

گزارش درس یادگیری ماشین آماری

**«پروژه اول»**

گردآورنده: سعید دادخواه (۹۲۳۱۰۶۶)

استاد: دکتر نیک‌آبادی

دی ۱۳۹۶

# مقدمه

پروژه با زبان R انجام شده است. برای تولید نمودارها از کتابخانه ggplot2، در مواردی که نیاز بود چند نمودار کنار هم رسم شود از کتابخانه gridExtra استفاده شده است. برای تولید نمودار همبستگی از کتابخانه reshape2 نیز استفاده شده است. در ابتدا از دستور rm(ls()) برای پاکسازی متغیرهای قبلی محیط استفاده شده است.

# پیشنیازهای اجرای کدها

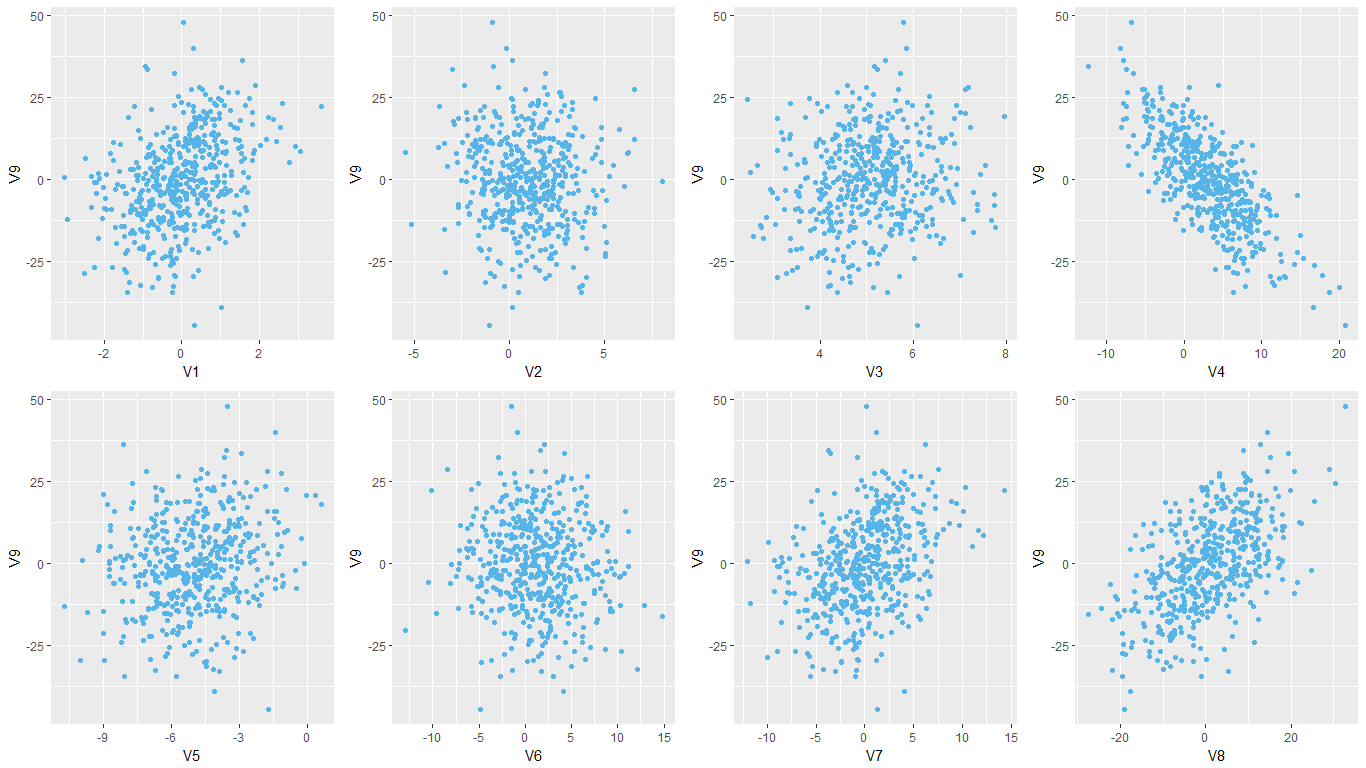
* زبان R
* کتابخانه‌های مورد نیاز
  + ggplot2
  + gridExtra
  + reshape2

دسته‌کدها به شکلی نوشته شده‌اند که هربار وجود کتابخانه‌های مورد نیاز را تست می‌کنند و در صورتی که کتابخانه مورد نظر وجود نداشت آن را دانلود و نصب می‌کنند.

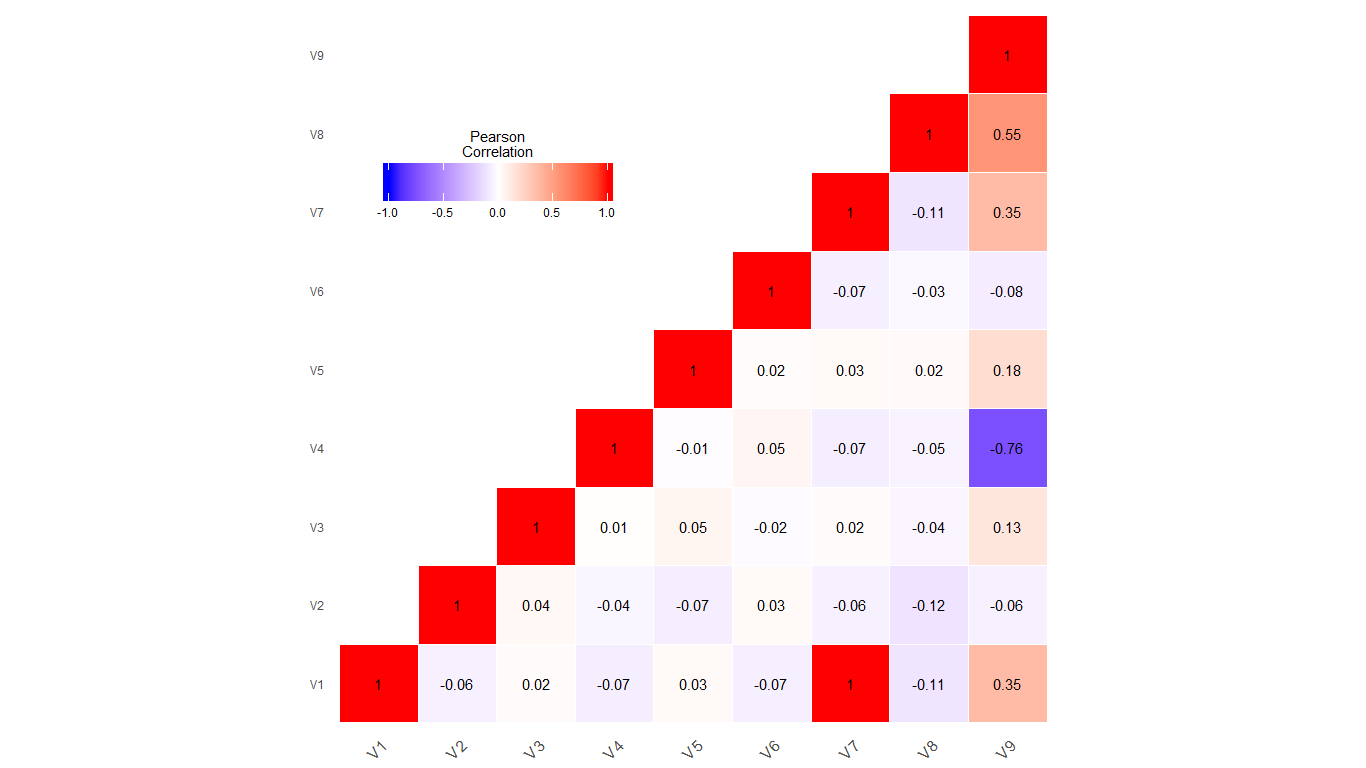
# مجموعه داده Dataset1.csv

## الف

نمودارهای زیر نحوه‌ی توزیع متغیرهای مستقل را در برابر متغیر وابسته نمایش می‌دهند.

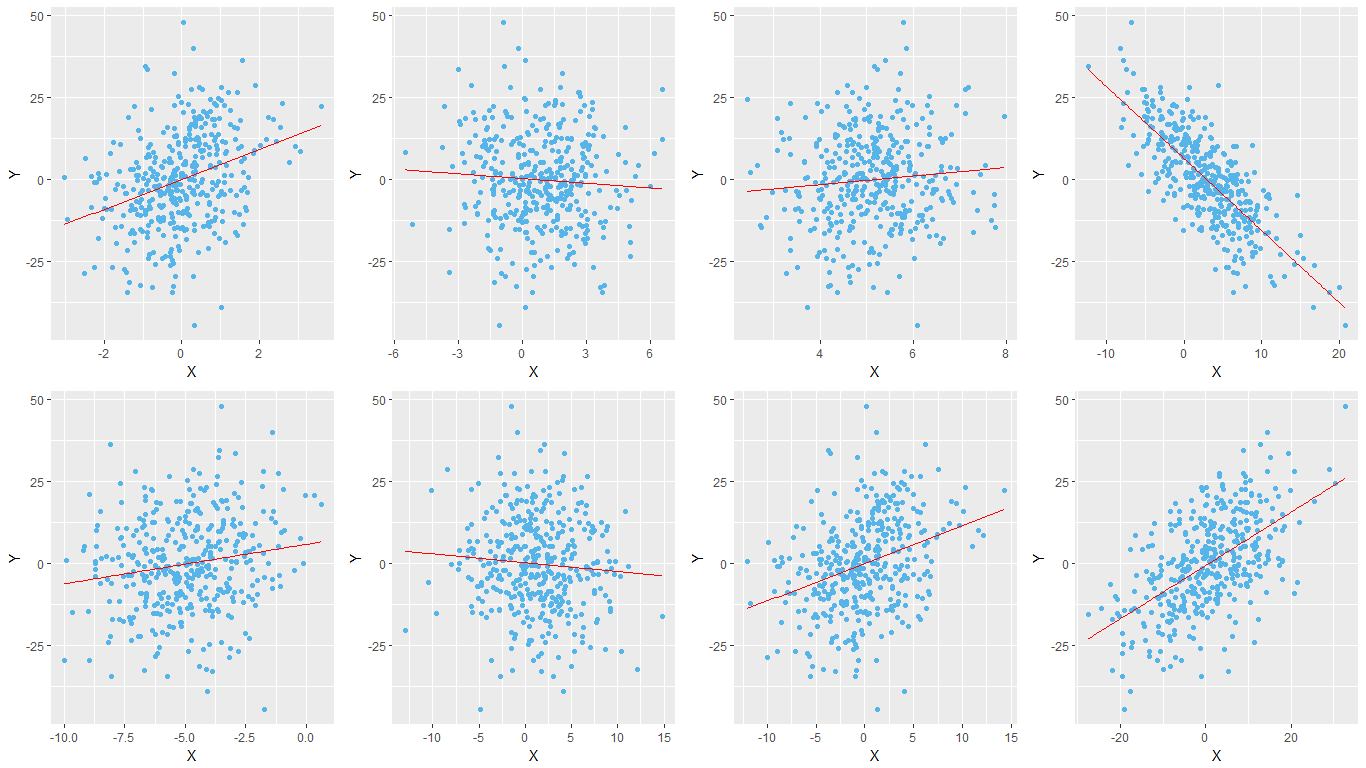


در نمودارهای توزیع نقاط نمایش دهنده این نکته هستند که با تغییر متغیر مستقل Vi متغیر وابسته در کل چه رفتاری نمایش می‌دهند. در نمودارهایی که نقاط دایره‌ای را تشکیل داده‌اند یعنی تقریبا رابطه‌ای بین این دو متغیر وجود ندارد. هرچه شکل به سمت بیضی باریک‌تر یا در بهترین حالت به خط میل کنند یعنی رابطه میان دو متغیر قوی‌تر است و تغییر متغیر مستقل تاثیر بیشتری در متغیر وابسته می‌گذارد. برای تعیین مقدار تاثیر این متغیرها در متغیر وابسته از مقدار همبستگی استفاده می‌کنیم. با نزدیک شدن مقدار همبستگی به یک یا منفی یک این تاثیر بیشتر می‌شود و با همبستگی نزدیک به صفر این تاثیر کاهش می‌یابد. نمودار زیر همبستگی متغیرها را نمایش می‌دهد. با توجه به نمودارهای فوق انتظار می‌رود مقادیر همبستگی متغیر چهارم و هشتم با متغیر وابسته زیاد باشد. از طرفی چون با افزایش متغیر چهارم مقدار متغیر وابسته کاهش می‌یابد انتظار داریم مقدار همبستگی آن نسبت به بقیه متغیرهای مستقل دیگر به منفی یک و با استدلالی مشابه مقدار همبستگی متغیر هشتم نسبت به بقیه متغیرهای مستقل دیگر به یک نزدیک باشد. نمودار زیر که یک نقشه حرارتی است نیز گزاره‌های بالا را تایید می‌کند.



## ب

داده‌ها و خطوط رگرسیون خطی ساده که محاسبه شده‌اند به شکل زیر هستند.



جدول زیر مقادیر خواسته شده برای مدل‌هایی که با متغیرهای مستقل متفاوت و متغیر وابسته آموزش داده شده‌اند ارائه می‌کند.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Independent Variable | Intercept | | Slope | |  |
|  |  |  |  |
| 1 | -0.0767 | 0.6731 | 4.5803 | 0.6510 | 181.0520 |
| 2 | 0.1784 | 0.7720 | -0.4812 | 0.3614 | 202.5584 |
| 3 | -6.9381 | 3.7519 | 1.3122 | 0.7196 | 201.7787 |
| 4 | 6.2154 | 0.5344 | -2.1902 | 0.0925 | 84.7082 |
| 5 | 5.7199 | 1.8971 | 1.2053 | 0.3575 | 197.8352 |
| 6 | 0.1191 | 0.7434 | -0.2693 | 0.1723 | 202.2213 |
| 7 | -0.0767 | 0.6731 | 1.1451 | 0.1628 | 181.0520 |
| 8 | -0.7270 | 0.5910 | 0.8112 | 0.0597 | 139.1683 |

یکی از نکاتی که می‌توان از جدول بالا به آن پی برد این است که هرقدر همبستگی میان متغیر مستقل و وابسته بیشتر به یک یا منفی نزدیک شود کاهش می‌یابد.

جدول زیر معیارهای RSS و را برای مدل‌های مختلف و با توجه به داده‌های آموزشی و آزمایشی نمایش می‌دهد.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | | |
|  |  |  | |  |  |  |
| 1 | 72420.7974 | 0.1123 | 3220.6645 | | 14463.9536 | 0.1921 | 787.2122 |
| 2 | 81023.3538 | 0.0069 | 3265.5621 | | 17831.5457 | 0.0040 | 808.1431 |
| 3 | 80711.4603 | 0.0107 | 3264.0193 | | 16988.5583 | 0.0511 | 803.3002 |
| 4 | 33883.2848 | 0.5847 | 2916.8358 | | 8751.9791 | 0.5111 | 736.9742 |
| 5 | 79134.0629 | 0.0301 | 3256.1245 | | 16991.6051 | 0.0509 | 803.3182 |
| 6 | 80888.5196 | 0.0086 | 3264.8959 | | 17770.2894 | 0.0074 | 807.7990 |
| 7 | 72420.8162 | 0.1123 | 3220.6646 | | 14463.9533 | 0.1921 | 787.2122 |
| 8 | 55667.3145 | 0.3177 | 3115.4244 | | 14081.0158 | 0.2135 | 784.5290 |

معیار هرمقدار که بیشتر باشد یعنی مدل بهتر می‌تواند تغییرات موجود در متغیر وابسته را با توجه به تغییرات در متغیر مستقل پیشبینی کند. مشاهده می‌شود که این مقدار برای متغیرهای چهارم و هشتم بسیار بالاتر از بقیه است.

## ج

در مسئله پیشبینی متغیر هدف هرقدر بتوان تغییرات آن را با توجه به تغییرات متغیر مستقل بیان کرد بهتر می‌توان آن را پیشبینی کرد. در حقیقت می‌توان این دو موضوع را هم‌ارز یکدیگر در نظر گرفت. با این توصیفات متغیر مستقل چهارم بهترین گزینه برای انتخاب به عنوان متغیر اول برای پیشبینی متغیر وابسته است.

حال متغیرهای دیگر به جز متغیر چهارم را در کنار این متغیر قرار داده و مدل خطی آموزش می‌دهیم. مقادیر خواسته شده به ترتیب زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 1 | 27465.4681 | 0.6625 | 2834.8389 | 6314.3603 | 0.6473 | 706.3289 |
| 2 | 33014.2795 | 0.5943 | 2908.4432 | 8786.8343 | 0.5092 | 739.3717 |
| 3 | 32880.9461 | 0.5960 | 2906.8245 | 7996.8246 | 0.5960 | 729.9507 |
| 5 | 31598.1886 | 0.6117 | 2890.9070 | 8201.4186 | 0.5419 | 732.4769 |
| 6 | 33807.3222 | 0.5846 | 2917.9381 | 8680.4800 | 0.5151 | 738.1539 |
| 7 | 27465.4572 | 0.6625 | 2834.8388 | 6314.3629 | 0.6473 | 706.3289 |
| 8 | 13576.6607 | 0.8332 | 2553.0080 | 3365.5051 | 0.8120 | 643.4040 |

مقدار معیار AIC برای حالتی که فقط از متغیر چهارم استفاده می‌شد برای داده‌های آموزشی ۲۹۱۶ و برای داده‌های آزمایشی ۸۷۲ بود. با اضافه کردن هرکدام متغیرهای فوق (به جز متغیر ششم) به مدل این مقدار کاهش می‌یابد یعنی مدل بهبود پیدا می‌کند. این موضوع در مورد معیار RSS (Residual Sum of Squares) نیز که در مدل پایه دارای مقادیر ۳۳۸۸۳ و ۸۷۵۱ به ترتیب برای داده‌های آموزشی و آزمایشی بود، صادق است. اما در مورد معیار به گونه‌ای دیگر تحلیل انجام می‌دهیم. در همه موارد این معیار بیشتر از مقدار قبلی یعنی ۵۸.۴۷ درصد و ۵۱.۱۱ درصد به ترتیب برای داده‌های آموزشی و آزمایشی می‌شود که یعنی متغیرهای مستقل بهتر می‌توانند تغییرات متغیر وابسته را بیان و در نتیجه آن را پیشبینی کنند. به این ترتیب بهترین متغیر متغیر هشتم خواهد بود.

با اضافه کردن متغیر هشتم بار دیگر مدل‌ها را بر اساس متغیر چهارم، هشتم و یکی از متغیرهای مستقل دیگر آموزش می‌دهیم و مقادیر خواسته شده را محاسبه می‌کنیم.

## د

با اضافه کردن هرکدام از متغیرها به مدل رگرسیون خطی مقادیر زیر را محاسبه می‌کنیم.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 1 | 3221.0619 | 0.9604 | 1979.5515 | 923.9905 | 0.9484 | 516.1409 |
| 2 | 13492.1484 | 0.8342 | 2552.5103 | 3345.7026 | 0.8131 | 644.8139 |
| 3 | 12084.8691 | 0.8515 | 2508.4488 | 2465.8933 | 0.8623 | 614.3016 |
| 5 | 11780.7226 | 0.8552 | 2498.2529 | 2727.0343 | 0.8477 | 624.3677 |
| 6 | 13547.5045 | 0.8335 | 2554.1481 | 3340.0931 | 0.8134 | 644.6461 |
| 7 | 3221.0613 | 0.9604 | 1979.5515 | 923.9860 | 0.9484 | 516.1404 |

با استدلالی مشابه استدلال مورد ج در این مرحله یکی از متغیرهای اول یا هفتم می‌توانند اضافه شوند چون مقادیر AIC برای آن‌ها یکسان است. متغیر اول را بر حسب ترتیب اضافه می‌کنیم و مرحله بعدی را اجرا می‌کنیم.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 2 | 3220.5836 | 0.9604 | 1981.4922 | 927.4676 | 0.9482 | 518.5165 |
| 3 | 1550.7808 | 0.9809 | 1689.1705 | 341.6127 | 0.9809 | 418.6385 |
| 5 | 1566.8952 | 0.9807 | 1693.3056 | 481.3489 | 0.9731 | 452.9299 |
| 6 | 3220.9866 | 0.9604 | 1981.5422 | 924.2005 | 0.9484 | 518.1636 |
| 7 | 3221.0403 | 0.9604 | 1981.5489 | 923.6638 | 0.9484 | 518.1055 |

متغیر بعدی که متغیر سوم است اضافه می‌شود. مدل‌های بعدی با متغیرهای چهارم، هشتم، اول، سوم و یکی از متغیرهای زیر آموزش داده می‌شوند و مقادیر زیر محاسبه می‌شوند.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 2 | 1545.3915 | 0.9810 | 1689.7780 | 337.5449 | 0.9811 | 419.4405 |
| 5 | 15.7468 | 0.9998 | -144.7791 | 4.8429 | 0.9997 | -4.9786 |
| 6 | 1548.1962 | 0.9810 | 1690.5033 | 344.2620 | 0.9808 | 421.4110 |
| 7 | 1548.7934 | 0.9810 | 1690.6576 | 340.4434 | 0.9810 | 420.2956 |

متغیر بعدی که متغیر پنجم است اضافه می‌شود. مدل‌های بعدی با متغیرهای چهارم، هشتم، اول، سوم، پنجم و یکی از متغیرهای زیر آموزش داده می‌شوند و مقادیر زیر محاسبه می‌شوند.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 2 | 15.6790 | 0.9998 | -144.5065 | 5.0428 | 0.9997 | 1.0659 |
| 6 | 15.7165 | 0.9998 | -143.5516 | 4.8879 | 0.9997 | -2.0536 |
| 7 | 15.6223 | 0.9998 | -145.9555 | 4.8549 | 0.9997 | -2.7308 |

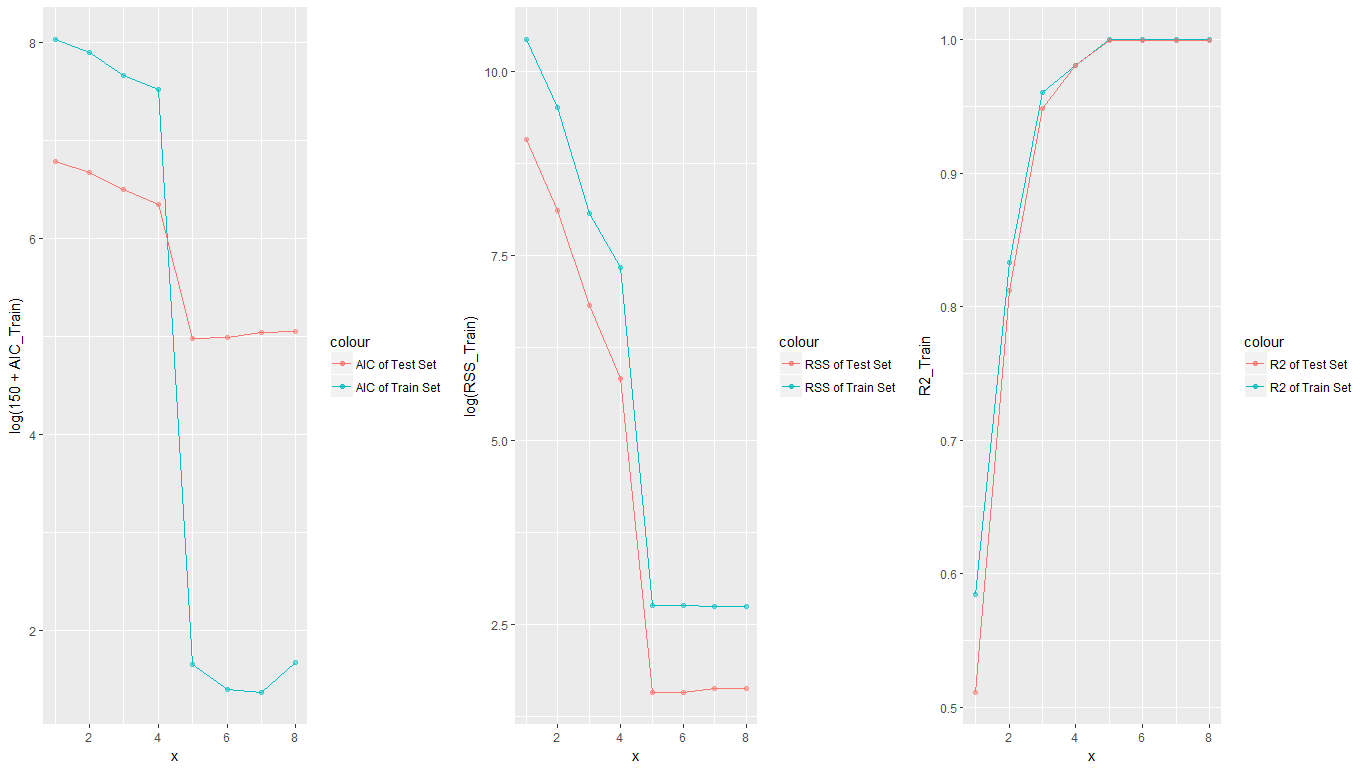
متغیر بعدی که متغیر هفتم است اضافه می‌شود. نکته‌ی قابل توجه در مورد این مرحله این است که مقادیر تغییر چشم‌گیری نداشته‌اند. مدل‌های بعدی با متغیرهای چهارم، هشتم، اول، سوم، پنجم، هفتم و یکی از متغیرهای زیر آموزش داده می‌شوند و مقادیر زیر محاسبه می‌شوند.

| Independent Variable | Train | | | Test | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| 2 | 15.5396 | 0.9998 | -146.0794 | 5.0849 | 0.9997 | 3.8984 |
| 6 | 15.6005 | 0.9998 | -144.5142 | 4.8920 | 0.9997 | 0.0307 |

متغیر بعدی که متغیر دوم است اضافه می‌شود. نکته‌ی قابل توجه در مورد این مرحله این است که مقادیر تغییر چشم‌گیری نداشته‌اند. مدل بعدی با تمامی متغیرها آموزش داده می‌شود و مقادیر محاسبه می‌شوند.

| Independent Variable | Train | | | Test | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| ALL | 15.5167 | 0.9998 | -144.6694 | 5.1241 | 0.9997 | 6.6664 |

در کل می‌توان نتیجه گرفت اگر معیار AIC را در نظر بگیریم به ترتیب متغیرهای چهارم، هشتم، اول، سوم، پنجم، هفتم، دوم و ششم به مدل اضافه می‌شوند. مقادیر ، RSS و AIC به ترتیب از چپ به راست برای داده‌های آموزشی و آزمایشی در طول اضافه کردن متغیرهای مستقل به مدل در نمودارهای زیر رسم شده‌اند.



مقادیر RSS و AIC تغییر بسیار زیادی دارند و نمایش آن‌ها در نمودار نحوه تغییر آن‌ها را نمی‌تواند نشان دهد به همین دلیل از توابعی اکیدا صعودی برای تغییر آن‌ها استفاده شده است. با استفاده از این توابع نحوه تغییر آن‌ها مشخص می‌شود و مشکلی در ترتیب آن‌ها به وجود نمی‌آید. در تصویر بالا مشخص است که مقادیر افزایش می‌یابند و از ویژگی پنجم به بعد تغییر قابل توجه ندارند. اما معیار RSS از همان ویژگی شروع به افزایش جزئی می‌کند که معنی آن پیچیده شدن بیش از حد مدل است. معیار AIC این پیچیده شدن را با تغییرات بیشتری نمایش می‌دهد. البته لازم به ذکر است که این نکته در معیارهای RSS و AIC با تغییر کمی خود را نشان می‌دهد ولی با توجه به اعمال تابع‌های اکیدا صعودی و این تغییرات به خوبی در نمودار مشهود هستند.

## ه

مراحل را برای معیار BIC انجام می‌دهیم ولی فقط در هر مرحله BICهای مربوط به داده‌های آموزشی و مقادیر مربوط به بهترین مدل بر اساس BIC را گزارش می‌کنیم، با اجرای کد باقی موارد قابل مشاهده خواهند بود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 1 | 3232.6389 |
| 2 | 3277.5365 |
| 3 | 3275.9937 |
| 4 | 2928.8102 |
| 5 | 3268.0989 |
| 6 | 3276.8703 |
| 7 | 3232.6390 |
| 8 | 3127.3988 |

متغیر مستقل چهارم انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 4 | 33883.2848 | 0.5837 | 2928.8102 | 8751.9791 | 0.5111 | 744.7897 |

در مرحله بعد متغیرهای مستقل با متغیر مستقل چهارم به ترتیب مدل اضافه می‌شوند و BIC مدل‌ها برای داده‌های آموزشی محاسبه می‌شود تا بهترین مدل انتخاب شود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 1 | 2850.8048 |
| 2 | 2924.4090 |
| 3 | 2922.7903 |
| 5 | 2906.8729 |
| 6 | 2933.9039 |
| 7 | 2850.8046 |
| 8 | 2568.9739 |

متغیر مستقل هشتم انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 8 | 13576.6607 | 0.8332 | 2568.9739 | 3365.5051 | 0.8120 | 653.8247 |

در مرحله بعد متغیرهای مستقل با متغیر مستقل چهارم و هشتم به ترتیب مدل اضافه می‌شوند و BIC مدل‌ها برای داده‌های آموزشی محاسبه می‌شود تا بهترین مدل انتخاب شود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 1 | 1999.5089 |
| 2 | 2572.4676 |
| 3 | 2528.4061 |
| 5 | 2518.2102 |
| 6 | 2574.1054 |
| 7 | 1999.5088 |

متغیر مستقل اول انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 1 | 3221.0619 | 0.9604 | 1999.5089 | 923.9905 | 0.9484 | 529.1667 |

در مرحله بعد متغیرهای مستقل با متغیر مستقل چهارم، هشتم و اول به ترتیب مدل اضافه می‌شوند و BIC مدل‌ها برای داده‌های آموزشی محاسبه می‌شود تا بهترین مدل انتخاب شود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 2 | 2005.4409 |
| 3 | 1713.1193 |
| 5 | 1717.2543 |
| 6 | 2005.4910 |
| 7 | 2005.4977 |

متغیر مستقل سوم انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 3 | 1550.7808 | 0.9809 | 1713.1193 | 341.6127 | 0.9809 | 434.2695 |

در مرحله بعد متغیرهای مستقل با متغیر مستقل چهارم، هشتم، اول و سوم به ترتیب مدل اضافه می‌شوند و BIC مدل‌ها برای داده‌های آموزشی محاسبه می‌شود تا بهترین مدل انتخاب شود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 2 | 1717.7183 |
| 5 | -116.8389 |
| 6 | 1718.4436 |
| 7 | 1718.5978 |

متغیر مستقل پنجم انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 5 | 15.7468 | 0.9998 | -116.8389 | 4.8429 | 0.9997 | 13.2576 |

در مرحله بعد متغیرهای مستقل با متغیر مستقل چهارم، هشتم، اول، سوم و پنجم به ترتیب مدل اضافه می‌شوند و BIC مدل‌ها برای داده‌های آموزشی محاسبه می‌شود تا بهترین مدل انتخاب شود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 2 | -112.5748 |
| 6 | -111.6199 |
| 7 | -114.0238 |

متغیر مستقل هفتم انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 7 | 15.6223 | 0.9998 | -114.0238 | 4.8549 | 0.9997 | 18.1106 |

در مرحله بعد متغیرهای مستقل با متغیر مستقل چهارم، هشتم، اول، سوم، پنجم و هفتم به ترتیب مدل اضافه می‌شوند و BIC مدل‌ها برای داده‌های آموزشی محاسبه می‌شود تا بهترین مدل انتخاب شود.

|  |  |
| --- | --- |
| Independent Variable | BIC |
| 2 | -110.1562 |
| 6 | -108.5910 |

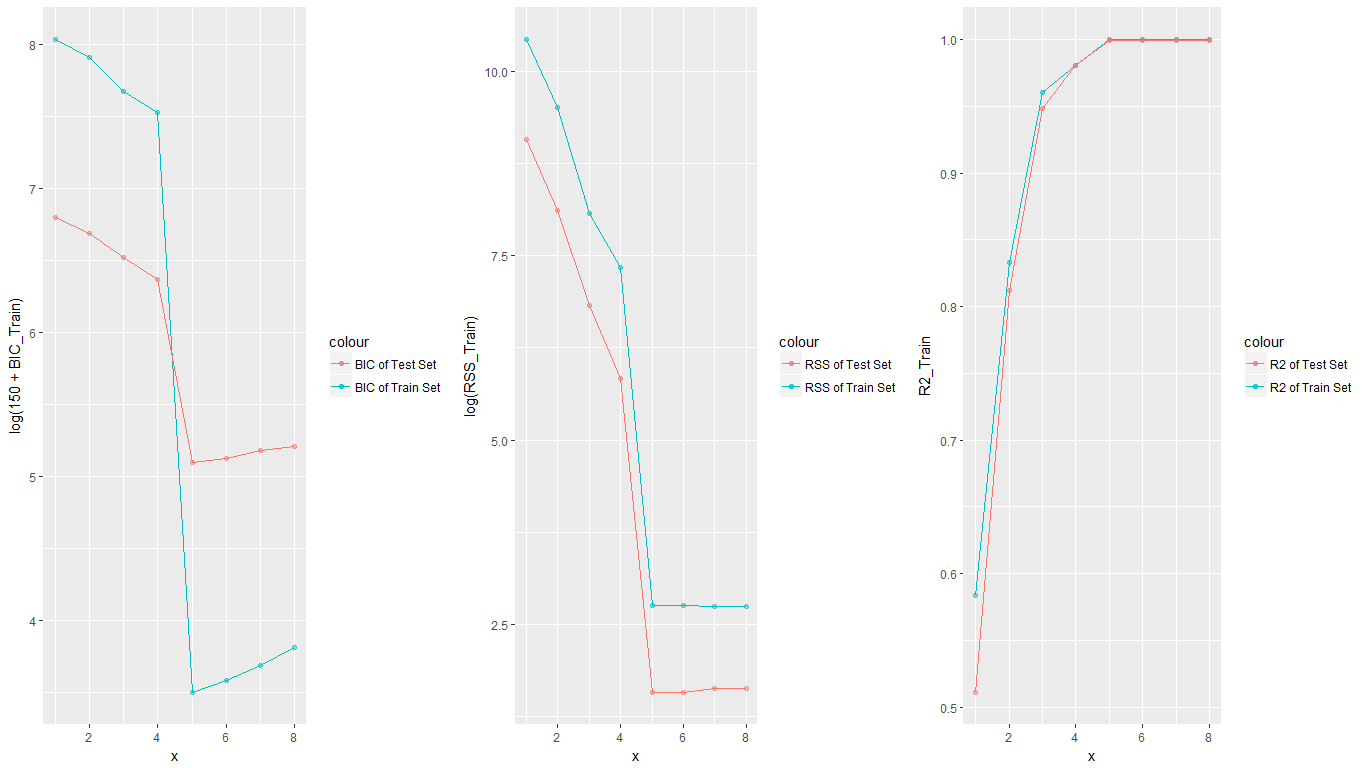
متغیر مستقل دوم انتخاب می‌شود و مقادیر مربوط به آن به شکل زیر گزارش می‌شود.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 2 | 15.5396 | 0.9998 | -110.1562 | 5.0849 | 0.9997 | 27.3450 |

در نهایت متغیر مستقل ششم اضافه می‌شود و مقادیر مربوط به آن به ترتیب زیر گزارش می‌شود.

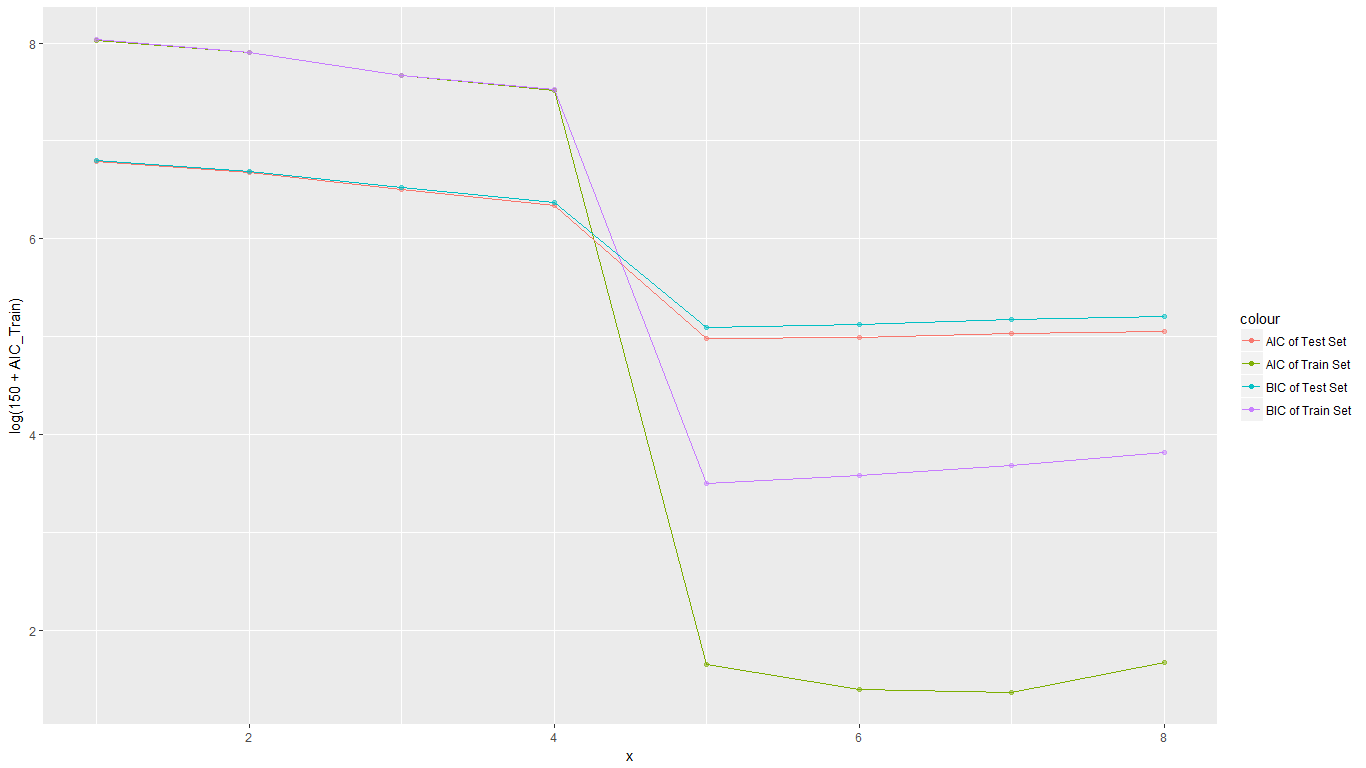
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Independent Variable | Train | | | Test | | |
|  |  |  |  |  |  |
| 6 | 15.5167 | 0.9998 | -104.7548 | 5.1241 | 0.9997 | 32.7181 |

در کل می‌توان نتیجه گرفت اگر معیار BIC را در نظر بگیریم به ترتیب متغیرهای چهارم، هشتم، اول، سوم، پنجم، هفتم، دوم و ششم به مدل اضافه می‌شوند. مقادیر ، RSS و BIC به ترتیب از چپ به راست برای داده‌های آموزشی و آزمایشی در طول اضافه کردن متغیرهای مستقل به مدل در نمودارهای زیر رسم شده‌اند.



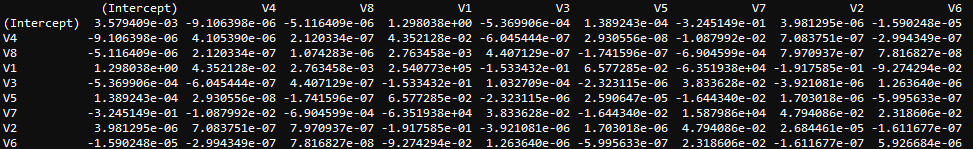
مقادیر RSS و BIC تغییر بسیار زیادی دارند و نمایش آن‌ها در نمودار نحوه تغییر آن‌ها را نمی‌تواند نشان دهد به همین دلیل از توابعی اکیدا صعودی برای تغییر آن‌ها استفاده شده است. با استفاده از این توابع نحوه تغییر آن‌ها مشخص می‌شود و مشکلی در ترتیب آن‌ها به وجود نمی‌آید. در تصویر بالا مشخص است که مقادیر افزایش می‌یابند و از ویژگی پنجم به بعد تغییر قابل توجه ندارند. اما معیار RSS از همان ویژگی شروع به افزایش جزئی می‌کند که معنی آن پیچیده شدن بیش از حد مدل است. معیار BIC این پیچیده شدن را با تغییرات بیشتری نمایش می‌دهد. البته لازم به ذکر است که این نکته در معیارهای RSS و BIC با تغییر کمی خود را نشان می‌دهد ولی با توجه به اعمال تابع‌های اکیدا صعودی و این تغییرات به خوبی در نمودار مشهود هستند.

با رسم AIC و BIC با استفاده از همان تابع اکیدا صعودی توضیح داده شده در نمودار فوق در هر مرحله می‌توان عملکردشان را در کنار یکدیگر با دقت بیشتری بررسی کرد. همانگونه که مشخص است دو معیار با مقادیری بسیار نزدیک به هم شروع می‌شوند ولی به مرور و با پیچیده‌تر شدن مدل مقادیرشان فاصله می‌گیرند. با پیچیده‌تر شدن مدل معیار BIC نسبت به معیار AIC بیشتر افزایش می‌یابد. این افزایش مقدار BIC‌به گونه‌ای است که در مرحله اضافه شدن متغیر ششم می‌توان به قطع گفت که اضافه کردن این متغیر تاثیر مثبتی در مدل ندارد و حتی تاثیر منفی دارد. در حالی که معیار AIC در اضافه شدن متغیر هشتم پیچیده‌تر شدن مدل را مضرر اعلام می‌کند.



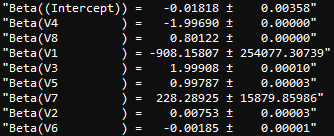
## و

ماتریس واریانس-کوواریانس مدل به شکل زیر خواهد بود.



تخمین بایاس نشده برابر ۰.۰۳۹۶۸۵ است.

مقدار بتا نیز به شکل زیر تخمین زده شده است.



با استفاده از روش‌های محاسبه Leave One Out Cross Validation دو تخمین برای R به دست می‌آید. در روش اول که n بار مدل آموزش داده می‌شود مقدار ۱۶.۲۱۰۳۲۳ و برای حالتی که فقط یک بار مدل آموزش داده می‌شود مقدار ۱۵.۸۲۲۸۱۸ به دست می‌آید. همانگونه که مشخص است مقدار هر دو تخمین بیشتر از است.

## ز

برای این قسمت از پروژه هر بار مدل را به ازای همه متغیرهای مستقل به جز یکی از آن‌ها آموزش می‌دهیم و متغیری را انتخاب و حذف می‌کنیم که ریسک با حذف آن کمینه می‌شود. در مرحله اول روی هر هشت متغیر به جز یکی از آن‌ها آموزش می‌دهیم.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 1 | 15.9569 | 15.6455 | 5.0915 |
| 2 | 15.8953 | 15.6005 | 4.8920 |
| 3 | 1581.5069 | 1551.2087 | 500.3209 |
| 4 | 39319.0308 | 38561.5103 | 9744.8849 |
| 5 | 1570.4173 | 1540.8336 | 339.2090 |
| 6 | 15.8396 | 15.5396 | 5.0849 |
| 7 | 15.9583 | 15.6469 | 5.0914 |
| 8 | 24230.9763 | 23729.7790 | 5581.8957 |

با توجه به اینکه معیار با حذف متغیر ششم کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مرحله بعد به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 1 | 15.9473 | 15.6775 | 5.0429 |
| 2 | 15.8756 | 15.6223 | 4.8549 |
| 3 | 1581.9258 | 1555.8561 | 494.4373 |
| 4 | 39374.2920 | 38707.7220 | 9692.5917 |
| 5 | 1570.0272 | 1543.8627 | 336.4822 |
| 7 | 15.9488 | 15.6790 | 5.0428 |
| 8 | 24185.4658 | 23754.0275 | 5596.8773 |

با توجه به اینکه معیار با حذف متغیر دوم کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مرحله بعد به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 1 | 15.9684 | 15.7454 | 4.8428 |
| 3 | 1588.6948 | 1566.0803 | 482.5020 |
| 4 | 39454.3159 | 38888.9869 | 9608.3739 |
| 5 | 1571.2260 | 1548.7934 | 340.4433 |
| 7 | 15.9698 | 15.7468 | 4.8429 |
| 8 | 24651.6117 | 24278.9288 | 5507.7943 |

با توجه به اینکه معیار با حذف متغیر اول کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مرحله بعد به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 3 | 1586.6359 | 1566.8988 | 481.3467 |
| 4 | 39452.0316 | 38956.4799 | 9565.7801 |
| 5 | 1569.6228 | 1550.7752 | 341.6106 |
| 7 | 10532.9289 | 10405.0560 | 1952.4567 |
| 8 | 24606.1049 | 24289.9243 | 5491.9844 |

با توجه به اینکه معیار با حذف متغیر پنجم کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مرحله بعد به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 3 | 3254.1231 | 3221.0613 | 923.9860 |
| 4 | 40798.9174 | 40406.5910 | 10145.9345 |
| 7 | 12204.9093 | 12084.8691 | 2465.8933 |
| 8 | 26664.8147 | 26386.9343 | 5743.1670 |

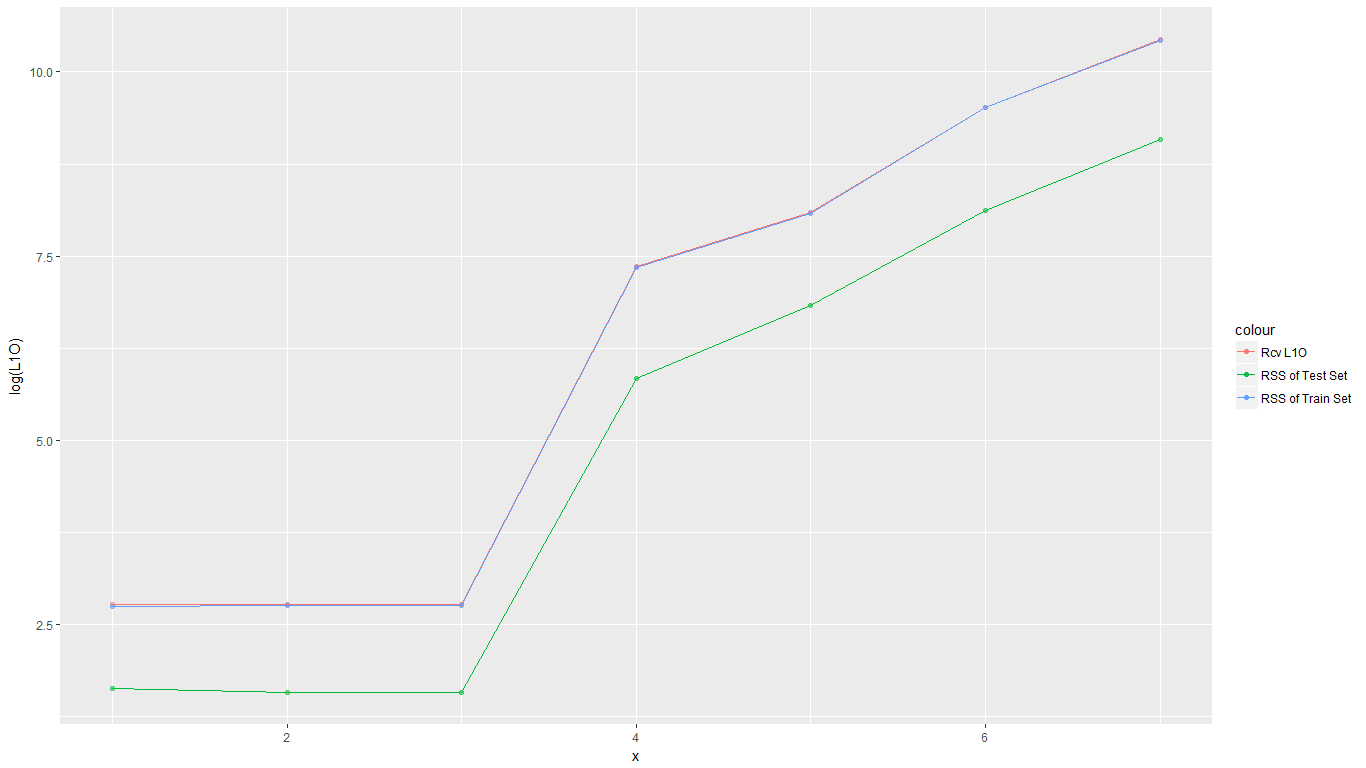
با توجه به اینکه معیار با حذف متغیر سوم کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مرحله بعد به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 4 | 42069.5430 | 41765.2664 | 10908.4726 |
| 7 | 13677.4352 | 13576.6607 | 3365.5051 |
| 8 | 27680.8314 | 27465.4572 | 6314.3629 |

با توجه به اینکه معیار با حذف متغیر هفتم کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مرحله بعد به شرح زیر است.

|  |  |  |  |
| --- | --- | --- | --- |
| Removed Variable |  |  |  |
| 4 | 55959.4675 | 55667.3145 | 14081.0158 |
| 8 | 34051.5154 | 33883.2848 | 8751.9791 |

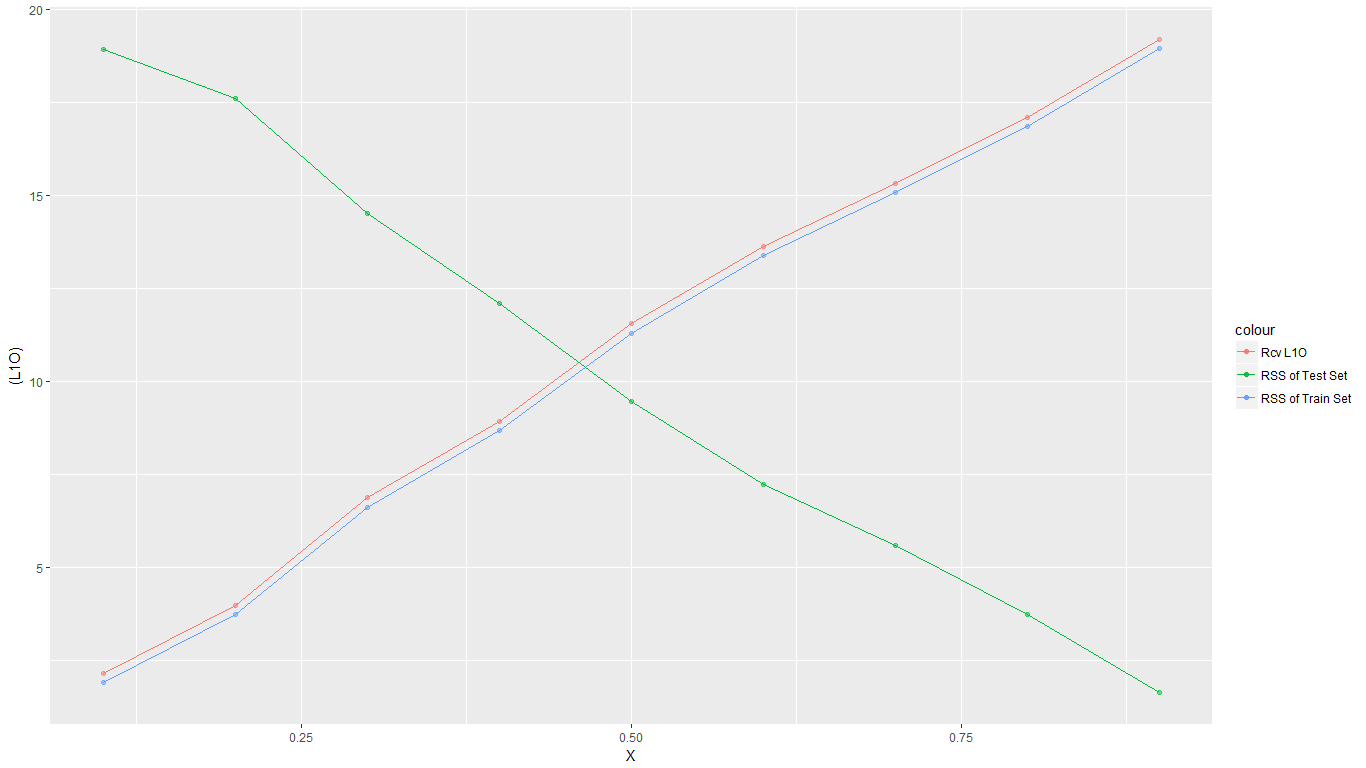
با توجه به اینکه معیار با حذف متغیر هشتم کمینه می‌شود پس این متغیر را حذف می‌کنیم. مقادیر محاسبه شده در مراحل طی شده گزارش شده‌اند و نمودار آن‌ها به شکل زیر است.



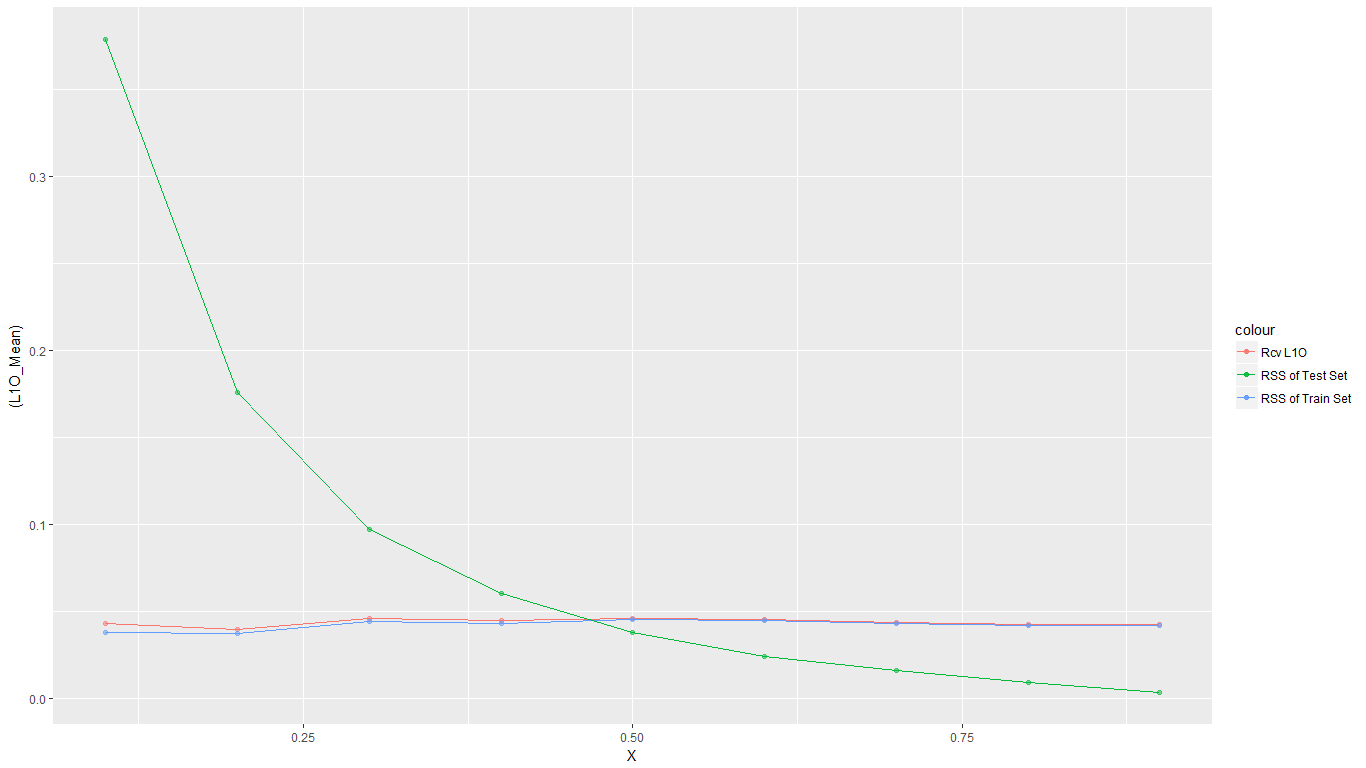
سه مشاهده مهم در مورد نمودار بالا وجود دارد. اول اینکه مقادیر محاسبه شده برای و شباهت زیادی به هم دارند؛ دلیل این شباهت هم این است که هر دو معیار تخمین زننده یک معیار هستند. نکته دوم این است که با حذف متغیرها افزایش می‌یابد ولی با حذف سومین متغیر مستقل افزایش چشمگیری دارد. نکته سوم این است که افزایش چشمگیر دقیقا همان جایی رخ می‌دهد که شروع به رشد می‌کند. قبل از حذف این متغیر با حذف متغیرهای مستقل مقدار کاهش می‌یابد و پس از حذف این متغیر مستقل شروع به رشد و با حذف متغیرهای دیگر رشد آن ادامه می‌یابد.

## ح

با تغییر نسبت داده‌های آموزشی به آزمایشی از ۱۰ درصد تا ۹۰ درصد با فاصله ۱۰ درصد مقادیر ، و به شکل زیر تغییر می‌کنند.



این مقادیر مجموع خطای پیشبینی داده‌های آموزشی و آزمایشی هستند و کاملا به تعداد داده‌ها در مجموعه داده آموزشی و آزمایشی وابسته هستند. از این رو نمی‌توان اطلاعات مناسبی از این نمودار به دست آورد. در عوض از نسبت این مقادیر به تعداد داده‌ها در مجموعه داده‌های آموزشی و آزمایشی می‌توان استفاده کرد. چون در این معیارها مقادیر قبل بر تعداد داده‌ها تقسیم می‌شوند تاثیر تعداد داده‌ها حذف می‌شود. نمودار به دست آمده به شکل زیر است.

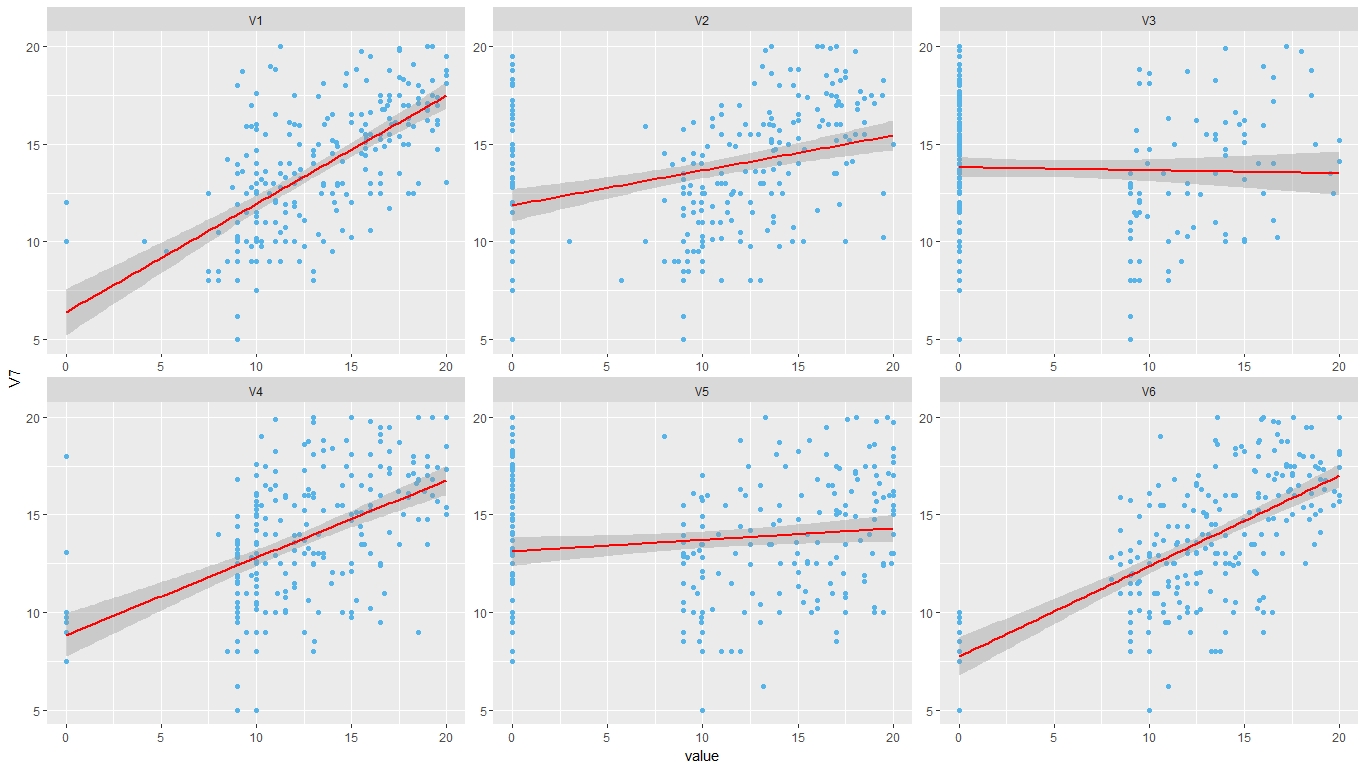


نکته اول در مورد نمودار بالا این است که تقریبا خطای و ثابت هستند و با تغییر تعداد داده‌ها تغییر نمی‌کنند. همچنین این دو مقدار با افزایش تعداد داده‌ها به هم نزدیک‌تر می‌شوند و می‌توان گفت که مقداری که تخمین می‌زنند هم نزدیک این مقدار است. نکته دوم در مورد خطای داده‌های آزمایشی است. با افزایش داده‌های آموزشی پارامترهای یادگیری شده تخمین بهتری از پارامترهای واقعی ارائه می‌کنند در نتیجه خطای داده‌های آزمایشی نیز کمتر می‌شود. در ابتدا با افزایش داده‌ها با شدت زیادی خطا کاهش می‌یابد ولی با ادامه روند کاهش خطا کندتر می‌شود تا جایی که تقریبا متوقف می‌شود و این مقدار را می‌توان خطای مدل در نظر گرفت. برای کاهش این خطا نیز باید از مدل‌های پیچیده‌تر استفاده کرد.

# مجموعه داده Dataset2.csv

## الف

نمودار نقطه‌ای مربوط به هرکدام از ویژگی‌ها و متغیر هدف به شکل زیر است.



خط قرمز خط مربوط به مدل رگرسیون خطی است که رابطه متغیر هدف و ویژگی مربوط را نمایش می‌دهد. خطوط قرمز نشان دهنده رابطه متغیرهای مستقل و متغیر وابسته هستند. می‌توان از هرکدام از خطوط دو گزاره نتیجه گرفت:

1. با توجه به اینکه طول و عرض نمودارها تقریبا برابر هستند، می‌توان گفت تقریبا این خطوط همان خطوط همبستگی هستند و مانند خطوط همبستگی تحلیل می‌شوند. شیب خطوط نمایش دهنده رابطه متغیر مستقل و وابسته هستند؛ یعنی هرقدر شیب بیشتر باشد با تغییر یک واحد در متغیر مستقل، می‌توان انتظار داشت متغیر مستقل بیشتر در جهت شیب تغییر کند. در همه نمودارهای بالا شیب مثبت است پس با در نظر گرفتن دو داده تصادفی و تفاضل مقدار متغیر مستقلشان می‌توان انتظار داشت به مقداری متناظر با آن تفاضل مقدار متغیر وابسته داده با متغیر مستقل بیشتر، بیشتر باشد.
2. نوار اطراف خطوط یک بازه اطمینان ۷۵٪ برای این خطوط است. یعنی مقدار واقعی همبستگی با احتمال ۷۵٪ در آن محدوده است و می‌توان بیانی مانند آنچه در مورد ۱ گفته شد ارائه داد.

## ب

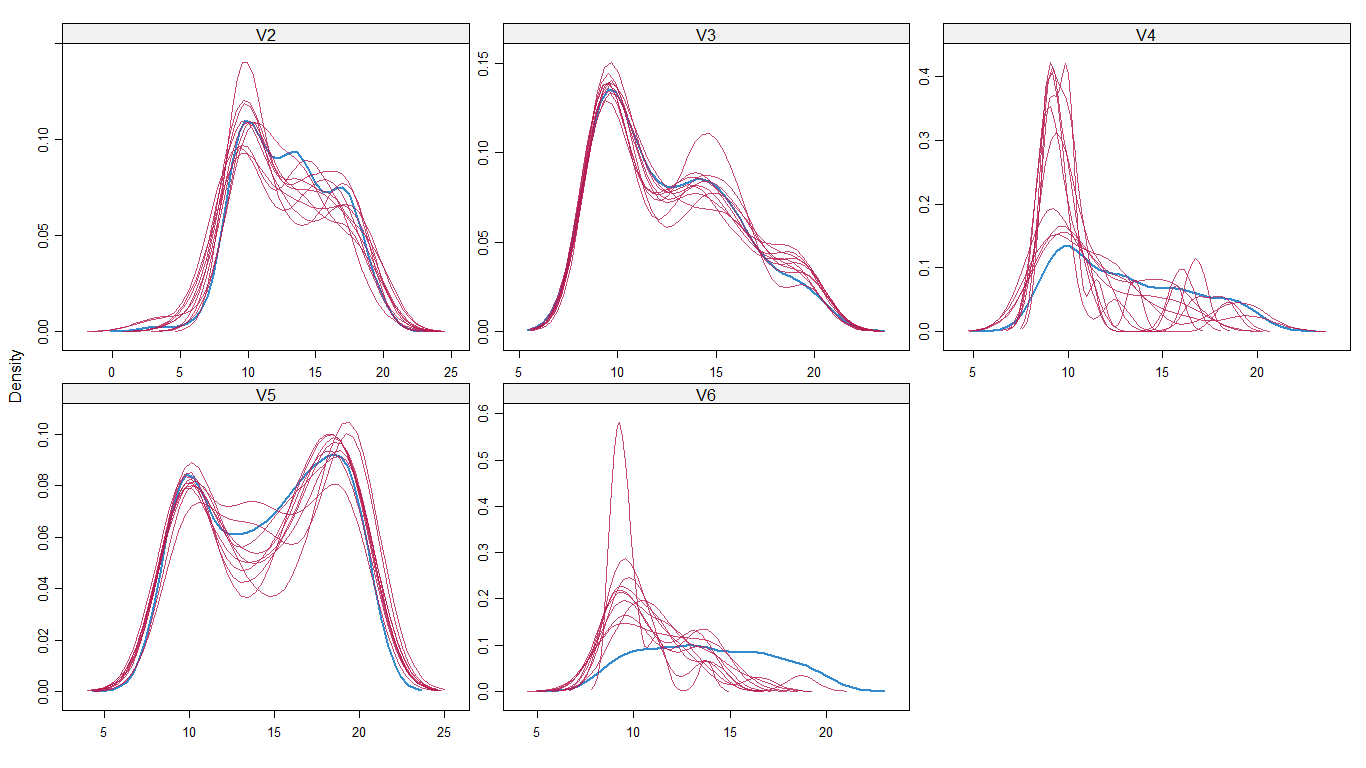
در ابتدا داده‌های اصلی را بدون داده‌های گم شده رسم می‌کنیم.



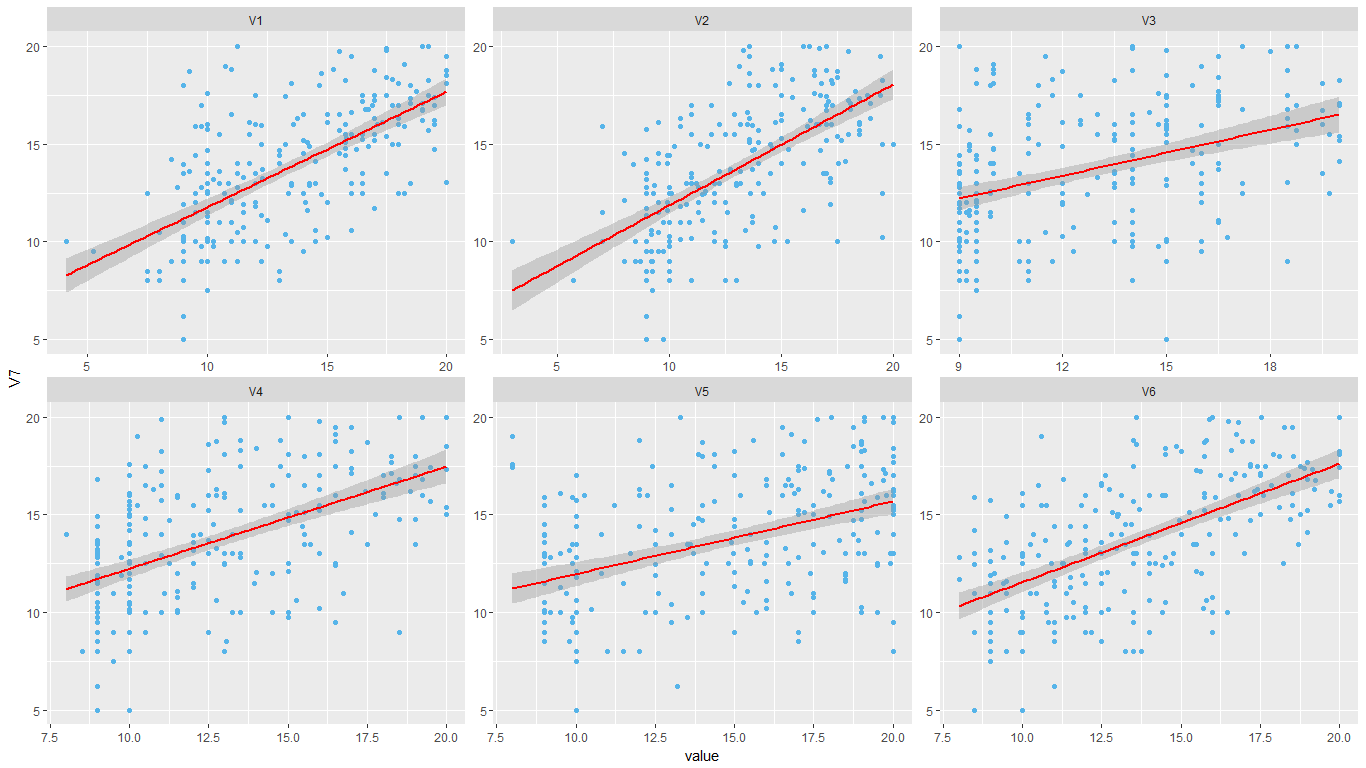
با حذف داده‌های گم شده و رسم دوباره نمودار می‌توان مشاهده کرد که رابطه برخی از متغیرهای مستقل با متغیر وابسته بیشتر از مقداری است که در نمودارهای اولیه دیده شد. تغییر شیب در نمودار متغیرهای دوم، سوم و پنجم محسوس است. حال باید داده‌های گم شده را به شیوه بهتری پر کنیم تا داده‌های گم شده تاثیر منفی بر مدل یاد گرفته شده از داده‌ها نگذارد و مدل آموزش داده شده قابل تعمیم به کل جامعه باشد.

* در حال حاضر داده‌های گم شده با مقدار صفر پر شده‌اند ولی این مقادیر باعث تاثیر منفی در توزیع تخمین زده شده برای جامعه می‌شوند. دلیل اصلی این است که وزن زیادی روی مقدار صفر قرار می‌دهند که باعث می‌شود توزیع تخمین زده شده به کلی متفاوت با توزیع اصلی جامعه باشد.
* روش بعدی استفاده از میانگین هر متغیر برای پر کردن داده‌های گم شده مربوط به آن متغیر است. این روش هم تاثیری مانند مورد قبل می‌گذارند.
* روش بعدی این است که با استفاده از داده‌های موجود یک توزیع برای متغیر پیدا کرده و مقدارهای تصادفی از آن را تولید و به جای داده‌های گم شده از آن‌ها استفاده کنیم. تعمیم بهتر برای این روش این است که به ازای هر داده به نحوی با توجه به داده‌ها احتمال تولید داده‌ی گم شده در صورت وجود را پیدا کنیم و حال فقط مقادیری که بیشترین احتمال تولید خواهند داشت را برای آن‌ها در نظر بگیریم. یعنی به طور کلی می‌توان با این مسئله به نحوی به عنوان یک مسئله پیش‌بینی برخورد کنیم.

در ادامه چارچوب الگوریتمی برای پر کردن داده‌های گم شده با استفاده از دو مورد آخر ذکر شده در موارد بالا بیان می‌شود. این الگوریتم که Multiple Imputation by Chained Equations (MICE) نام دارد در ابتدا یک مقدار ساده (مانند میانگین) را برای داده‌های گم شده در نظر می‌گیرد. حال با استفاده از داده‌های کامل شده مدل‌هایی را آموزش می‌دهد و داده‌های گم شده که پر شده بودند را پاک می‌کند. حال با استفاده از داده‌های موجود داده‌های گم شده را پیش‌بینی می‌کند. این عمل تا زمان همگرایی تکرار می‌شوند و آخرین مقدار به دست آمده به عنوان مقادیر پر شده معرفی می‌شوند. تصویر زیر توزیع مربوط به داده‌ها و داده‌ها همراه با داده‌های پر شده پس از هر تکرار را نمایش می‌دهند.



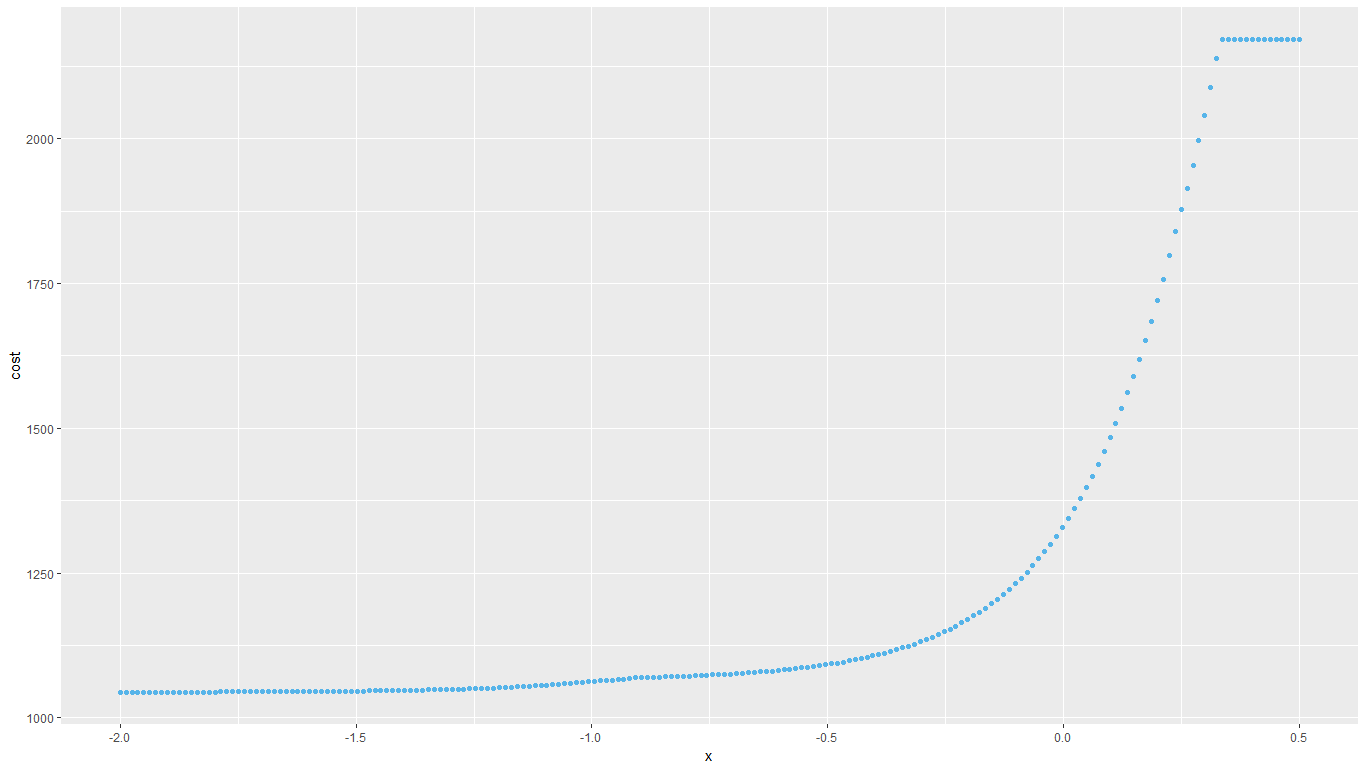
با پر کردن داده‌ها با روش بالا به مجموعه داده کاملی دست پیدا می‌کنیم. رابطه متغیرهای مستقل و متغیر وابسته به شکل زیر است.



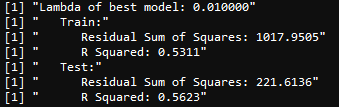
خطوط همبستگی شباهت زیادی با حالتی دارند که داده‌های ناقص حذف شده بودند؛ یعنی شیب و عرض از مبداشان نزدیک به حالت قبلی است؛ این نکته نشانه روش مناسب برای پر کردن داده‌های ناقص است. با مقایسه بازه اطمینان به دست آمده برای خطوط کوچک‌تر شده است که نشانه دقیق‌تر شدن تخمین همبستگی دو متغیر است.

## د

انجام مورد ج ارزش تحلیلی ندارد و در واقع جزئی از موارد د و بعدی است.



نمودار معیار Lasso با توجه به به شکل بالا است. کاملا مشخص است بهترین مدل مدلی است که در آن به صفر نزدیک باشد. دلیل این که مدلی با این ویژگی بهتر است کمبود ویژگی‌ها است. مقادیر مورد نظر به ترتیب زیر محاسبه شدند.



## ه

این داده‌ها نیز مقادیر گم شده دارند. برای پر کردن مقادیر گم شده این داده‌ها، داده‌های قبلی را که پر کرده بودیم در کنار این داده‌ها قرار می‌دهیم و به روش قبلی داده‌های گم شده را پر می‌کنیم. برای خروجی گرفتن از مدل ابتدا مدل را روی کل داده‌ها آموزش می‌دهیم و پس از آن با استفاده از این مدل نهایی پیش‌بینی‌ها را انجام می‌دهیم. خروجی‌ها در فایلی ضمیمه شده‌اند.

## و