**Data Science in Healthcare**

**Problem Set #3**

이름: 김연주

학과: 통계학과

학번: 2021250461

**PART A: Conceptual Questions**

**1. We understand that causation requires an association between two variables, but an association alone does not imply causation. Explain the difference between causation and association, using relevant equations discussed in the lecture.**

연관 (association)은 두 변수 간의 관계를 의미하며, 한 변수의 변화와 다른 변수의 변화가 동시에 발생하는 경향을 나타낸다. 이를 수식으로 나타내면 E(Y∣A)!=E(Y)이며, 이는 A의 값이 달라짐에 따라 Y의 기대값이 달라질 수 있음을 보여준다. 반면, 인과 (causation)는 한 변수의 변화가 다른 변수의 변화를 직접적으로 초래함을 의미한다. 즉, 인과적 효과는 평균처리효과 (Average Treatment Effect)로 정의되며 ATE=E(Y1​−Y0​)이다. 이때 Y1, Y0는 각각 treatment가 있을 때와 없을 때의 결과를 나타낸다. 즉, 연관은 두 변수 간의 관계만을 보여주며 인과를 보장하지 않는다. 인과를 추론하기 위해서는 RCT (모든 혼란 변수를 통제하여 A의 변화가 Y에 미치는 순수한 인과 효과를 확인하는 무작위 대조군 실험)과 수식으로 편향을 보정한 인과적 추정이 필요하다. 후자의 경우에는 E(Y1​−Y0​)=E(Y∣A=1)−E(Y∣A=0)+Bias terms로 나타내어 지며, 이때 bias term은 혼란 변수의 영향을 반영하여 인과 추론애서 조정되어야 하는 항이다.

**2. Give one reason a randomized experiment would be preferred to an observational study.**

무작위 실험이 관찰 연구보다 선호되는 이유 중 하나는 선택 편향을 제거할 수 있기 때문이다. 무작위 실험에서는 대상자를 무작위로 배정하여 실험군과 대조군이 치료 이외의 다른 모든 특성에서 유사하도록 만든다. 이렇게 하면 관찰 연구에서 발생할 수 있는 confounder의 영향을 최소화할 수 있다, 예를 들어 COVID-19 백신 효과를 연구한다고 가정할 때, 관찰 연구에서는 백신 접종 여부가 개인의 건강 상태, 연령, 소득 수준 등과 연관될 수 있어 결과에 편향을 초래할 가능성이 있다. 그러나 무작위 실험에서는 이러한 요인이 두 그룹에 고르게 분포되도록 설계되어 더 정확한 인과적 추론이 가능하다. 관찰 연구에서는 무작위 배정이 없기 때문에 실험군과 대조군 간에 혼란 변수가 존재할 가능성이 높아, 결과를 단순한 연관으로 오도할 위험이 존재하지만, 무작위 실험은 해당 위험을 최소화하는 것이다.

**3. Give two reasons an observational study might be preferred to a randomized experiment.**

관찰 연구가 무작위 실험보다 선호될 수 있는 이유는 다음과 같다.

첫번째로 윤리적 문제를 해결할 수 있다. 무작위 실험에서는 특정 집단에 치료를 의도적으로 제공하지 않거나, 비효율적인 치료를 제공할 가능성이 존재하므로 윤리적 문제가 존재한다. 예를 들어 생명을 위협하는 질병 치료 연구에서 치료를 무작위로 배제하는 것은 비윤리적인 행위인 것이다.

두번째로 현실적인 제약이 덜하다. 무작위 실험은 실행에 시간이 많이 걸리고 비용이 높을 가능성이 있다. 관찰 연구는 이미 존재하는 데이터를 사용하므로, 시간과 비용 측면에서 훨씬 효율적이다. 예를 들어, 강의자료에 따르면 무작위 실험의 비용은 한 명당 수십만 달러에 달할 수 있다. 그러나 이와 달리 관찰 연구는 대규모 인구 데이터를 이용해 상대적으로 적은 비용으로 실험을 수행할 수 있다. 따라서 관찰 연구는 윤리적 문제를 피하고 비용 및 시간 측면에서 효율성을 제공하는 장점이 있어 무작위 실헝을 수 행할 수 없는 상황에서 중요한 대안으로 사용된다.

**4. Suppose that we estimate the effect of health insurance on hospitalization. When relying on a randomized experiment, we find that having health insurance increases hospitalization by 10 percentage points. When relying on an observational study, however, we find that health insurance is associated with a 26 percentage points increase in hospitalization. Describe how to explain the discrepancy in these two findings. If possible, explain potential sources of bias.**

무작위 실험과 관찰 연구에서 건강보험이 입원율에 미치는 영향에 대한 추정치가 다른 이유는 편향과 방법론적 차이에서 기인할 수 있다.

우선 혼란 변수의 존재를 이유로 들 수 있다. 관찰 연구에서는 무작위 실험과 달리 참가자들이 스스로 건강보험 여부를 선택하기 때문에 건강보험을 가진 집단과 그렇지 않은 집단 간에 차이가 있을 수 있다. 예를 들어 건강보험을 가진 사람들은 건강 상태가 더 나쁘거나 의료 서비스를 더 자주 필요로 하는 경향이 있을 수 있다. 이러한 차이를 조정하지 않으면 관찰 연구의 결과가 과대 추정될 수 있다.

두번째 이유로 선택 편향이 존재한다. 무작위 실험은 대상자를 임의로 배정하여 실험군과 대조군 간의 모든 요인이 동일해지도록 설정 가능하다. 그러나 관찰 연구는 선택 편향의 영향을 받을 수 있다. 예를 들어 건강보험이 없는 사람들은 경제적으로 취약한 상태일 가능성이 크며, 입원을 피하려는 경향이 있을 수 있다. 이러한 편향은 관찰 연구에서 건강보험이 입원율에 미치는 영향을 실제보다 과대평가할 수 있다.

세번째로 측정 편향이 존재한다. 관찰 연구에서는 데이터를 수집하는 방식이난 질적 차이로 인해 측정 편향이 발생할 수 있다. 예를 들어 무작위 실험에서는 입원율 데이터를 더 체계적으로 수집할 수 있지만, 관찰 연구에서는 환자 스스로 보고한 데이터나 병원 기록을 기반으로 하기 때문에 정확성이 내려갈 수 있는 것이다.

마지막으로 도구 변수의 사용 가능성이 있다. 무작위 실험에서는 인과적 추론을 위한 설계가 포함되므로, 건강보허밍 입원율에 미치는 순수한 인과적 효과를 측정한다. 반면, 관찰 연구는 상관관계만 보여줄 가능성이 높은 것이다.

따라서 두 연구 결과 간의 차이는 **혼란 변수**, **선택 편향**, 그리고 **측정 편향** 때문일 가능성이 높다. 무작위 실험에서는 건강보험이 입원율에 미치는 **순수한 인과적 효과**를 추정할 수 있다. 이는 무작위 배정을 통해 치료군과 대조군 간의 혼란 변수를 제거하기 때문이다. 반면, 관찰 연구는 개인이 건강보험 여부를 스스로 선택하는 과정에서 발생하는 **선택 편향**과 **혼란 변수**의 영향을 받기 쉬워 결과를 과대평가할 가능성이 있다. 그렇기에 관찰 연구의 결과(26%p)는 건강보험의 효과를 과대 추정한 것일 수 있으며, 무작위 실험의 결과(10%p)가 건강보험의 실제 효과를 더 정확히 반영할 가능성이 높다.

**5. When analyzing data from an observational study, give four methods we can make the observational study more similar to a randomized experiment.**

관찰 연구를 무작위 실험에 더 가깝게 만들기 위해 사용할 수 있는 네가지 방법은 다음과 같다.

1. Matching: 매칭은 처치군과 대조군을 비슷한 특성을 가진 사람들로 구성하여 혼란 변수의 영향을 최소화하는 방법이다. 예를 들어 나이, 성별, 소득 수준 등 관찰 가능한 변수들을 기준으로 두 집단을 일치시키는 것이다. 이러한 매칭을 통해 E(X1∣A=1)−E(X1|A=0) 와 같은 관찰 가능한 혼란 변수의 차이를 줄일 수 있다.
2. 성향 점수 매칭: 성향 점수 (propensity score)는 처치군에 속할 확률을 예측하는 점수로, 이를 이용해 유사한 성향 점수를 가진 처치군과 대조군을 매칭할 수 있다. 이렇게 하면 고차원의 혼란 변수 문제를 단일 점수로 단순화하여 효과적으로 조정 가능하다. 강의자료에 관찰 연구의 데이터에서 성향점수를 계산하여 매칭을 수행하는 방법이 설명되어 있다.
3. 차분-차분법 (difference-in-differences): DiD는 시간에 따른 변화를 이용하여 처치군과 대조군 간의 결과 차이를 비교하는 방법이다. 이 방법은 관찰 연구에서도 자연 실험 또는 정책 변화와 같은 상황에서 인과적 효과를 추정하는 데 유용하다. DiD는 평행 추세 가정 (parallel trends assumption)을 전제로 하며, 전후 결과 변화를 비교하여 혼란 요인을 통제한다.
4. 도구 변수 접근법 (instrumental variables): 도구 변수는 처치 변수와는 상관이 있지만 결과 변수에는 직접적인 영향을 미치지 않는 변수이다. 이를 이용하여 인과적 효과를 추정할 수 있다. 예를 들어 건강보험 연구에서 정책 변화나 지리적 요인 등을 도구 변수로 사용할 수 있는 것이다. 이는 관찰 연구에서 혼란 변수와 선택 편향을 해결하는 데 효과적이다.

**6. Read the following research brief entitled “The Health Insurance Experiment” and answer to the following questions.**

**a. Explain whether the RAND study relies on an observational study or not.**

RAND의 건강보험 실험은 관찰 연구가 아닌 무작위 실험에 해당한다. 연구 참가자 7700명을 다섯가지 건강보험 유형 (무료 플랜 및 서로 다른 수준의 본인 부담 플랜)중 하나에 무작위로 배정하여 실험을 설계했기 때문이다. 이를 통해 관찰 연구에서 흔히 발생하는 혼란 변수의 영향을 최소화하였다.

**b. Describe key findings from the Health Insurance Experiment.**

실험의 주요 발견은 다음과 같다.

1. 의료 서비스 이용 감소: 본인 부담 비용을 설정한 그룹은 의료 서비스 사용이 감소하였다. 연간 병원 방문 횟수는 20% 감소하였고 비용 분담률 (25%, 50% ,95%)에 따라 지출도 20%~30% 감소하였다.
2. 의료 서비스의 적합성과 품질: 본인 부담 비용이 효과적인 치료와 덜 효과적인 치료 모두에서 서비스 이용을 줄이는 데 동일한 영향을 미쳤다. 그러나 치료의 품질 (의료 서비스의 적합성)은 무료 플랜과 비용 분담 플랜 간에 유의미한 차이를 보이지 않았다.
3. 건강 결과에 대한 영향: 일반적으로 본인 부담 비용으로 인해 의료 서비스가 줄어든 것이 참가자들의 전반적인 건강에 악영향을 미치지 않았다. 예외적으로, 무료 플랜은 저소득층 및 만성 질환자에게서 고혈압, 치과 치료, 시력 개선과 같은 더 나은 건강 결과를 보여주었다.

**c. State whether these findings can be interpreted causally or not. If so, provide detailed explanations.**

위와 같은 실험 결과는 인과적으로 해석 가능하다. 그 이유는 다음과 같다.

1. Randomization: 참가자를 무작위로 배정하여 처치군과 대조군 간에 혼란 변수를 제거하였다. 이는 결과를 인과적으로 해석할 수 있는 강력한 근거를 제공한다.
2. 대규모 실험 설계: 7700명 이상을 포함한 대규모 실험으로, 결과의 외적 타당성이 높다.
3. 측정 방법의 체계성: RAND는 연구 초기에 신체검사를 실시하고, 연구 종료 시 동일한 절차를 적용하여 데이터를 수집해 결과의 신뢰도를 높였다.

**7. Suppose that a new policy was implemented to improve access to care for the elderly population (i.e., treatment group). We assume that the non-elderly population was not affected by the policy (i.e., control group). Then, we want to conduct difference-indifferences analysis to estimate the effect of the policy implementation on health care spending among the elderly population. Each number in Figure A measures health care spending in terms of US dollars. For example, the elderly population spent $50 on health care before the policy was implemented**

**a. Calculate the effect of the policy implementation on health care spending among the elderly population using the difference-in-differences framework (10 points).**

처치군: 정책 시행 전 $50, 정책 시행 후 $85

대조군: 정책 시행 전 $35, 정책 시행 후 $55

DiD 공식:

정책효과=(처치군의 사후 지출-처치군의 사전 지출)-(대조군의 사후 지출-대조군의 사전 지출)=35-20=15

* 정책 시행으로 인해 노인 인구의 의료비 지출이 $15 증가한 것으로 추정된다.

**b. What could be a potential problem if we conduct difference-in-differences analysis? Choose an answer (5 points).**

**i. There may be observed time-varying factors that affect the outcome**

**ii. The parallel trends assumption is unlikely to hold as pre-treatment trends are different**

**iii. The outcomes for the treatment group are initially higher than the outcomes for the control group**

**iv. The observed outcomes in the pre-period are different such that the estimates are confounded by this initial difference.**

두번째 선택지가 DiD 분석의 잠재적 문제로 뽑힐 수 있다. Figure A를 보면, 정책 시행 전 대조군과 처치군의 지출 증가 추세가 다를 수 있다. 대조군의 지출 변화는 완만한 증가를 보이는 반면, 처치군은 더 가파른 추세를 보이기 때문이다. 이는 평행 추세 가정이 충족되지 않을 가능성을 나타내며 분석 결과에 편향을 초래할 수 있다.

**8. A government introduced a new tax on sugary drinks in 2015 to combat obesity. How could the synthetic control method be used to estimate the causal effect of this policy? Outline the key steps.**

합성 통제법 (synthetic control method)은 정책 시행 전후의 데이터의 비교를 통해 특정 정책의 인과적 효과를 추정하는 데에 효과적이다. 분석의 주요 단계는 다음과 같다.

1. 분석 대상 정의

* 처치군: 세금이 도입된 국가 (또는 지역)
* 잠재적 대조군: 세금이 도입되지 않은 국가 (또는 지역)

1. 정책 전 기간 데이터 수집

* 결과 변수: 비만율, 당분 음료 소비량, 건강 관련 지표 등
* 공변량 (covatiates): 비만율에 영향을 미칠 수 있는 경제적, 사회적, 인구학적 요인 (예: GDP, 인구, 교육 수준 등)
* 데이터는 정책 시행 전후의 시계열 데이터로 구성되어야 한다.

1. 합성 대조군 구축: 합성 통제군은 대조군 중 여러 단위를 조합하여 정책 시행 이전에 처치군과 가장 유사한 경향을 보이는 가상의 대조군이다. 이러한 합성 통제군은 가중치 (w1, w2, …) 를 요구하며 다음과 같이 표현된다.

합성 통제군=w1​×대조군1​+w2​×대조군2​+…

1. 정책 효과 추정: 정책 시행 후 처치군과 합성 통제군 간 결과 변수의 차이를 비교한다. 정책효과=Y(처치군,사후)-Y(합성통제군,사후)로 나타낼 수 있으며 해당 값이 당분 음료 세금의 비만율 감소 효과를 추정한다.
2. 유효성 테스트: Placebo test 를 통해 세금이 도입되지 않은 대조군에 대해 동일한 분석을 수행하여 정책 효과가 우연이 아닌지 확인 가능하다. 또한, Sensitivity Aalysis를 통해 결과가 특정 가중치나 모델링 선택에 얼마나 민감한지 평가할 수 있다.

이러한 합성 통제법은 다수의 대조군을 결합해 정책 전 처치군과 유사한 비교 대상을 만들기 때문에 정책 효과의 신뢰도 높은 인과 추정을 가능하게 한다. 이 방법은 특히 하나의 처치군과 다수의 대조군을 포함하는 정책 평가에 유용하다.

**PART B: Analytical Questions**

**9. The purpose of this analysis is to evaluate the impact of mental illness on health care expenditures while accounting for demographic, socioeconomic, and health-related factors. For simplicity, we assume that no other factors influence health care expenditures beyond those specified below. While causal interpretation is typically limited, for the purposes of this analysis, we interpret the relationship between mental illness and health care expenditures as causal. Use the MEPS HC-233: 2021 Full Year Consolidated Data File for this analysis.**

**• Outcome variable: Total health care expenditure (TOTEXP21)**

**• Primary independent variable: Symptom of mental illness (identified using PHQ242 and K6SUM42)**

* **Based on prior research, individuals are defined as experiencing mental illness if their Patient Health Questionnaire (PHQ242) score is 3 or higher or their Kessler 6 Psychological Distress Scale (K6SUM42) score is 13 or higher**
* **Individuals with no missing data in both variables and with a PHQ242 score below 3 and a K6SUM42 score below 13 are defined as not experiencing mental illness**

**• Control variables: Age (AGE21X), Sex (SEX), Race/ethnicity (RACETHX), Health insurance (INSCOV21), Family income (POVCAT21), Diagnosis of heart disease (MIDX), Diagnosis of stroke (STRKDX), Diagnosis of hypertension (HIBPDX), Diagnosis of diabetes (DIABDX\_M18), Diagnosis of asthma (ASTHDX), Diagnosis of cancer (CANCERDX), Diagnosis of hyperlipidemia (CHOLDX)**

**Before the analysis, restrict the study population to adults aged 18 and older. If any variable contains missing values, exclude those observations from the analysis. Treat health care expenditures and age as continuous variables, while using all other variables as categorical in the analysis. For variables indicating specific diagnoses, recode them as '1' for diagnosed cases and '0' for non-diagnosed cases. Use all other variables as provided in the dataset. To ensure representative results, include the weight variable (PERWT21F) in the analysis.**

**a. Construct the dataset.**

코드 설명:

2021년 FYC 데이터를 d2021로 저장한 후 필요한 변수만을 선택한 d\_21을 생성하였다. 해당 데이터셋에 mental illness를 나타내는 변수 MTILL을 추가하였다. 이후 결측값을 제거하였고 질병 진단과 관련된 변수들의 경우 NO가 2로 기입되어 있었기에 이를 0으로 변경하였다. 또한, 연령 변수인 AGE가 18 이상인 경우만 데이터셋에 포함시켰다.

**b. We aim to estimate the impact of mental illness on health care expenditures. Assume that no other factors, beyond the demographic, socioeconomic, and health-related factors described above, influence health care expenditures. First, perform a Generalized Linear Model (GLM) analysis (using a gamma family with a log link) to estimate the difference in health care expenditures based on the presence or absence of mental illness.**

코드 설명:

종속변수가 TOTEXP21, 독립변수가 MTILL인 glm모델 fit1을 생성하였다. 이때 종속변수에 0이 포함되면 분석이 되지 않으므로 0을 매우 작은 값인 0.0001로 바꿔 분석하였다.

결과:

Fit1의 summary 결과 정신질환이 없는 경우 총 의료비의 로그값은 9.02856이다. 즉 정신질환이 없는 사람들의 평균 의료비는 exp(9.02856)=약 8327이다. 또한, 정신질환이 있는 경우 총 의료비의 로그값은 0.52892 증가한다. 즉 정신질환이 있는 사람들의 의료비가 exp(0.52892)=1.696배라는 것을 의미한다. MTILL변수의 회귀계수는 통계적으로 유의하였고 AIC값은 3013802417, residual deviance는 970737442인 것으로 나타났다. 또한, dispersion parameter값은 100387.9로 이는 감마 분포의 분산값을 나타낸다.

**c. Perform a logistic regression analysis to estimate propensity scores related to the presence or absence of mental illness. Compare the distribution of estimated propensity scores between the group with mental illness and the group without mental illness and then describe the results. Based on the findings, explain whether it is feasible to conduct a matching analysis using the propensity scores.**

코드 설명:

MTILL이 종속변수, 이외의 다른 변수들이 독립변수인 glm fit2를 생성하였다. 이를 바탕으로 추정된 propensity score 값을 prop\_score변수로 해당 데이터셋에 저장하였다. 이후 ggplot을 이용하여 정신질환의 여부에 따른 propensity score값을 비교하였다.

결과:

정신질환이 없는 그룹은 propensity score가 낮은 쪽에 집중되어 있고 분포가 0에 가깝다. 반면, 정신질환이 있는 그룹은 더 넓은 분포를 가지며 높은 propensity score를 가진 관측치도 존재한다. 두 그룹의 분포가 겹치는 부분이 존재하므로 매칭 분석이 가능함을 시사한다.

정신질환 여부에 따른 propensity score가 겹치는 부분이 존재하므로 매칭 분석이 가능하다. 겹치는 영역은 매칭 분석의 공통지원영역 (common support region)을 충족함을 의미한다. 따라서, propensity score를 사용해 정신질환이 있는 그룹과 없는 그룹의 관측치를 매칭할 수 있다. 다만, 분포가 겹치지 않는 부분도 존재하여 일부 관측치, 특히 높은 propensity score를 지닌 MTILL=1의 관측치에서는 매칭이 힘들다는 한계가 있다.

**d. Then, replicate the analysis described in 'b' using propensity scores. Specifically, estimate the impact of mental illness on health care expenditures using propensity scores based matching and calculate both the absolute and relative differences compared to the results from the GLM. Compare and analyze the results across these approaches.**

코드 설명:

MatchIt 패키지를 사용하여 propensity score 기반 매칭을 수행하였다. MTILL을 처리변수로 놓은 fit4를 생성하였고 가장 가까운 propensity score 값을 가진 처리군과 통제군을 매칭하였다. 매칭 결과를 확인하고 매칭된 데이터를 이용하여 glm을 통해 TOTEXP21을 종속변수로 하는 분석을 실행하였다. 이러한 fit\_psm은 독립변수인 정신질환 여부가 의료비에 미치는 영향을 분석한다. 이후 기존 glm과 매칭 데이터 기반 glm의 회귀계수를 비교하고 절대, 상태 차이를 계산하였다.

결과:

Fit4를 분석한 결과 매칭 전에는 std. Mean Diff.가 여러 공변량에서 큰 값을 지녔 (HIBPDX에서 0.2768, ASTHDX에서 0.2786등). 반면, 매칭 후에는 해당 값들이 거의 0에 가까워졌다. 또한, propensity score의 분포는 매칭 후 처리군과 통제군의 값이 거의 일치했다. 이는 매칭이 성공적이었음을 나타낸다.

기존 glm에서 MTILL의 회귀계수는 0.52892로 정신질환이 있는 경우 의료비가 약 69.6%높았다. 또한 모델 적합성을 나타내는 AIC값은 3013802417이었다. 반면, 매칭 기반 glm에서 MTILL의 회귀계수는 0.39565로, 정신질환이 있는 경우 의료비가 약 48.5% 높았다. 그리고 이때 AIC값은 529933093으로 감소하였다. 두 회귀계수의 절대차이는 0.13327, 상대차이는 25.2%로, 이는 propensity score 매칭 후 정신질환의 여부가 의료비에 미치는 영향의 추정치가 기존의 glm보다 약 25.2% 감소했음을 의미한다. 이는 매칭 전 처리군과 통제군 간의 차이를 보정한 결과이다.

그래프를 통해서도 매칭이 잘 이루어졌음을 확인할 수 있다. 매칭 후 propensity score의 분포가 가의 일치하였고, 매칭 전과 매칭 후의 결과를 비교한 그래프에서도 같은 결과를 확인할 수 있기 때문이다.

**10. The purpose of this analysis is to estimate the impact of the minimum wage policy introduced in New Jersey on full-time and part-time employment in the franchise industry. Specifically, on April 1, 1992, New Jersey's minimum wage was increased from $4.25 to $5.05 per hour.**

**• Outcome variable: Full time employment (empft), Part time employment (emppt)**

**• Control variables: Franchise brand (chain), Ownership type (co-owned)**

**To more precisely estimate the policy's effects, the population in the neighboring state of Pennsylvania, where the policy was not implemented, will be used as a control group. Use the file 'CK1994.dta' from Blackboard.**

**a. Construct the dataset.**

해당 데이터는 stata에서 작성된 .dta 파일이기 때문에 haven 패키지를 이용해 데이터를 읽었다. 이후 필요한 변수만을 선택하고 결측치를 제거한 data\_1을 생성하였다.

**b. Explain conceptually why the results from a Difference-in-Differences analysis can provide unbiased estimates of the impact of the minimum wage policy.**

Difference-in-Differences는 정책 개입 전후의 변화와 비교군 및 처리군 간의 차이를 분석하여 정책 효과를 추정하는 방법이다. 이러한 DiD는 정책 전후의 시간 변화를 처리군과 통제군 모두에서 관찰한다. 또한, 처리군과 통제군을 비교함에 있어서 통제군은 정책의 영향을 받지 않았으므로 처리군에서 정책의 효과를 추출하는 데 사용할 수 있다. 두 집단 간의 차이는 정책 되의 다른 요인들의 영향을 제거하는 데 기여한다. 그리고 DiD는 평행 추세를 가정하므로 정책이 없었을 경우 두 집단이 동일한 추세를 따랐을 것이라는 가정 하에 정책의 순수한 효과를 추정할 수 있다.

**c. Before conducting the Difference-in-Differences analysis, verify whether there were changes in hourly wages in New Jersey around the implementation of the minimum wage policy on April 1, 1992. Perform a visual comparison with the neighboring state of Pennsylvania, where the policy was not implemented.**

코드 해석:

결측값을 제거하고 필요한 변수만 선택한 wage\_data 데이터셋을 생성하였다. 이후 wage\_st의 분포를 상자그림으로 시각화하였다.

결과:

정책 시행 이전에는 뉴저지와 펜실베니아의 시간당 임금 분포가 비슷하며 중앙값과 IQR도 서로 유사했다. 이는 정책 시행 이전 두 주의 시간당 임금 수준이 비슷했음을 나타낸다. 정책 시행 아후에는 뉴저지의 시간당 임금 중앙값이 크게 상승하였지만 펜실베니아의 임금 분포에는 큰 변화가 없었다. 이는 정책이 시행되지 않은 통제군에서의 임금 수준에 변화가 없었음을 나타낸다.

**d. To make causal inferences from the results obtained from the Difference-in-Differences analysis, it is necessary to verify whether certain assumptions are satisfied. Can these assumptions be tested with the given data? If not, what additional data would be required?**

DiD 분석을 통해 인과관계를 추론하려면 평행 추세 가정, 외생적 개입의 부재, 정책 효과의 동일성 가정이 충족되어야 한다. 첫번째로 평행 추세 가정의 경우 현재 데이터로 검증이 불가능하다. time변수가 정책 전과 정책 후로만 존재하며 정책 시행 이전의 다중 시점 데이터가 없기 때문이다. 그에 따라 정책 전 두 그룹 간의 변화 추세를 비교할 수 없다. 정책 이전의 다중 시점 데이터를 확보하면 평행 추세 가정을 검증 가능하다. 두번째로 외생적 개입의 여부 또한 현재 데이터로 검증 불가능하다. 최저임금 정책 외 다른 정책적, 경제적 변화에 대한 정보가 필요하다. 세번째로 정책 효과의 동일성의 경우 chain, co\_owned 변수를 포함하여 각 지점에서 정책 효과 차이를 통제 가능하다. 그러나 최저임금을 실제로 준수했는지에 대한 정보가 부재하여 각 지점의 실제 임금 준수 여부를 나타내는 변수가 있다면 정확하게 검증이 가능할 것으로 예상된다.

**e. Perform a Difference-in-Differences analysis to estimate the impact of the minimum wage policy. As discussed in class, first examine the outcome variable values for the treatment and control groups before and after the policy implementation. Then, interpret and describe the results obtained through regression analysis**

코드 설명:

우선 데이터를 state, time 별로 그룹화하여 그룹별 풀타임과 파트타임 고용의 평균을 계산하였다. 이후 풀타임 고용과 파트타임 고용의 평균을 정책 전후로 비교하기 위해 시각화하였다. Empft, emppt를 각각 종속변수로 한 DiD 분석을 실행하였고 이때 time과 state의 상호작용항도 고려하였다.

결과:

정책 시행 이전 펜실베니아의 풀타임 고용 임금의 평균은 10.2, 파트타임 평균은 19.5, 뉴저지의 풀타임 평균은 7.72, 파트타임 평균은 18.7이었다. 정책 시행 이후에는 펜실베니아의 풀타임 고용 임금의 평균은 7.56, 파트타임 평균은 20.0, 뉴저지의 풀타임 평균은 8.43, 파트타임 평균은 18.3인 것으로 나타났다.

DiD 결과 정책 이후 풀타임 고용은 2.64 감소하였고, 뉴저지의 고용 수준이 펜실베이아에서보다 2.48 낮았다. 또한, 뉴저지에서 정책이 풀타임 고용에 약 3.34명 증가의 효과를 미쳤다. 세 회귀계수 모두 통계적으로 유의하였다. 파트타임 고용애서는 정책 이후 고용은 0.47 증가하였고, 뉴저지의 고용 수준이 펜실베이아에서보다 0.8 낮았다. 또한, 뉴저지에서 정책이 풀타임 고용에 약 0.85명 감소의 효과를 미쳤다. 그러나 세 회귀계수 모두 통계적으로 유의하지 않았다.