Practical regression and Anova using R

Введение

Регрессионный анализ используется для объяснения и моделирования отношений между зависимой переменной *Y* (response, output, dependent variable) и независимыми предикторами (predictor, input, independent or explanatory variable) $X_1, \ldots, X_p$

Зависимая переменная должна быть непрерывной, предикторы могут быть непрерывными, дискретными или категориальными величинами

Цели регрессии:

1. Прогноз будущих наблюдений
2. Оценка эффекта (или отношения) предикторов на зависимую переменную
3. Общее описание структуры данных

Описание модели

Рассмотрим простой пример, который послужит описанию регрессионной модели.

Пусть наши переменные – это:

- Y – расход топлива автомобиля
- X_1 – вес автомобиля
- X_2 – количество лошадиных сил
- X_3 – число цилиндров

Тогда линейная модель регрессионного анализа будет выглядеть следующим образом:

Линейная модель регрессионного анализа:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Та же модель на данных будет выглядеть в виде системы линейных уравнений

$$y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i2} + \varepsilon_i$$

Или то же самое в матричной форме:

$$y = X\beta + \varepsilon$$

где $y = (y_1 \ldots y_n)^T$, $\varepsilon = (\varepsilon_1 \ldots \varepsilon_n)^T$, $\beta = (\beta_0 \ldots \beta_3)^T$, $X = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{pmatrix}$

Наша задача состоит в нахождении «хорошей» оценки β. Одной из таких оценок является оценка по методу наименьших квадратов:

$$\sum \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta) \to \min$$

Решением данной оптимизационной задачи является:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Эта оценка является несмещенной: $E\hat{\beta} = \beta$

Почему оценка методом наименьших квадратов «хорошая»?

На это существует как минимум две причины:

- Если ошибки независимые одинаково распределенные нормальные случайные величины, то она является оценкой максимального правдоподобия

- Это лучшая линейная несмещенная оценка при условии, что ошибки независимые и имеют одинаковую дисперсию, $var\varepsilon = \sigma^2 I$. (Теорема Гаусса-Маркова)

На основании данной оценки коэффициентов мы имеем следующие оценки:

- $\hat{y} = X\hat{\beta}$ - прогноз зависимых величин
- $\hat{\varepsilon} = y - X\hat{\beta}$ - ошибки
- $RSS = \hat{\varepsilon}^T \hat{\varepsilon} = (y - X\hat{\beta})(y - X\hat{\beta})$ - остаточная сумма квадратов
- $\hat{\sigma}^2 = \dfrac{RSS}{n-p}$ - оценка σ^2

Объясненная дисперсия

Основным параметром соответствия модели данным является коэффициент детерминации или уровень объясненной дисперсии, который вычисляется следующим образом:

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{RSS}{SS}$$

Свойства:

1. $0 \le R^2 \le 1$
2. для простой регрессии: $R^2 = r^2$

Важно отметить, что *говорить о коэффициенте детерминации имеет смысл только в том случае, когда свободный член включен в регрессионное уравнение.*

Предположение нормальности

Когда говорится о регрессионном анализе, обычно предполагается, что мы имеем дело с ошибками, распределенными по нормальному закону распределения:

$$\varepsilon \overset{d}{=} N(0, \sigma^2 I)$$

Тогда:

$$\hat{y} \overset{d}{=} N(X\beta, \sigma^2 i)$$

$$\hat{\beta} \overset{d}{=} N(\beta, (X^T X)^{-1}\sigma^2)$$

То есть, имея предположение нормальности для ошибок, мы можем строить доверительные интервала для независимых величин и для параметров, оценивающих эффекты предикторов.

Сравнение моделей

Обычной процедурой для статистического анализа является сравнение двух или более моделей.

В случае с регрессией мы можем делать это следующим образом:

Пусть у нас имеется две модели – большая Ω и маленькая ω. Нулевая гипотеза заключается в выборе ω.

В этом случае F-статистика:

$$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega / df_\Omega} \overset{d}{=} F_{df_\omega, df_\Omega}$$

Гипотеза отклоняется, если:

$$F > F_{df_\omega, df_\Omega}{}^{\alpha}$$

Мы можем подвести итоги:

Мы описали линейную регрессионную модель $y = X\beta + \varepsilon$. Параметр β может быть оценен при помощи метода наименьших квадратов. Далее, предположив, что $\varepsilon \overset{d}{=} N(0, \sigma^2 I)$ мы можем строить доверительные интервала как для предсказаний, так и для параметров, а также сравнивать модели по F-критерию.

Проблемы с моделью и анализом

Сталкиваясь я реальными данными, существуют основные трудности, которые можно систематизировать следующим образом:

Источники и качество данных

1. Смещенная выборка
2. Важные предикторы не включены
3. Проблемы с ортогональностью
4. Диапазон и размер данных

Ошибки

1. Коррелируют или имеют неодинаковую дисперсию – GLM
2. Распределение с тяжелыми хвостами – робастные методы
3. Предикторы сильно коррелируют – ridge regression
4. Ненормальность ошибок – нелинейные методы

Структура

1. Неправильная структура
2. Будущее не выводимо из прошлого
3. Нет арпиорной идеи

Проблемы с интерпретацией

Что значит β ?

1. Изменение X на единицу ведет к изменению Y на β ?
2. Изменение X на единицу ведет к изменению Y на β при неизменных остальных предикторах?
3. Еще?

Проблемы:

1. Не включена переменная Z, влияющая на модель
2. Зависимость предикторов

Предсказание более стабильны, чем оценка параметров.

Регрессия в R

Формулы для регрессии в статистической среде R:

model <- lm(Y~X_1+ X_2+…+X_n, data = data)

Чтобы посмотреть результаты анализа:

summary(model)

Сравнение моделей:

anova(model1, model2)

Примеры:

Результаты регрессионного анализа.

```
> g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
> summary(g)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.566087   7.354516    3.88  0.00033
pop15       -0.461193   0.144642   -3.19  0.00260
pop75       -1.691498   1.083599   -1.56  0.12553
dpi         -0.000337   0.000931   -0.36  0.71917
ddpi         0.409695   0.196197    2.09  0.04247

Residual standard error: 3.8 on 45 degrees of freedom
Multiple R-Squared: 0.338,      Adjusted R-squared: 0.28
F-statistic: 5.76 on 4 and 45 degrees of freedom,       p-value: 0.00079
```

Результаты регрессионного анализа с исключенной переменной pop15.

```
> g2 <- lm(sr ~ pop75 + dpi + ddpi, data=savings)
> summary(g2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.487494   1.427662    3.84  0.00037
pop75       0.952857   0.763746    1.25  0.21849
dpi         0.000197   0.001003    0.20  0.84499
ddpi        0.473795   0.213727    2.22  0.03162

Residual standard error: 4.16 on 46 degrees of freedom
Multiple R-Squared: 0.189,      Adjusted R-squared: 0.136
F-statistic: 3.57 on 3 and 46 degrees of freedom,       p-value: 0.0209
```

Результаты регрессионного анализа с двумя исключенными переменными:

```
> g3 <- lm(sr ~ pop75 + ddpi, data=savings)
> summary(g3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.470      1.410    3.88  0.00033
pop75          1.073      0.456    2.35  0.02299
ddpi           0.464      0.205    2.26  0.02856

Residual standard error: 4.12 on 47 degrees of freedom
Multiple R-Squared: 0.188,      Adjusted R-squared: 0.154
F-statistic: 5.45 on 2 and 47 degrees of freedom,       p-value: 0.00742
```

Результаты регрессионного анализа с тремя исключенными переменными:

```
> g4 <- lm(sr ~ pop75, data=savings)
> summary(g4)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.152      1.248    5.73  6.4e-07
pop75          1.099      0.475    2.31    0.025

Residual standard error: 4.29 on 48 degrees of freedom
Multiple R-Squared:  0.1,       Adjusted R-squared: 0.0814
F-statistic: 5.34 on 1 and 48 degrees of freedom,       p-value: 0.0251
```

Результаты Anova.

```
> anova(g2,g)
Analysis of Variance Table

Model 1: sr ~ pop75 + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     46        798
2     45        651  1    147    10.2 0.0026
```