به نام خدا

آزمایشگاه R درس تحلیل رگرسیون

دانشگاه صنعتی شریف

دانشکده ریاضی و علوم کامپیوتر

مدرس : دکتر میرصادقی - دکتر کدیور

تمرین سری اول

علی نوریان

۹۸۱۰۲۵۲۷

نیمسال ۲-۰۲

۱ بخش اول - پیش پردازش و آماده سازی دیتاست

در ابتدا کتابخانه‌های مورد نیاز و سپس دیتاست را لود می‌کنیم:

```
# ========================= Import Libraries & Data ========================
Sys.setlocale(locale = 'persian')

library(data.table)
library(ggplot2)

d = fread('iranprovs_mortality_monthly.csv', encoding = 'UTF-8')
```

سپس متغیر ym_num را توسط مقادیر سال و ماه به صورتی میسازیم که برای هر ماه هر سال یک مقدار یکتا داشته باشد. از این متغیر برای مرتب سازی و رسم نمودار استفاده می‌کنیم. همچنین دیتاست جدیدی را با حذف گروه سنی و جنسیت می‌سازیم. محتوای دیتاست جدید را آمار ۶ سال قبل تا به امروز قرار می‌دهیم زیرا داده‌های قبل ۶ سال با توجه به تغییر شرایط کشور و وضعیت بهداشت و جمعیتی ممکن است مناسب نباشند. دو متغیر $provs$ و $months$ را نیز برای داشتن مقادیر یکتای استان‌ها و ماه‌ها تعریف می‌کنیم:

```
# ============================ pre-processing ==============================
d$ym_num = d$y + d$m / 12 - 1/24            # create year-month feature


ym_num_covid = 1398 + 10/12 - 1/24          # set start date of COVID-19
ym_num_start = ym_num_covid - 6             # set start date of task valid
    data

ds = d[, .(n = sum(n)), .(y, m, ym_num, prov)]# remove age & sex features
ds = ds[ym_num > ym_num_start]          # remove invalid data for this task

provs = unique(ds$prov)                     # provinces
months = unique(ds$m)                       # months
```

۲ بخش دوم - مدل خطی برای پیشبینی فوت عادی

برای بدست آوردن مقدار فوت اضافه ابتدا نیاز است مقدار فوت در شرایط عادی (بدون حضور بیماری) را پیش‌بینی کنیم. برای اینکار یک مدل خطی برای پیش‌بینی میزان فوت بدون حضور کرونا در بازه زمانی شیوع کرونا در نظر می‌گیریم. مدل خطی را با توجه به داده‌های قبل کرونا بدست آورده و برای بازه کرونا پیش‌بینی را انجام می‌دهیم.

با توجه به اینکه میزان فوتی در ماه‌های مختلف سال متغیر و وابسته به ماه است، مدل خطی را به تفکیک هر ماه برای هر استان بدست می‌آوریم. همچنین با توجه به مقدار p-value اگر مدل خطی مناسب نباشد از میانگین فوتی‌های قبل کرونا برای پیش‌بینی استفاده می‌کنیم.

برای هر مدل یک بازه اطمینان برای متغیر پاسخ (معادل با دو برابر انحراف معیار داده‌ها) در نظر می‌گیریم که اگر مقدار فوتی در زمان کرونا خارج از این بازه قرار گرفت به معنی تاثیر کرونا بر فوتی و ایجاد فوت اضافه است.

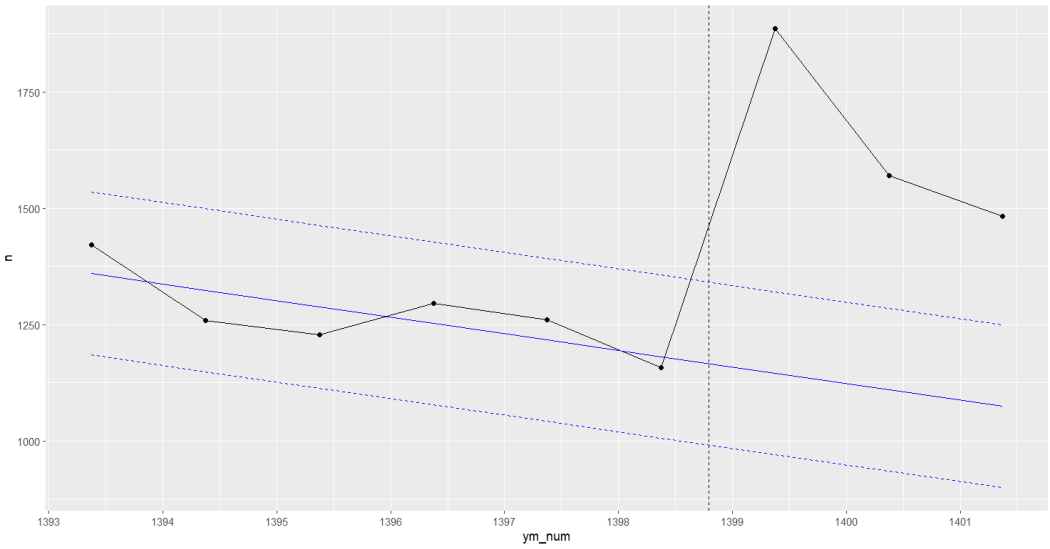
کد این قسمت در صفحه بعد آمده است.

```r
# ====== fit linear model and obtaining the number of excess death ======

ds$n_predicted = 0                    # model predicted value for death
ds$upper_thresh = 0                   # upper threshold for normal death
ds$lower_thresh = 0                   # lower threshold for normal death

# fit model for each month of each province
for (this_prov in provs) {
  for (this_month in months) {
    condition = ds$prov == this_prov & ds$m == this_month

    this_ds = ds[condition]
    this_train_ds = this_ds[ym_num < ym_num_covid]

    fit = lm(n ~ ym_num, this_train_ds)
    pvalue = summary(fit)$coefficients[2,4]  # calculating p-value

    # choose model prediction for p-value < 0.1
    # otherwise, choose average of samples instead of model prediction
    if (pvalue < 0.1) {
      n_predicted = pmax(predict(fit ,this_ds), 0)
      # if prediction gets negative then use min num of death instead
      n_predicted[n_predicted == 0] = min(this_train_ds$n)
    } else {
      n_predicted = rep(mean(this_train_ds$n),length(this_ds$n))
    }

    sigma = sd(this_train_ds$n)       # calculating standard-deviation

    ds[condition]$n_predicted = n_predicted              # save
        predictions
    # set upper threshold
    ds[condition]$upper_thresh = n_predicted + 2 * sigma
    # set lower threshold
    ds[condition]$lower_thresh = n_predicted - 2 * sigma
  }
}
```

سپس با توجه به پیش‌بینی مقدار فوت در غیاب کرونا و همچنین آستانه بالا برای در نظر گرفتن تغییرات عادی در فوتی‌ها بدین صورت عمل می‌کنیم: اگر مقدار فوتی از آستانه بالا بیشتر شود، آن مقدار فوتی نامعقول بوده و اختلاف آن با مقدار فوت پیش‌بینی شده، مقدار فوت اضافه خواهد بود.

```r
# calculating excess death according to predictions and upper thresholds
    of death
ds$excess_death = round((ds$n - ds$n_predicted)*(ds$n > ds$upper_thresh))
```

نمونه‌ای از این اعمال مدل خطی در شکل زیر آماده است:



شکل ۱: فوتی‌های استان آذربایجان‌غربی برای ماه مرداد از سال ۱۳۹۳ تا ۱۴۰۱

داده‌های مشکل مقدار فوت ثبت شده در مرکز آمار، خط آبی، مدل خطی بدست آمده توسط داده‌های قبل کرونا، خط چین مشکی مشخص کننده زمان شروع کرونا و دو خط چین آبی نشان‌دهنده آستانه بالا و پایین برای میزان فوتی می‌باشد.

همانطور که در نمودار بالا نیز مشخص است بعد از شروع کرونا، میزان فوتی در ماه مرداد سال‌های ۱۳۹۹ تا ۱۴۰۱ تفاوت فاحشی با سال‌های قبل از کرونا دارد.

۳ بخش سوم - رسم نمودارهای فوت اضافه استان‌ها

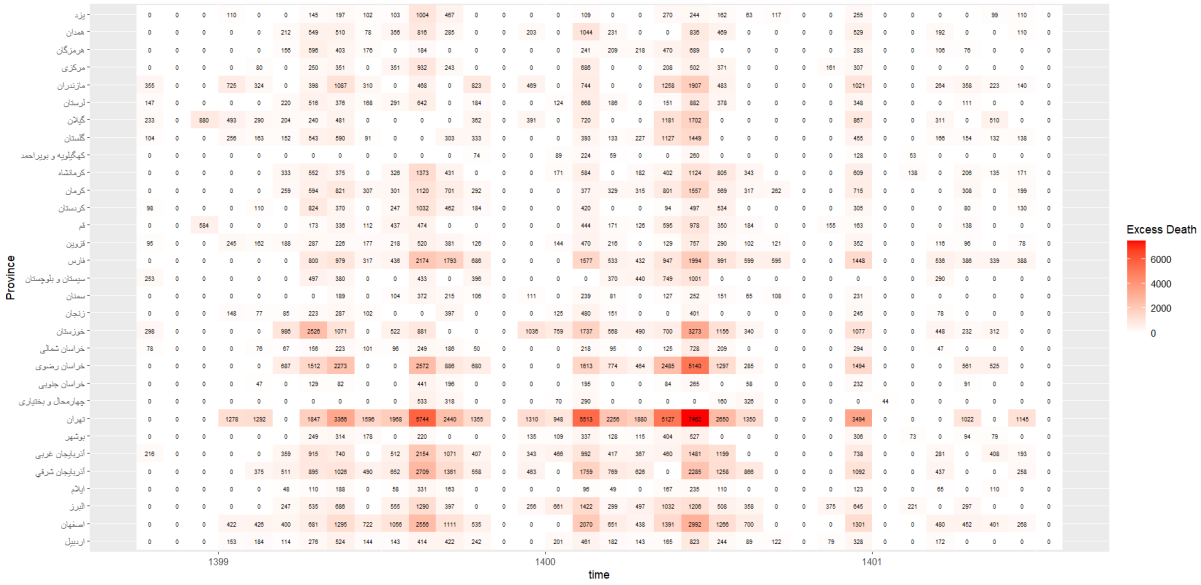
در قسمت قبل مقدار فوت اضافه برای هر استان به تفکیک ماه‌های سال را بدست آوردیم. در این قسمت جدول مربوط به آمار فوتی اضافه در هر ماه از استان‌ها و همچنین نقشه حرارتی را برای نشان دادن شدت و ضعف کرونا در ماه‌های مختلف هر استان رسم می‌کنیم:

```
# ========================== COVID-19 excess death ==========================

COVID19_ds = ds[ym_num >= ym_num_covid]          # COVID-19 data
COVID19_ds$norm_excess_death = COVID19_ds$excess_death / COVID19_ds$n_
    predicted * 100

# plot heat map: percentage of excess death for each month of each
    province
ggplot(COVID19_ds, aes(x = ym_num, y = prov)) +
  geom_raster(aes(fill = norm_excess_death)) +
  scale_fill_gradient(low = "white", high = "red")+
  labs(title = "COVID-19 Precentage of Excess Death", x = "time", y = "
      Province")
```

پس از بدست آوردن مقدار فوت اضافه در هر ماه از هر استان، با توجه به اینکه جمعیت استان‌ها با یکدیگر متفاوت است، در نقشه حرارتی برای نشان دادن شدت و ضعف کرونا از نسبت میزان فوت اضافه به فوت معقول (مقدار پیش‌بینی شده) استفاده می‌کنیم. بنابراین مقادیر موجود در نقشه حرارتی درصد تغییر فوتی‌ها را در اثر کرونا نشان می‌دهد. البته در جدول صفحه بعد مقدار فوت اضافه در هر خانه جدول نمایش داده شده است.

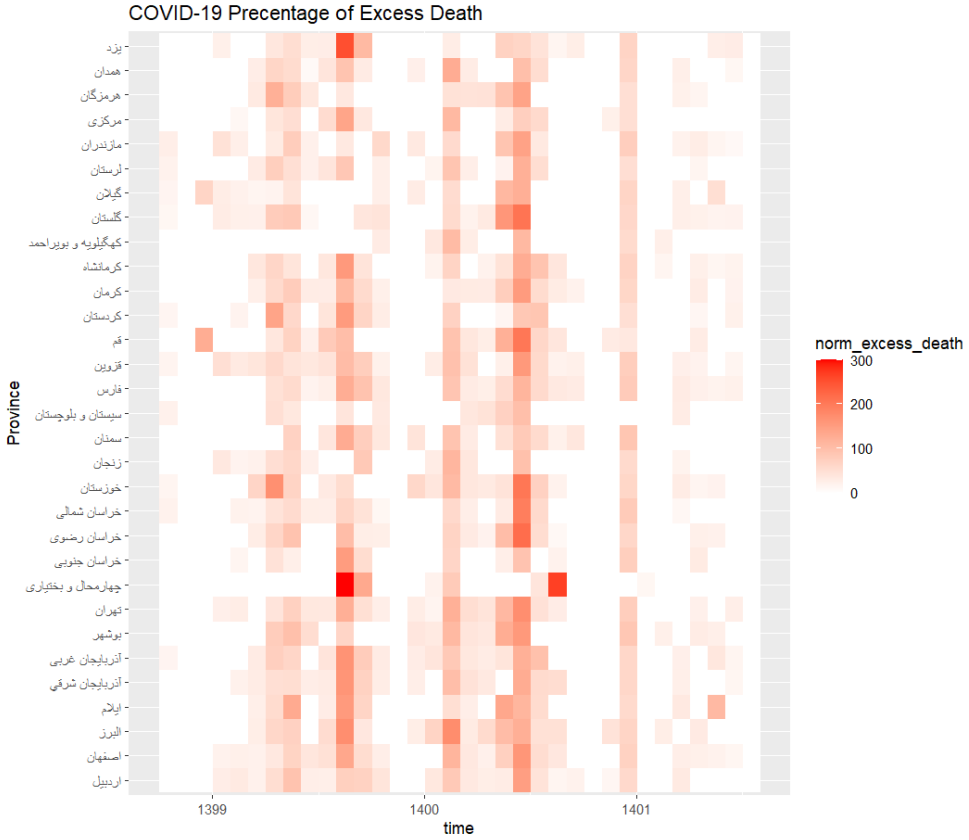
خروجی کد فوق در ادامه آمده است.

جدول فوتی‌های اضافه:



شکل ۲: مقدار فوت اضافه در ماه‌های مختلف در استان‌های کشور

نقشه حرارتی شدت و ضعف کرونا:



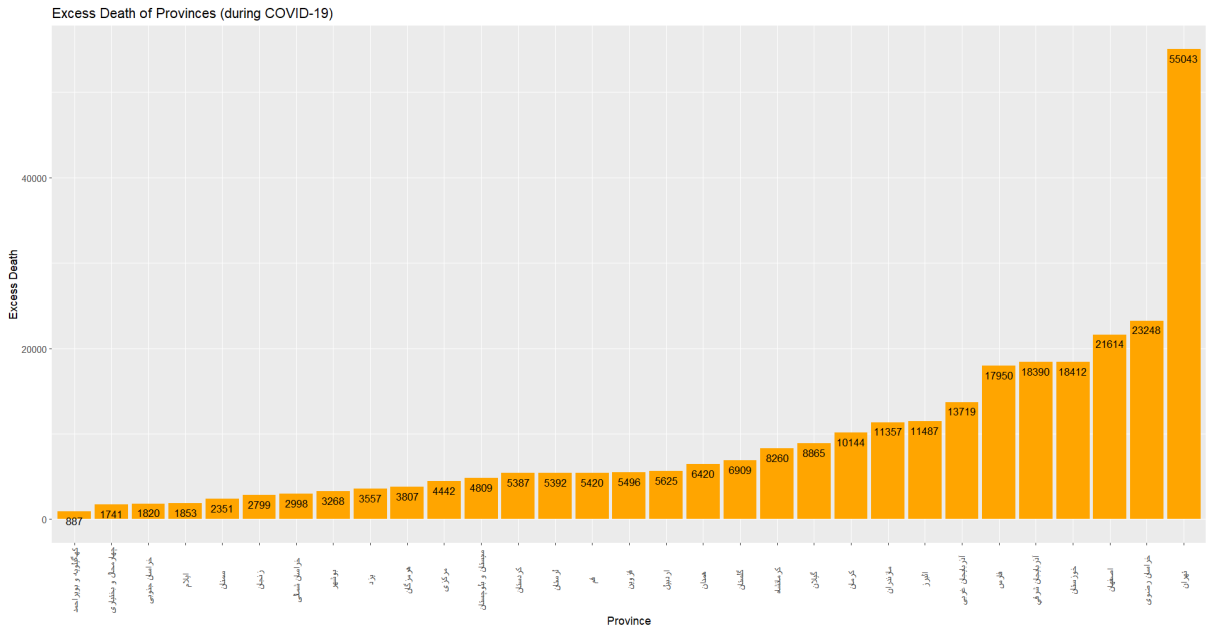
شکل ۳: نقشه حرارتی شدت و ضعف کرونا در استان‌های کشور

۴ بخش چهارم - مجموع تعداد فوت اضافه در استان‌ها و کشور

در قسمت قبل مقدار فوت اضافه را برای هر ماه از هر استان حساب کردیم. در این قسمت نمودار مجموع فوتی‌های اضافه استان‌ها را نشان می‌دهیم:

```
# calculate the total excess death for each province
prov_excess_death = COVID19_ds[, .(excess_death=sum(excess_death)), .(y,
    prov)]

# plot the total excess death of provinces
  ggplot(prov_excess_death, aes(x = reorder(prov, excess_death), y =
      excess_death, fill=factor(y))) +
  geom_bar(stat="identity")+
  labs(title = "Excess Death of Provinces (during COVID-19)", x = "
      Province", y = "Excess Death")+
  scale_fill_manual(values=rainbow(20))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5))

all_excess_death = sum(prov_excess_death$excess_death)
sistan_excess_death =sum(prov_excess_death[prov==provs[16]]$excess_death)
yazd_excess_death = sum(prov_excess_death[prov==provs[31]]$excess_death)
```

نتایج کد فوق به صورت زیر است:



شکل ۴: فوتی‌های اضافه هر استان

تعداد فوتی‌های اضافه کل کشور و تعداد فوتی‌های استان سیستان و بلوچستان و استان یزد (خواسته سوال) به صورت زیر است:

تعداد فوتی‌های اضافه کل کشور: ۲۹۳۴۷۰
تعداد فوتی‌های اضافه استان سیستان و بلوچستان: ۴۸۰۹
تعداد فوتی‌های اضافه استان یزد: ۳۵۵۷

۵ بخش پنجم - مقایسه عملکرد استان‌ها برای کنترل کرونا

اکنون میخواهیم عملکرد استان‌های مختلف در برابر کنترل و مهار کرونا را بررسی کنیم. ایده اصلی بدست آوردن سیر تغییر فوتی‌ها در هر استان است. بدین صورت که هر استانی کاهش فوت بیشتری را تجربه کرده باشد یا به عبارت دیگر شیب فوتی‌های آن استان کمتر از سایر استان‌ها باشد، در نتیجه عملکرد بهتری در مهار داشته است.

با توجه به آنکه میزان فوتی در سنین مختلف متفاوت است یا به بیان دیگر تاثیر کرونا بر سنین مختلف متفاوت بوده است، ابتدا جمعیت را به سه بازه سنی بازه سنی ۰ تا ۲۵ سال، ۲۵ سال تا ۶۵ سال، بالای ۶۵ سال تقسیم می‌کنیم و ابتدا عملکرد استان‌ها را در سه بازی سنی معرفی شده به تفکیک بررسی می‌کنیم.

البته توجه به نکته خوب است که بازه سنی زیر ۱۰ سال در برابر این ویروس مقاومت بیشتری داشته‌اند و درگیری این سنین با کرونا بسیار کم گزارش شده است. همچنین در سنین بسیار پایین دلایل فوت دلایلی غیر از کرونا دارد. بنابراین بازه سنی زیر ۱۰ سال را نیز از دیتا حذف کرده‌ایم.

با توجه به اینکه توزیع جمعیتی در استان‌های مختلف متفاوت است، داده‌های هر استان را نرمالایز می‌کنیم تا شیب‌های بدست آمده برای استان‌های مختلف قابل مقایسه با یکدیگر باشند. برای اینکار داده‌های قبل کرونا را جدا کرده و مشابه همان‌کاری که در بخش ۲ انجام دادیم، میزان فوت عادی در ماه‌های مختلف هر استان را بدست می‌آوریم. سپس داده‌های فوتی بعد از کرونا را با مقدار فوتی عادی متنطار با آن نرمالایز می‌کنیم.

موارد فوق را برای هر بازه سنی به صورت جداگانه اعمال شده است. سپس بر روی داده‌های بدست آمده که نرمال شده‌ی آمار فوتی بعد از کرونا هستند، مدل خطی اعمال می‌کنیم و شیب آن را برای آن استان دخیره می‌کنیم. بنابراین در نهایت ۳ شیب برای هر استان که هر کدام متعلق به یک گروه سنی هستند بدست می‌آید.

برای نمونه کد مربوط به عملیات برای یک گروه سنی در زیر آماده است:

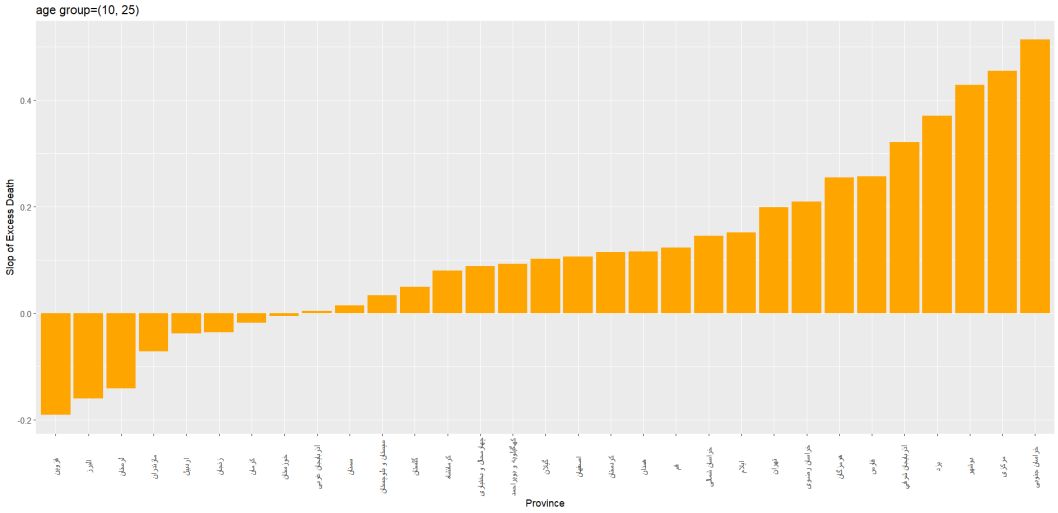
```
# ============== Best Provinces in Controlling COVID-19 ==============
new_ds = d[, .(n = sum(n)), .(ym_num, y, m, prov, age_group)]

# "bc" stands for "before COVID19"
before_COVID19_ds = new_ds[ym_num >= ym_num_start & ym_num < ym_num_covid]

bc_young_ds = before_COVID19_ds[age_group >= "10-14" & age_group <= "20-24"]
bc_young_ds = bc_young_ds[, .(n=mean(n)), .(m, prov)]

after_COVID19_ds = new_ds[ym_num >= ym_num_covid]

young_ds = after_COVID19_ds[age_group>="10-14" & age_group<="20-24"]
young_ds = young_ds[, .(n=sum(n)), .(ym_num, y, m, prov)]

young_ds$n2 = 0

# normalizing the number of deaths for each months of each province
for (this_prov in provs) {
  for (this_month in months) {
    young_ds[m == this_month & prov == this_prov]$n2 =
    young_ds[m == this_month & prov == this_prov]$n / bc_young_ds[m == this_month & prov == this_prov]$n
  }
}
```

ادامه کد صفحه قبل:

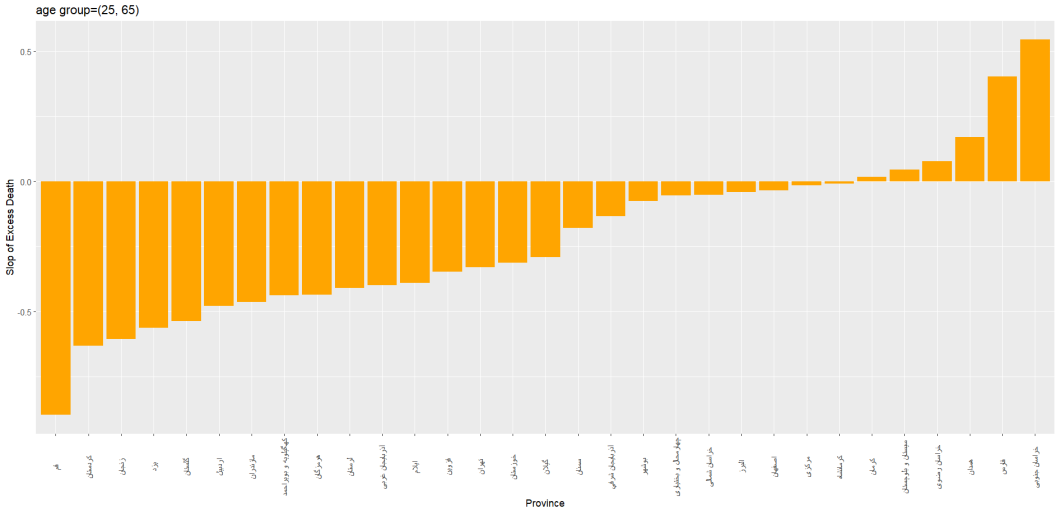
```
  young_ds$n_preidicted = 0
  slop_result = new_ds[, .(n=sum(n)), .(prov)]
  slop_result$young_slop = 0
  for (this_prov in provs) {
    condition = young_ds$prov == this_prov
    this_young_ds = young_ds[condition]
    fit = lm(n2 ~ ym_num, this_young_ds)
    slop_result[prov==this_prov]$young_slop=summary(fit)$coefficients[2,1]
    young_ds[condition]$n_preidicted = predict(fit, this_young_ds)
  }
```

مشابه آنچه برای گروه سنی جوان آمده است و در کنار همین عملیات، عملیات مربوط دو رده سنی بزرگسال و پیر نیز انجام می‌شود.

خروجی مقایسه کاهش فوتی‌ها برای سه گروه سنی جوان، بزرگسال و پیر به تفکیک در ادامه آمده است:

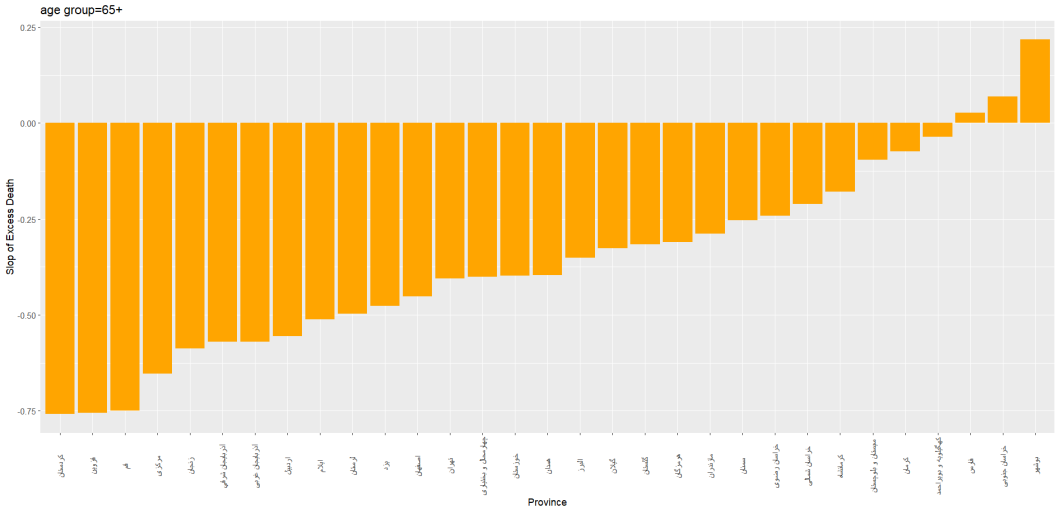


شکل ۵: شیب تغییر فوتی‌ها در استان‌های مختلف در گروه سنی جوان



شکل ۶: شیب تغییر فوتی‌ها در استان‌های مختلف در گروه سنی بزرگسال

ذکر این نکته خوب است که استان قم عملکرد خوب قابل توجهی نسبت به سایر استان‌ها در بازه سنی بزرگسال دارد.

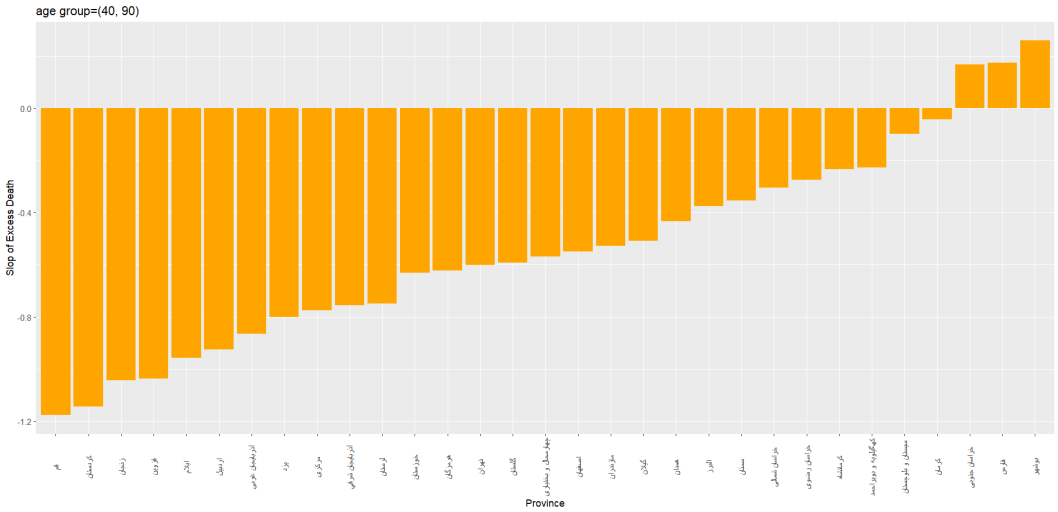


شکل ۷: شیب تغییر فوتی‌ها در استان‌های مختلف در گروه سنی پیر

همانطور که نمودارها مشخص است بهترین استان‌ها در کنترل کرونا در بازه‌های سنی مشخص شده به صورت زیر است:
بازه سنی جوان (۱۰ تا ۲۵ سال): قزوین - البرز - لرستان - مازندران - اردبیل
بازه سنی بزرگسال (۲۵ تا ۶۵ سال): قم - کردستان - زنجان - یزد - گلستان
بازه سنی پیر (بالای ۶۵ سال): کردستان - قزوین - قم - مرکزی - زنجان
عملکرد سایر استان‌ها نیز در نمودارها مشخص است.

اکنون با توجه به اینکه سنین بالاتر بیشتر در معرض این ویروس بودند بنابراین عملکرد هر استان در سنین بالاتر بیشتر از سنین پایین شایان توجه است. همچنین افراد با سن بسیار بالا (بالای ۹۰ سال) بسیار آسیب‌پذیر هستند و کادر درمان نمی‌تواند تاثیر جدی‌ای برای درمان این افراد داشته باشد. ازین رو استان‌هایی که میانگین سنی بالاتری دارند فوتی بیشتری را تجربه می‌کنند. بنابراین بازه سنی ۴۰ تا ۹۰ سال را برای بررسی عملکرد کلی یک استان درنظر می‌گیریم.

کد این قسمت کاملا مانند کد ذکر شده در ابتدای این قسمت است. نتیجه عملکرد استان‌ها به صورت کلی در شکل زیر قابل مشاهده است:



شکل ۸: شیب تغییر فوتی‌ها در استان‌های مختلف

میتوان ترکیبی از سه شیب بدست آمده از گروه سنی‌های مختلف را ارائه داد. نتایج این کار نیز با تقریب خوبی مشابه نمودار فوق است. در نتیجه ۵ استان دارای بهترین عملکرد در کنترل کرونا قم، کردستان، زنجان، قزوین و ایلام می‌باشند.