|  |
| --- |
| REPORT |
| **과제 2**  LOGIT & MDA |





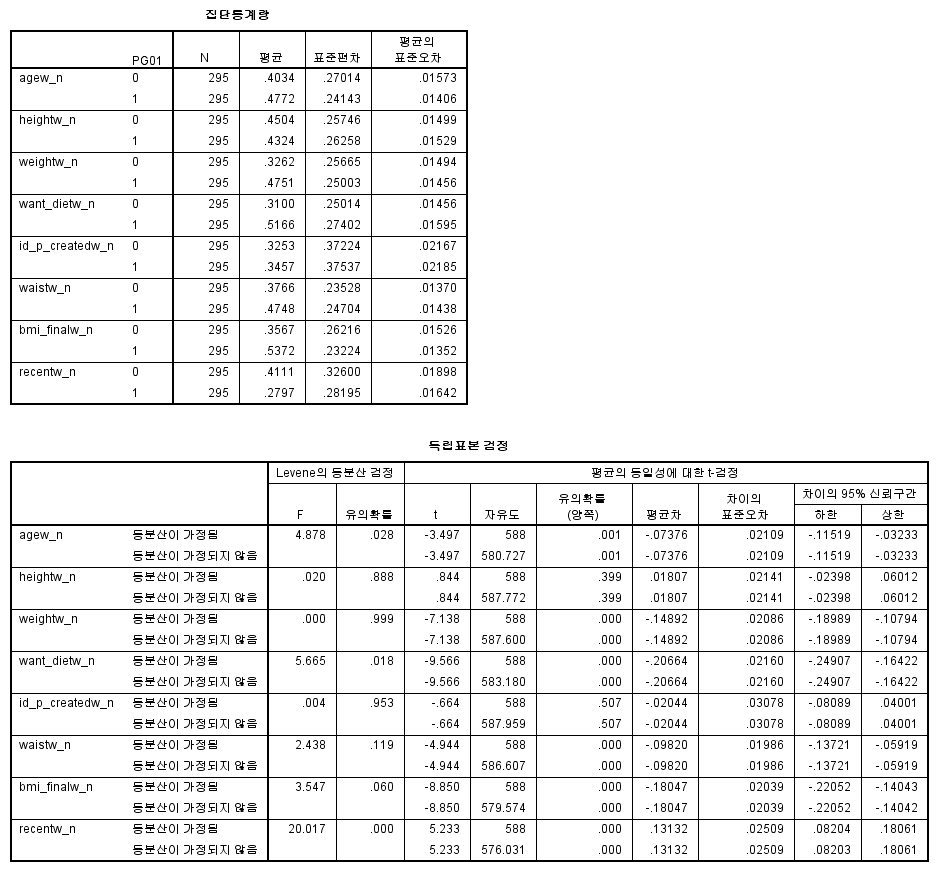
|  |  |  |
| --- | --- | --- |
| 제출일 | 2020. 04. 18 |  |
| 과목명 | 비즈니스애널리틱스개론 |  |
| 담당교수 | 안현철 |  |
| 전공 | 경영정보전공 |  |
| 학번 | 20175288 |  |
| 이름 | 최광삼 |  |

**목차**

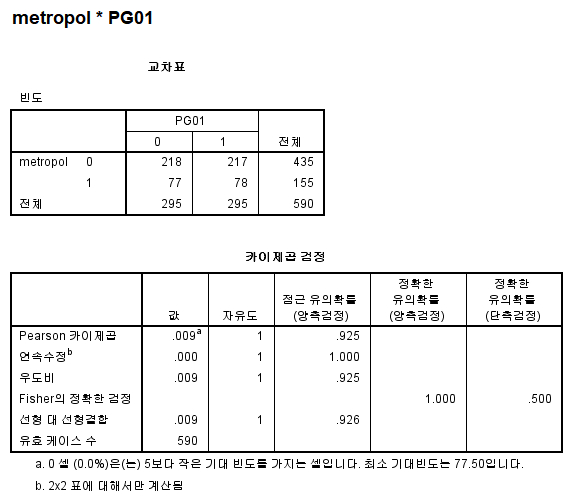
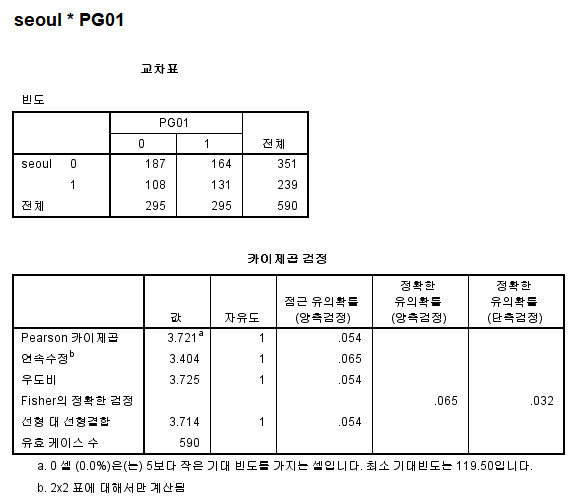
1. **LOGIT**
2. 독립표본t검정
3. 카이제곱검정
4. 로지스틱스 회귀분석
5. **MDA**
6. **LOGIT**

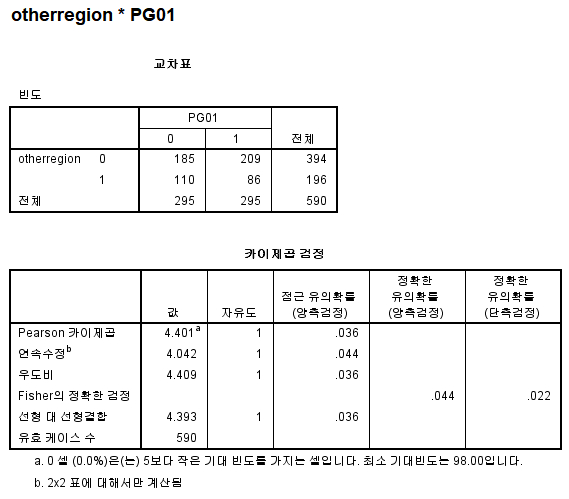
로지스틱스 회귀분석을 위해서 먼저 유의미한 변수를 필터링 할 필요가 있다. 이를 위해서 먼저 데이터 전처리 과정을 거친 SPSS 파일 ‘online\_diet.sav’를 사용하여 각각의 변수에 대해 전략제품 구매여부에 따른 독립표본t검정과 카이제곱검정을 실시하였다. 연속형 변수의 경우에는 독립표본t검정을 적용하였으며, 이분형 변수에는 카이제곱검정을 적용하였다.

1. 독립표본t검정

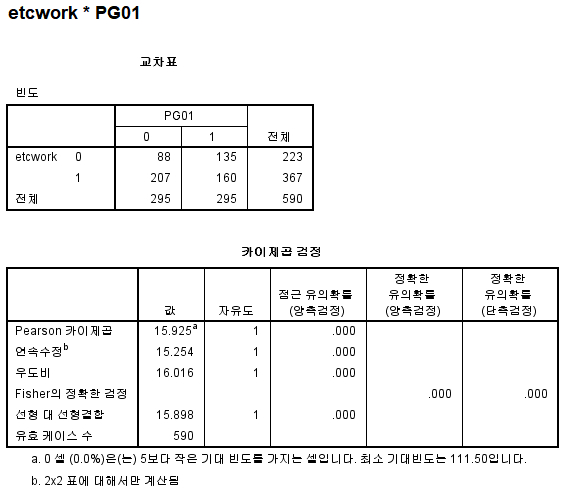
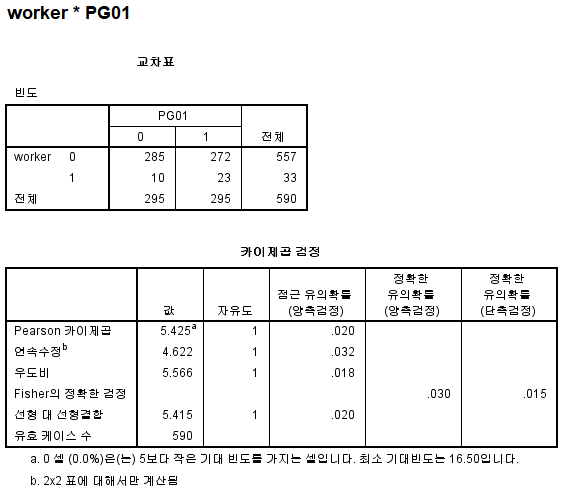
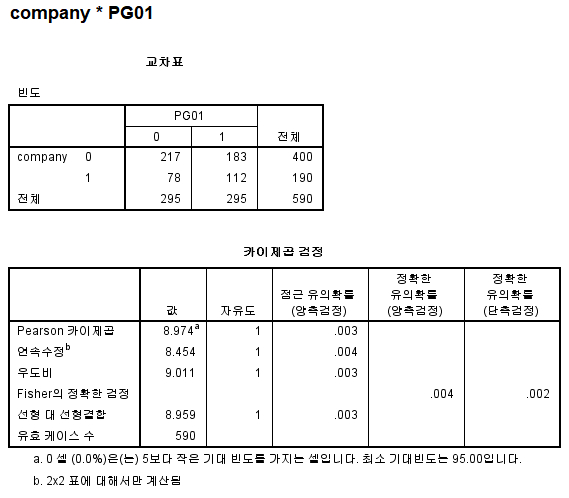


1. **agew\_n**(나이)의 경우, Levene 등분산 검정에서 유의확률이 0.028로 95% 신뢰수준 하에서 유의하며 귀무가설이 기각된다. 따라서 등분산이 가정되지 않았을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.001로 95% 신뢰수준 하에서 유의하므로 전략제품을 구매한 집단과 사지 않은 집단 사이에 유의미한 차이가 있음을 알 수 있다.
2. height\_n(키)의 경우, Levene 등분산 검정에서 유의확률이 0.888로 귀무가설이 기각되지 않는다. 따라서 등분산이 가정됐을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.399로 역시 95% 신뢰수준 하에서 귀무가설이 기각되지 않으므로 키에 대한 변수는 로지스틱스 회귀분석에 적합하지 않다고 볼 수 있다.
3. **weightw\_n**(몸무게)의 경우, Levene 등분산 검정에서 유의확률이 0.999로 귀무가설이 기각되지 않는다. 따라서 등분산이 가정됐을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.000으로 95% 신뢰수준 하에서 유의미하다. 이로써 몸무게와 전략제품 구매여부는 유의미한 상관관계가 있음을 알 수 있다.
4. **want\_dietw\_n**(감량목표)의 경우, Levene 등분산 검정에서 유의확률이 0.018로 95% 신뢰수준 하에서 귀무가설이 기각된다. 따라서 등분산이 가정되지 않았을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.000으로 95% 신뢰수준 하에서 유의미하므로 감량목표에 따라 전략제품을 구매한 집단과 그렇지 않은 집단 사이에 유의미한 차이가 있음을 알 수 있다.
5. id\_p\_createdw\_n(아이디 생성일로부터 구매일)의 경우, Levene 등분산 검정에서 유의확률이 0.953으로 귀무가설이 기각되지 않는다. 따라서 등분산이 가정됐을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.507로 역시 95% 신뢰수준 하에서 귀무가설이 기각되지 않으므로 로지스틱 회귀분석에 적합하지 않은 변수라고 할 수 있다.
6. **waistw\_n**(허리둘레)의 경우, Levene 등분산 검정에서 유의확률이 0.119로 귀무가설이 기각되지 않는다. 따라서 등분산이 가정됐을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.000으로 95% 신뢰수준 하에서 유의미하다. 이로써 허리둘레와 전략제품 구매여부에 따른 유의미한 차이가 있음을 알 수 있다.
7. **bmi\_finalw\_n**(BMI지수)의 경우, Levene 등분산 검정에서 유의확률이 0.060으로 95% 신뢰수준 하에서 아슬아슬하게 귀무가설이 기각되지 않는다. 따라서 등분산이 가정됐을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.000으로 95% 신뢰수준 하에서 유의미하므로 BMI지수와 전략제품 구매여부 사이에는 유의미한 관련이 있음을 알 수 있다.
8. **recentw\_n**(최근 구매일)의 경우, Levene 등분산 검정에서 유의확률이 0.000으로 95% 신뢰수준 하에서 귀무가설이 기각된다. 따라서 등분산이 가정되지 않았을 경우의 유의확률을 살펴보아야 한다. 이때의 유의확률은 0.000으로 역시 95% 신뢰수준 하에서 유의미하므로 최근 구매일과 전략제품을 구매 여부에 따른 의미한 차이가 있음을 알 수 있다.
9. 카이제곱검정

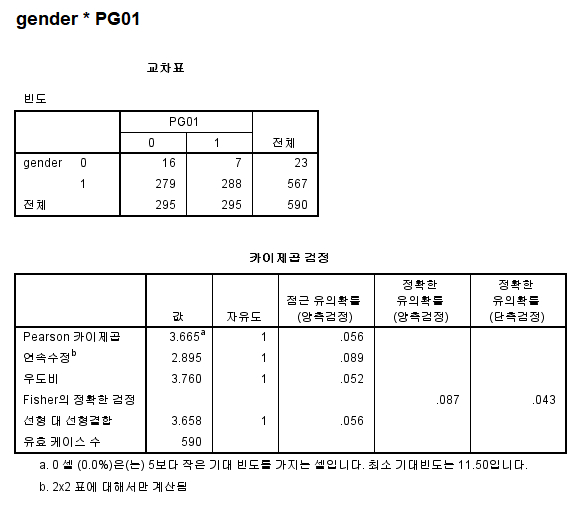




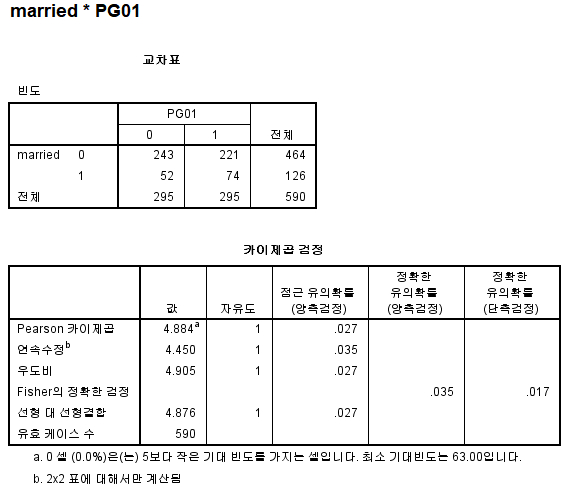
1. seoul: 서울 거주여부와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.054%이므로 95% 신뢰수준 하에서 유의미한 차이가 없음을 알 수 있다.
2. metropol: 수도권 거주여부와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.925%로 유의미한 차이가 확연히 없음을 알 수 있다.
3. **otherregion**: 기타지역 거주여부와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.036%로 95% 신뢰수준 하에서 관련이 있음을 알 수 있다.



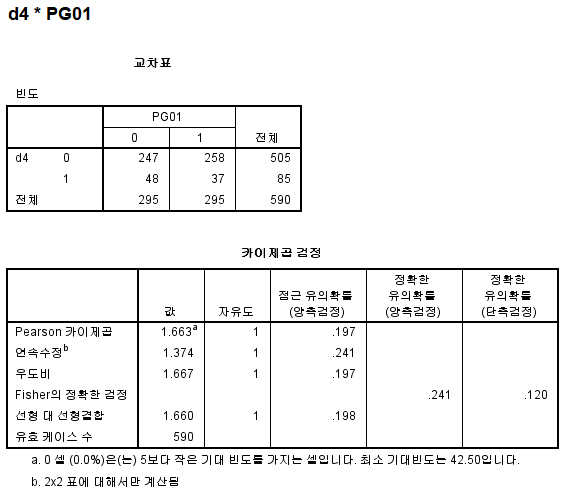
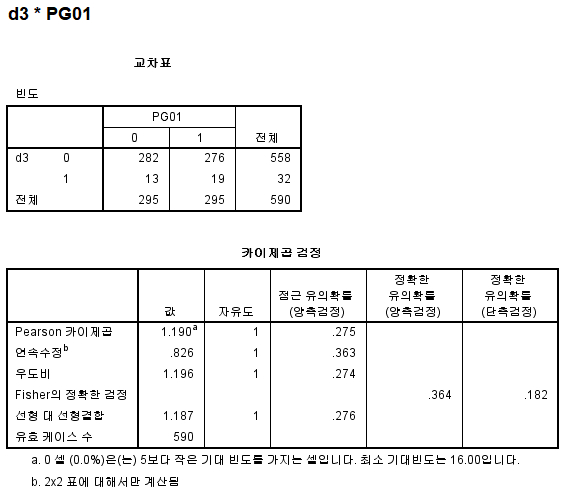
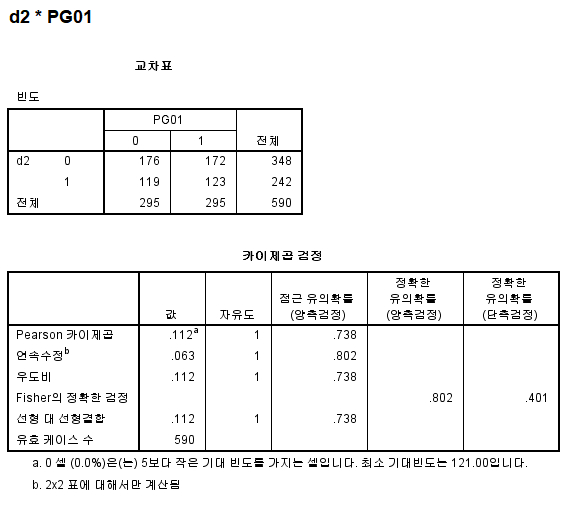
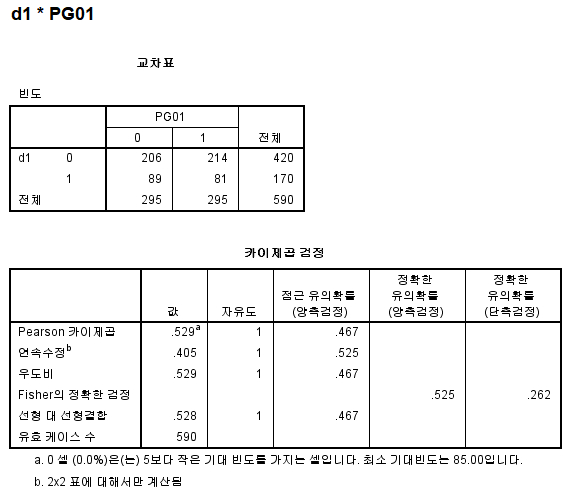
1. **company**: 직업이 직장인인 경우와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.003%로 95% 신뢰수준 하에서 관련성이 있음을 알 수 있다.
2. **worker**: 직업이 개인사업자인 경우와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.020%로 95% 신뢰수준 하에서 관련성이 있음을 알 수 있다.
3. **etcwork**: 기타 직업을 가진 경우와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.000%로 95% 신뢰수준 하에서 관련성이 크다고 할 수 있다.



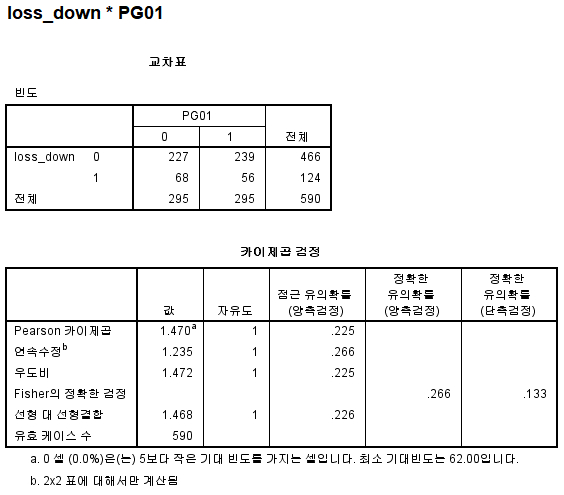
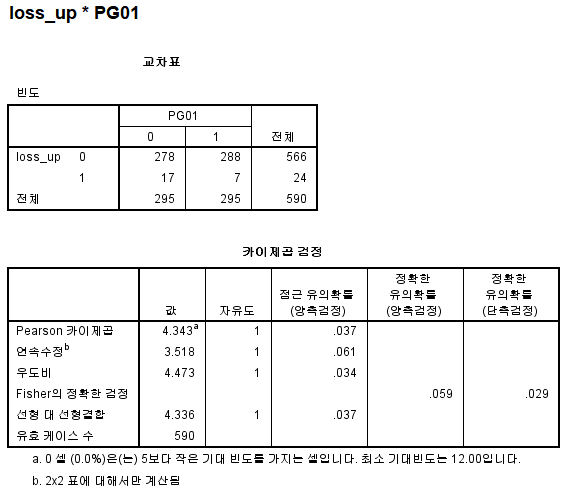
1. gender: 성별과 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.056%로 95% 신뢰수준 하에서 관련이 없다고 볼 수 있다.

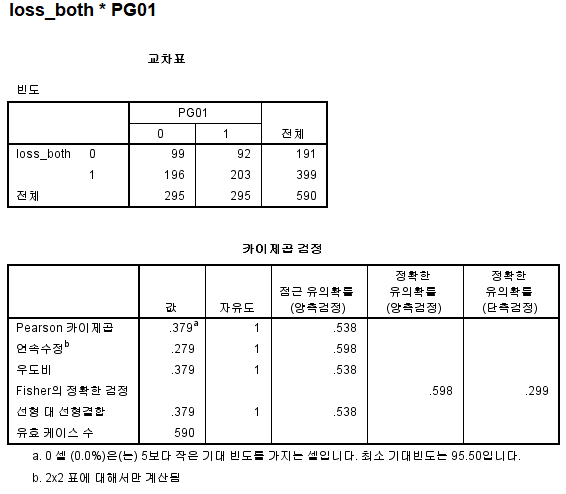


1. **married**: 결혼여부와 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.027%로 95% 신뢰수준 하에서 관련성이 있음을 알 수 있다.

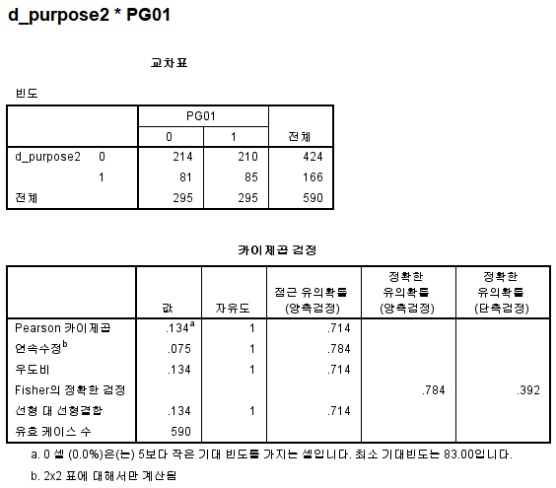
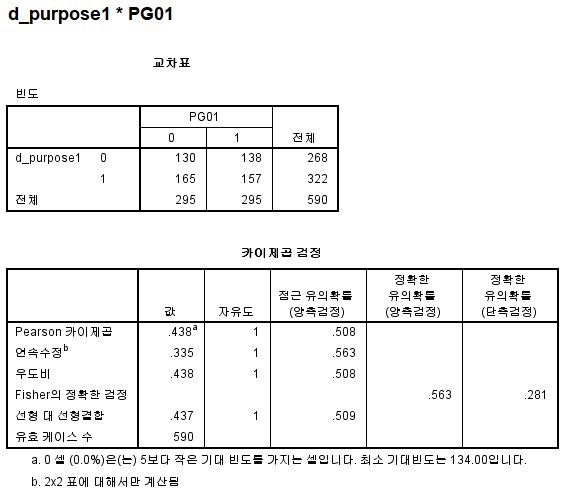


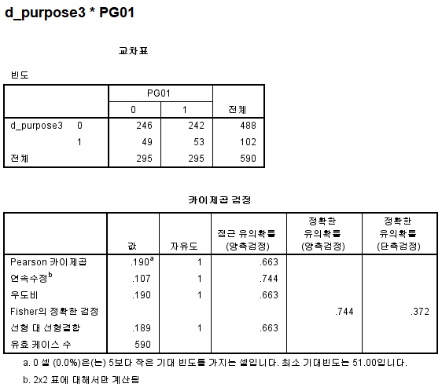
1. d1, d2, d3, d4: 전에 앓았던 질병과 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 각각 0.467%, 0.738%, 0.275%, 0.197%로 앓았던 질병과 전략제품 구매여부에는 큰 관련이 없다고 할 수 있다.



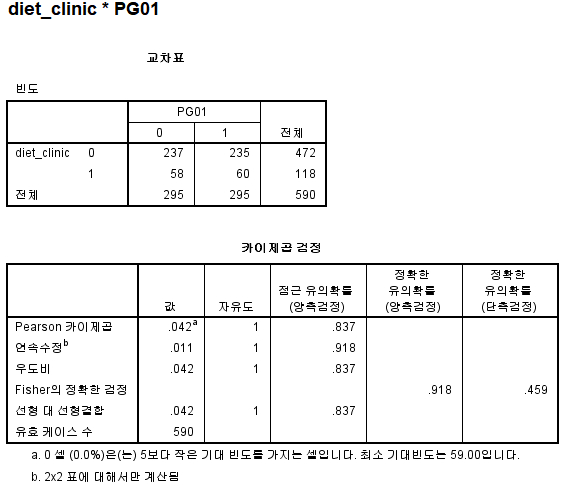
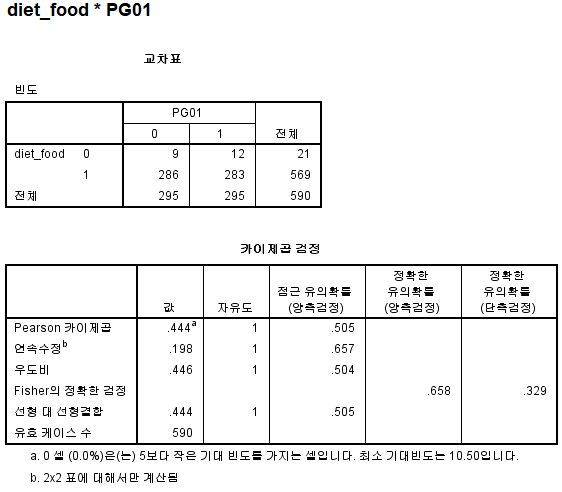


1. **loss\_up**: 상체살을 빼고 싶은 집단과 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.037%로 95% 신뢰수준 하에서 관련성이 있음을 알 수 있다.
2. loss\_down: 하체살을 빼고 싶은 집단과 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.225%로 95% 신뢰수준 하에서 관련성이 없다고 할 수 있다.
3. loss\_both: 상하체살 모두를 빼고 싶은 집단과 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 0.538%로 95% 신뢰수준 하에서 관련성이 없다고 할 수 있다.





1. d\_purpose1, d\_purpose2, d\_purpose3: 다이어트 목적과 전략제품 구매여부 사이의 상관관계는 점근 유의확률 결과 각각 0.508%, 0.714%, 0.663%로 별다른 관련성이 없다고 할 수 있다.



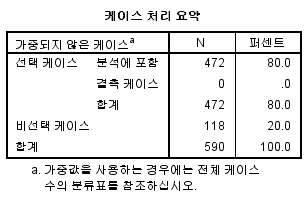
1. diet\_food, diet\_clinic: 전에 경험해봤던 다이어트 방법과 전략제품 구매여부 사이의 관련성은 점근 유의확률 결과 각각 0.505%, 0.837%로 큰 관련성이 없다고 할 수 있다.

독립표본t검정과 카이제곱검정을 실시한 결과 로지스틱스 회귀분석에 사용될 수 있는 변수는 agew\_n, weightw\_n, want\_dietw\_n, waistw\_n, bmi\_finalw\_n, recentw\_n, otherregion, company, worker, etcwork, married, loss\_up로 총 12개이다.

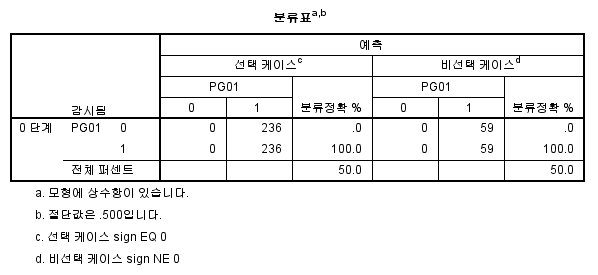
1. 로지스틱스 회귀분석



SPSS에서 로지스틱스 회귀분석을 위해서 위 이미지와 같이 설정을 하고 분석을 진행하였다.



케이스 처리 요약 결과를 확인해보면 표본추출을 하면서 sign 변수에 지정하였던 바와 같이 80%에 해당하는 472건의 데이터가 로지스틱스 회귀분석을 하는데 사용되었고, 나머지 20%에 해당하는 118건의 데이터는 검증을 위해 남겨진 것을 알 수 있다.



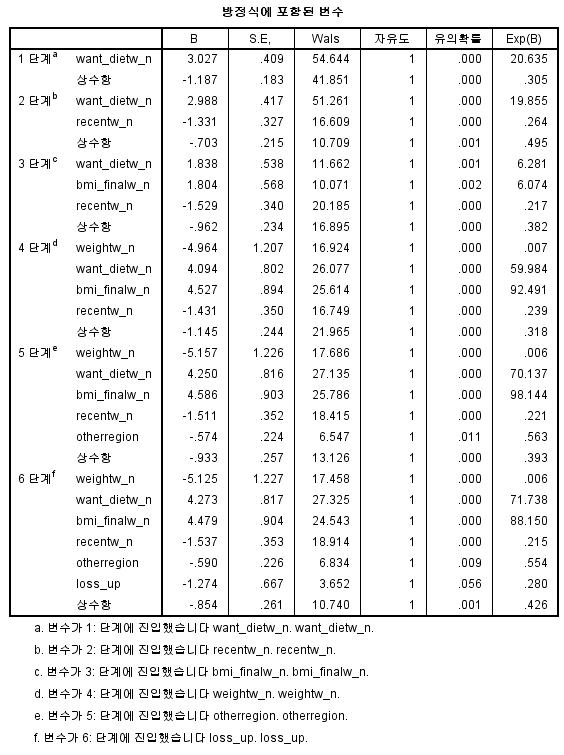
실제 관측된 결과를 보면 분석에 포함된 470건의 데이터 가운데 PG1의 값이 0인 데이터가 236건, 1인 데이터가 236건이고, 분석에 포함되지 않은 나머지 118건의 데이터 가운데에서도 PG1의 값이 0인 데이터가 59건, 1인 데이터가 59건으로 데이터가 어느 한쪽에 쏠리지 않고 균질하게 추출되어 있음을 알 수 있다.



이 모형의 경우 점진 단계선택으로 총 6단계까지 진행되었으며 총 6개의 변수가 분석하는데 활용되었다는 것을 알 수 있다.



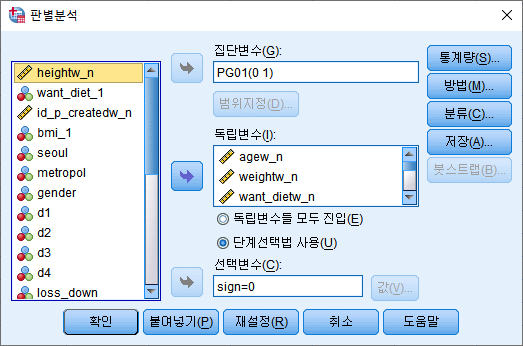
예측 정확도 측면에서 확인해보면, 최종 6단계에서 학습용 데이터를 가지고 0을 0으로 맞춘 것이 169건, 1을 1로 맞춘 것이 167건으로 대략 71.2%의 정확도를 가지고 있음을 볼 수 있다. 검증용 데이터에 대해서도 77.1%의 정확도를 보인 것을 확인할 수 있다.



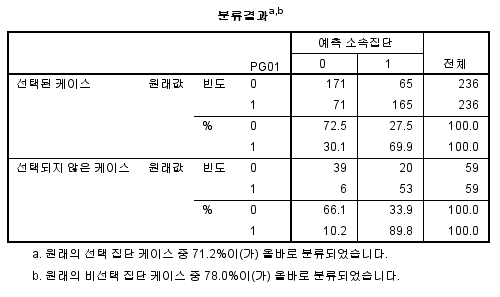
로지스틱스 회귀분석에서 최종 단계인 6단계에서의 베타값과 Odds 비율(Exp(B))을 확인해보면, 다음과 같은 결론을 내릴 수 있다.

1. loss\_up의 베타값이 -1.274, Odds 비율이 0.28로 상체살을 빼고 싶은 사람일수록 전략제품을 구매할 가능성은 낮아지고 상체살을 뺴고 싶은 사람은 그렇지 않은 사람과 대비하여 전략제품을 구매할 가능성이 0.28배로 감소한다.
2. otherregion의 베타값이 -0.590, Odds 비율이 0.554로 서울 및 수도권 이외에 거주하는 사람일수록 전략제품을 구매할 확률은 낮아지고 기타 지역에 거주하는 사람은 그렇지 않은 사람보다 전략제품을 구매할 가능성이 0.554배로 감소한다.
3. recentw\_n의 베타값이 -1.537, Odds 비율이 0.215로 최근 제품을 구매한 사람일수록 전략제품을 구매할 확률은 낮아지고 최근 제품을 구매한 사람은 그렇지 않은 사람보다 전략제품을 구매할 확률이 0.215배로 감소한다.
4. Bmi\_finalw\_n의 베타값이 4.479, Odds 비율이 88.150으로 BMI지수가 높은 사람일수록 전략제품을 구매할 가능성은 높아지고 BMI 지수가 높은 사람이 그렇지 않은 사람보다 전략제품을 구매할 확률이 88.15배 높아진다.
5. Want\_dietw\_n의 베타값이 4.273, Odds 비율이 71.738로 체중을 많이 감량하고 싶은 사람일수록 전략제품을 구매할 가능성은 높아지고 감량목표가 높은 사람이 그렇지 않은 사람보다 전략제품을 구매할 확률이 71.738배 증가한다.
6. Weightw\_n의 베타값이 -5.125, Odds 비율이 0.006으로 몸무게가 높은 사람일수록 전략제품을 구매할 확률은 낮아지고 몸무게가 높은 사람이 그렇지 않은 사람보다 전략제품을 구매할 가능성이 0.006배로 감소한다.

**II. MDA**



SPSS에서의 다중판별분석은 위 이미지와 같은 설정을 따랐다. 다중판별분석은 로지스틱스 회귀분석에서 사용된 데이터와 동일한 데이터로 진행하였다.



다중판별분석 결과 PG1 값 0을 0으로 맞춘 확률은 총236건의 데이터 중 171건에 대해 정확히 맞춰 72.5%, PG1 값 1을 1로 맞춘 확률은 총 236건의 데이터 중 165건에 대해 정확히 맞춰 69.9%의 정확도를 보였다. 종합해 볼 때 약 71.2%의 정확도를 보였다.

검증용 데이터(선택되지 않은 케이스)에서는 PG1 값 0을 0으로 맞춘 확률은 총59건의 데이터 중 39건에 대해 정확히 맞춰 66.1%, PG1 값 1을 1로 맞춘 확률은 총 59 건의 데이터 중 53건에 대해 정확히 맞춰 89.8%의 정확도를 보였다. 종합해 볼 때 약 78.0%의 정확도를 보였다.

보통 확률이 70% 이상이면 안정적인 모델이라고 여겨지기 때문에 이 데이터는 어느 정도 신뢰할 수 있는 모델이라고 볼 수 있다.