Влияние переезда в крупный город (SMSA central) на уровень заработной платы

Пункт 1.

Нам необходимо вывести таблицу с балансом средних значений в 4 группах для следующих показателей: размер семьи, способности, образование, образование родителей, уверенность в себе, размер фирмы, отношение к риску.

Сначала мы определили группу для каждого человека. Группа «Всегда в central» означает, что у человека показатель $SMSA\_central = 1$ всегда; группа «Никогда в central» - что показатель $SMSA\_central = 0$ всегда; группа «Переехал в central» - что показатель $SMSA\_central$ изменился с 0 на 1; группа «Уехал из central» - что показатель SMSA_central изменился с 1 на 0.

Далее мы убрали тех людей, у которых есть N/A по исследуемым показателям (пропуски в данных мешают при подсчете, лучше их убрать). Осталось 1176 человек из 1721, или 18816 наблюдений. Это достаточно много, поэтому мы продолжили исследование.

Для каждого из показателей мы сделали попарные сравнения между каждой парой групп. Получились такие результаты:

```
Results for: fam_size
              Group1             Group2  adj_p_value Balance
1   Всегда в central  Никогда в central 3.998594e-01     Yes
2   Всегда в central Переехал в central 1.261517e-09      No
3   Всегда в central   Уехал из central 3.998594e-01     Yes
4   Никогда в central Переехал в central 1.621400e-09      No
5   Никогда в central   Уехал из central 8.258914e-01    Yes
6   Переехал в central   Уехал из central 8.134700e-08    No


Results for: education
              Group1             Group2  adj_p_value Balance
1   Всегда в central  Никогда в central 7.204273e-03      No
2   Всегда в central Переехал в central 4.759304e-25      No
3   Всегда в central   Уехал из central 7.138557e-10      No
4   Никогда в central Переехал в central 1.160700e-35     No
5   Никогда в central   Уехал из central 2.913820e-32     No
6   Переехал в central   Уехал из central 4.086594e-12    No


Results for: AFQT2
              Group1             Group2  adj_p_value Balance
1   Всегда в central  Никогда в central 1.611225e-52      No
2   Всегда в central Переехал в central 8.409801e-31      No
3   Всегда в central   Уехал из central 1.316064e-17      No
4   Никогда в central Переехал в central 7.705698e-06     No
5   Никогда в central   Уехал из central 7.705698e-06     No
6   Переехал в central   Уехал из central 4.269526e-10    No


Results for: HGT_father
              Group1             Group2  adj_p_value Balance
1   Всегда в central  Никогда в central 0.0213912191      No
2   Всегда в central Переехал в central 0.0750449947     Yes
3   Всегда в central   Уехал из central 0.0750449947     Yes
4   Никогда в central Переехал в central 0.0009351207      No
5   Никогда в central   Уехал из central 0.8096883134    Yes
6   Переехал в central   Уехал из central 0.0020441966    No
```

Results for: HGT_mother

	Group1	Group2	adj_p_value	Balance
1	Всегда в central	Никогда в central	8.695476e-03	No
2	Всегда в central	Переехал в central	4.675911e-06	No
3	Всегда в central	Уехал из central	8.695476e-03	No
4	Никогда в central	Переехал в central	4.699051e-04	No
5	Никогда в central	Уехал из central	3.523734e-11	No
6	Переехал в central	Уехал из central	3.267400e-11	No

Results for: self_conf

	Group1	Group2	adj_p_value	Balance
1	Всегда в central	Никогда в central	2.279590e-17	No
2	Всегда в central	Переехал в central	7.069937e-01	Yes
3	Всегда в central	Уехал из central	4.865539e-03	No
4	Никогда в central	Переехал в central	1.821781e-07	No
5	Никогда в central	Уехал из central	1.454323e-34	No
6	Переехал в central	Уехал из central	1.414410e-01	Yes

Results for: size_of_firm

	Group1	Group2	adj_p_value	Balance
1	Всегда в central	Никогда в central	0.71639640	Yes
2	Всегда в central	Переехал в central	0.23484658	Yes
3	Всегда в central	Уехал из central	0.71639640	Yes
4	Никогда в central	Переехал в central	0.09505768	Yes
5	Никогда в central	Уехал из central	0.24013470	Yes
6	Переехал в central	Уехал из central	0.71639640	Yes

Results for: risk

	Group1	Group2	adj_p_value	Balance
1	Всегда в central	Никогда в central	2.881366e-01	Yes
2	Всегда в central	Переехал в central	1.771572e-25	No
3	Всегда в central	Уехал из central	9.741322e-05	No
4	Никогда в central	Переехал в central	3.175765e-28	No
5	Никогда в central	Уехал из central	4.395780e-05	No
6	Переехал в central	Уехал из central	1.165642e-14	No

Balance summary across all treatment pairs

Type	Max.Diff.Un	M.Threshold.Un	
Contin.	0.2860	Not Balanced, >0.1	education
Contin.	0.5911	Not Balanced, >0.1	AFQT2
Contin.	0.5413	Not Balanced, >0.1	
HGT_father	Contin.	0.1642	Not Balanced, >0.1
HGT_mother	Contin.	0.2985	Not Balanced, >0.1
self_conf	Contin.	0.2903	Not Balanced, >0.1
size_of_firm	Contin.	0.0959	Balanced, <0.1 risk
Binary	0.2257	Not Balanced, >0.1	

Balance tally for mean differences

count Balanced, <0.1	1
Not Balanced, >0.1	7

Variable with the greatest mean difference

Variable	Max.Diff.Un	M.Threshold.Un	
0.5911	Not Balanced, >0.1	education	

Sample sizes

	Всегда в central	Никогда в central	Переехал в central	Уехал из central
All	2032	13728	752	2304

Как мы заметили, при попарных сравнениях баланса нет ни в одной из пары групп по показателям education, AFQT2, HGT_mother. Отсутствие баланса в этих показателях может указывать на то, что принадлежность к той или иной группе тесно связана с различиями в образовании и умственных способностях, а также наследственными факторами.

По показателям risk, self_conf, HGT_father, fam_size баланс есть не более, чем по трем попарным сравнениям из шести. Это говорит о том, что представители разных групп хоть и могут быть схожи по степени уверенности в себе и отношению к риску, все же они различаются в значительной степени по оставшимся показателям.

И заметно выделился показатель size_of_firm с балансом по всем сравнениям. Это подтверждает и Balance Summary, где при пороговом значении 10% баланс наблюдается только у size_of_firm. Это может указывать на то, что размер фирмы распределен более равномерно по группам и фактор переезда в центральный город не оказывает значительного влияния на эту характеристику.

Например, если компании предлагают аналогичные возможности для роста и работы независимо от того, где они расположены, это способно привести к меньшим различиям в размерах фирм среди различных групп. Это связано с различными технологическими изменениями, позволившими компаниям функционировать более эффективно вне зависимости от местонахождения.

Пункт 2.

Мы сделали оценку эффекта переезда на зарплату с помощью парной регрессии, оставив только 2 группы: $SMSA\_central = 0$ всегда, т. е. и в 1979 г., и в 1994 г. (человек жил в маленьком городе и не переехал никогда – контрольная группа). И $SMSA\_central$ изменился с 0 в 1979 г. на 1 в 1994 г. (человек переехал в крупный город – treatment-группа).

Для этой регрессии мы создали новый dataframe, где для каждого человека взяли данные по всем характеристикам за 1979 г., а показатель заработной платы cpi_w за 1994 г.

Люди разделены на две группы, контрольную (1036 человек) и treatment (64 человека). Теперь с помощью парной регрессии мы оценили, как переезд человека в крупный город повлиял на его зарплату:

```
Call:
lm(formula = cpi_w ~ Treatment, data = data_2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1276.6  -451.3  -174.0   211.7  8437.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1090.00      24.22  45.008  < 2e-16 ***
Treatment     279.59     100.40   2.785  0.00545 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 779.5 on 1098 degrees of freedom
Multiple R-squared:  0.007013, Adjusted R-squared:  0.006109 
F-statistic: 7.755 on 1 and 1098 DF,  p-value: 0.005449
```

Итак, оценка эффекта переезда равна 279.59. Она означает, что переезд человека в central повышает его зарплату на 279.59 у. е. по сравнению с теми, кто остался жить в не central. Но эта оценка является смещенной и ей нельзя доверять!

Во-первых, treatment-группа и контрольная группа не являются сопоставимыми по всем характеристикам, которые могут влиять на заработную плату (например, уровень образования, навыки и т. д.). Баланс ковариат показал, то группы отличаются по важным признакам, это может привести к смещению оценки эффекта переезда.

Кроме того, в модели не учитываются важные переменные, которые могут влиять на зарплату, это тоже может привести к смещению результатов.

И, наконец, вероятность попасть в treatment-группу. Если люди, которые переезжали в крупные города, изначально имели более высокую зарплату или лучшие карьерные возможности, выбор в treatment-группу происходит на основе предшествующих характеристик, это также может вызвать смещение.

Пункт 3.

Чтобы исправить вышеупомянутые недостатки, мы сделали оценку эффекта переезда на зарплату мэтчингом на основе сопоставления ковариат. Ковариаты мы взяли из первого пункта (размер семьи, способности, образование, образование родителей, уверенность в себе, размер фирмы, отношение к риску).

Для начала мы удалили тех людей, у которых нет данных по интересующим нас ковариатам. В итоге осталось 850 человек. Далее, мы провели процедуру мэтчинга через поиск ближайшего соседа.

У нас осталось по 43 человека-«близнеца» с сопоставимыми характеристиками в контрольной и treatment группах. У этих наблюдений веса равняются 1, а у остальных наблюдений – 0.

Множественная регрессия по тем ковариатам, которые мы рассмотрели, и Treatment показала следующий результат:

```
Call:
lm(formula = cpi_w ~ Treatment + fam_size + AFQT2 + education +
HGT_mother + HGT_father + self_conf + size_of_firm + risk, data =
data_3, weights = match_1$weight)

Weighted Residuals:
   Min     1Q Median     3Q    Max
 -1065      0      0      0   1600

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.179e+02   5.754e+02   -0.726    0.4699
Treatment      6.306e+01   1.237e+02    0.510    0.6116
fam_size       7.616e-01   3.279e+01    0.023    0.9815
AFQT2          3.597e+00   2.707e+00    1.329    0.1879
education      6.117e+01   3.075e+01    1.990    0.0502
. HGT_mother    2.331e+01   3.122e+01    0.747    0.4575
HGT_father     1.393e+01   2.366e+01    0.589    0.5579
self_conf      1.594e+01   1.550e+01    1.028    0.3071
size_of_firm   2.717e-03   4.212e-03    0.645    0.5207
risk           -2.086e+02   1.298e+02   -1.607    0.1122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 555.7 on 76 degrees of freedom
```

```
Multiple R-squared:  0.2673, Adjusted R-squared:  0.1805
F-statistic:  3.08 on 9 and 76 DF,  p-value: 0.003344
```

Итак, оценка эффекта переезда не является статистически значимой. Она означает, что переезд человека в central статистически не повышает его зарплату по сравнению с теми, кто остался жить в не central. Значимым показателем в данном случае, влияющим на cpi_w, является только образование человека. В следующем пункте мы показали, почему и этим результатам не стоит доверять.

Пункт 4.

Мы вывели таблицу с балансом ковариатов после процедуры мэтчинга:

```
Balance Measures
                Type Diff.Adj        M.Threshold
fam_size     Contin.   0.1452 Not Balanced, >0.1
education    Contin.  -0.0630     Balanced, <0.1 AFQT2
Contin.  -0.0591     Balanced, <0.1
HGT_father   Contin.  -0.0110     Balanced, <0.1
HGT_mother   Contin.   0.0222     Balanced, <0.1
self_conf    Contin.  -0.1876 Not Balanced, >0.1
size_of_firm Contin.   0.2653 Not Balanced, >0.1 risk
Binary   0.0698     Balanced, <0.1

Balance tally for mean differences
count Balanced, <0.1         5
Not Balanced, >0.1     3

Variable with the greatest mean difference
Variable Diff.Adj        M.Threshold  size_of_firm
0.2653 Not Balanced, >0.1

Effective sample sizes
           Control Treatment
Unadjusted     807        43
Adjusted        43        43
```

При пороговом значении 10% у нас получились 5 сбалансированных переменных из 8, баланс явно улучшился (до мэтчинга было 1 сбалансированная ковариата из 8). По оставшимся 3 ковариатам наблюдается явный дисбаланс, что негативно сказывается на результатах. Общий баланс еще не достигнут.

Также здесь обязательно надо отметить, что после мэтчинга размер контрольной группы и группы обработки (treatment) стали одинаковым (по 43 наблюдения), но число наблюдений в контрольной группе значительно уменьшилось — с 1036 до 43. Это означает, что возможность экстраполировать результаты существенно снизилась (вплоть до того, что обобщать результаты нельзя). Нужны дальнейшие корректировки данных.

Пункт 5.

Если в прошлой процедуре мэтчинга у наблюдений были веса 0 и 1, что снижало точность модели, здесь мы сделали оценку эффекта переезда на заработную плату методом inverse probability weighting на основе propensity score.

```
Summary of weights
```

```
- Weight ranges:

          Min                                       Max treated
2.1770  |-------------------------| 66.6206 control 1.0014
||                          1.4851

- Units with the 5 most extreme weights by group:

             734     794     735      44     108
 treated 45.8819 46.2116 48.5273 58.2108 66.6206
411     596     764       8     847  control  1.3738
1.3969  1.4085  1.4294  1.4851

- Weight statistics:

        Coef of Var   MAD Entropy # Zeros treated
0.854 0.699   0.326       0 control       0.056
0.034   0.001       0

- Effective Sample Sizes:

           Control Treated Unweighted
807.     43.
Weighted    804.45   25.11
```

Для treatment группы веса варьируются от 2.18 до 66.62, что указывает на значительную для них вариацию. В контрольной группе веса варьируются гораздо меньше: от 1.00 до 1.49.

Это подтверждает коэффициент вариации (Coef of Var). Для группы treatment он равен 0.854, что говорит о неравномерности распределения. В группе control коэффициент вариации очень низкий (0.056) – веса равномерные. Средняя абсолютная разница (MAD) также значительно выше для группы treatment (0.699) по сравнению с контрольной (0.034). Отсутствие нулевых весов в обеих группах означает, что у всех наблюдений есть вес – ни одно из наблюдений не было исключено из анализа.

Эффективный размер выборки для контрольной группы с учетом весов (804.45) близок к ее е взвешенному размеру (807) – добавление весов не привело к значительным изменениям.

В treatment группе эффективный размер выборки значительно ниже (25.11 по сравнению с 43) из-за высокой дисперсии весов. Несмотря на то, что все наблюдения участвуют в анализе, небольшая их часть вносит значительный вклад в оценку эффектов. Это может снижать надежность результата.

В данном случае множественная регрессия по тем ковариатам, которые мы рассмотрели, и Treatment показала следующий результат:

```
Call:
lm(formula = cpi_w ~ Treatment + fam_size + AFQT2 + education +
HGT_mother + HGT_father + self_conf + size_of_firm + risk,      data
= data_3, weights = ipw_1$weight)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-6652.2  -426.3  -120.8   245.1  8637.0

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   -1.773e+02   1.776e+02   -0.998   0.31857
Treatment      1.880e+02  4.049e+01   4.642 4.00e-06 ***
fam_size         2.346e+00   1.009e+01    0.233   0.81614
AFQT2            1.546e+00   9.725e-01    1.590   0.11220
education      3.582e+01  1.265e+01   2.831  0.00475 **
HGT_mother     2.056e+01  1.035e+01    1.987  0.04720 *
HGT_father     3.474e+01  8.221e+00   4.226 2.64e-05 ***
self_conf      1.042e+01  4.991e+00    2.087  0.03719 *
size_of_firm  -7.058e-04   1.043e-03   -0.677   0.49890
risk          -1.219e+02  4.185e+01  -2.912  0.00369 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 823.8 on 840 degrees of freedom
Multiple R-squared:  0.1847, Adjusted R-squared:  0.1759  F-
statistic: 21.14 on 9 and 840 DF,  p-value: < 2.2e-16
```

Некоторые переменные, не только Treatment, но и education, HGT_mother, HGT_father, self_conf и risk имеют статистически значимое влияние на переменную cpi_w. А вот AFQT2, fam_size и size_of_firm, не показали значимого влияния.

Это может указывать на то, что факторы, влияющие на cpi_w, не обязательно коррелируют с размером семьи или размером компании. А все остальное важно: переезд в central, образование человека и его родителей, уверенность в себе и склонность рисковать и пробовать что-то новое.

Несмотря на значимость некоторых переменных, R-квадрат (0.1847) показывает, что модель объясняет лишь небольшую часть вариации зависимой переменной. Это может указывать на то, что есть другие факторы, не учтённые в данной модели, которые также могут влиять на cpi_w.

Итак, оценка эффекта переезда является статистически значимой. Она означает, что переезд человека в central статистически повышает его зарплату на 188 у. е. по сравнению с теми, кто остался жить в не central.

Пункт 6.

В данном пункте надо было применить метод Double Lasso.

Этот метод помогает отобрать те контрольные переменные, которые вносят наибольший вклад в объяснение зависимой переменной.

Поэтому мы взяли используемые ранее 8 переменных ("fam_size", "education", 'AFQT2', 'HGT_father', 'HGT_mother', "self_conf", "size_of_firm") и решили применить метод LASSO, привлекая их.

При помощи функции rlassoEffect, предварительно создав матрицу коварит, вектор из значений зависимой переменной cpi_w и вектор тритмента, мы оценили модель и вот что получили:

```
[1] "Estimates and significance testing of the effect of target variables"
     Estimate. Std. Error t value Pr(>|t|) [1,]
115.4      102.4   1.126     0.26
```

Значение коэффициента при переменной Treatment получилось равным 115.4, однако сам результат, как видим, не значим.

При этом модель отобрала ровно три регрессора:

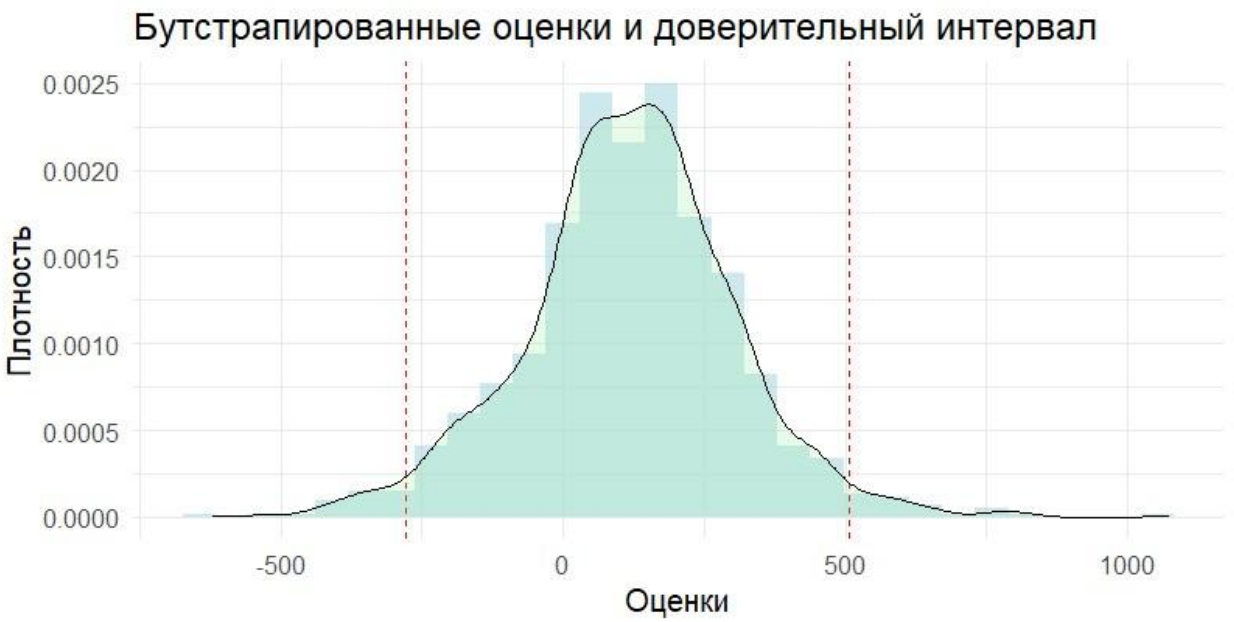
```
(Intercept)     fam_size    education         AFQT2    HGT_father    HGT_mother
self_conf size_of_firm         risk
      FALSE         FALSE         TRUE          TRUE          TRUE         FALSE
FALSE         FALSE         FALSE
```

Модель оставила AFQT2 - оценку за тест на навыки и умственные способности, HGT_father, HGT_mother - высшее образование родителей.

Оценка от переезда в большой город в данном случае ниже, чем при мэтчинге, и несостоятельна. Это можно объяснить тем, что в Double Lasso просто отбираются самые значимые переменные, но не учитывается эндогенность воздействия и эффект самоотбора.

Пункт 7.

Мы построили бутстраповский доверительный интервал для оценки из 5 пункта. Для этого мы сделали 1000 бутстраповских подвыборок с повторениями. Получился такой график:



Мы получили 1000 бутстрапированных оценок для коэффициента перед Treatment. Эти оценки показывают, как варьируется эффект переменной "Treatment" на зависимую переменную "cpi_w" при различных подвыборках из нашего исходного набора данных.

Доверительный интервал, представленный красными пунктирными линиями на графике, показывает диапазон значений, в котором с уровнем значимости 95%) находится истинное значение коэффициента. Интервал достаточно широкий, и включает 0, это может указывать на то, что эффект от Treatment статистически не значим.