统计学与 R 语言

第 20 讲回归分析 II

张敬信

2022 年 4 月 25 日

哈尔滨商业大学

四. 多元线性回归实例

1. 准备数据与简单探索

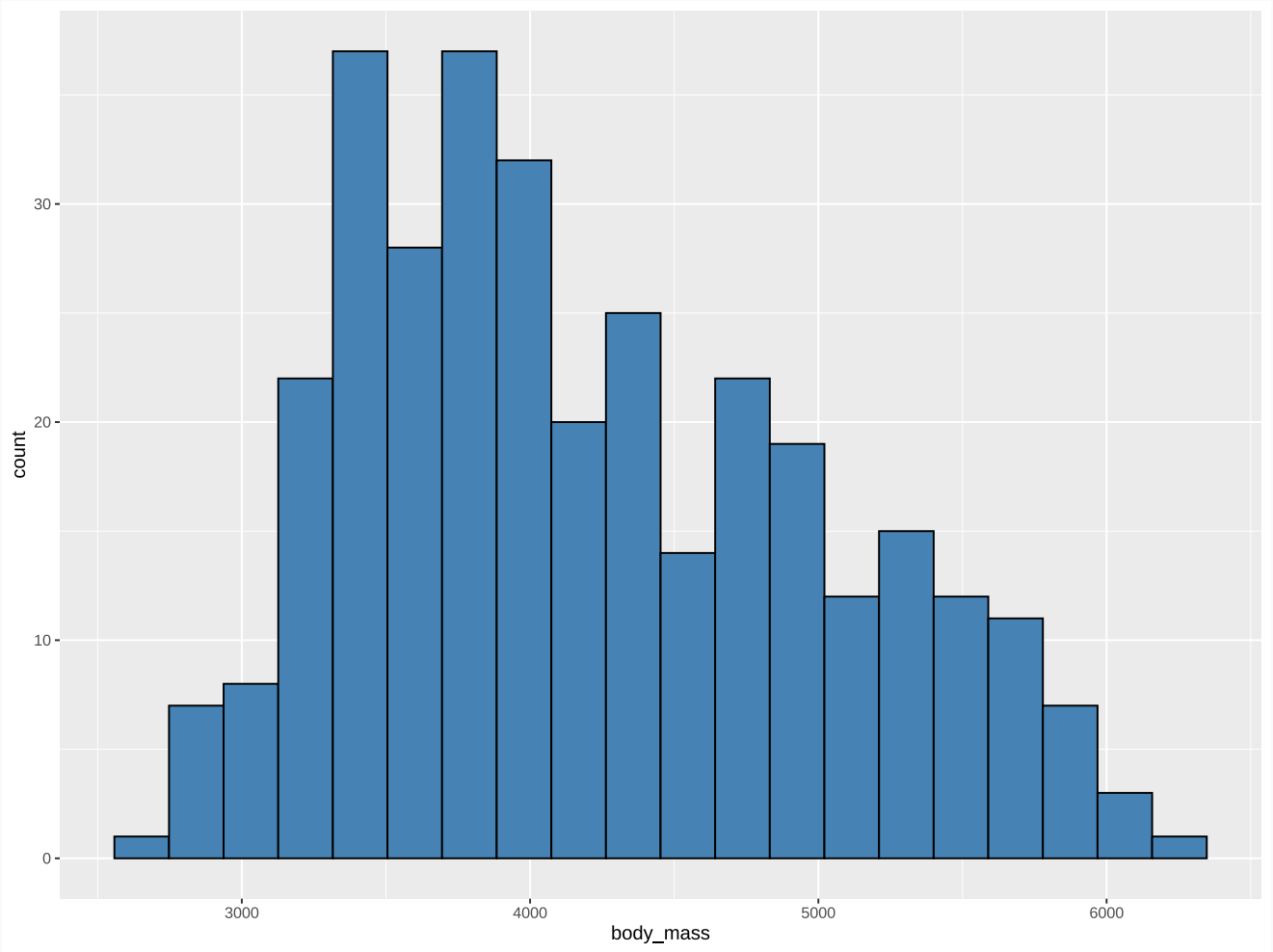
企鹅的数据集 penguins，包含 333 个样本，是有关企鹅的特征信息，包括种类、岛屿、嘴长、嘴宽、鳍长、性别。想确定企鹅体重与这些特征的关系。

```
penguins = read_csv("datas/penguins.csv") %>%
  mutate(species = factor(species))
penguins
#> # A tibble: 333 x 7
#>   species island    bill_length bill_depth flipper_length
#>   <fct>   <chr>           <dbl>      <dbl>           <dbl>
#> 1 Adelie  Torgersen        39.1       18.7             181
#> 2 Adelie  Torgersen        39.5       17.4             186
#> 3 Adelie  Torgersen        40.3       18               195
#> # ... with 330 more rows
```

先探索因变量 body_mass（体重）的分布[1]：

```
ggplot(penguins, aes(body_mass)) +
  geom_histogram(bins = 20, fill = "steelblue",
                 color = "black")
```

[1]若因变量是右偏分布，可以尝试做对数变换变成近似正态分布，这里不做变换.


count
30
20
10
0
3000
4000
5000
6000
body_mass

2. 构建多元线性回归模型

`lm(formula, data, ...)`: 拟合多元线性回归模型

- `formula` 为要拟合的回归模型的形式，例如：`y ~ x1 + x2`, 对应模型 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, 默认包含截距项，若不想包含截距项用 `y ~ x1 + x2 -1`
- 返回值列表包含回归系数、统计量、拟合值、残差等，用 `summary()` 查看汇总模型结果，或者用 `broom` 包提供的 `tidy()`, `glance()`, `augment()` 将模型结果变成整洁数据框。

- `formula` 设定模型公式，遵从 `Wilkinson` 表示规则，更多常用写法：
  - `y ~ .`：包含所有自变量的主效应
  - `x1:x2`：交互效应，即 x_1x_2 项
  - `x1*x2`：包含全部主效应和交互效应，`x1 + x2 + x1:x2` 的简写
  - `I()`：打包式子作为整体
  - `y ~ poly(x, 2, raw = TRUE)`：一元二次多项式回归，同 `y ~ x + I(x^2)`
  - `y ~ polym(x1, x2, degree = 2, raw = TRUE)`：二元二次多项式回归
  - `log(y) ~ x`：对 `y` 做对数变换

- 先把自变量都用上，构建初始多元线性回归模型（往往不是成功模型）：

```
mdl0 = lm(body_mass ~ ., penguins)
summary(mdl0)
#>
#> Call:
#> lm(formula = body_mass ~ ., data = penguins)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -779.2 -167.4   -3.2  179.4  914.3
#>
#> Coefficients:
#>                   Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      -1500.03     575.82   -2.61  0.00961 **
#> speciesChinstrap  -260.31      88.55   -2.94  0.00352 **
#> speciesGentoo      987.76     137.24    7.20  4.3e-12 ***
#> islandDream        -13.10      58.54   -0.22  0.82303
```

```
library(bruceR)
print_table(mdl0, file = " 线性回归系数表.docx")
#> √ Table saved to "C:/Users/zhjx_/Desktop/2021-22-2 学期/统
```

	Estimate	S.E.	t	p	
(Intercept)	–1500.029	(575.822)	–2.605	.010	**
speciesChinstrap	–260.306	(88.551)	–2.940	.004	**
speciesGentoo	987.761	(137.238)	7.197	< .001	***
islandDream	–13.103	(58.541)	–0.224	.823	
islandTorgersen	–48.064	(60.922)	–0.789	.431	
bill_length	18.189	(7.136)	2.549	.011	*
bill_depth	67.575	(19.821)	3.409	< .001	***
flipper_length	16.239	(2.939)	5.524	< .001	***
sexmale	387.224	(48.138)	8.044	< .001	***

整洁模型结果

- tidy(): 模型系数估计及其统计量

```
library(broom)
tidy(mdl0)
#> # A tibble: 9 x 5
#>   term              estimate std.error statistic  p.value
#>   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)        -1500.     576.      -2.61 9.61e- 3
#> 2 speciesChinstrap    -260.      88.6     -2.94 3.52e- 3
#> 3 speciesGentoo        988.     137.       7.20 4.30e-12
#> # ... with 6 more rows
```

- glance(): 模型诊断信息

```
glance(mdl0)
#> # A tibble: 1 x 12
#>   r.squared adj.r.squared sigma statistic   p.value    df
#>       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>
#> 1     0.875         0.872  288.      284. 1.85e-141     8
#> # ... with 3 more variables: deviance <dbl>, df.residual <
```

- augment(): 增加预测值列、残差列等

```
augment(mdl0)
#> # A tibble: 333 x 13
#>   body_mass species island   bill_length bill_depth flippe
#>       <dbl> <fct>   <chr>          <dbl>      <dbl>
#> 1      3750 Adelie  Torgers~        39.1       18.7
#> 2      3800 Adelie  Torgers~        39.5       17.4
#> 3      3250 Adelie  Torgers~        40.3       18
#> # ... with 330 more rows, and 5 more variables: .resid <db
#> #   .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

3. 共线性诊断与逐步回归

用 `car::vif()`[2]诊断回归模型的多重共线性：

```
car::vif(mdl0)
#>                    GVIF Df GVIF^(1/(2*Df))
#> species           63.52  2             2.82
#> island             3.73  2             1.39
#> bill_length        6.10  1             2.47
#> bill_depth         6.10  1             2.47
#> flipper_length     6.80  1             2.61
#> sex                2.33  1             1.53
```

只有分类变量 `species` 的 VIF 值较大，其余均小于 10，说明不存在共线性。

[2] `mctest::imcdiag()` 诊断回归模型的多重共线性更全面，除了计算 VIF 值外，还计算其他诊断指标值.

处理该共线性，可以剔除相对不那么重要的变量，或者用 `step()` 做逐步回归，它可以剔除不显著的自变量，顺便剔除共线性的自变量。

逐步回归是以 AIC 值（越小越好）作为加入和剔除变量的判别条件，参数 `direction` 设置逐步选择的方法："both", "backward"（逐步剔除），"forward"（逐步加入）。

Akaike 信息准则（AIC）常用来比较不同回归模型的拟合效果，优点是既考虑模型的拟合效果又对模型参数过多施加一定惩罚，其定义为：

$$\text{AIC} = 2(p+1) - 2\ln(L)$$

其中，p 为回归模型中自变量的个数，L 为回归模板的对数似然。

```
mdl1 = step(mdl0, direction = "backward", trace = 0)
summary(mdl1)
#>
#> Call:
#> lm(formula = body_mass ~ species + bill_length + bill_dept
#>     flipper_length + sex, data = penguins)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -779.7 -173.2   -9.1  186.6  914.1
#>
#> Coefficients:
#>                   Estimate Std. Error t value Pr(>|t|)
#> (Intercept)      -1460.99     571.31   -2.56  0.01100 *
#> speciesChinstrap  -251.48      81.08   -3.10  0.00209 **
#> speciesGentoo     1014.63     129.56    7.83  6.9e-14 ***
#> bill_length         18.20       7.11    2.56  0.01086 *
```

结果给出了回归系数的标准误、显著性、回归模型的标准误等，基于理论的回归系数的置信区间，可用 confint() 来提取：

```
confint(mdl1)
#>                       2.5 % 97.5 %
#> (Intercept)       -2584.91 -337.1
#> speciesChinstrap   -410.98  -92.0
#> speciesGentoo       759.75 1269.5
#> bill_length           4.22   32.2
#> bill_depth           28.38  106.1
#> flipper_length       10.23   21.7
#> sexmale             295.76  484.0
```

该模型基本上是成功的模型，回归系数都是显著的，模型的调整 R^2 为 0.873。

要计算模型的均方根误差：

```
library(modelr)
rmse(mdl1, penguins)
#> [1] 284
```

4. 关于回归模型中的分类变量

分类变量，取值是有限的类别值，如性别：男、女。分类变量是不能直接用到回归模型中的，即使用 1 表示男，用 0 表示女，这个 1 和 0 仍然只能是起类别区分的作用，如果不加处理让它们当数值 1 和 0 使用了，那么整个模型的逻辑和结果都是不正确的！

分类变量要想正确地用到回归模型，必须处理成虚拟变量[3]。

[3] R 中分类变量只要是因子型或字符型，当加入回归模型时，不需要做任何额外操作将自动处理成虚拟变量用进模型.

企鹅数据 `species` 列是分类变量，包含 3 个类别：“Adelie”, “Gentoo”, “Chinstrap”.

```
table(penguins$species)
#>
#>    Adelie Chinstrap    Gentoo
#>       146        68       119
```

虚拟变量是一种二值变量（0-1），只表示是否。二分类或多分类变量，可以转化为多个二值变量：

> species 是否为 Adelie, species 是否为 Gentoo, species 是否为 Chinstrap

比如第 1 个样本，其 `species = Adelie`，要用上述 3 个二值变量表示的话，就是分别为 `1, 0, 0`.

每个样本都做这样的处理，这就是分类变量转化为虚拟变量，可用 modelr::model_matrix() 函数实现，其参数 data 为数据，formula 为模型公式。

若给 formula 参数提供用于 lm() 的模型公式，则返回真正用于回归模型的分类变量处理成虚拟变量的自变量数据：

```
# species 变成虚拟变量的效果
model_matrix(penguins, ~ species - 1)
#> # A tibble: 333 x 3
#>   speciesAdelie speciesChinstrap speciesGentoo
#>           <dbl>            <dbl>         <dbl>
#> 1             1                0             0
#> 2             1                0             0
#> 3             1                0             0
#> # ... with 330 more rows
```

不是将原 `species` 列，而是换成新的虚拟变量列用到回归模型，注意：这 3 个虚拟变量列是线性相关的：每一列都能用其余 2 列线性表示（1 减去其余 2 列），即有一列是冗余的，这是线性回归所不允许的。

故需要任意去掉一列，再线性回归建模。去掉哪一列都可以，去掉哪一列，做回归建模就相当于以谁为参照列。

比如去掉 `species` 是否为 `Adelie` 列，就相当于"Adelie" 组是参照组，另外 2 组"Gentoo"、"Chinstrap" 与参照组做比较。

去掉冗余列，再增加截距列（一列 1），才是将 species 列真正用于回归模型的转化为虚拟变量后的数据：

```
model_matrix(penguins, ~ species)
#> # A tibble: 333 x 3
#>   `(Intercept)` speciesChinstrap speciesGentoo
#>           <dbl>            <dbl>         <dbl>
#> 1             1                0             0
#> 2             1                0             0
#> 3             1                0             0
#> # ... with 330 more rows
```

冗余列默认是去掉第一水平，若想去掉另一水平（该组作为参照组），可以借助 relevel() 修改第一水平，再处理成虚拟变量：

```
penguins$species = relevel(penguins$species, ref = "Gentoo")
```

根据逐步回归得到的 `mdl1` 的回归系数估计，可以写出拟合的回归方程：

$$\begin{aligned}\text{body\_mass} = &-1460.995 - 251.477 * \text{speciesChinstrap} \\ &+ 1014.627 * \text{speciesGentoo} + 18.204 * \text{bill\_length} \\ &+ 67.218 * \text{bill\_depth} + 15.950 * \text{flipper\_length} \\ &+ 389.892 * \text{sexmale}\end{aligned}$$

连续变量的回归系数好解释，比如 `bill_length` 的系数 `18.204`, 表示嘴长每增加 1 个单位（毫米），体重将增加 `18.204` 个单位（克）。

分类变量回归系数的解释

原二分类变量 `sex`，变成虚拟变量去掉冗余列后只剩一列 `sexmale` (是否为雄性，1 是 0 否)，代入模型来看：

- 若性别不是雄性，`SexMale = 0`

$$\text{body\_mass} = \cdots + 389.892 * 0 + \cdots$$

- 若性别是雄性，`SexMale = 1`

$$\text{body\_mass} = \cdots + 389.892 * 1 + \cdots$$

即雌性则 `+ 0`, 雄性则 `+ 389.892`, 这就相当于是以雌性为参照组，雄性的体重平均比雌性重 `389.982` 克，这就是该回归系数的解释。

原多分类变量 `species`（3 分类），变成虚拟变量去掉冗余列 `speciesAdelie` 后剩下 2 列，若种类是 `Adelie`，这 2 列均为 0，即回归模型不包含这 2 项，此时是参照组；若种类是任一非参照的种类，比如 `Gentoo`，则 `speciesGentoo = 1`，此时回归模型多了一项：

$$\text{body\_mass} = \cdots + 1014.627 * 1 + \cdots$$

这就相当于是以 `Adelie` 为参照组，`Gentoo` 组相对于参照组 `Adelie` 平均体重要重 `1014.627` 克。

分类变量用于回归模型，所起的作用就是分组之间做比较，也只能是起分组比较的作用。这实际上也等效于分别对各分组建立线性回归模型，再做比较。

切记：分类变量用于建模时，始终是起分类的作用，绝对不能因为表示为数值形式，就直接当数值使用。

5. 模型改进

自变量又称为特征，利用原有自变量构造新的自变量，就是特征工程。特征工程是改进模型的重要手段，也是数据挖掘/机器学习中的关键步骤。

多元线性回归相当于是用 1 次多项式去逼近真实的函数关系，如果提高的 2 次，即把所有二次项包括交互项[4]：$x_1^2, x_2^2, x_1x_2, \cdots$ 都加入模型，拟合效果大概率会有提升。但新加入的项，可能会有不显著或产生共线性。解决办法，就是用逐步回归进行变量筛选。

常用的构建特征方法还有：对特征做各种变换，连续特征离散化，比如年龄相差 1 岁影响不一定显著，但年龄段的差异，比如从青年到中年到老年，很可能会显著。

---

[4]关于交互项 x1:x2 的解释：x1 对 y 的影响受 x2 的调节，反之亦同，其回归系数相当于 y 对 x1 和 x2 的二阶偏导.

将三个数值变量的二次项，以及交互项 `sex:island` 加入模型，再接逐步回归剔除不显著项：

```
mdl2 = lm(body_mass ~ species + sex * island + bill_length
          + I(bill_length^2) + bill_depth + I(bill_depth^2)
          + flipper_length + I(flipper_length^2),
          penguins) %>%
  step(direction = "backward", trace = 0)
```

```
model_summary(list(mdl0, mdl1, mdl2),
              file = " 线性回归汇总结果.docx")
#> √ Table saved to 'C:/Users/zhjx_/Desktop/2021-22-2 学期/统
```

Table X. Regression Models.

	(1) body_mass	(2) body_mass	(3) body_mass
(Intercept)	−1500.029**	−1460.995*	1089.742*
	(575.822)	(571.308)	(429.571)
speciesChinstrap	−260.306**	−251.477**	−1252.954***
	(88.551)	(81.079)	(129.436)
speciesGentoo	987.761***	1014.627***	
	(137.238)	(129.561)	
islandDream	−13.103		94.231
	(58.541)		(67.438)
islandTorgersen	−48.064		17.315
	(60.922)		(76.168)
bill_length	18.189*	18.204*	18.103*
	(7.136)	(7.106)	(7.083)
bill_depth	67.575***	67.218***	68.820***
	(19.821)	(19.742)	(19.670)
flipper_length	16.239***	15.950***	
	(2.939)	(2.910)	
sexmale	387.224***	389.892***	481.063***
	(48.138)	(47.848)	(57.192)
speciesAdelie			−997.958***
			(136.468)
flipper_length^2			0.039***
			(0.007)
sexmale:islandDream			−209.140**
			(68.104)
sexmale:islandTorgersen			−124.313
			(94.746)
R^2	0.875	0.875	0.879
Adj. R^2	0.872	0.873	0.875
Num. obs.	333	333	333

Note. Unstandardized regression coefficients are displayed, with standard errors in parentheses.
* $p < .05$. ** $p < .01$. *** $p < .001$.

二次项 `flipper_length^2` 项和交互项 `sexmale:islandDream` 都非常显著的，模型的修正 R^2 比 `mdl1` 稍有提高 (0.0028)。

这说明 `mdl2` 相比 `mdl1` 有所改进，但同时也增加了模型的复杂度（多了 4 项)。那么，接受哪个模型更好呢?

基本原则是在模型没有显著差异的情况下，优先选择更简单的模型。

可用似然比检验 `lmtest::lrtest()` 或方差分析 `anova()` 比较两个模型有无显著差异:

```
anova(mdl1, mdl2)
#> Analysis of Variance Table
#>
#> Model 1: body_mass ~ species + bill_length + bill_depth +
#>     sex
#> Model 2: body_mass ~ species + sex + island + bill_length
#>     I(flipper_length^2) + sex:island
#>   Res.Df      RSS Df Sum of Sq    F Pr(>F)
#> 1    326 26915647
#> 2    322 26000518  4    915128 2.83  0.025 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

检验 P 值 = 0.025 小于 0.05, 说明两个模型有显著差异，应该选择 mdl2.

6. 回归诊断

- **残差检验**

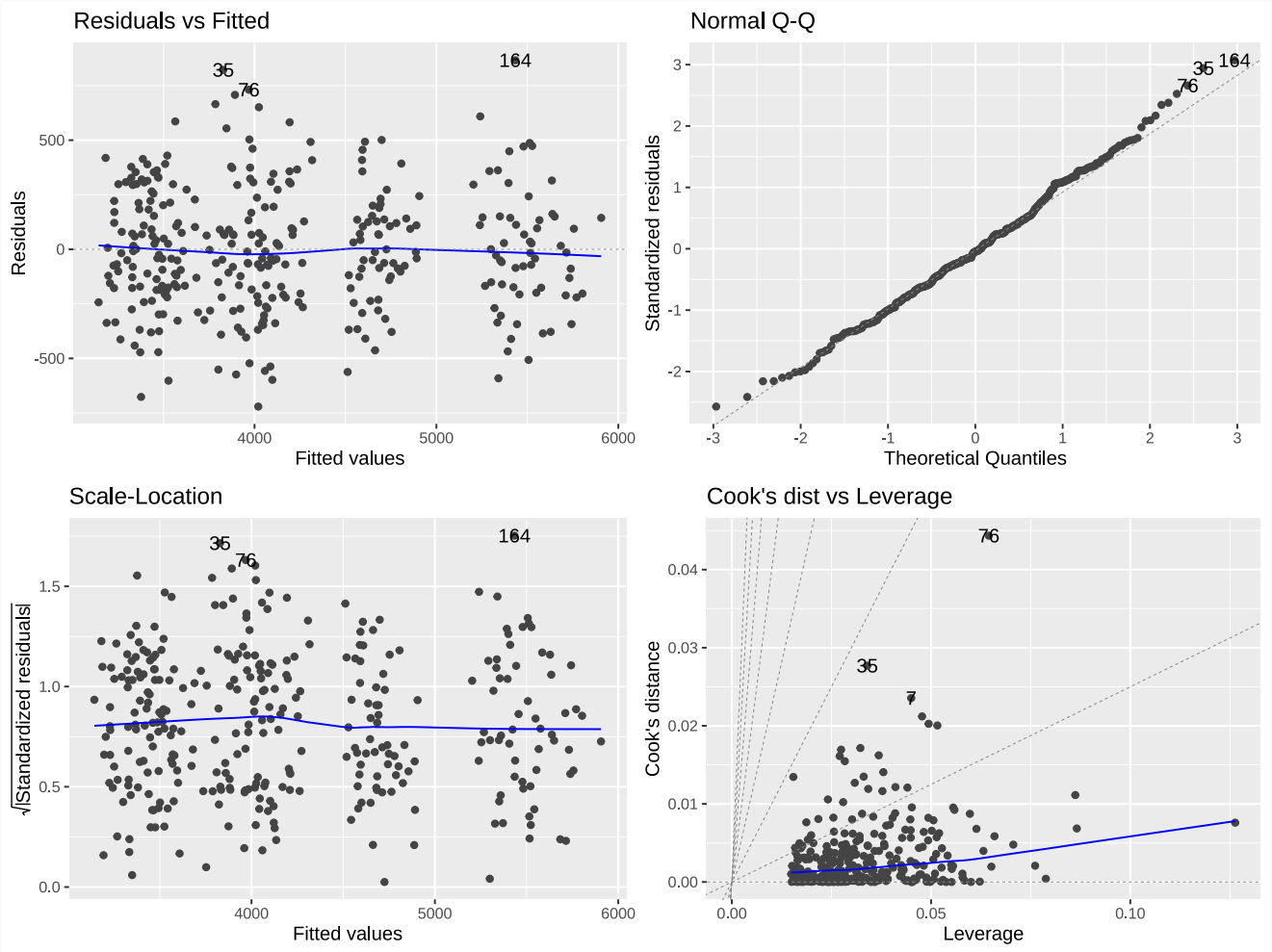
理想的模型（标准化）残差应服从"0 均值小方差"（标准）正态分布，对于残差，通常是绘制（标准化）残差图、残差 QQ 图、残差直方图，或者对（标准化）残差的正态性、独立性、异方差性做统计检验。

- **强影响分析**

对参数估计或预测值有异常影响的数据，称为强影响数据。回归模型应当具有一定的稳定性，若个别样本数据对估计有异常大的影响，剔除后将得到与原来差异很大的回归方程，从而有理由怀疑原回归方程是否真正描述了变量间的客观存在的关系。这些强影响样本是异常值，应当识别出来剔除之后，再重新拟合回归模型。

用 `ggfortify::autoplot()` 绘制回归诊断图，包括：残差图、残差 QQ 图、标准化残差图、强影响图等，还能同时标记强影响样本。

```
library(ggfortify)
autoplot(mdl2, which = c(1:3,6))   # 6 个图形可选
```



```
shapiro.test(mdl2$residuals)        # 残差正态性检验
#>
#>  Shapiro-Wilk normality test
#>
#> data:  mdl2$residuals
#> W = 1, p-value = 0.4
```

```
library(lmtest)
dwtest(mdl2)                                    # 残差独立性检验
#>
#>  Durbin-Watson test
#>
#> data:  mdl2
#> DW = 2, p-value = 0.9
#> alternative hypothesis: true autocorrelation is greater th
```

```
bptest(mdl2)                                    # 残差异方差检验
#>
#>  studentized Breusch-Pagan test
#>
#> data:  mdl2
#> BP = 15, df = 10, p-value = 0.1
```

可见，mdl2 能通过残差正态性、独立性检验。

回归模型预测

通过检验的回归模型，提供新的自变量数据框，用 predict() 就可以预测因变量值。

```
newdat = slice_sample(penguins[,-6], n = 5)
predict(mdl2, newdat, interval = "confidence")
#>    fit  lwr  upr
#> 1 4856 4779 4934
#> 2 3391 3291 3492
#> 3 3463 3331 3595
#> 4 4209 4098 4320
#> 5 4576 4494 4658
```

本篇主要参阅 (张敬信, 2022), 以及包文档，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

黄湘云 (2021). *Github: R-Markdown-Template*.