

Описание работы

Владимир Димитров, Владислав Балабаев гр. 19701

28 сентября 2022 г.

1 Отчет по заданию

Первая часть задания направлена на создания новых переменных. В таблице 1 представлены новые переменные.

Переменная	Информация о переменной
empl	Создается из rj1. 1 - человек работает или находится в отпуске; 0 - человек не работает;
lwage	Создается из rj10 (берется логарифм);
lincome	Создается из rj60 (добавляется 1 к доходу и берется логарифм);
educ0	Фиктивная переменная из r_educ. 1 - не закончил школу; 0 - иначе;
educ1	Фиктивная переменная из r_educ. 1 - закончил школу и/или ПТУ/техникум; 0 - иначе;
educ2	Фиктивная переменная из r_educ. 1 - получил высшее образование; 0 - иначе;
lhourwage	Логарифм почасовой зарплаты;
lnonlabinc	Логарифм нетрудового дохода;
hours	Рабочий день в часах;
married	Фиктивная переменная из r_marst. 1 - состоит в браке; 0 - иначе;
kids18	Фиктивная переменная из rj72.173. 1 - имеет детей до 18; 0 - иначе;
female	Фиктивная переменная из rh5. 1 - женщина; 0 - иначе.

Таблица 1: новые переменные

В таблице 2 выведены описательный статистики.

	vars	n	mean	sd	min	max	range	se
empl	1	12001	0.75	0.43	0.00	1.000000e+00	1.000000e+00	0.00
lwage	2	8223	9.73	0.72	3.91	1.290000e+01	8.990000e+00	0.01
lincome	3	11579	8.57	3.13	0.00	1.463000e+01	1.463000e+01	0.03
educ0	4	12008	0.09	0.29	0.00	1.000000e+00	1.000000e+00	0.00
educ1	5	12008	0.63	0.48	0.00	1.000000e+00	1.000000e+00	0.00
educ2	6	12008	0.27	0.45	0.00	1.000000e+00	1.000000e+00	0.00
lhourwage	7	7538	4.60	0.70	-1.21	8.590000e+00	9.810000e+00	0.01
lnonlabinc	8	8197	2.08	3.72	0.00	1.460000e+01	1.460000e+01	0.04
hours	9	8666	9.47	3.76	1.00	2.400000e+01	2.300000e+01	0.04
married	10	12010	0.55	0.50	0.00	1.000000e+00	1.000000e+00	0.00
kids18	11	8626	0.65	0.48	0.00	1.000000e+00	1.000000e+00	0.01
female	12	12011	0.51	0.50	0.00	1.000000e+00	1.000000e+00	0.00

Таблица 2: описательные статистики

Была оценена модель предложения труда для женщин. В качестве зависимой переменной выступали часы работы, а независимой переменной были: логарифм почасовой зарплаты и нетрудового дохода, возраст, возраст в квадрате, фиктивные переменные для образования, семейное положение и наличие детей до 18 лет. Результаты представлены в таблице 3. На уровне значимости 5 процентов, значимыми фактами оказались: логарифм почасовой зарплаты, получение высшего образования и семейное положение.

```
Call:
lm(formula = hours ~ lhourwage + lnonlabinc + r_age + I(r_age^2) +
    educ1 + educ2 + married + kids18, data = df[df$rh5 == 2,
])

Residuals:
    Min       1Q   Median       3Q      Max
-7.380 -1.831 -0.743  0.792 16.841

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.4318288   1.4611403   7.824 7.05e-15 ***
lhourwage    -0.4883894   0.0997631  -4.895 1.03e-06 ***
lnonlabinc   -0.0177876   0.0165349  -1.076  0.282
r_age        0.0761836   0.0735814   1.035  0.301
I(r_age^2)   -0.0013046   0.0009442  -1.382  0.167
educ1        -0.0430476   0.3039799  -0.142  0.887
educ2       -1.2589046   0.3176616  -3.963 7.57e-05 ***
married     -0.5955883   0.1342864  -4.435 9.53e-06 ***
kids18      -0.2922923   0.1923652  -1.519  0.129
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.542 on 2991 degrees of freedom
(3170 пропущенных наблюдений удалены)
Multiple R-squared:  0.05546,    Adjusted R-squared:  0.05293
F-statistic: 21.95 on 8 and 2991 DF,  p-value: < 2.2e-16
```

Таблица 3: Наивная оценка регрессии

В результате последовательного удаления незначимых факторов, была получена регрессионная модель, где все переменные значимы на 5%-ом уровне. На уровне значимости 5 процентов модель оказалась значимой.

```
Call:
lm(formula = hours ~ lhourwage + educ1 + educ2 + married, data = df[df$rh5 ==
  2, ])

Residuals:
    Min       1Q   Median       3Q      Max
-7.6592 -1.7626 -0.6311  0.6826 16.9694

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.26440    0.45190   27.139  < 2e-16 ***
lhourwage    -0.49353    0.08747   -5.642 1.80e-08 ***
educ1        -0.33243    0.26408   -1.259  0.208
educ2        -1.48995    0.27428   -5.432 5.92e-08 ***
married      -0.49051    0.11333   -4.328 1.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.469 on 3769 degrees of freedom
(2396 пропущенных наблюдений удалены)
Multiple R-squared:  0.05187, Adjusted R-squared:  0.05086
F-statistic: 51.55 on 4 and 3769 DF, p-value: < 2.2e-16
```

Таблица 4: Оценка регрессии после удаление признаков

Рассмотрим модель, которая является устойчивой к гетероскедастичности.

```
Call: rlm(formula = hours ~ lhourwage + lnonlabinc + r_age + I(r_age^2) +
  educ1 + educ2 + married + kids18, data = df[df$rh5 == 2,
  ])
Residuals:
    Min       1Q   Median       3Q      Max
-6.5037 -1.0288 -0.2819  1.3046 16.6138

Coefficients:
              Value Std. Error t value
(Intercept)  10.5845  0.8537   12.3986
lhourwage    -0.2112  0.0583   -3.6240
lnonlabinc   -0.0419  0.0097   -4.3405
r_age         0.0108  0.0430    0.2508
I(r_age^2)   -0.0004  0.0006   -0.7347
educ1        -0.1894  0.1776   -1.0666
educ2        -0.8719  0.1856   -4.6976
married      -0.5172  0.0785   -6.5919
kids18       -0.1781  0.1124   -1.5845

Residual standard error: 1.612 on 2991 degrees of freedom
(3170 пропущенных наблюдений удалены)
```

Таблица 5: Наивная робустная модель

Многие факторы являются незначимыми. Проведем процедуру исключения незначимых факторов.

```
Call: rlm(formula = hours ~ lhourwage + lnonlabinc + I(r_age^2) + educ1 +
educ2 + married, data = df[df$rh5 == 2, ])
Residuals:
    Min       1Q   Median       3Q      Max
-6.6819 -0.9423 -0.2368  1.2843 16.5336

Coefficients:
            Value Std. Error t value
(Intercept) 10.3393  0.2586   39.9777
lhourwage   -0.2188  0.0490   -4.4680
lnonlabinc  -0.0334  0.0082   -4.0530
I(r_age^2)  -0.0001  0.0000   -2.2468
educ1       -0.2635  0.1494   -1.7645
educ2       -0.9168  0.1544   -5.9362
married     -0.4053  0.0640   -6.3314

Residual standard error: 1.484 on 3759 degrees of freedom
(2404 пропущенных наблюдений удалены)
-> |
```

Таблица 6: Робустная модель с значимыми признаками

Интерпретация полученных моделей:

- наличие высшего образования у женщины уменьшает продолжительность ее рабочего дня на 54 минуты.
- продолжительность рабочего дня женщины в среднем составляет 10 часов.
- при увеличении нетрудового дохода на 1% женщины работают на 0,0003 часа в день меньше.

Оцененный коэффициент эластичности равен -0,023. Следовательно, при росте платы на 1% количество рабочих часов в день снижается на 0,023%.

Для того чтобы оценки были состоятельны, нам необходимо учесть тех людей, которые не работают. Построим регрессионную модель, которая предсказывает значение заработной платы для неработающего населения.

```
> summary(model)

Call: rlm(formula = lhourwage ~ female + r_age + I(r_age^2) + educ1 +
educ2, data = df, na.action = na.exclude)
Residuals:
    Min       1Q   Median       3Q      Max
-5.5736099 -0.4072507  0.0006092  0.4026688  4.0517370

Coefficients:
            Value      Std. Error t value
(Intercept)  3.7938      0.1050   36.1242
female       -0.3020      0.0147  -20.4893
r_age        0.0428      0.0056   7.6869
I(r_age^2)   -0.0006      0.0001  -8.4076
educ1        0.1327      0.0292   4.5498
educ2        0.5392      0.0307  17.5812

Residual standard error: 0.6018 on 7531 degrees of freedom
(4474 пропущенных наблюдений удалены)

> |
```

Таблица 7: Модель для неработающего населения

Были созданы новые переменные, lhourwage_p (предсказанные значения заработной платы на основе работающей выборки), hours_p (0 для неработающего населения). В таблице 8 оценена модель устойчивая к гетероскедастичности.

```
Call: rlm(formula = hours_p ~ lhourwage_p + lnonlabinc + r_age + I(r_age^2) +
educ1 + educ2 + married + kids18 + 0, data = df[df$rh5 ==
2, ])
Residuals:
    Min       1Q   Median       3Q      Max
-7.2337 -1.1323 -0.1981  1.3132 16.8663

Coefficients:
            Value      Std. Error t value
lhourwage_p -0.0358      0.0588  -0.6081
lnonlabinc  -0.0296      0.0097  -3.0425
r_age       0.4970      0.0180  27.6375
I(r_age^2)  -0.0064      0.0003 -23.5925
educ1       0.2520      0.1770   1.4236
educ2      -0.6387      0.1868  -3.4190
married     -0.3989      0.0793  -5.0302
kids18     -0.1568      0.1155  -1.3574

Residual standard error: 1.704 on 3227 degrees of freedom
(2935 пропущенных наблюдений удалены)
```

Таблица 8: Предложения труда для женщин на полной выборке

Интерпретация полученных результатов:

- наличие высшего образования у женщины уменьшает продолжительность ее рабочего дня на 38 минут

- если женщина состоит в зарегистрированном браке, то рабочее время сокращается в среднем 24 минуты
- с возрастом женщины работают в среднем на полчаса в день больше

Коэффициент эластичности предложения труда женщин по заработной плате в полной выборке составил -0,01. Интерпретация такова: при росте почасовой заработной платы на 1% продолжительность рабочего дня снижается на 0,01%.

Как можно понять по данной работе, для более правдивых результатов необходимо использовать большую часть данных.

КОД