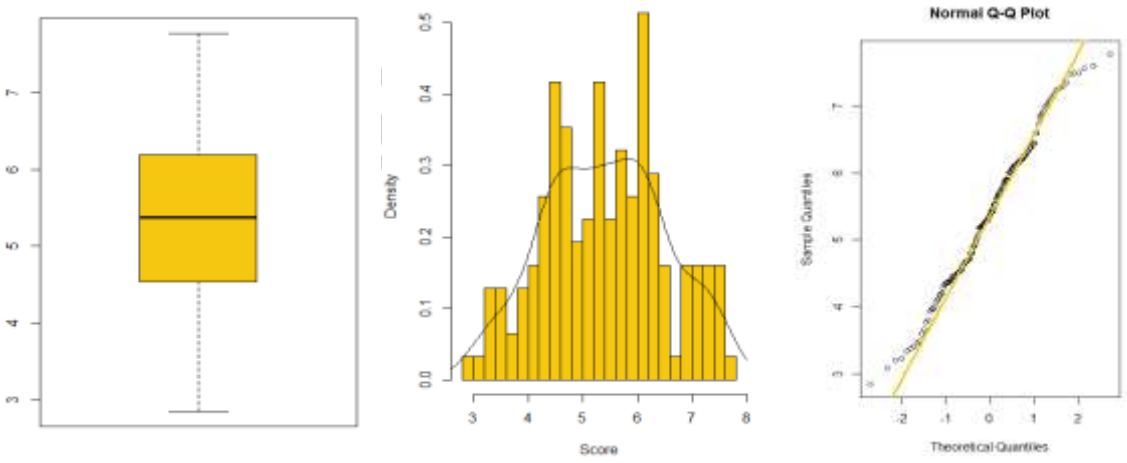
روش های تست نرمال بودن به صورت گرافیکی

آزمون گرافیکی برای نرمال بودن یک روش بصری برای استنتاج اطلاعات از نمودار داده ها است. روش گرافیکی نتایج دقیقی ارائه نمی دهد و فقط بر اساس قضاوت تحلیل گر است. بنابراین، این روش غیرقابل اعتماد است و وجود توزیع نرمال را برای یک متغیر تضمین نمی کند. علاوه بر این، این روش برای حجم نمونه بسیار بزرگ نیز مناسب نیست. نرمال بودن داده ها را می توان به روش های مختلف به صورت گرافیکی آزمایش کرد.

بررسی نرمال بودن متغیر score با استفاده از روش های گرافیکی

نمودار هیستوگرام ،نمودار جعبه ای و qqline وqqnorm متغیر scoreرا رسم میکنیم.

بااستفاده از نمودار های رسم شده استنباط می کنیم که توزیع داده های این متغیر نرمال است به دلیل آن که نمودار جعبه ای متقارن است هم چنین نمودار هیستوگرام شکل زنگوله ای دارد و نمودار qq داده ها تقریبا(بااغماض) بر روی خط قرار دارند که این هم گواهی بر نرمال بودن توزیع داده هاست.



روش گرافیکی تست نرمال بودن برای قضاوت اولیه مناسب است اما قابل اعتماد نیست. از نظر بصری می توان تشخیص داد که منحنی زنگی شکل تشکیل شده است یا مقدار مشاهده شده نزدیک به خط توزیع شده عادی است، اما نتایج واقعی را نمی توان ایجاد کرد. بنابراین انجام آزمون تجربی نرمال بودن با استفاده از نرم افزارهای آماری حائز اهمیت است پس حالا بااستفاده از روش های آماری نرمال بودن متغیر هدف را بررسی میکنیم.

بااستفاده ازتست شاپیرو و k.s testرا برروی متغیر پاسخ اعمال میکنیم که باتوجه به مقدار خروجیp-value = 0.1633 و p-value< 2.2e-16که یکی فرض را رد و دیگری پذیرش کرد پس فرض نرمال بودن پذیرفته می شود .

فرضیه های زیر مطرح شد:

Ho:نمونه نرمال به نظر نمی رسد

Ha :نمونه نرمال به نظر می رسد

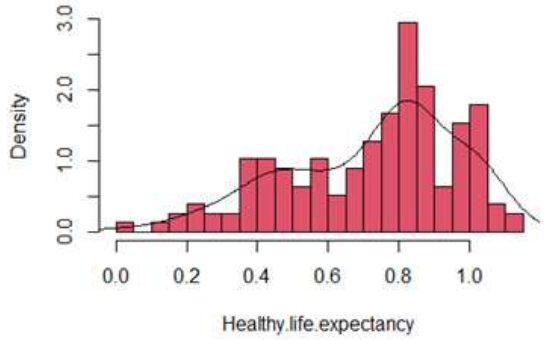
علاوه بر این باانجام این تست برروی GDP ، و همین طور در ابتدا ر سم نمودار های تست نرمالیتی تایید می شود که تولید ناخالص داخلی به طور مساوی در سراسر جهان توزیع شده است،و توزیع نرمال دارد.

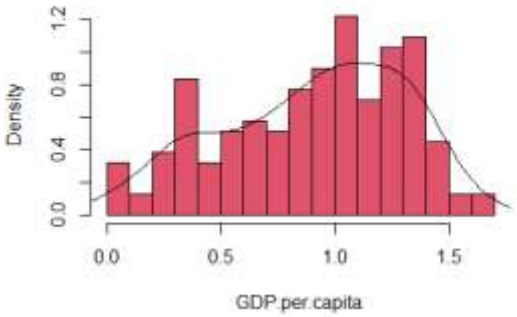
و در میانگین داده های آن که ۰.۹ است متمرکز شده است.

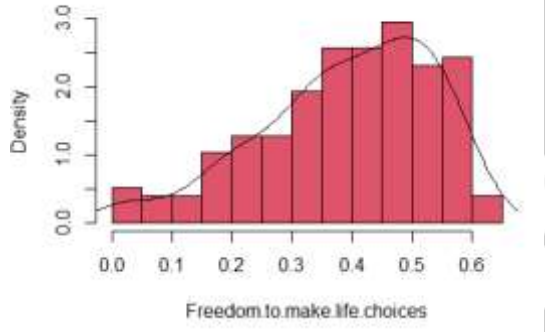
متغیرهای حمایت اجتماعی، امید به زندگی سالم وآزادی در انتخاب زندگی، نامتقارن است.

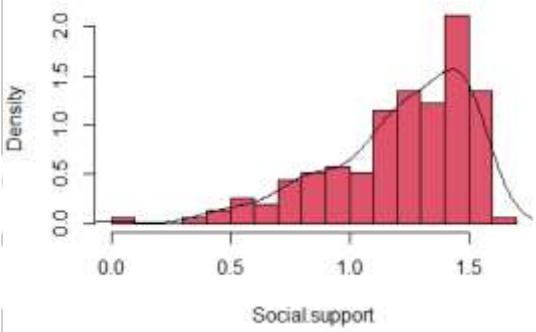
نگاهی به توزیع متغیر های توضیحی

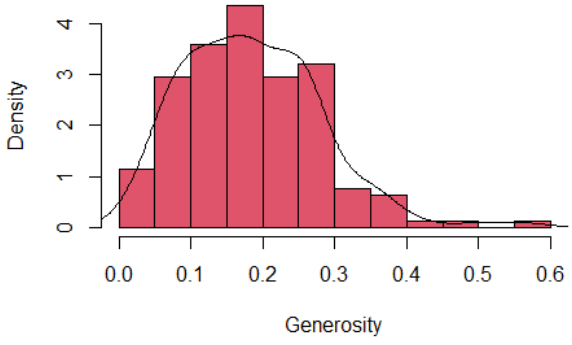
نمودار هیستوگرام متغیر ها

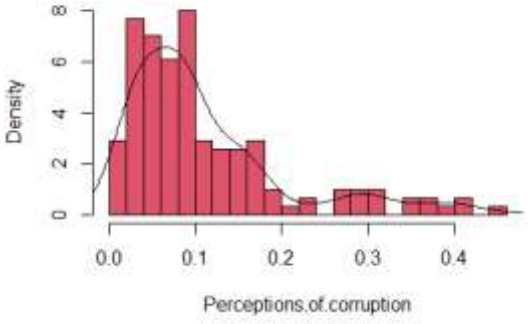




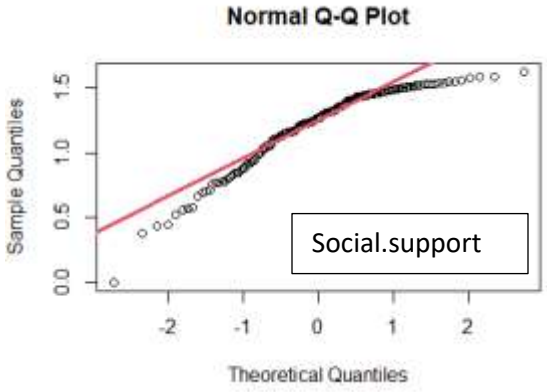


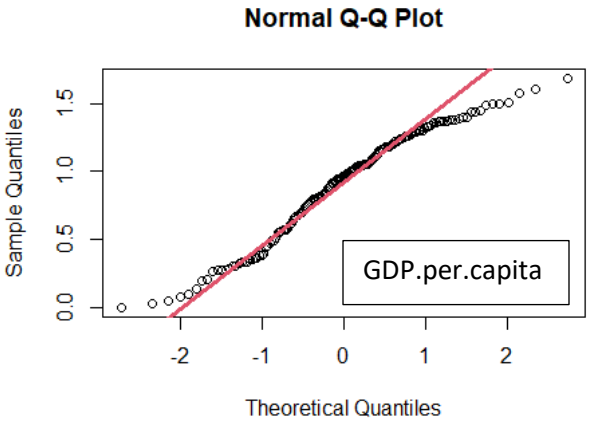


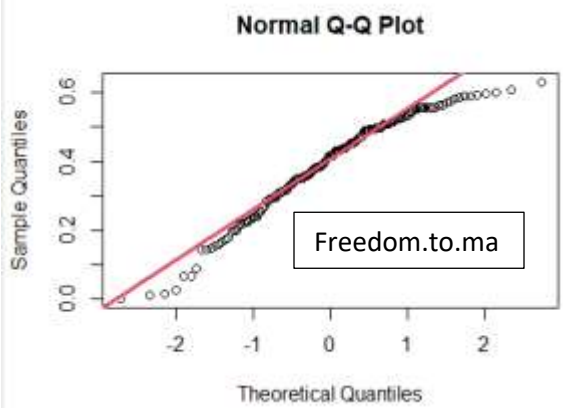


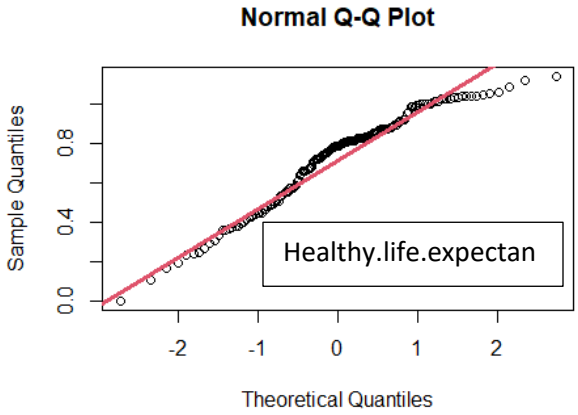


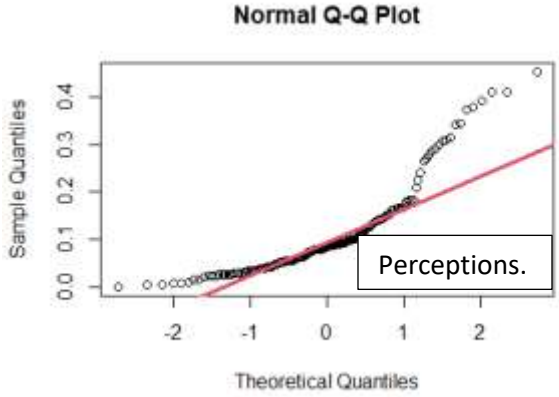
نمودار های qqplotو qqnorm

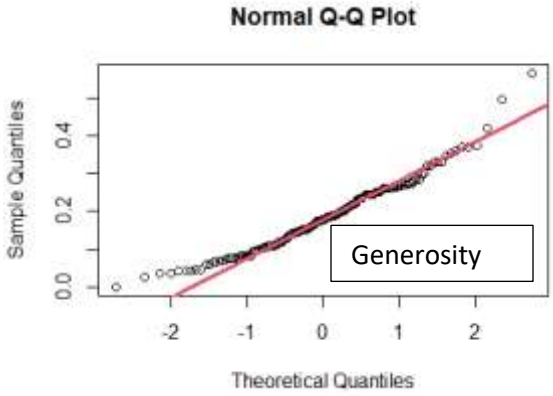












برآورد یابی

برآورد یابی نقطه ای

کد شماره۳

متغیر	میانگین	میانه	واریانس	چولگی	کشیدگی
Score	۵.۴۰۷۰۹۶	۵.۳۸۰	۱.۲۳۹۰۳۶	۰.۰۱۱۳۳۹۵۶	۲.۳۷۲۷۳۳
Healthy.life.expectancy	۰.۷۲۵۲۴۳۶	۰.۷۸۹۰	۰.۰۵۸۶۲۴۰۳	-۰.۶۰۷۹۲۲۴	۲.۶۶۸۵۱
GDP.per.capita	۰.۹۰۵۱۴۷۴	۰.۹۶۰۰	۰.۱۵۸۷۱۴۲	-۰.۳۸۱۵۱۸۲	۲.۲۱۶۳۳۷
Social.support	۱.۲۰۸۸۱۴	۱.۲۷۲	۰.۰۸۹۵۱۵۴۹	-۱.۱۲۳۷۸۷	۴.۱۵۱۷۵
Freedom.to.make.life.choices	۰.۳۹۲۵۷۰۵	۰.۴۱۷۰	۰.۰۲۰۵۳۱۸۷	-۰.۶۷۹۰۲۵۳	۲.۸۹۵۱۱۴
Generosity	۰.۱۸۴۸۴۶۲	۰.۱۷۷۵	۰.۰۰۹۰۷۳۴۰۸	۰.۷۳۸۷۴۹۹	۴.۰۹۷۷۰۷
Perceptions.of.corruption	۰.۱۱۰۶۰۲۶	۰.۰۸۵۵	۰.۰۰۸۹۳۷۴۰۲	۱.۶۳۴۴۹۸	۵.۳۰۱۸۳۷

Generosity	مقدار اصلی	مقدار تخمین زده شده پارامتر با بیشینه درست نمایی با دستور nlm	مقدار تخمین زده شده پارامتر با بیشینه درست نمایی با دستور optim	مقدار تخمین زده شده با برآورد گشتاوری
میانگین	۰.۱۸۴۸۴۶۲	۰.۱۸۴۸۴۵۸۷۴	۰.۱۸۴۸۳۹۲۷۰	۰.۱۸۴۸۴۶۲
واریانس	۰.۰۰۹۰۷۳۴۰۸	۰.۰۰۹۰۱۵۳۱۶	۰.۰۰۹۰۱۳۵۹۳	۰.۰۰۹۰۷۳۴۰۸

برآورد بیشینه درست نمایی و گشتاوری

کد شماره۴

بررسی هادر بخش توزیع متغیر ها نشان می دهد متغیر Generosity دارای توزیع نرمال است.

مقادیر اصلی مختص به میانگین و واریانس این برآورد گر در جدولی در قسمت ابتدای این بخش آورده شده است.

حالا میخواهیم برآوردی برای پارامتر های این متغیر بدست آوریم .

هم چنین فاصله اطمینان ۹۵ درصدی برای میانگین این متغیر بدست می آوریم.

مقدار خروجی مشاهده شده نشان دهنده ان است که برآورد های بدست آمده از روش های گشتاوری و بیشینه درست نمایی گرچه که با مقادیر اصلی برابر نیستند اما تخمین ها بسیار نزدیک به مقادیر اصلی هستند.

فاصله اطمینان ۹۵ درصدی برای برآورد ماکسیمم درست نمایی این متغیر ۰.۱۹۹۷۹۳۷و ۰.۱۶۹۸۹۸۶

تابع فاصله اطمینان را برای میانگین وقتی حجم نمونه زیاد است تعریف شده است و فاصله اطمینان برای متوسط بخشندگی در سطح اطمینان ۹۵ درصد است.

این بدان معناست که از ۱۰۰ فاصله ۹۵ درصد در این بازه شامل پارامتر های جامعه هستند.

آزمون فرض و فاصله اطمینان و پاسخ به سوالات و چالش ها

آزمون فرض و فاصله اطمینان

کد شماره۵

در پژوهش‌های انجام شده درباره نرخ شادی و عوامل تاثیر گذار براآن ، اعتقاد بر این است که سطح درآمد لزوماً بر شادی افراد تأثیر نمی‌گذارد. بنابراین، به دنبال تایید این موضوع است که کشورهایی با شادی بالا اما نرخ تولید ناخالص داخلی پایین وجود دارند.

بررسی با سطح معناداری ۰.۰۵ آزمون می شود.هم چنین میخواهیم یک فاصله اطمینان ۹۵ درصدی برای این شاخص بیابیم.

برای این منظور میانگین μ تولید ناخالص داخلی به عنوان مرجع در نظر گرفته شد که با توجه به تحلیل ها در قسمت برآورد یابی و جداول پیوست شده ۰/۹ نشان داده شده و همچنین در قسمت چک کردن نوع توزیع متغیر ها بااستفاده از تست های گرافیکی و آماری استنباط کردیم که دارای توزیع نرمال است.

چون فرض نرمالیتی برقرار است می توان طبق آزمون تی.تست آزمون میانگین را برای متغییر تولید ناخالص ملی آزمون فرض را برقرار کرد.

فرضیه های زیر مطرح شد:

Ho:همه کشورهای شاد دارای نرخ تولید ناخالص داخلی بزرگتر از میانگین هستند GDP rate>μ

Ha:کشورهای شادی وجود دارند که نرخ تولید ناخالص داخلی آنها کمتر از میانگین است GDP rate<μ

از آزمون t استفاده شد که در آن مشخص شد 0.05<p_value است.

شواهدبرای ردنشدن HO وجود دارد که اشاره می کند همه کشورهای شاد دارای نرخ تولید ناخالص داخلی بزرگتر از میانگین هستند(GDP > μ)هستند.و هیچ کشوری وجود ندارد که شاد باشد ولی تولید ناخالص ملی پایینی داشته باشد.

تابع فاصله اطمینان را برای میانگین وقتی حجم نمونه زیاد است تعریف شده

بازه اطمینان بدست آمده ۹۵ درصدی برای این شاخص ۰.۹۶۷۶۶۳۸ و ۰.۸۴۲۶۳۱۱ است.

سوالات چالش برانگیز پروژه :

کد شماره ۶

سوال ۱

از جنبه اجتماعی برای جوامع این موضوع قابل اهمیت است که آیا میزان حمایت دولت از افراد آن جامعه تاثیری در پایین آمدن نرخ امید به زندگی در افراد آن جوامع دارد یا خیر؟

از این رو میخواهیم بررسی کنیم که در جوامعی که مقدار حمایت دولت در آنها از میانگین بالاتر است و جوامعی که دارای حمایت دولت پایین تر از حد معمول هستند چه میزان تفاوت در نرخ امید به زندگی درافراد آن جامعه مشاهده می شود.

بااستفاده از کد های اعمال شده نتیجه میگیریم میانگین نرخ امید به زندگی در جوامعی با میزان حمایت دولت بیشتر از میانگین،۰.۸۶۷۹۲۵۵ است و در جوامعی با میانگین حمایت دولت کمتر از میانگین ، ۰.۵۰۸۹۱۹۴ است.

که اختلاف ۰.۳۵۹۰۰۶۱ از هم دارند.

این اندازه گیری ها نشانگر آن است که در جوامعی که از میزان حمایت بالایی از طرف دولت برخوردار هستند بااختلاف چشم گیری از امید به زندگی بالاتری نسبت به گروه دیگر برخوردار هستند.

سوال ۲

در پژوهش ها مختلف انجام شده از لحاظ اقتصادی ، اجتماعی همواره پژوهشگران بااین سوال روبرو بوده اند که میزان ثروتمند بودن کشور از لحاظ اقتصادی چه میزان بر امید به زندگی و تمایل افراد جامعه برای بقا و هم چنین مقدار شاد بودن افراد تاثیر گذار است و آیا تفاوت چشم گیری بین سطح شادی و امید به زندگی بین این دو گروه وجود دارد یا خیر؟

بااستفاده ازخروجی های زیر را نتیجه گرفتیم که نشان دهنده تاثیر به سزای ثروتمندی افراد برروی میزان امید به زندگی آنها و سطح شادی آنهاست.

Healthy.life.expectancy	Score	
۰.۸۷۸۸۶۳۶	۶.۰۱۳۲۰۵	GDP rate>μ
۰.۵۲۶۴۴۱۲	۴.۶۲۲۷۲۱	GDP rate<μ

تحلیل رگرسیونی

مدل رگرسیونی

در مدل‌های آماری، تحلیل رگرسیون، یک فرایند آماری برای تخمین روابط بین متغیرها می‌باشد. این روش شامل تکنیک‌های زیادی برای مدل‌سازی و تحلیل متغیرهای خاص و منحصر بفرد، با تمرکز بر رابطه بین متغیر وابسته و یک یا چند متغیر مستقل، می‌باشد. تحلیل رگرسیون خصوصاً کمک می‌کند به فهم اینکه چگونه مقدار متغیر وابسته با تغییر هرکدام از متغیرهای مستقل و با ثابت بودن دیگر متغیرهای مستقل تغییر می‌کند.

اگر چه از تکنیک های بسیاری برای تحلیل و پیش بینی میتوان استفاده کرد اما در این پروژه رویکرد مااستفاده از مدل های رگرسیونی خطی برای تحلیل و پیش بینی است.

مدل رگرسیون خطی

پیش فرض های استفاده از مدل رگرسیونی خطی ساده

معمولاً چند پیش‌فرض برای استفاده از رگرسیون خطی در نظر گرفته می‌شود. اگر اختلاف بین متغیر وابسته و پیش‌بینی مدل را «خطا» یا «مانده» بنامیم، آنگاه مفروضات زیر باید در مدل‌سازی رگرسیون خطی برقرار باشند.

مانده ها از توزیع نرمال پیروی کنند

مانده ها از هم مستقل باشند

واریانس مانده ها ثابت باشد

تخمین پارامترها

فرق رگرسیون خطی با سایر مدل‌های رگرسیون در این است که در این مدل رابطه بین متغیرهای مستقل و متغیر وابسته یک رابطه خطی فرض می‌شود. رگرسیون خطی، که خود نوعی تابع پیش‌بینی‌کننده خطی است، پیش‌بینی متغیر وابسته را از حاصل‌جمع ضرب متغیرهای مستقل در یک سری ضرایب به‌دست می‌آورد. در رگرسیون خطی ساده که تنها یک متغیر مستقل وجود دارد، پیش‌بینی متغیر وابسته شکل یک خط مستقیم به خود می‌گیرد؛ در رگرسیون خطی با دو متغیر شکل پیش‌بینی یک صفحه خواهد بود، و در رگرسیون خطی با

بیش از دو متغیر مستقل پیش‌بینی متغیر وابسته به صورت یک اَبَرصفحه خواهد بود.

برای تخمین این مدل رگرسیون باید سه پارامتر تخمین زده بشوند: دو ضریب β 0 1 βو مانده ها (e i). روش رایج برای به‌دست‌آوردن پارامترها، روش کمترین مربعات است. در این روش پارامترها را با کمینه‌کردن مجموع مربعات خطا به دست می‌آورند.

فرمول های مربوطه :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, N$$

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i$$

$$e_i = y_i - \hat{y}_i$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\mathrm{SSE} = \sum_{i=1}^{N} e_i^2$$

مدل شماره یک

مدل رگرسیونی خطی ساده

کد شماره۷

متغیر scoreکه متعلق به میزان نرخ شادی است همانطور که در ابتدای گزارش گفته شد متغیر هدف ما در این پروژه است.

مدل رگرسیونی را با در نظر گرفتن تمامی متغیر های پاسخ اعمال می شود. تا متغیر توضیحی مناسب و معنا دار برای مدل انتخاب شود

خروجی مدل اعمال شده بصورت زیر می باشد.

	تخمین پارامتر ها	Std. Error	t value	Pr(>|t|)	سطح معناداری	F value	Pr(>F)	سطح معناداری
(Intercept)	۱.۷۹۵۲	۰.۲۱۱۱	۸.۵۰۵	1.77E-14	۰.۰۰۱			
Healthy.life.expectancy	۱.۰۷۸۱	۰.۳۳۴۵	۳.۲۲۳	۰.۰۰۱۵۶۰	۰.۰۱	۴۱۰.۳۶۹	2.2E-16	۰.۰۰۱
GDP.per.capita	۰.۷۷۵۴	۰.۲۱۸۲	۳.۵۵۳	۰.۰۰۰۵۱۰	۰.۰۰۱	۴۵.۲۵۱۹	3.47E-10	۰.۰۰۱
Social.support	۱.۱۲۴۲	۰.۲۳۶۹	۴.۷۴۵	4.38E-.06	۰.۰۰۱	۳۴.۴۳۶۱	2.748E-.08	۰.۰۰۱
Freedom.to.make.life.choices	۱.۴۵۴۸	۰.۳۷۵۳	۳.۸۷۶	۰.۰۰۰۱۵۹	۰.۰۰۱	۳۰.۰۹۶۹	1.722E-.07	۰.۰۰۱
Generosity	۰.۴۸۹۸	۰.۴۹۷۷	۰.۹۸۴	۰.۳۲۶۷۰۹	۱	۲.۳۴۰۱	۰.۱۲۸۲۰	۱
Perceptions.of.corruption	۰.۹۷۲۳	۰.۵۴۲۴	۱.۷۹۳	۰.۰۷۵۰۵۳	۰.۱	۳.۲۱۳۷	۰.۰۷۵۰۵	۰.۱

خطای استاندارد باقی مانده	۰.۵۳۳۵
شاخص R*2	۰.۷۷۹۲
p-value	< 2.2e-16
آماره F	۸۷.۶۲

مانده ها	کمترین مقدار	بیشترین مقدار
	-۱.۷۵۳۰۴	۱.۱۹۰۵۹

بابررسی شاخصه R*2 متوجه می شویم چقد از داده ها به خط برازش مدل نزدیک هستند

در اینجا ۷۷ درصد از داده ما با مدل برازش شده قابل توجیح هستند که نشان دهنده برازش خوبی برای مجموعه داده است.

حالا pvalueرا بررسی میکنیم تا متوجه شویم کل مدل معنا دار است یانه؟

این شاخصه یک اندازه گیری آماری است که برای تایید یک فرضیه در برابر داده های مشاهده شده استفاده می شود. یک مقدار p احتمال به دست آوردن نتایج مشاهده شده را با فرض صحت فرضیه صفر اندازه گیری می کند. هرچه مقدار p کمتر باشد، اهمیت آماری تفاوت مشاهده شده بیشتر است.

در اینجا مقدار pvalueمشاهده شده نشانگر معنادار بودن مدل است .چون فرض صفر که مبتنی بر این است که متغیر ها رابطه خطی با متغیر پاسخ ندارد رد شده و نتیجه میگیریم متغیر های توضیحی بر روی متغیر هدف تاثیر گذار هستند.

نتیجه گیری کلی از مدل اعمال شده:

مدل بر روی داده های آموزشی با استفاده از همه متغیرها برازش می شود F-Statistic . با p پایین آن نشان می دهد که مدل قابل توجه است. مقدار R2 به ما می گوید که ۷۷ درصد از تغییرات در نرخ شادی را می توان با مدل توضیح داد. تعدادی از متغیرها معنی دار به نظر می رسند، بنابراین فرضیه صفر را می توان برای آنها رد کرد. با این حال، سودمندی کلی مدل را استنباط میکنیم.

در این مرحله به دلیل آن که هدف از انجام پروژه اعمال رگرسیون خطی ساده برروی داده هاست

باید معنادار ترین متغیر را از بین متغیر های توضیحی انتخاب و ادامه فرآیند را بااین متغیر ادامه دهیم.

همانطور که در جدول بالا مشخص است متغیر های تولید ناخالص ملی و امید به زندگی و حق آزادی انتخاب و حمایت دولت از لحاظ معنادار بودن برای مدل از سطح معناداری یکسان برخوردار هستند پس باید از راه های دیگر بررسی شود تاتنها یکی از معنادار ترین متغیرها برای مدل را انتخاب کنیم.

همانطور که در پیوست جدول ها و تصاویر قابل مشاهده است R2برای این متغیر ها نسبت به متغیر هدف محاسبه شده .

شاخص R2نشانگر آن است که چقدر از داده ها با خط برازش نزدیک هستند

پس بابررسی این شاخص به این نتیجه میرسیم معنادار ترین متغیر GDP.per.capita

است البته راه های از جمله بررسی شاخص t نیز میتواند بما در بررسی و انتخاب معنادار ترین متغیر کمک کند .

متغیر	R2
GDP.per.capita	۰.۶۳۰۳
Social.support	۰.۶۰۳۸
Freedom.to.make.life.choices	۰.۳۲۱۲
Healthy.life.expectancy	۰.۶۰۸۲

در ادامه مدل رگرسیونی با متغیر توضیحی انتخاب شده و متغیر هدف اعمال می شود.

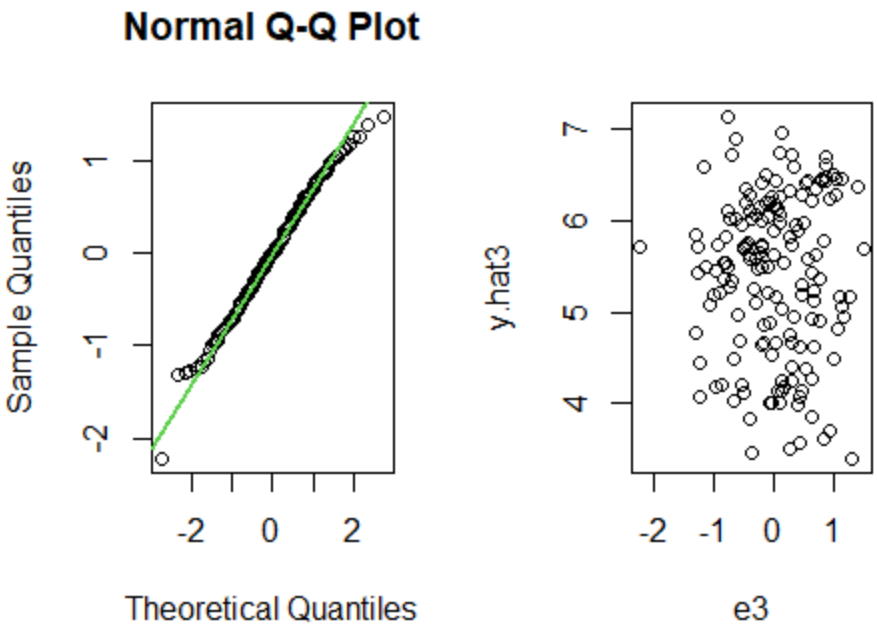
	Estimate	Std. Error	t value	Pr(>|t|)	سطح معناداری	F value	Pr(>F)	سطح معناداری
(Intercept)	۳.۳۹۹۳	۰.۱۳۵۳	۲۵.۱۲	2e-16	۰.۰۰۱			
GDP.per.capita	۲.۲۱۸۱	۰.۱۳۶۹	۱۶.۲۰	2e-16	۰.۰۰۱	۲۶۲.۵	2e-16	۰.۰۰۱

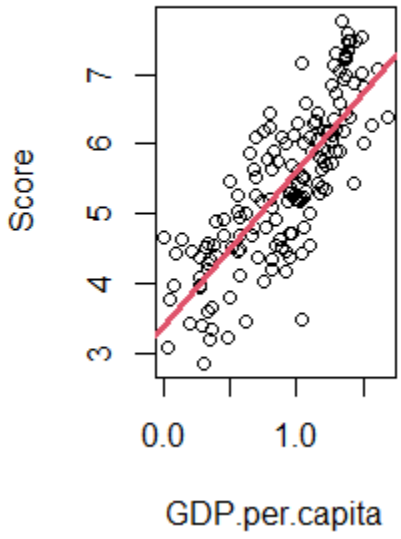
مدل اعمال شده دارای R2 ۶۳ درصد است که مقداری را نشان می دهد که مدل باان توجیح می شود

علاوه برا این دارای مقدار Pکمتر از ۰.۰۵ است که نشان دهنده معناداری مدل برازش شده می باشد.

اما باید به این نکته توجه داشت که این مقادیر نشان گر یک برازش مناسب نیست ومانده ها طبق مطالب گفته شده در ابتدای این فصل باید بررسی شوند و فرضیات مدل باید مورد ارزیابی قرار گیرد.

نمودار های لازم برای بررسی نرمالیتی و ثابت بودن واریانس را رسم میکنیم تا توزیع مانده ها بررسی شوند


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
y.hat3
e3


Score
GDP.per.capita

همانطور که از شکل پیداست مانده ها توزیع نرمال دارند و واریانس نیز غیر ثابت بنظر میرسد از بررسی رابطه بین پیش بینی متغیر توضیحی و مانده ها نیز این استنباط را خواهیم داشت که رابطه خطی موجود و مدل برازش شده مناسب است . این گویای این مسئله است که در سطوح بعدی پیش بینی ها قابل اعتماد خواهند بود.

هم چنین فاصل اطمینان ۹۵ درصدی بدست آمده برای مدل برازش شده بازه ۱.۹۴۷۶۸۹ و۲.۴۸۸۶۰۷ می باشد.

حالا به نسبت ۸۰به ۲۰ داده های آموزش و آزمایش را جدا کرده و مدل را ارزیابی می کنیم.

مدل برازش شده با متغیر توضیحی انتخاب شده را هم برای داده های تست و هم آموزش اعمال کرده سپس با پیش بینی انجام گرفته مقایسه می کنیم تا ببینیم مدل مناسب برای داده ها انتخاب شده یا نه؟

R2مدل برازش شده با داده تست ۰.۷۳۳۷ و مربعات خطای مانده ۰.۶۷۷۷و

R2مدل برازش شده با داده آموزشی ۰.۶۱۱۶و مربعات خطای مانده ۰.۶۶۸۱

است که این خود نشانگر آن است که مدل انتخاب شده احتمالا مدل مناسبی برا مجموعه داده است.

حالا برای اطمینان یک فاکتور دیگر را بررسی میکنیم

Mseبدست آمده از اختلاف داده های آموزشی و پیش بینی ما ۰.۵۰۵۲۸۲

است این شاخصه اندازه گیری نزدیکی یک خط برازش به نقاط داده است. که در اینجا با توجه به مجموعه داده ما عدد مناسبی بنظر می رسد چون نشانگر آن است که انحراف مدل از داده های واقعی تنها ۰.۰۵۰۲۸۲ درصد است.

مدل شماره دو

رگرسیون خطی چند گانه

کد شماره۸

در این مدل میخواهیم از ابتدا داده های آموزشی و آزمایشی را جدا کنیم تا داده های تست و آموزش کاملا ایزوله باشند تا نتایج دقیق تری برای پیش بینی بدست آید.

مجموعه داده شادی برای آموزش و آزمایش به ۲ تقسیم شده است:

۸۰درصد از مجموعه داده، عملکرد آن برای مدل آموزش

۲۰درصد از مجموعه داده، عملکرد آن برای آزمایش مدل

حالا یک مدل رگرسیونی با در نظر گرفتن score به عنوان هدف و اعمال ۶ متغیر توضیحی ایجاد می کنیم.

بعد از اعمال مدل رگرسیونی نتایج بدست آمده به شکل زیر است.

خطای استاندارد باقی مانده	۰.۵۳۷۳
شاخص R*2	۰.۷۶۶۸
p-value	2.2E-16
آماره F	۸۰.۲۲

از سطح معناداری متغیر ها و بررسی AIC آنها نتیجه میگیریم دو متغیر Perceptions.of.corruptionو Generosity به دلیل معناداری پایین برای مدل حذف می شوند. و مدلی جدید مجددا بدون حضور این دو متغیر اعمال می شود.

بر اساس مدل آموزشی ، اطلاعاتی به دست می آوریم:

۱ .بهترین (کمترین) امتیاز ۱۵۱.۹۵ - AIC است.

۲ .مقدار R2 تنظیم شده ۰.۷۵۳ یا ۷۵.۳٪ است، مدل میانگین آن می تواند داده های تغییرات را از متغیر هدف توضیح دهد .

۳ .مقدار p هر متغیر پیش‌بینی کننده کمتر از ۰.۰۵ است، که نشانگر معناداری مدل است ، میانگین هر متغیر پیش‌بینی کننده آن معنادار است یا بر متغیر هدف تأثیر می‌گذارد.

خروجی مدل جدید را که بعد حذف دو متغیر توضیحی کمتر معنادار اعمال شد بصورت زیر می باشد که همانطور که مشاهده می شود تغییر جزئی در شاخصه های آن ایجاد شده و تنها آماره Fتغییر فاحشی داشته که نشانگر تغییرات در پراکندگی داده ها از خط می باشد .

آماره F به سادگی نسبتی از دو واریانس است. واریانس ها معیاری برای پراکندگی یا میزان پراکندگی داده ها از میانگین هستند. مقادیر بزرگتر نشان دهنده پراکندگی بیشتر است. واریانس مربع انحراف معیار است.

از این جا متوجه شدیم که برازش در مدل دوم یعنی بعد از حذف متغیر ها افت داشته در ادامه به ارزیابی مدل می پردازیم.

خطای استاندارد باقی مانده	۰.۵۴۱۹
p-value	2.2E-16
آمارهF	۹۷.۸۱
شاخص R2	۰.۷۵۳

ارزیابی مدل

مدلی بر اساس داده های تست اعمال و پیش بینی را انجام می دهیم

با بررسی خطا با استفاده از RMSE (ریشه میانگین مربعات خطا)، مدل نتیجه به اندازه ۰.۵۳۲۴۳۸۱ از داده های واقعی منحرف می شود. که این مقدار نشان دهنده پیش بینی خوب مدل است.

فرضیات مدل

حالا نوبت به بررسی فرضیات مدل است تا بتوانیم صحت مدل پیش بینی شده را تایید کنیم.

همانطور که در ابتدای فصل گفته شد چند شرط برای بررسی فرضیات باید چک شوند .

استقلال

ابتدا استقلال متغیر ها را با شاخصه ای به نام VIFرا بررسی میکنیم

این شاخصه بیانگر ضریب تورم واریانس است و میزان همبستگی بین یک پیش‌بینی‌کننده و دیگر پیش‌بینی‌کننده‌ها را در یک مدل اندازه گیری می‌کند. برای تشخیص همخطی /

چند خطی استفاده می شود. مقادیر بالاتر نشان می دهد که ارزیابی دقیق سهم پیش بینی کننده ها در یک مدل دشوار تا غیرممکن است.

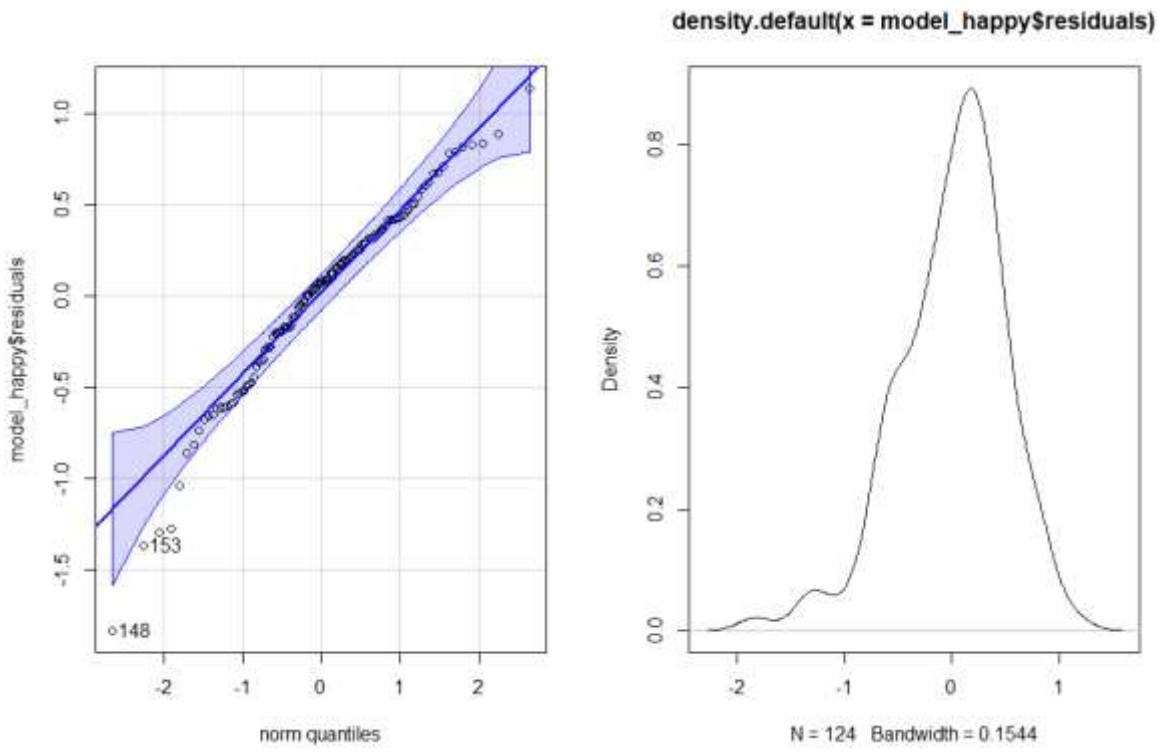
بطور کلی VIFکمتر از ده مقدار قابل قبولی برای تشخیص استقلال متغیر هاست.

Freedom.to.make.life.choices	Healthy.life.expectancy	Social.support	GDP.per.capita
۱.۲۲۶۳۶۶	۳.۳۷۱۵۲۶	۲.۴۰۸۰۵۹	۳.۶۹۵۶۳۲

مقدار این شاخصه برای تمامی متغیر های توضیحی مدل کمتر از ده است که نشانگر استقلال این متغیر هااست.

نرمالیتی

نرمال بودن مانده ها را بااستفاده از روش های گرافیکی چک میکنیم.

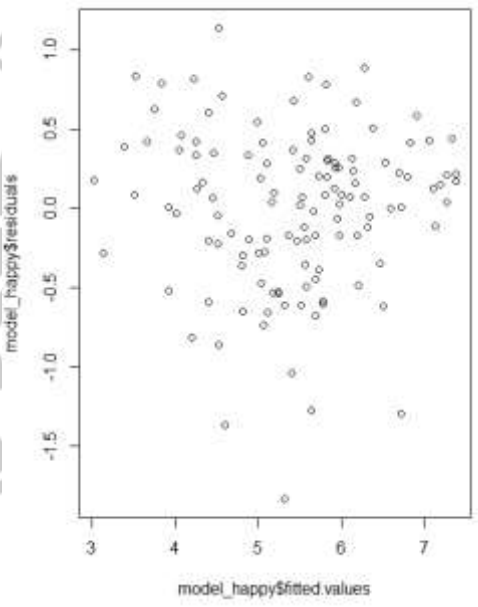


همانطور که از شکل پیداست مانده ها توزیع نرمال دارند. پس دومین فرضیه برای ایکه بتواینم مدل رگرسیونی را بپذیریم نیز تایید شد

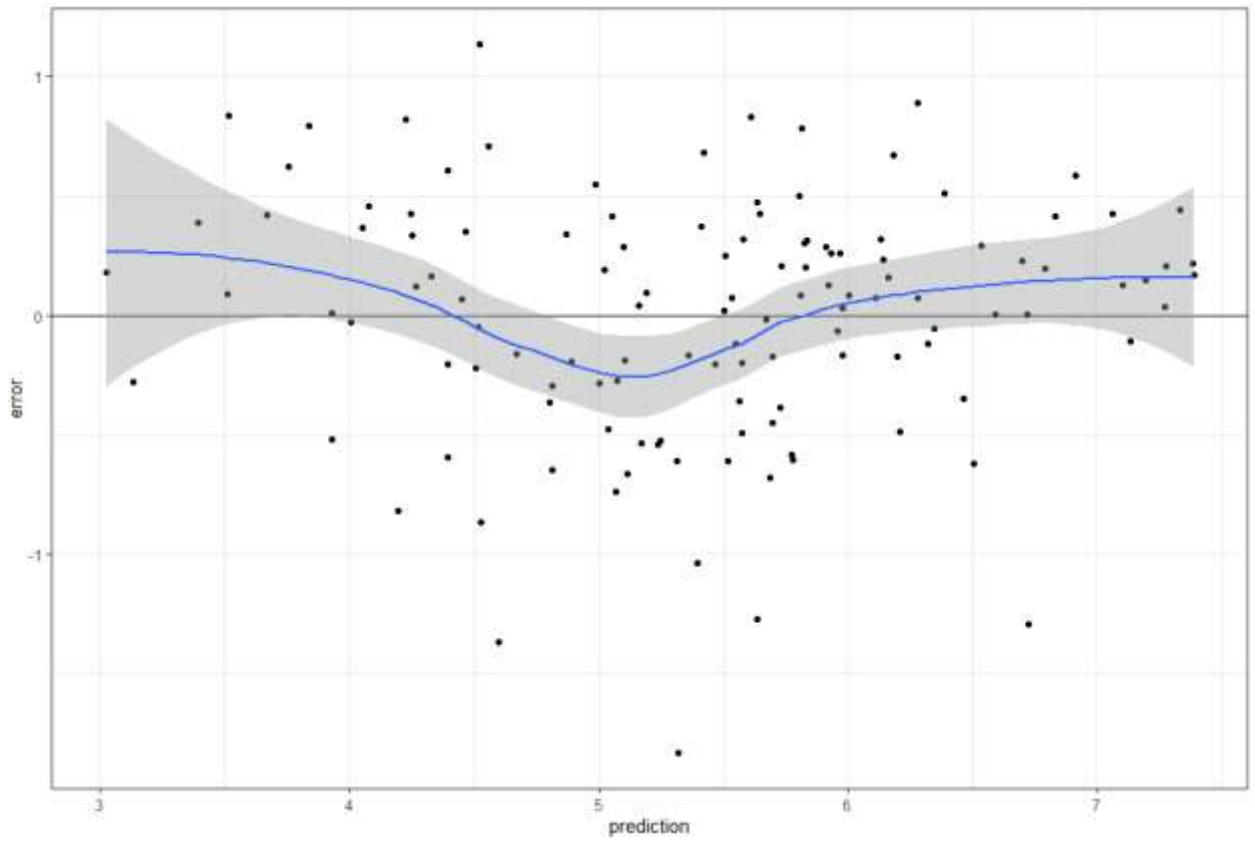
دگرگونی یا ثابت نبودن واریانس

شاخصه بعدی برای چک کردن فرضبات مدل ثابت نبودن واریانس است.

در شکل زیر متوجه می شویم که شکل خاصی ندارند پس ان فرضیه نیز تایید شد .



خطی بودن مانده ها نسبت به پیش بینی



از نمودار بالا، در نمودار باقیمانده ما هیچ الگوی قابل تشخیصی وجود ندارد، می‌توان نتیجه گرفت که مدل ما خطی است.

در اینجا تمامی فرضیات مدل بررسی شد . و مدل رگرسیونی تحلیل و ارزیابی شده ما قابل قبول است. و می توان نتیجه گرفت که در تعیین شادکامی، متغیرهای مهم عبارتند از: سرانه تولید ناخالص داخلی، حمایت اجتماعی، امید به زندگی سالم، انتخاب آزادی برای ساختن زندگی.

نتیجه گیری

در طول پروژه هدف این بود که فاکتورهای مختلف تاثیر گذار در سطح شادی عمومی بررسی شود ، فاکتورهای مختلف را از روش های آماری بررسی کردیم تا استنباط کنیم که کشورها و جوامع با سطح شادی بالا چه تفاوتی از لحاظ اخلاقی فرهنگی و اقتصادی با کشورها و جوامع با سطح شادی پایین داشته اند.

با بررسی های انجام شده از طریق نرخ هم بستگی هر متغیر با متغیر پاسخ به ترتیب ارجحیت تاثیرگذاری به شرح زیر هستند.(جدول در قسمت پیوست)

۱-میزان تولید ناخالص داخلی

۲-میزان امید به زندگی

۳-میزان حمایت دولت

۴-میزان حق انتخاب و آزادی مردم

۵-پایین بودن نرخ فساد

۶-حمایت افراد جامعه از یکدیگر

هم چنین خروجی آزمون فرض انجام شده بر این قضیه دلالت دارد که هیچ یک از کشورهایی با نرخ شادی بالا میزان تولید ناخالص ملی پایینی ندارند درکنار مدل رگرسیونی اعمال شده این ادعا ثابت می شود که میزان ثروتمندی جامعه بیشترین نقش را در میزان شاد مردم آن جامعه ایفا می کند ،بطوری که به ازای هر واحد تغییر در تولید ناخالص ملی ۰.۲۰۸ تغییر مثبت در نرخ شادی اندازه گیری شده را مشاهده کرده ایم.

برای مشاهده تاثیر تغییرات هر واحد از هر متغیرها برروی متغیر پاسخ در قسمت پیوست، معادله تمامی این متغیر ها نسبت به متغیر هدف را مشاهده خواهید کرد.

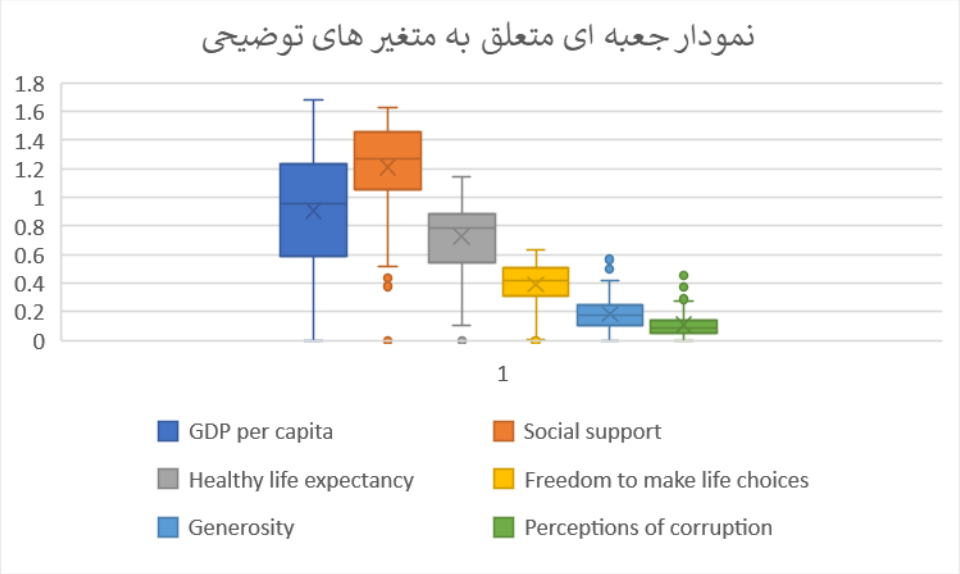
هم چنین بررسی شد که در جوامعی که دولت حمایت بیشتر از میانگین نسبت به مردم دارد نرخ امید به زندگی و شادی در مردم آن کشور به طرز قابل توجهی بالاتر از جوامعی است که دولت حمایت پایین تر از حد میانگین از مردم داشته .

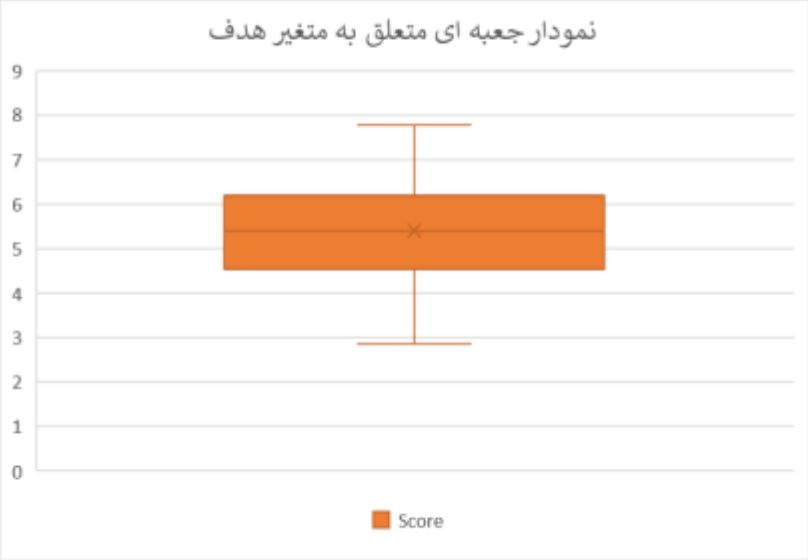
نمودار ها و جداول

پراکندگی داده ها در هر متغیر مستقل

با استفاده از نمودارهای جعبه، همانطور که در ذیل نشان داده شده است، تجزیه و تحلیل شده مشاهده میشود که متغیر های آزادی انتخاب ،سخاوت ،و فساد دارای کمترین پراکندگی،و شاخص تولید ناخالص ملی دارای بیشترین پراکندگی است.

نمودار های جعبه ای

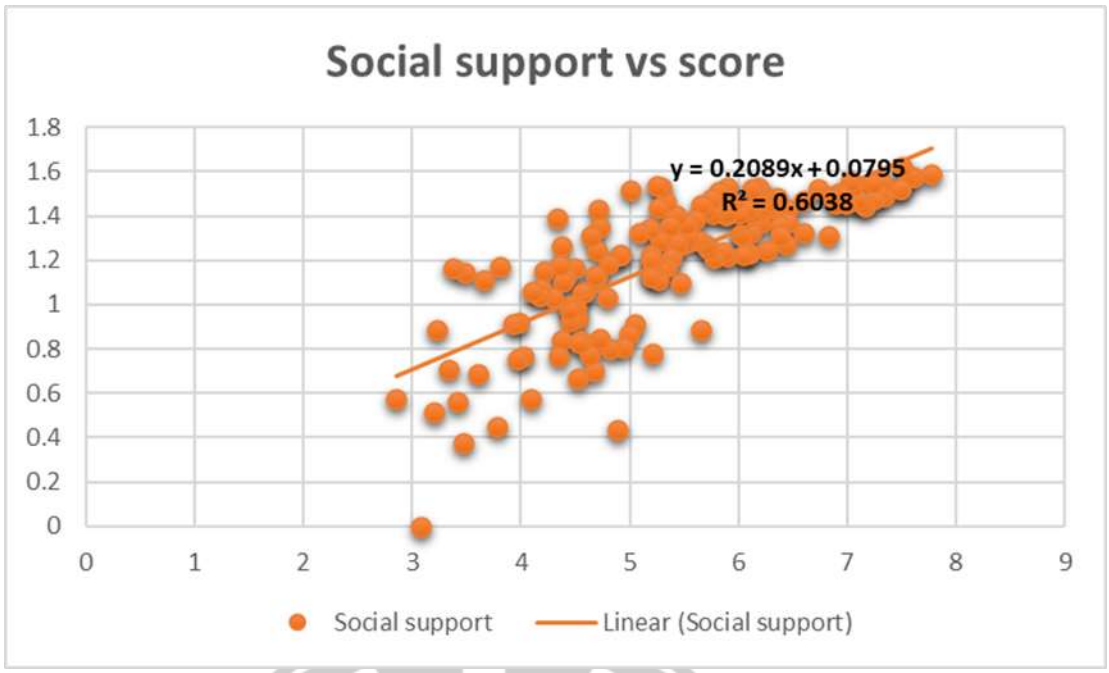


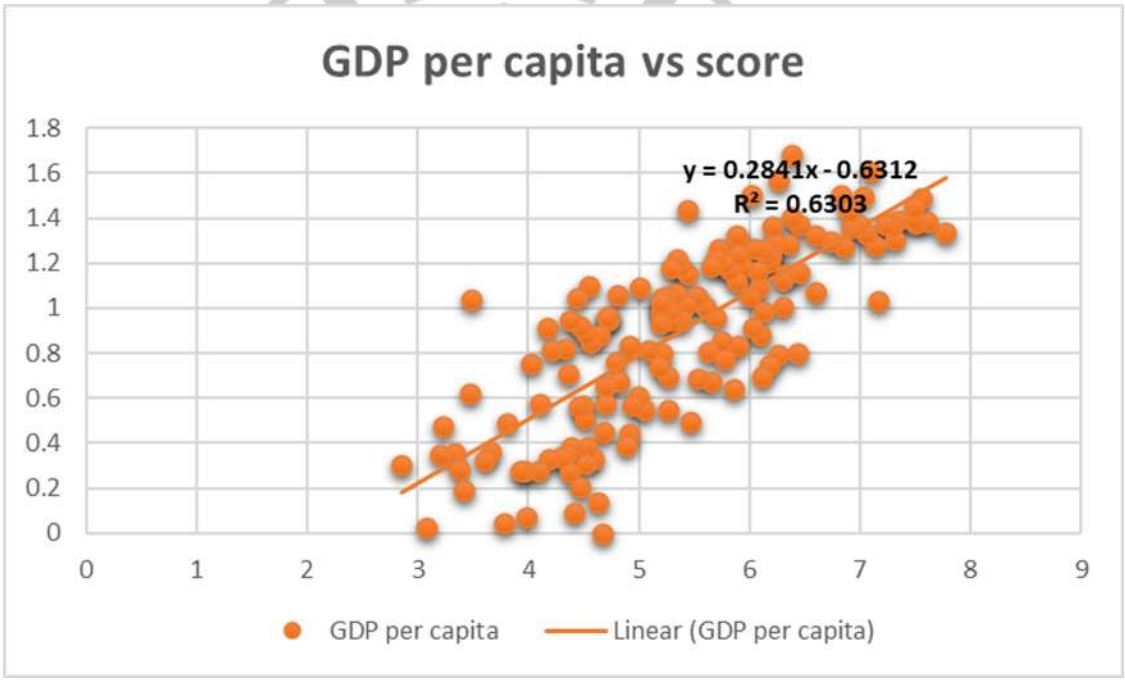


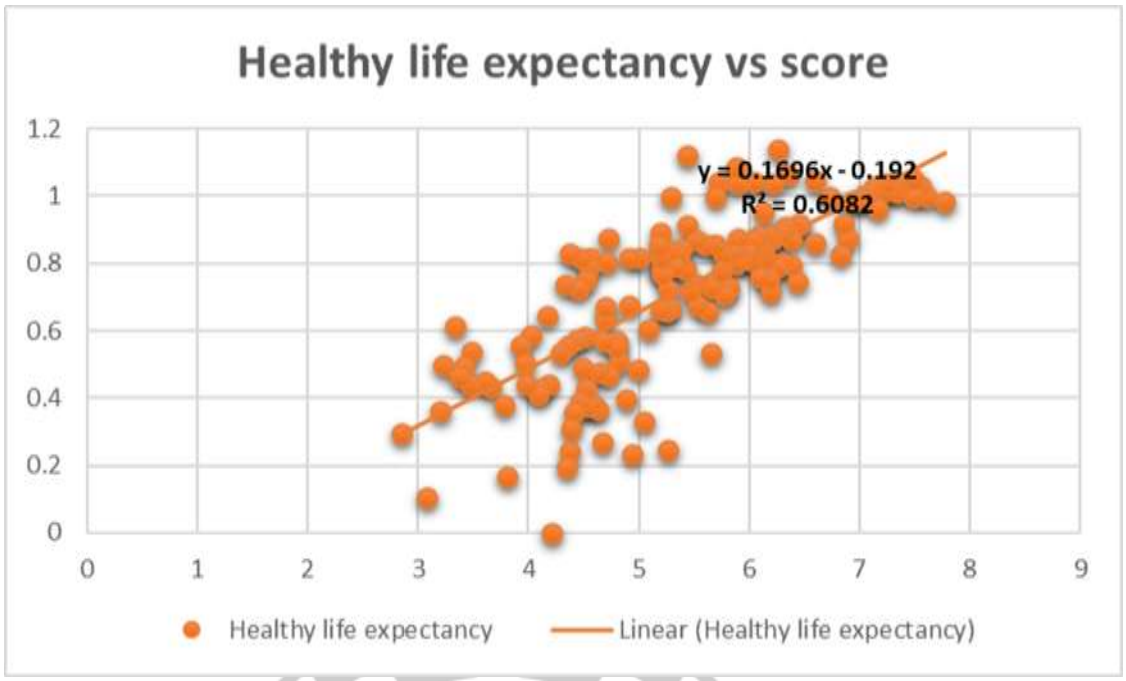
نمودار های پراکنش

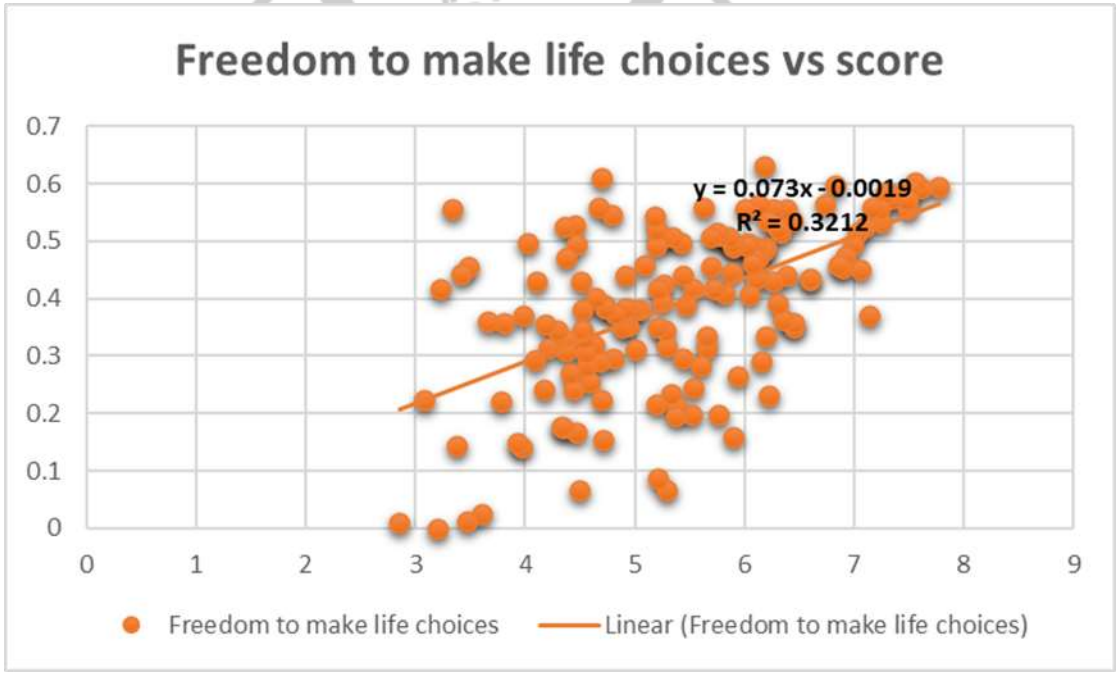
این نمودار روشی بصری برای تجزیه و تحلیل رابطه بین

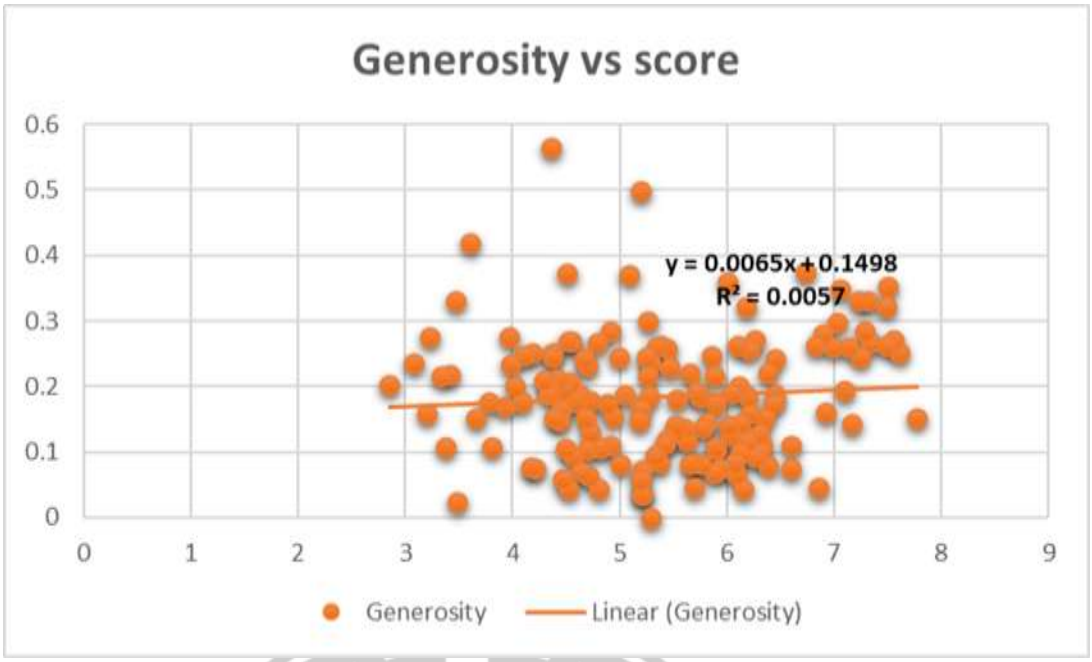
متغیرها از طریق یک نمودار رگرسیون خطی است.

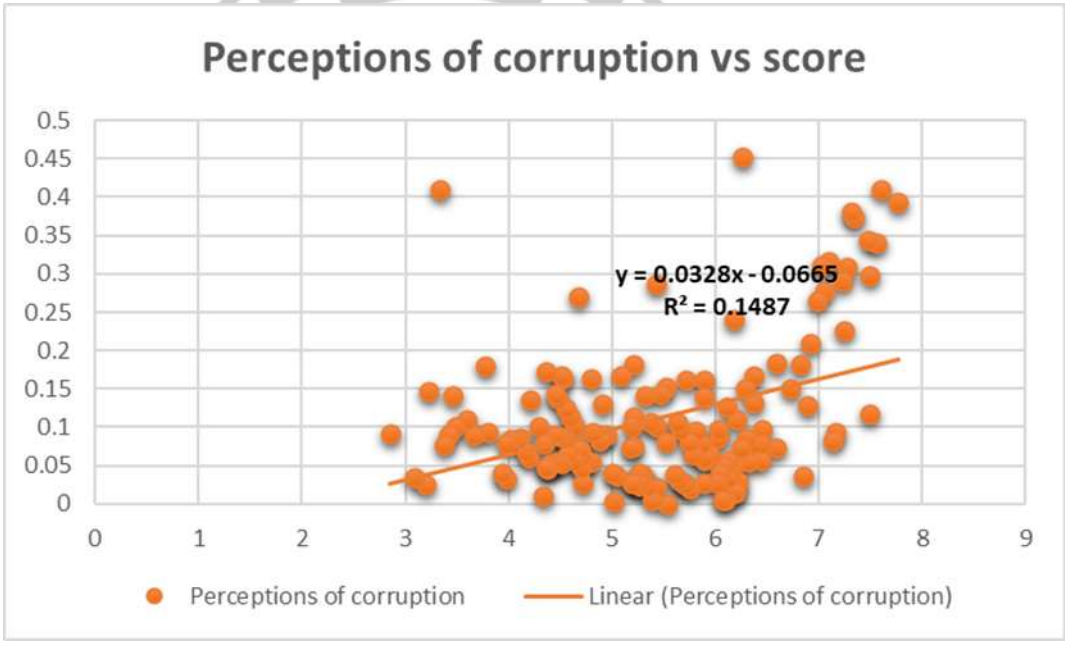

Social support vs score
y = 0.2089x + 0.0795
R² = 0.6038
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9
Social support
Linear (Social support)


GDP per capita vs score
y = 0.2841x - 0.6312
R² = 0.6303
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9
GDP per capita
Linear (GDP per capita)

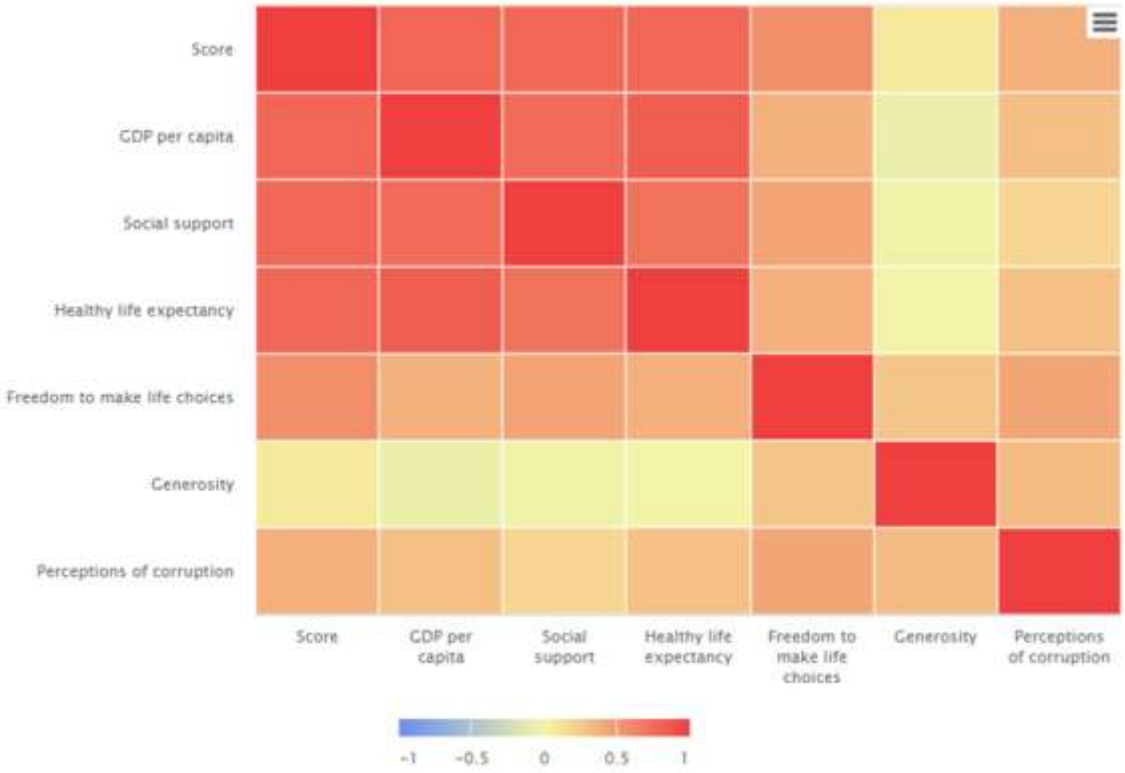

Healthy life expectancy vs score
y = 0.1696x - 0.192
R² = 0.6082
1.2
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8 9
Healthy life expectancy
Linear (Healthy life expectancy)

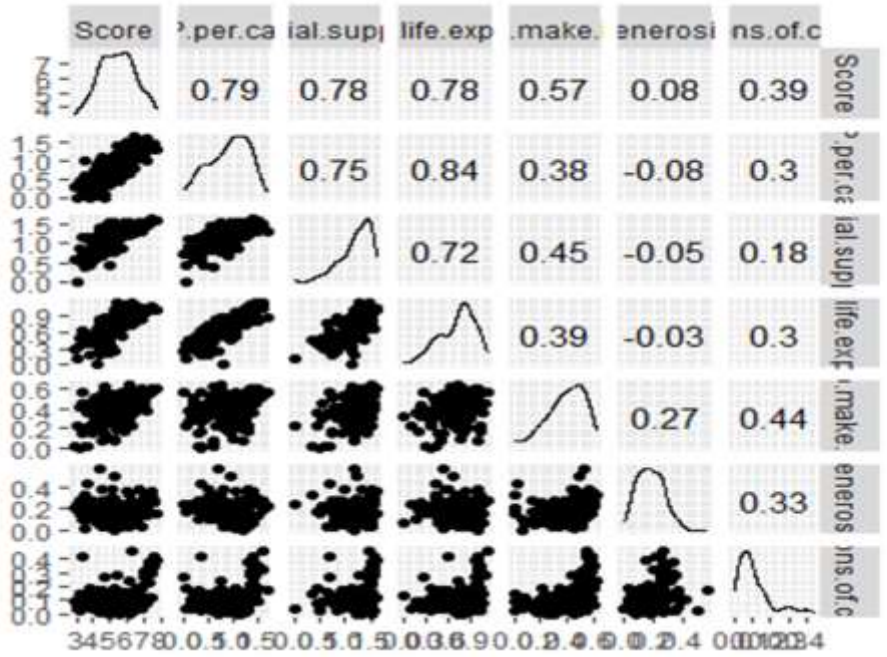

Freedom to make life choices vs score
y = 0.073x - 0.0019
R² = 0.3212
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 1 2 3 4 5 6 7 8 9
Freedom to make life choices
Linear (Freedom to make life choices)


Generosity vs score
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
y = 0.0065x + 0.1498
R² = 0.0057
Generosity
Linear (Generosity)


Perceptions of corruption vs score
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
9
y = 0.0328x - 0.0665
R² = 0.1487
Perceptions of corruption
Linear (Perceptions of corruption)

نمودار های هم بستگی

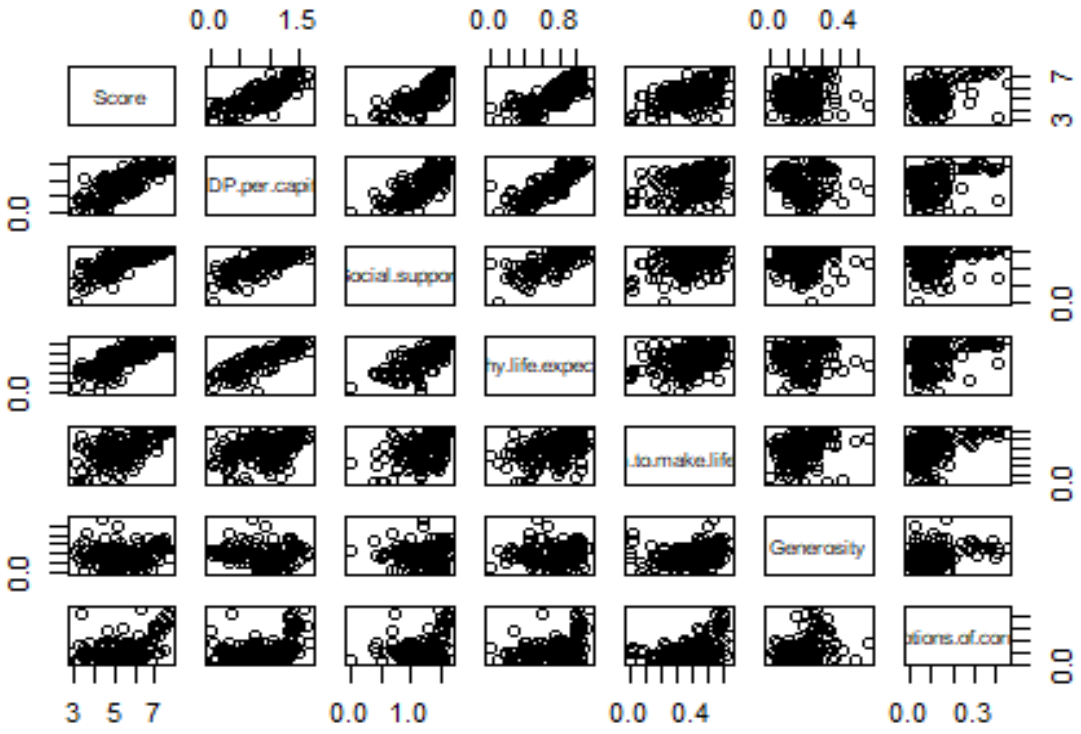




به نظر می رسد همه متغیرها به جز Generosity رابطه خطی با متغیر هدف دارند البته متغیر Perceptions.of.corruption نیز نسبت به سایر متغیرها هم بستگی کمتری به متغیر هدف دارد.

. تولید ناخالص داخلی، حمایت اجتماعی و زندگی سلامت به همبستگی شدیدی با نرخ شادی دارند.

همه متغیرهای مستقل سطوح مختلفی از چولگی را در توزیع خود نشان می دهند



نرخ هم بستگی بین متغیر ها

corruption	Generosity	Freedom	support	GDP	healthy	score	corr
.386	.076	.567	.777	.794	.780	1	Score
.295	-0.030	.390	.719	.835	1	.780	healthy
.299	-0.080	.379	.755	1	.835	.794	GDP
.182	-0.048	.447	1	.755	.719	.777	support
.439	.270	1	.447	.379	.390	.567	Freedom
.327	1	.270	-0.048	-0.080	-0.030	.076	Generosity
1	.327	.439	.182	.299	.295	.386	corruption

تصاویر خروجی بیشینه درست نمایی

```
> fn = function(x , theta){
+   loglik = (-1)*sum((-0.5)*(x - theta[1])^2 /theta[2] -  0.5 * log(theta[2]) - log(2* pi)/2)
+   loglik
+ }
> nlm(fn , x=Generosity      , theta <- c(0,1) , hessian = TRUE)
$minimum
[1] -145.935

$estimate
[1] 0.184845874 0.009015316

$gradient
[1] -0.004838048 -0.642028696

$hessian
            [,1]         [,2]
[1,] 17303.88567    -94.38578
[2,]   -94.38578 918307.76896

$code
[1] 2

$iterations
[1] 21
```

```
> optim(theta<-c(-1,2) , fn , x=Generosity       , hessian = TRUE)
$par
[1] 0.184839270 0.009013593

$value
[1] -145.935

$counts
function gradient
      93       NA

$convergence
[1] 0

$message
NULL

$hessian
             [,1]          [,2]
[1,] 17307.19429 1.338219e+01
[2,]    13.38219 1.035422e+06
```

کد ها

کد مربوط به پیش پردازش شماره ۱

```
library(fastR2)
library(graphics)
rawdata=read.csv(file.choose(),header=T,sep=",",na.string=c(""," ","?"," ?"))
rawdata
str(rawdata)
nrow(rawdata)
ncol(rawdata)
summary(rawdata)
rawdata=read.csv(file.choose(),header=T,sep=",",na.string=c(""," ","?"," ?"))
view(rawdata)
summary(rawdata)
head(rawdata)
str(rawdata)
```

```
happy <-
rawdata[,c("Score","GDP.per.capita","Social.support","Healthy.life.expectancy","F
reedom.to.make.life.choices","Generosity","Perceptions.of.corruption"
)]
str(happy)
glimpse(happy)
boxplot(x)
View(happy)
str(happy)
which(duplicated(happy)==TRUE)
attach(happy)
```

کد های مربوط به رسم شکل ها شماره ۲

```
lines(density(x),col=1)
qqnorm( x ,col=9 )
qqline( x ,col=2 , lwd=3)
boxplot(x,col=2)
shapiro.test( x)
boxplot(x)
```

کد های مربوط به برآورد های نقطه ای شماره ۳

```
library(moments)
mean(x )
var(x )
skewness(x   )
```

```
kurtosis(x   )
```

کد شماره مربوط به برآورد گشتاوری شماره ۴

```
x=Generosity
mu.mom <-function(x){
  x.bar= mean(x)
  mu.hat = x.bar
  return(mu.hat)
}
mu.mom(Generosity)
sigma.mom <- function(x){
  n = length(x)
  s2n=(mean(x^2)-mean(x)^2)
  s2 = s2n * n /(n-1)
  return(s2)
}
sigma.mom(Generosity)
```

کد مربوط به ماکسیمم درست نمایی شماره ۴

```
fn = function(x , theta){
  loglik = (-1)*sum((-0.5)*(x - theta[1])^2 /theta[2] -  0.5 * log(theta[2]) - log(2* pi)/2)
  loglik
}
nlm(fn , x=Generosity      , theta <- c(0,1) , hessian = TRUE)
nlm
```

```
optim(theta<-c(-1,2) , fn , x=Generosity      , hessian = TRUE)
optim
```

کد مربوط به فاصله اطمینان شماره۴

```
Z.conf.int <- function(x, conf.level) {
  alpha = 1 - conf.level
  n = length(x)
  sd<-sd(x)
  zstar <- qnorm(alpha / 2)
  interval <- mean(x) + c(-1,1) * zstar * sd /sqrt(n)
  return(list(conf.int = interval, estimate.mu = mean(x),estimate.sigma2=var(x)))
}
Z.conf.int(Generosity,.95)
```

کد مربوط به آمون فرض شماره ۵

```
m=mean(GDP.per.capita)
t.test(GDP.per.capita,mu!=m, alternative="greater") ### H0:mu!  =.9 vs H1:mu< .9
Z.conf.int <- function(x, conf.level) {
  alpha = 1 - conf.level
  n = length(x)
  sd<-sd(x)
  zstar <- qnorm(alpha / 2)
  interval <- mean(x) + c(-1,1) * zstar * sd /sqrt(n)
  return(list(conf.int = interval, estimate.mu = mean(x),estimate.sigma2=var(x)))
  }
Z.conf.int(GDP.per.capita,.95)
```

کد مربوط به مقایسه امید به زندگی در افراد با حمایت بالا و حمایت پایین شماره ۶

پرسش اول

```
m=mean(Social.support)
group1=mean(Healthy.life.expectancy[which(Social.support >m)])
group2=mean(Healthy.life.expectancy[which(Social.support <m)])
ekhtelaf_Healthy.life.expectancy=group2-group1
ekhtelaf_Healthy.life.expectancy
```

کد مربوط به نرخ شادی و امید به زندگی در بین افراد ثروتمند و فقیر شماره ۶

پرسش دوم

```
m=mean( GDP.per.capita )
group1=mean(Healthy.life.expectancy[which( GDP.per.capita  >m)])
group1
group2=mean(Healthy.life.expectancy[which( GDP.per.capita  <m)])
group2
```

پرسش دوم

```
m=mean( GDP.per.capita )
group1=mean(Score [which( GDP.per.capita  >m)])
group1
group2=mean(Score [which( GDP.per.capita  <m)])
group2
```

کد مربوط به مدل رگرسیونی اول شماره ۷

```
ggscatmat(happy, columns = 1:ncol(happy), corMethod = "pearson")


plot(happy)

attach(happy)

str(happy)

model1<-lm(Score ~ Healthy.life.expectancy+GDP.per.capita +Social.support
+Freedom.to.make.life.choices+Generosity +Perceptions.of.corruption)

summary(model)

#نموونه گیری

index = sample(1:nrow(happy), nrow(happy), replace = FALSE)

#تقسیم داده ها به تستو ترین و گذاشتن درسقف ۹۰درصد

train_index = index[1:floor(0.9*length(index))]

#داده های تست جدا میشه و ترین ازش کم میشه

train_data = happy[train_index,]

test_data = happy[-train_index,]

#---------- Now we can train our model using the lm() function

model = lm(Score~., data=train_data)

#---------- To view the coefficients only

model

summary(model)

#---------- To shuffle the data

index = sample(1:nrow(happy), nrow(happy), replace = FALSE)

#---------- To divide the set of observations we put 80% of the data in train set

#---------- and the rest in the test set

train_index = index[1:floor(0.8*length(index))]

#---------- We put each observation with an index from the train_index set in the training data

#---------- and then we put the remainder in the test data
```

```
train_data = happy[train_index,]
test_data = happy[-train_index,]
#---------- Now we can train our model using the lm() function
modell = lm(Score~., data=train_data)
#---------- To view the coefficients only
modell
#---------- To dive in deeper and view some interesting stuff
summary(modell)
modelll = lm(Score~GDP.per.capita , data=train_data)
modelll
summary(modelll)
modellll = lm(Score~GDP.per.capita , data=test_data)
modellll
summary(modelll)
#---------- With the model in our hands, we can predict our test data
ytest_pred = predict(modelll, test_data[,-1])
ytest_pred
summary(modelll)
#---------- Now we can evaluate our prediction with metrics such as mse (Mean Squared Error)
mse = mean((test_data[,1] - ytest_pred)^2)
mse
```

کد مربوط به مدل رگرسیونی دوم شماره ۸

```
library(GGally) #check correlation
library(dplyr) #piping daya
library(rsample) #sampling data
library(tidyverse) #wrangling data
```

```
library(lmtest) #check assumption

library(car) #check vif

library(MLmetrics) #calculate error

rawdata=read.csv(file.choose(),header=T,sep=",",na.string=c("","  ","?"," ?"))

view(rawdata)

summary(rawdata)

head(rawdata)

str(rawdata)

happy <-
rawdata[,c("Score","GDP.per.capita","Social.support","Healthy.life.expectancy","Freedom.to.make.life.c
hoices","Generosity","Perceptions.of.corruption"

)]

str(happy)

glimpse(happy)

RNGkind(sample.kind = "Rounding")

set.seed(1616)

init <- initial_split(happy,

                      prop = 0.8,

                      strata = Score)

happy_train <- training(init)

happy_test <- testing(init)

rawmodelhappy <- lm(Score~.,happy_train)

model_rawmodelhappy <- step(rawmodelhappy, direction = "backward")

summary(model_rawmodelhappy)

model_linear_happy <- lm(Score~ . ,happy_train)

model_happy <- step(model_linear_happy, direction = "backward")

summary(model_happy)

pred_test <- predict(model_happy, newdata = happy_test)

head(pred_test)

RMSE(pred_test, happy_test$Score)
```

```
vif(model_happy)
qqPlot(model_happy$residuals)
plot(density(model_happy$residuals))
plot(model_happy$fitted.values, #prediksi
    model_happy$residuals) #eror
data.frame(prediction=model_happy$fitted.values,
         error=model_happy$residuals) %>%
  ggplot(aes(prediction,error)) +
  geom_hline(yintercept=0) +
  geom_point() +
  geom_smooth() +
  theme_bw()
```


پایان