# RLab5

Eunjin Park

date

**DELETE ANYTHING FROM THIS TEMPLATE BELOW THAT IS NOT PART OF YOUR SOLUTION.**

(0) Instructions for installing tinytex for PDF rendering: **https://yihui.org/tinytex/**

```
install.packages('tinytex')
tinytex::install_tinytex()
```

**I.Logistic Regression**

```
otters = read.csv("River_Otters_-_High_Mountain_Lakes_[ds813].csv")
head(otters)
```

```
##           X       Y OBJECTID   Region Site_Name Waterbody Elevation_m Timeframe
## 1 -13501910 4947965        1 Cascades     Butte      Lake        1844    August
## 2 -13501910 4947965        2 Cascades     Butte      Lake        1844    August
## 3 -13518222 4939619        3 Cascades    Dersch     Marsh        2012    August
## 4 -13501910 4947965        4 Cascades     Butte      Lake        1844    August
## 5 -13501910 4947965        5 Cascades     Butte      Lake        1844    August
## 6 -13501910 4947965        6 Cascades     Butte      Lake        1844    August
##   Year Otters_Found Rank       Source      Lat      Long   UTME    UTMN
## 1 2006            1    2 LVNP Records 40.56210 -121.2897 644789 4491553
## 2 2006            1    2 LVNP Records 40.56210 -121.2897 644789 4491553
## 3 2006            1    2 LVNP Records 40.50511 -121.4362 632496 4484997
## 4 2003            1    3 LVNP Records 40.56210 -121.2897 644789 4491553
## 5 2003            1    3 LVNP Records 40.56210 -121.2897 644789 4491553
## 6 2007            1    3 LVNP Records 40.56210 -121.2897 644789 4491553
##               DATUM
## 1 UTM NAD83 Zone 10
## 2 UTM NAD83 Zone 10
## 3 UTM NAD83 Zone 10
## 4 UTM NAD83 Zone 10
## 5 UTM NAD83 Zone 10
## 6 UTM NAD83 Zone 10
```

(1) Read the data in River_Otters_-*High_Mountain_Lakes*[ds813].csv into a dataframe called otters. Provide Executable Code (PEC) that creates a new variable in the otters dataframe called Detected which is TRUE when the total number of otters observed is greater than 0, and FALSE when the total number of otters observed is exactly 0. Check that you have the same number of TRUE values as I do:

```r
otters$Detected = otters$Otters_Found > 0

#otters$Detected = with(otters, ifelse(otters$Otters_Found == 0, FALSE, TRUE))
sum(otters$Detected )
```

```
## [1] 72
```

```r
#why is different when using length
```

(2) [3 pts] Make a plot of the binary variable Detected as a function of Year. From looking at the figure, do you suspect there is a significant relationship between these two variables? Why/why not? (Hint: You might find it the jitter() function in R useful for spreading out overlapping data points.)

```r
#b = jitter(Detected$Year)
#x = Detected$Year
#plot(x,b)
```

(3) [5 pts] Fit a logistic regression model with Detected as a response and Region, Waterbody, Elevation_m, and Year as predictors (PEC). Which variable(s) appear(s) to be the most important for predicting the probability of detecting otters at a given site?

```r
fit_glm = glm(Detected ~ Region + Waterbody + Elevation_m
+ Year, data = otters, family= binomial(link = "logit"))
summary(fit_glm)
```

```
##
## Call:
## glm(formula = Detected ~ Region + Waterbody + Elevation_m + Year,
##     family = binomial(link = "logit"), data = otters)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.53190   0.00007   0.30138   0.45615   1.29499
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -9.204e+01  2.511e+03  -0.037   0.9708
## RegionKlamath      -1.776e+01  2.511e+03  -0.007   0.9944
## RegionSierra       -1.785e+01  2.511e+03  -0.007   0.9943
## WaterbodyMarsh     -7.788e-01  1.104e+04   0.000   0.9999
## WaterbodyReservoir  5.158e-01  1.713e+00   0.301   0.7634
## WaterbodyStream     1.810e+01  5.345e+03   0.003   0.9973
## Elevation_m         1.974e-03  1.566e-03   1.260   0.2077
## Year                5.404e-02  2.248e-02   2.404   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 56.511  on 80  degrees of freedom
## Residual deviance: 40.994  on 73  degrees of freedom
```

```
## AIC: 56.994
##
## Number of Fisher Scoring iterations: 18
```

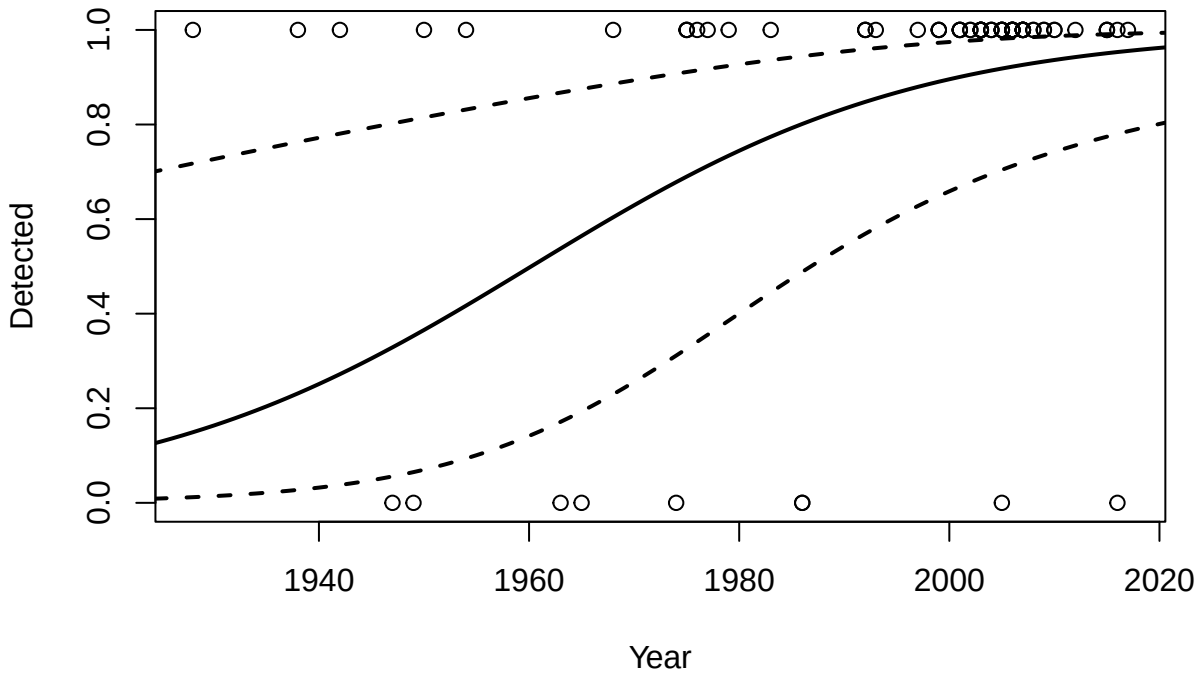
(4) [3 pts] Write a function called expit that computes the inverse-logit of each element in an input vector x. Make sure your function output matches mine:

```
#x = predict(fit_glm, type = "response")

expit = function(x=NULL){
    exp(x)/(1 + exp(x))}
```

(5) [5 pts] For all years 1900-2022, use your fitted model to make predictions for the probability of detection at a Lake in the Sierra region at an elevation of 2000m. Plot your predictions as a function of years and add curves representing 90% pointwise confidence intervals (PEC). (Hint: You may find it helpful to use your expit() function.)

```
year_seq = 1900:2022
plot(Detected ~  Year, data = otters)
p_curve = predict(fit_glm, se.fit = T, newdata = data.frame( Year = year_seq, Region = 'Sierra', Waterb
lines(year_seq, expit(p_curve$fit),lwd = 2)
lines(year_seq, expit(p_curve$fit + qnorm(0.95)*p_curve$se.fit),lwd = 2, lty = 2)
lines(year_seq, expit(p_curve$fit + qnorm(0.05)*p_curve$se.fit),lwd = 2, lty = 2)
```



```
#plot(p_curve, otters$Year, xlab = "predict", ylab = "Year")
```

(6) [2 pts] Based on your plot, at about what year is the predicted probability of observing an otter 0.5?

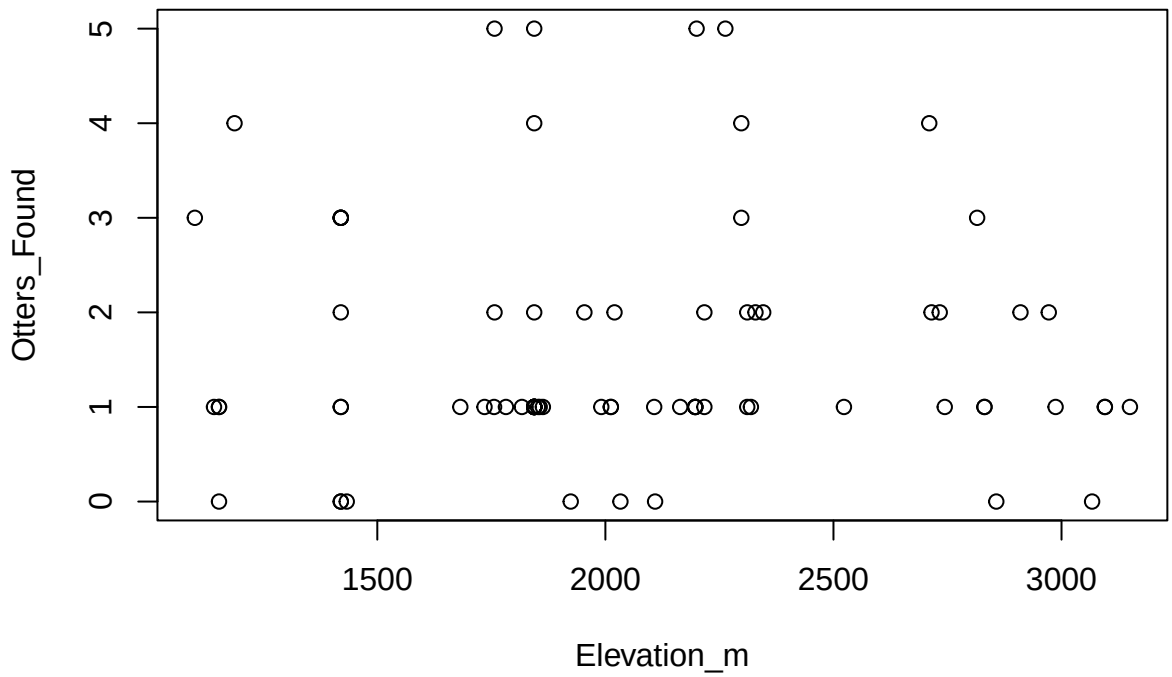
->About 1965

**II. Poisson regression**

Another way to use these data would be to treat the counts as reliable and use a generalized linear model (GLM) for counts to understand the relationship between the predictors and abundance.

(7) [3 pts] Make a plot of the number of otters observed as a function of elevation. From looking at the figure, do you suspect there is a significant relationship between these two variables? Why/why not?

```
plot(Otters_Found ~ Elevation_m, data= otters)
```




(8) [5 pts] Fit a GLM with a Poisson distribution for the response variable, Otters_Found, and include Region, Waterbody, Elevation_m, and Year as predictors (PEC). Which variable(s) appear(s) to be the most important for predicting the number of otters at a given site?

```
#The Poisson model assumes that the variance is equal to the mean, which is not always a fair assumptio

pois = glm(Otters_Found ~ Region + Waterbody + Elevation_m + Year, data=otters, family=poisson)
summary(pois)
```

```
##
## Call:
## glm(formula = Otters_Found ~ Region + Waterbody + Elevation_m +
##     Year, family = poisson, data = otters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9505  -0.6273  -0.1811   0.3549   2.4102
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.652e+01  1.162e+01  -1.422    0.155
## RegionKlamath      -2.999e-01  3.372e-01  -0.889    0.374
## RegionSierra        1.330e-01  3.103e-01   0.429    0.668
## WaterbodyMarsh     -4.406e-01  1.022e+00  -0.431    0.666
## WaterbodyReservoir  1.388e-01  4.177e-01   0.332    0.740
## WaterbodyStream     3.399e-01  4.351e-01   0.781    0.435
## Elevation_m         1.054e-04  3.100e-04   0.340    0.734
## Year                8.351e-03  5.686e-03   1.469    0.142
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 76.691  on 80  degrees of freedom
## Residual deviance: 71.248  on 73  degrees of freedom
## AIC: 257.15
##
## Number of Fisher Scoring iterations: 5
```

(9) [4 pts] Use your fitted model to make 95% confidence interval for the number of otters at a Marsh in the Cascades region at an elevation of 2000m in 2022. What important feature of the data could explain why the interval for this prediction is so wide?

```
pred = predict(pois, newdata = data.frame(Region = "Cascades", Elevation_m = 2000,Year = 2022, Waterbody
se.fit = T)

CI = pred$fit+ qnorm(c(0.025, 0.975))* pred$fit
CI
```

```
## [1] -0.1270593  0.3917762
```

(10) [4 pts] Fit a new GLM with a quasi-Poisson response distribution and the same predictors. Report the estimated dispersion parameter. What does its value suggest about the level of overdispersion present in the data?

```
#The Quasi-Poisson Regression is a generalization of the Poisson regression and is used when modeling a

quasips = glm(Otters_Found ~ Region + Waterbody + Elevation_m + Year, data=otters, family= quasipoisson
summary(quasips)
```

```
##
## Call:
## glm(formula = Otters_Found ~ Region + Waterbody + Elevation_m +
```

```
##     Year, family = quasipoisson, data = otters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9505  -0.6273  -0.1811   0.3549   2.4102
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.652e+01  1.172e+01  -1.411    0.163
## RegionKlamath      -2.999e-01  3.399e-01  -0.882    0.381
## RegionSierra        1.330e-01  3.128e-01   0.425    0.672
## WaterbodyMarsh     -4.406e-01  1.030e+00  -0.428    0.670
## WaterbodyReservoir  1.388e-01  4.211e-01   0.330    0.743
## WaterbodyStream     3.399e-01  4.385e-01   0.775    0.441
## Elevation_m         1.054e-04  3.125e-04   0.337    0.737
## Year                8.351e-03  5.731e-03   1.457    0.149
##
## (Dispersion parameter for quasipoisson family taken to be 1.015935)
##
##     Null deviance: 76.691  on 80  degrees of freedom
## Residual deviance: 71.248  on 73  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

-> Dispersion parameter is estimated to 1.015935.

RegionKlamath and WaterbodyMarsh are the negative coefficient for Otters_found (if the RegionKlamath and WaterbodyMarsh are increased the #of found otters is smaller )

(11) [4 pts] Create a new 95% CI for the same prediction as in (9), but based on the quasi-Poisson response model. How similar/different is yous CIs the the one in (9)? Why do you think that might be?

```
pred_quasi = predict(quasips, newdata = data.frame(Region = "Cascades", Elevation_m = 2000,Year = 2022,
se.fit = T)

CI_quasi = pred_quasi$fit+ qnorm(c(0.025, 0.975))* pred_quasi$fit
CI_quasi
```

```
## [1] -0.1270593  0.3917762
```