Лабораторная работа №2. Регрессионные модели. Вариант №8

Андрей Владимиров

дек 18 20

# РАЗДЕЛ 0. Выгрузка данных из файла

Первое, что необходимо сделать, это выгрузить данные из файла, который расположен в папке с проектом и имеет расширение RData (labs\_VladimirovAA\_data.Rdata).

# РАЗДЕЛ 1. Регрессионная модель для нелогарифмированных данных

#### Подготовка данных

С помощью корреляционной модели из прошлой лабораторной работы, а именно “Лабораторная работа №1”, мы узнали, что для нашей регрессионной модели подходят только следующие переменные: Корреляционная модель из прошлой лабораторной работы (лабораторная работа №1)

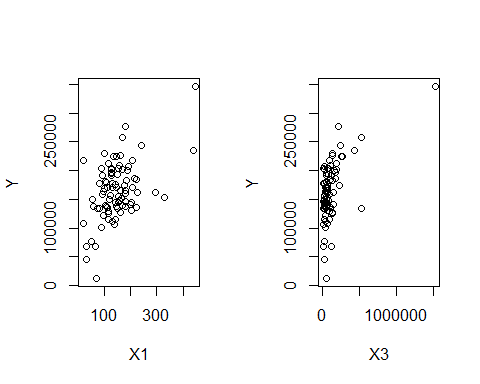
* Y - оборот розничной торговли на душу населения в 2016;
* X1 - число малых предприятий на 10000 человек населения в 2016;
* X3 - расходы консолидированных бюджетов субъектов Российской Федерации (всего) в 2015. Количество наблюдений: 85.

Модель: .

#### Множественная регрессионная модель

Построим модель множественной регрессии и выведем график зависимости Y от X1 и X3, который представлен ниже.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 123871.395 | 10374.740 | 11.940 | 0.000 |
| X1 | 190.306 | 73.873 | 2.576 | 0.012 |
| X3 | 0.108 | 0.029 | 3.685 | 0.000 |



Гипотеза : коэффициенты при X1 равен 0 в генеральной совокупности, то есть является незначимым; гипотеза :коэффициенты при X1 не равен 0 в генеральной совокупности, то есть является значимым; Проверка будет происходить за счет сравнения с P-значением при уровне значимости, равному 0.05.

*Проверка параметра X1*

P-значение при коэффициенте X1: . P-значение X1 меньше, чем уровень значимости, то есть принимается гипотеза .

*Проверка параметра X3*

P-значение при коэффициенте X3: . P-значение X3 меньше, чем уровень значимости, то есть принимается гипотеза .

Коэффицент детерминации 0.35, тое есть составляет 35%. Это означает, что 35% разброса переменной Y, которая является зависимой, объясняется переменными X1 и X3.

Вид модели после проверки параметров: .

*Вывод*: все коэффициенты значимы, означающие, что исключать регрессоры не требуется.

#### Модель с переменной структурой

Под моделью с переменной структурой подразумивается модель, в которой учитывается, что каждый регион относится к одному из федеральных округов. При этом включим включим фиктивные переменны в константы и коэффициенты.

Теперь постороим модель и выведем описательную статистику коэффициентов.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 114693.712 | 34140.965 | 3.359 | 0.001 |
| Federal\_districtПФО | -55921.145 | 50410.738 | -1.109 | 0.272 |
| Federal\_districtСЗФО | 77920.744 | 40686.174 | 1.915 | 0.060 |
| Federal\_districtСКФО | -49386.870 | 47538.418 | -1.039 | 0.303 |
| Federal\_districtСФО | -53191.183 | 41294.105 | -1.288 | 0.203 |
| Federal\_districtУФО | 26218.407 | 54801.420 | 0.478 | 0.634 |
| Federal\_districtЦФО | 85396.429 | 45785.033 | 1.865 | 0.067 |
| Federal\_districtЮФО | -44538.781 | 54743.638 | -0.814 | 0.419 |
| X1 | 242.358 | 231.599 | 1.046 | 0.299 |
| X3 | 0.440 | 0.174 | 2.525 | 0.014 |
| Federal\_districtПФО:X1 | 366.897 | 333.186 | 1.101 | 0.275 |
| Federal\_districtСЗФО:X1 | -441.828 | 272.384 | -1.622 | 0.110 |
| Federal\_districtСКФО:X1 | -115.513 | 563.216 | -0.205 | 0.838 |
| Federal\_districtСФО:X1 | 104.106 | 288.264 | 0.361 | 0.719 |
| Federal\_districtУФО:X1 | -398.317 | 381.000 | -1.045 | 0.300 |
| Federal\_districtЦФО:X1 | -521.360 | 316.588 | -1.647 | 0.105 |
| Federal\_districtЮФО:X1 | 294.841 | 415.793 | 0.709 | 0.481 |
| Federal\_districtПФО:X3 | -0.430 | 0.188 | -2.286 | 0.026 |
| Federal\_districtСЗФО:X3 | -0.153 | 0.217 | -0.705 | 0.483 |
| Federal\_districtСКФО:X3 | 0.597 | 0.415 | 1.437 | 0.156 |
| Federal\_districtСФО:X3 | -0.285 | 0.251 | -1.136 | 0.260 |
| Federal\_districtУФО:X3 | 0.056 | 0.273 | 0.206 | 0.838 |
| Federal\_districtЦФО:X3 | -0.260 | 0.180 | -1.440 | 0.155 |
| Federal\_districtЮФО:X3 | -0.239 | 0.230 | -1.040 | 0.302 |

Данная модель является незначима, в ней много незначимых параметров. Коэффициент детерминации 0.69, то есть 69%. Необходимо исключить данные параметры. Для этого мы будем использовать функция ‘removeFactorsByPValue’ (ссылка: <https://raw.githubusercontent.com/aksyuk/R-Practice-basics/master/user_functions/removeFactorsByPValue.R>).

Ниже представлена описательная статистика коэффициентов после исключения незначимых коэффициентов без учета P-значения:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 98516.435 | 11053.597 | 8.913 | 0 |
| Federal\_districtСЗФО | 90060.901 | 23711.931 | 3.798 | 0 |
| Federal\_districtСФО | -48164.148 | 10933.898 | -4.405 | 0 |
| Federal\_districtЦФО | 122541.048 | 29697.659 | 4.126 | 0 |
| X1 | 382.367 | 78.692 | 4.859 | 0 |
| X3 | 0.221 | 0.036 | 6.062 | 0 |
| Federal\_districtСЗФО.X1 | -527.733 | 124.539 | -4.237 | 0 |
| Federal\_districtЦФО.X1 | -827.039 | 186.963 | -4.424 | 0 |
| Federal\_districtПФО.X3 | -0.228 | 0.060 | -3.804 | 0 |

Теперь у данной модели все значимые коэффициенты, при этом коэффициент детерминации 0.61, то есть 61%.

Константы оказались значимы для СЗФО, СФо и ЦФО. Коэффициент X1 значимы для СЗФО, ЦФО. Коэффициент X3 значимы для ПФО.

#### Модель с поправкой Бонферрони:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 89722.856 | 11315.591 | 7.929 | 0 |
| Federal\_districtСЗФО | 99845.245 | 24378.531 | 4.096 | 0 |
| Federal\_districtЦФО | 139756.345 | 30693.203 | 4.553 | 0 |
| X1 | 414.879 | 82.870 | 5.006 | 0 |
| X3 | 0.237 | 0.038 | 6.278 | 0 |
| Federal\_districtСЗФО.X1 | -573.523 | 129.976 | -4.413 | 0 |
| Federal\_districtЦФО.X1 | -926.093 | 196.034 | -4.724 | 0 |
| Federal\_districtПФО.X3 | -0.229 | 0.062 | -3.692 | 0 |
| Federal\_districtСФО.X3 | -0.383 | 0.102 | -3.762 | 0 |

Как видно из описательной статистики, коэффициент X3 является значимым. Коэффициент детерминации 0.59, то есть 59%.

Явный вид модели: .

#### Сравнение моделей

В данном части будет произведено сравнение трех моделей, которые были описаны выше, а именно: модель множественной регрессии, модель с переменной структурой и модель с поправкой Бонферрони.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | corr\_coeff\_det | calc\_value\_F | standard\_err |
| multiple regression model | 0.337 | 22.376 | 40661.87 |
| variable structure model | 0.573 | 15.098 | 32634.08 |
| bondferroni amendment model | 0.548 | 13.745 | 33570.60 |

*Результаты сравнения*:

* Сравнение показало, что более предпочтительная вторая модель (модель с переменной структурой);
* Сравнение F статистик показало, что более предпочтительная третья модель (модель с поправкой Бонферрони);
* Сравнение стандартных ошибок показало, что более предпочтительная вторая модель (модель с переменной структурой);
* Среднее значение Y составляет 123115.9 (по итогам описательных статистик трех моделей)

Таким образом, наиболее предпочтительная модель - модель по федеральным округам без поправки Бенферрони.

*Конечный вид модели*:

# РАЗДЕЛ 2. Регрессионная модель для логарифмированных данных

#### Подготовка данных

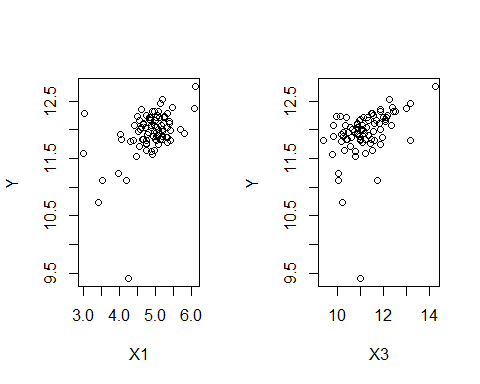
С помощью корреляционной модели из прошлой лабораторной работы, а именно “Лабораторная работа №1”, мы узнали, что для нашей регрессионной модели подходят только следующие переменные: Корреляционная модель из прошлой лабораторной работы (лабораторная работа №1)

* Y - оборот розничной торговли на душу населения в 2016;
* X1 - число малых предприятий на 10000 человек населения в 2016;
* X3 - расходы консолидированных бюджетов субъектов Российской Федерации (всего) в 2015. Количество наблюдений: 85.

Модель: .

#### Множественная регрессионная модель

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 9.305 | 0.542 | 17.174 | 0.000 |
| X1 | 0.280 | 0.081 | 3.462 | 0.001 |
| X3 | 0.114 | 0.050 | 2.274 | 0.026 |



Гипотеза : коэффициенты при X1 равен 0 в генеральной совокупности, то есть является незначимым; гипотеза :коэффициенты при X1 не равен 0 в генеральной совокупности, то есть является значимым; Проверка будет происходить за счет сравнения с P-значением при уровне значимости, равному 0.05.

*Проверка параметра X1*

P-значение при коэффициенте X1: . P-значение X1 меньше, чем уровень значимости, то есть принимается гипотеза .

*Проверка параметра X3*

P-значение при коэффициенте X3: . P-значение X3 меньше, чем уровень значимости, то есть принимается гипотеза .

Коэффицент детерминации 0.26, тое есть составляет 26%. Это означает, что 26% разброса переменной Y, которая является зависимой, объясняется переменными X1 и X3.

*Вывод*: проверка коэффициентов показала, что коэффициент при X3 незначим, поэтому он исключается.

Вид модели после проверки параметров: .

#### Модель с переменной структурой

Под моделью с переменной структурой подразумивается модель, в которой учитывается, что каждый регион относится к одному из федеральных округов. При этом включим включим фиктивные переменны в константы и коэффициенты.

Теперь постороим модель и выведем описательную статистику коэффициентов.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 10.185 | 1.497 | 6.802 | 0.000 |
| Federal\_districtПФО | -1.174 | 2.356 | -0.498 | 0.620 |
| Federal\_districtСЗФО | 2.061 | 1.882 | 1.095 | 0.277 |
| Federal\_districtСКФО | 1.128 | 1.755 | 0.643 | 0.523 |
| Federal\_districtСФО | -2.283 | 1.757 | -1.299 | 0.198 |
| Federal\_districtУФО | 0.897 | 2.578 | 0.348 | 0.729 |
| Federal\_districtЦФО | 0.212 | 1.947 | 0.109 | 0.914 |
| Federal\_districtЮФО | -1.491 | 2.458 | -0.606 | 0.546 |
| X1 | 0.394 | 0.302 | 1.305 | 0.196 |
| Federal\_districtПФО:X1 | 0.187 | 0.473 | 0.396 | 0.694 |
| Federal\_districtСЗФО:X1 | -0.422 | 0.374 | -1.128 | 0.263 |
| Federal\_districtСКФО:X1 | -0.302 | 0.387 | -0.778 | 0.439 |
| Federal\_districtСФО:X1 | 0.367 | 0.356 | 1.032 | 0.305 |
| Federal\_districtУФО:X1 | -0.177 | 0.517 | -0.343 | 0.732 |
| Federal\_districtЦФО:X1 | -0.058 | 0.389 | -0.148 | 0.882 |
| Federal\_districtЮФО:X1 | 0.284 | 0.511 | 0.555 | 0.581 |

Данная модель является незначима, в ней много незначимых параметров. Коэффициент детерминации 0.44, то есть 44%. Необходимо исключить данные параметры. Для этого мы будем использовать функция ‘removeFactorsByPValue’ (ссылка: <https://raw.githubusercontent.com/aksyuk/R-Practice-basics/master/user_functions/removeFactorsByPValue.R>).

Ниже представлена описательная статистика коэффициентов после исключения незначимых коэффициентов без учета P-значения:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 10.618 | 0.360 | 29.460 | 0.000 |
| Federal\_districtСФО | -2.716 | 0.970 | -2.799 | 0.006 |
| X1 | 0.285 | 0.074 | 3.870 | 0.000 |
| Federal\_districtСФО.X1 | 0.476 | 0.199 | 2.393 | 0.019 |

Теперь у данной модели все значимые коэффициенты, при этом коэффициент детерминации 0.37, то есть 37%.

Константы оказались значимы для СФО. Коэффициент X1 значимы для СФО.

#### Модель с поправкой Бонферрони:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 10.301 | 0.345 | 29.882 | 0 |
| Federal\_districtСФО | -0.408 | 0.108 | -3.775 | 0 |
| X1 | 0.350 | 0.070 | 4.976 | 0 |

Как видно из описательной статистики, коэффициент X1 является значимым. Коэффициент детерминации 0.33, то есть 33%.

Явный вид модели: .

#### Сравнение моделей

В данном части будет произведено сравнение трех моделей, которые были описаны выше, а именно: модель множественной регрессии, модель с переменной структурой и модель с поправкой Бонферрони.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | corr\_coeff\_det | calc\_value\_F | standard\_err |
| multiple regression model | 0.237 | 14.080 | 0.364 |
| variable structure model | 0.347 | 15.881 | 0.337 |
| bondferroni amendment model | 0.309 | 19.816 | 0.347 |

*Результаты сравнения*:

* Сравнение показало, что более предпочтительная вторая модель (модель с переменной структурой);
* Сравнение F статистик показало, что более предпочтительная третья модель (модель с поправкой Бонферрони);
* Сравнение стандартных ошибок показало, что более предпочтительная вторая модель (модель с переменной структурой); 2
* Среднее значение Y составляет 123115.9 (по итогам описательных статистик трех моделей)

Таким образом, наиболее предпочтительная модель - модель по федеральным округам без поправки Бенферрони.

*Конечный вид модели*: