# Regression : Final Report

248STG01 김다빈

June 9, 2024

## 1 Question 1

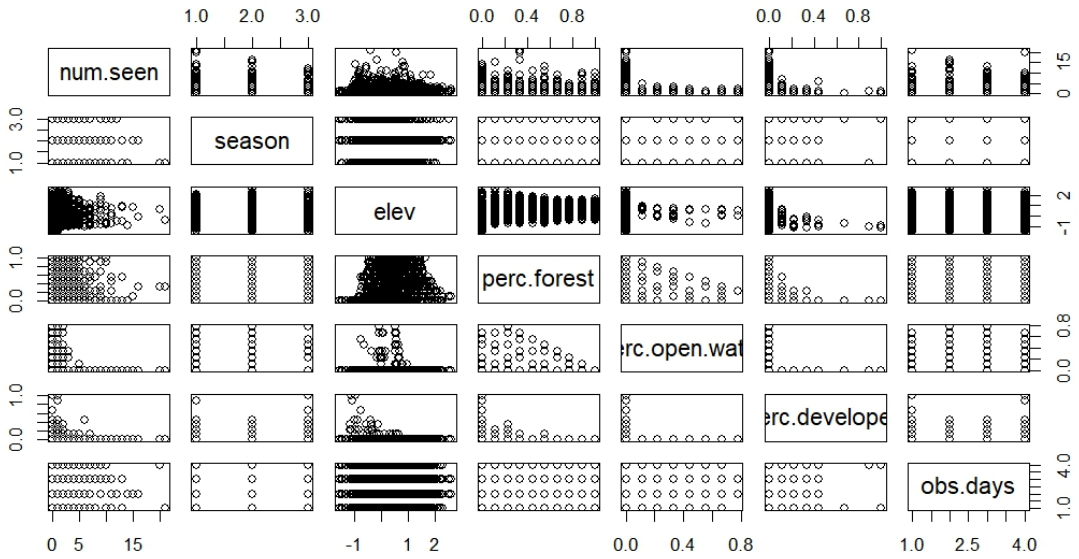


Figure 1: **Plot with pairs option in R**

### 1.1 technical statistics

모델을 적합시키기 전에, 먼저 총 7개의 변수에 대한 paired plot을 그려보았습니다. Figure 1에서 볼 수 있듯이, 뚜렷한 상관관계나 다른 변수 간 특징은 나타나지 않음을 확인할 수 있습니다. 그리고, Y = num.seen에 대해 각각의 변수 하나씩 분포 및 특징을 확인하였습니다. Figure1의 Paired plot에서 계절별 데이터 수가 다른 것을 확인하여 계절에 따른 평균 Elk 수를 구해보았습니다. Parturition일 때 평균 1.16, Summer of Fall 일 때 평균 0.975, 마지막으로 Winter에 평균 0.972로 겨울에 가장 적은 수가 관찰 된 것을 알 수 있었습니다.

아래의 Figure2에서 확인할 수 있듯이 num.seen의 변수는 대부분 0과 1로 구성되어있으며 right-skewed 한 분포를 띄고 있음을 확인할 수 있었습니다. Y가 count response이고, Y의 분포가 skewed되어 있으므로 적합시킬 모델을 **Poission Regression**을 사용하기로 결정하였습니다.

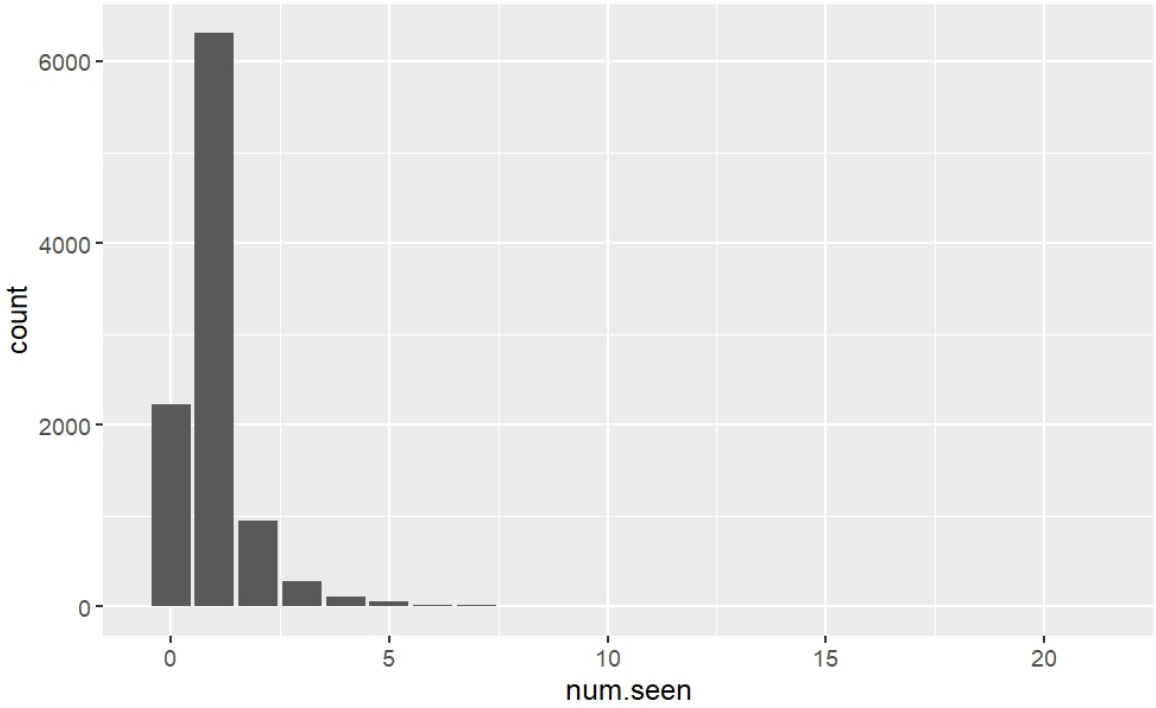


Figure 2: **Count bar plot of Y**

## 1.2 model and assumptions

```
Call:
glm(formula = num.seen ~ ., family = "poisson", data = ElkData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7113  -0.2668  -0.0566   0.1072   8.9426

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.19728    0.03188   6.189 6.06e-10 ***
seasonSF        -0.10229    0.02335  -4.381 1.18e-05 ***
seasonW         -0.18798    0.02784  -6.753 1.45e-11 ***
elev            -0.15637    0.01702  -9.189  < 2e-16 ***
perc.forest      0.01681    0.03169   0.531    0.596
perc.open.water -0.87946    0.20772  -4.234 2.30e-05 ***
perc.developed  -1.08615    0.19165  -5.667 1.45e-08 ***
obs.days        -0.01622    0.01018  -1.594    0.111
---
```

Figure 3: **First trial model with all variables**

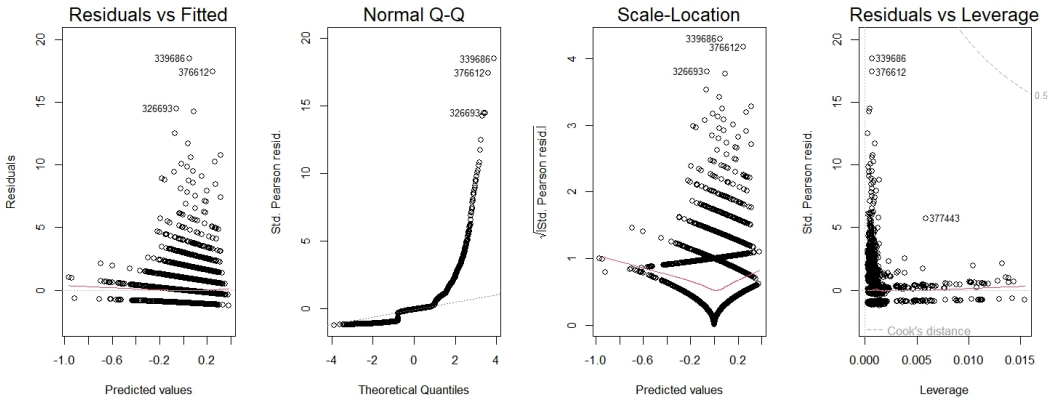


Figure 4: **Residual analysis for trial Model**

$TrialModel$ : $g(E[num.seen_i]) = \beta_0 + \beta_1 seasonSF + \beta_2 seasonW + \beta_3 elev + \beta_4 perc.forest + \beta_5 perc.open.water + \beta_6 perc.developed + \beta_6 obs.days$

다음으로, 전체 변수를 모두 사용하여 glm possion regression을 적합시켜보았습니다.ㄴseasonW와 seasonSF 변수의 계수가 음수 인 것을 보아, Parturtion일 때 관찰되는 수가 많고, SF, Winter일 수록 점점 수가 적어질 것이라고 판단할 수 있었습니다. 또한, perc.forest 변수에 대한 Pvalue가 0.596으로 상당히 높게 나온 것을 보아 , 후에 이 변수는 제외하고 모델을 적합시키는 것이 예측력을 더 높일 수 있을 것이라 가정하였습니다. 상단의 Figure4에서 보듯, residual, multicollinearity and outlier 등을 확인해보았을 때 influential point는 없다고 판단하였습니다. 해당 모델의 AIC값은 25011이었으며 더 예측력이 좋은 모델을 구축하기 위해, 다양한 시도를 해보았습니다.

### 1.2.1 Using AIC

$TrialModel : g(E[num.seen_i]) = 0.20285 - 0.15355 seasonSF - 0.10119 seasonW - 0.18832 elev - 0.88345 perc.open.water - 1.09807 perc.developed - 0.01626 obs.days$

첫째로, trial Model를 AIC 지표를 기반으로 단계적 회귀 분석을 통해 모델을 잘 설명할 수 있는 독립변수들을 선택하였습니다. 첫번째 trial Model에서 가정한 것처럼**perc.forest**변수가 제외된 변수들로 적합시킨 모델이 AIC가 25000.29로 trial Model보다 소폭 감소한 AIC를 기록하였습니다.

```
Call:
glm(formula = num.seen ~ elev + season + perc.open.water + perc.developed +
    obs.days, family = "poisson", data = ElkData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7129  -0.2697  -0.0532   0.1065   8.9396

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.20285    0.03010   6.740 1.58e-11 ***
elev            -0.15355    0.01614  -9.511  < 2e-16 ***
seasonSF        -0.10119    0.02326  -4.351 1.36e-05 ***
seasonW         -0.18832    0.02783  -6.767 1.31e-11 ***
perc.open.water -0.88345    0.20753  -4.257 2.07e-05 ***
perc.developed  -1.09807    0.19037  -5.768 8.02e-09 ***
obs.days        -0.01626    0.01018  -1.598     0.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

### 1.2.2 Adding Interaction term

$$g(E[num.seen_i]) = 0.16915 - 0.11013 seasonSF - 0.19494 seasonW - 0.26922 elev - 0.81128 perc.open.water - 1.70571 perc.developed - 0.01611 obs.days 0.37664 elev * perc.forest - 1.23741 elev * perc.developed$$

```
Call:
glm(formula = num.seen ~ season + elev + perc.open.water + perc.developed +
    obs.days + I(elev * perc.forest) + I(elev * perc.developed),
    family = "poisson", data = ElkData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8139  -0.3124  -0.0485   0.1163   8.9977

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.16915    0.03043   5.559 2.71e-08 ***
seasonSF                  -0.11013    0.02330  -4.726 2.29e-06 ***
seasonW                   -0.19494    0.02784  -7.001 2.54e-12 ***
elev                      -0.26922    0.02197 -12.256  < 2e-16 ***
perc.open.water           -0.81128    0.20753  -3.909 9.26e-05 ***
perc.developed            -1.70571    0.34866  -4.892 9.97e-07 ***
obs.days                  -0.01611    0.01017  -1.583   0.1134
I(elev * perc.forest)      0.37664    0.04625   8.143 3.85e-16 ***
I(elev * perc.developed) -1.23741    0.59064  -2.095   0.0362 *
---
```

두번째 시도로 , intercation term을 추가해보았습니다. Trial Model에서 polynomial term보다 interaction term이 적합하다고 판단하여 I(elev*perc.developed) 와 I(elev * perc.forest) 두 가지 interaction term을 추가하여 시도하였습니다. 이때, 앞선 AIC stepwise regression결과를 반영하여, **perc.forest** 변수를 제외하고 모델을 적합시켰습니다. AIC가 24940.94로 앞선 두 모델보다 AIC가 소폭 감소한 것을 확인할 수 있었습니다 .

### 1.2.3 Lasso Regression

$g(E[num.seen_i]) = 1.0345 - 1.01e^-37season - 1.4865eleve^-37 - 7e^-38perc.foest - 7.4151e^-37 - 6.679e^-37 - 1.5607e^-38$

마지막으로 Lasso Regression을 시도해 보았습니다. 앞선 3가지의 모델들의 결과와 달리 intercept term을 제외하고 다른 모든 변수들에 대해 매우 작은 coef를 할당한 것을 확인할 수 있었습니다.

```
7 x 1 sparse Matrix of class "dgCMatrix"
                              s1
(Intercept)      1.034600e+00
season          -1.061222e-37
elev            -1.486517e-37
perc.forest     -7.005089e-38
perc.open.water -7.415126e-37
perc.developed  -6.679202e-37
obs.davs        -1.560746e-38
```

## 1.3 Final model

| Model | Trial | AIC | Interaction | Lasso |
|---|---|---|---|---|
| MSPE | 1.058219 | 1.058201 | 1.050274 | 1.059444 |

Table 1: MSPE of each Model

$FinalModel : g(E[num.seen_i]) = log(\hat{\mu})) = 0.16915 - 0.11013seasonSF - 0.19494seasonW - 0.26922elev - 0.81128perc.open.water - 1.70571perc.developed - 0.01611obs.days + 0.37664elev * perc.forest - 1.23741elev * perc.developed$

시도한 4가지의 모델에 대해 MSPE를 구해본 결과는 위의 Table1과 같습니다. AIC와 마찬가지로 perc.forest 변수를 제외하고 interaction term을 추가한 모델에서 가장 작은 AIC,MSPE를 가졌습니다. 최종 모델에서 각 변수의 $\beta$값들을 살펴보았습니다. 다른 변수들이 통제되었을 때 elev가 한 단위 증가하면 num.seen의 값은 $e^-0.26922$만큼 변화하는 것을 알 수 있습니다.

따라서, MSPE가 가장 작은 해당 모델을 최종모델로 선택하고, 잔차 분석 및 적합이 잘 되었는지 검정하였습니다. 아래 Figure5와 같이 simulated model과 final model의 plot이 유사한 것으로 판단하여, 적합이 잘 되었다고 결론지었습니다.

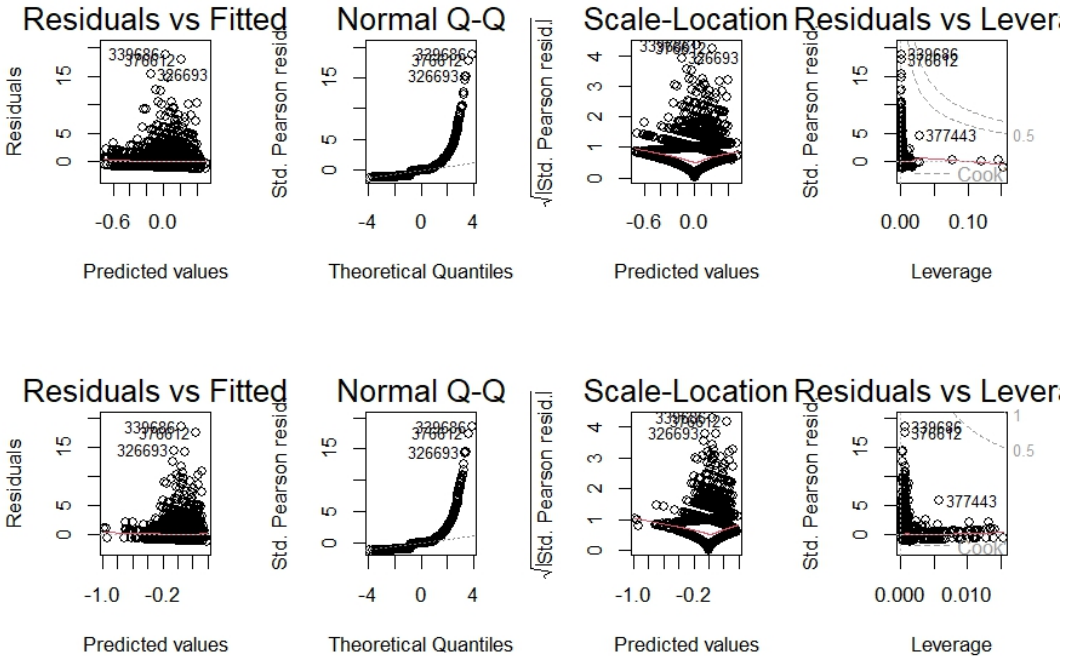


Figure 5: **Comparison of residual plot between final model(above) and simulated model(below)**

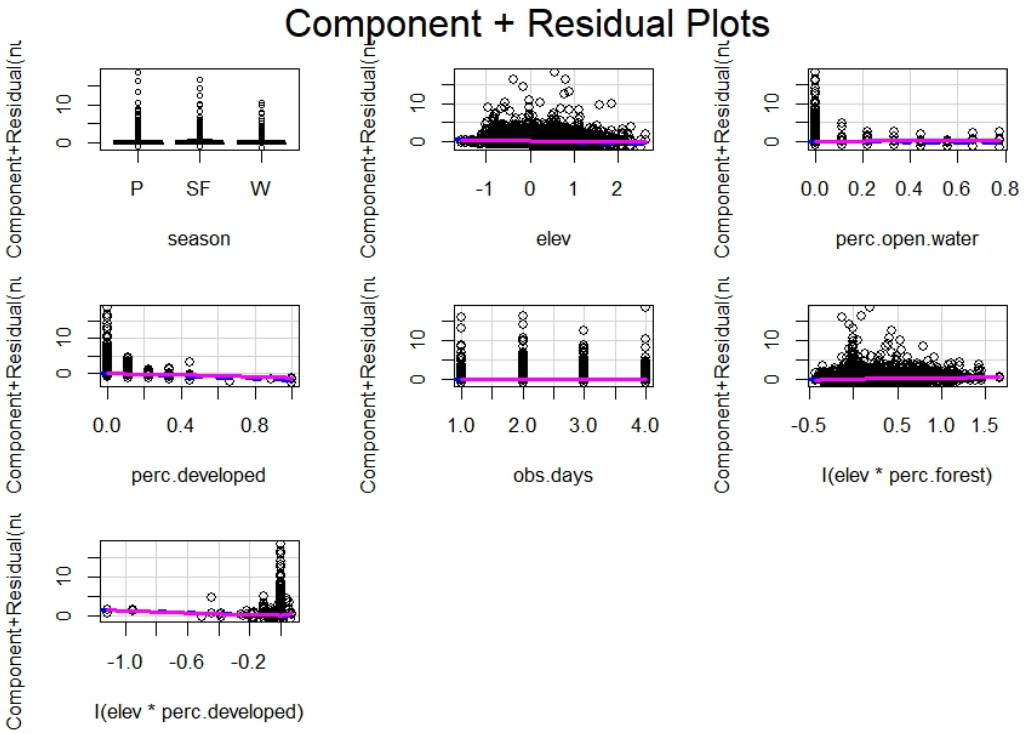


Figure 6: **Residual analysis for final model**

## 2 Question 2

```
   P   SF    W
3242 4732 2026
```

먼저, 각 계절별로 데이터의 갯수를 세어보면 Summer of Fall이 4732로 가장 많고, Winter가 2026으로 가장 적음을 확인할 수 있었습니다. 각 계절별로 인덱스를 추출해 계절별로 데이터를 분리하여, 앞서 Question1에서 선정한 best model을 이용하여 각 계절별 예측값을 구해보았습니다.

각 계절별 예측한 값과 실제 number of Elk의 수를 가지고 MSPE를 구했을 때 Parturition 때 MSPE가 가장 큰 것을 보아, 정확히 예측하지 못함을 알 수 있었습니다. 또한, 계절이 Winter인 경우 가장 정확히 예측함을 확인할 수 있었습니다. $\hat{\lambda}$의 값으로 simulated y를 구해 각 계절별 값을 비교해본 결과 Parturtition이 가장 높은 값을 보였고, Winter에 수가 적은 것으로 나타났습니다. 결론적으로, Parturtition때 데이터들이 가장 정확하지 않은 예측을 하는 것으로 보아, 추가적인 관찰이 필요하다고 결론 지을 수 있었습니다.

| | P | SF | W |
|---|---|---|---|
| MSPE | 1.453919 | 0.8934626 | 0.7706149 |

Figure 7: **MSPE for each season**

| | P | SF | W |
|---|---|---|---|
| y | 1.165638 | 0.964497 | 0.9590326 |

Figure 8: **Y for each season**

# 3 Question 3

두 개의 site에 대한 data 값들을 생성해주었고, 모델에 적합시키기 위해 obs.days에 대한 값은 임의로 가장 횟수가 많았던 2로 동일하게 추가하였습니다. 두 site를 비교해보면, site1은 elev가 낮고, perc.forest와 perc.open.water의 값이 0이며 site1은 0.209로 site1과 비교해서 높은 elev값을 가짐을 확인할 수 있었습니다. 또한 site1의 perc.forest의 값이 0이었던 것과 달리 0.0222이며, 두 site 모두 perc.developed 값은 0이었습니다. 따라서 elev와 perc.forest 값의 차이에 따른 영향을 가정해볼 수 있었습니다. Question2와 마찬가지로, Question1에서 결론지은 Final model을 이용하여 두 site에 대한 예측을 진행하였습니다.

```
  season   elev perc.forest perc.open.water perc.developed obs.days
1      W -0.954      0.0000            0.00              0        2
2      W  0.209      0.0222            0.01              0        2
```

두 site의 예측값, confidence interval of expected number of observed elk 은 다음과 같습니다. site1에서 expected number of observed elk가 큰 것으로 예측하였습니다. 해당 결과는 두 site의 elev와 perc.forest의 값의 차이로 인해 나타난 것으로 볼 수 있습니다. 임의로 넣은 obs.days값에 대해서는 작아질 수록 예측값 $\hat{\lambda}$값이 작아졌습니다. 따라서 결론적으로 Winter 계절동안 site1에서 더 많은 elk가 관찰될 것으로 예측되기 때문에 site2로 변경하기 보다 site1에 먹이를 주는 것이 더 좋다고 판단할 수 있었습니다.

| | muhat | LOWER | UPPER | exp.Lower | exp.Upper |
|---|---|---|---|---|---|
| **site1** | 1.2199765 | 1.1516141 | 1.2883388 | 3.163295 | 3.626757 |
| **site2** | 0.8863546 | 0.8428658 | 0.9298434 | 2.323015 | 2.534112 |

Figure 9: **Confidence interval for each site**