Lecture 4

Logistic Regression + Residual Analysis

Colin Rundel

1/29/2018

Background

Today we'll be looking at data on the presence and absence of the short-finned eel (Anguilla australis) at a number of sites in New Zealand.

These data come from

- Leathwick, J. R., Elith, J., Chadderton, W. L., Rowe, D. and Hastie, T. (2008), Dispersal, disturbance and the contrasting biogeographies of New Zealand's diadromous and non-diadromous fish species. Journal of Biogeography, 35: 1481–1497.

Species Distribution

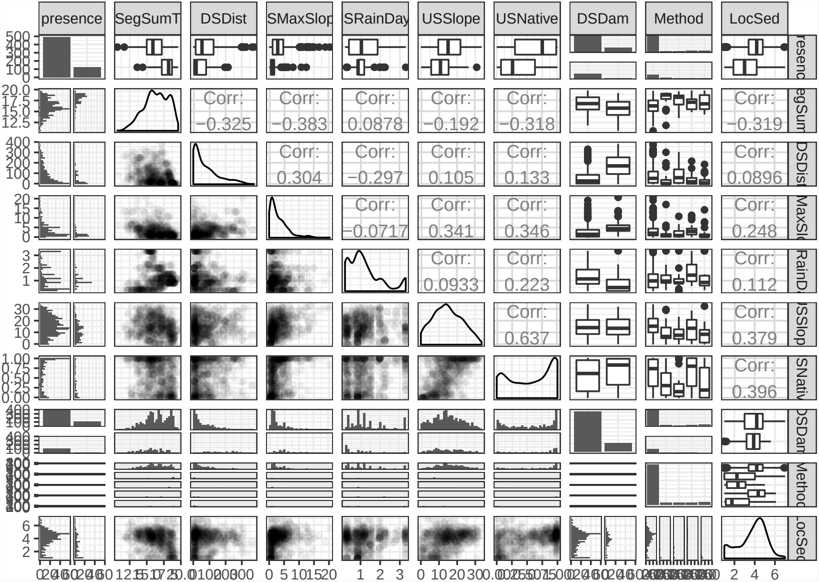
Codebook:

- `presence` - presence (`1`) or absence (`0`) of Anguilla australis at the sampling location
- `SegSumT` - Summer air temperature (degrees C)
- `DSDist` - Distance to coast (km)
- `DSMaxSlope` - Maximum downstream slope (degrees)
- `USRainDays` - days per month with rain greater than 25 mm
- `USSlope` - average slope in the upstream catchment (degrees)
- `USNative` - area with indigenous forest (proportion)
- `DSDam` - Presence of known downstream obstructions, mostly dams
- `Method` - fishing method (`electric`, `net`, `spot`, `trap`, or `mixture`)
- `LocSed` - weighted average of proportional cover of bed sediment
    1. mud
    2. sand
    3. fine gravel
    4. coarse gravel
    5. cobble
    6. boulder
    7. bedrock

Data

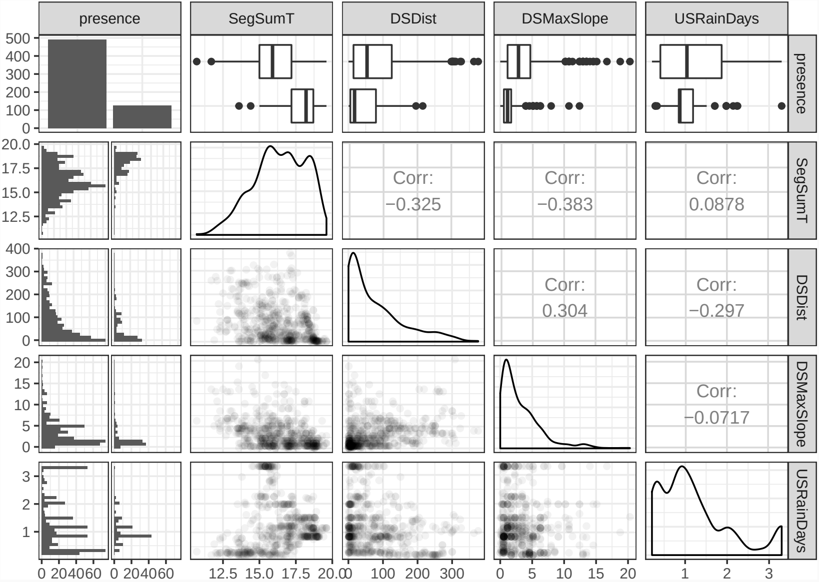
```
anguilla
## # A tibble: 617 x 10
##    presence SegSumT DSDist DSMaxSlope USRainDays USSlope USNative DSDam
##  *    <int>   <dbl>  <dbl>      <dbl>      <dbl>   <dbl>    <dbl> <int>
##  1        1    18.7 133         1.15       1.15    8.30    0.340      0
##  2        0    18.3 107         0.570      0.847   0.400   0          0
##  3        0    16.7 167         1.72       0.210   0.400   0.220      1
##  4        0    15.1  11.2       1.72       3.30   25.7     1.00       0
##  5        0    12.7  42.4       2.86       0.430   9.60    0.0900     0
##  6        1    18.2  94.4       3.43       0.847  20.5     0.920      0
##  7        1    18.3  91.9       1.72       0.861   6.70    0.580      1
##  8        1    17.1   6.80      0.520      0.620   0.700   0          0
##  9        0    13.4 190         3.43       0.770  20.1     0.990      0
## 10        0    13.1 224         6.84       0.290   9.80    0.980      0
## # ... with 607 more rows, and 2 more variables: Method <fct>, LocSed <dbl>
```

EDA

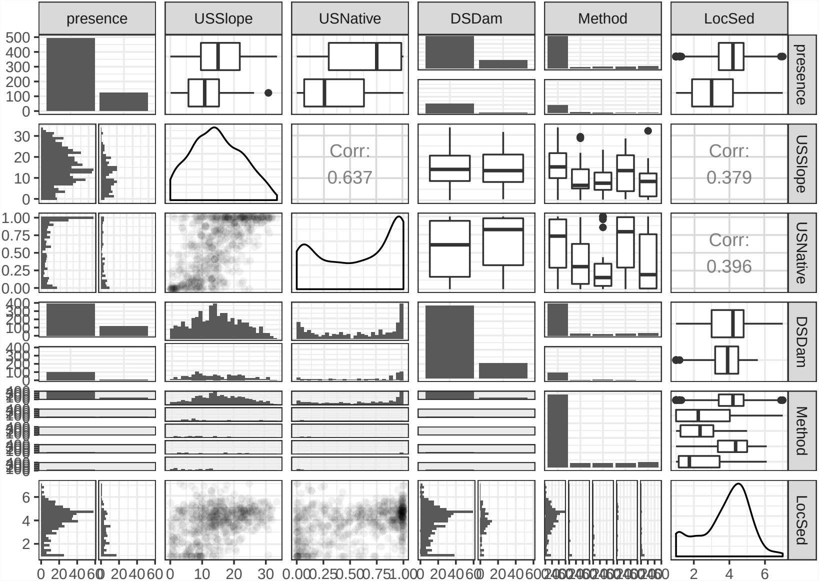


presence
SegSumT
DSDist
SMaxSlop
SRainDay
USSlope
USNative
DSDam
Method
LocSed
Corr: −0.325
Corr: −0.383
Corr: 0.0878
Corr: −0.192
Corr: −0.318
Corr: −0.319
Corr: 0.304
Corr: −0.297
Corr: 0.105
Corr: 0.133
Corr: 0.0896
Corr: −0.0717
Corr: 0.341
Corr: 0.346
Corr: 0.248
Corr: 0.0933
Corr: 0.223
Corr: 0.112
Corr: 0.637
Corr: 0.379
Corr: 0.396

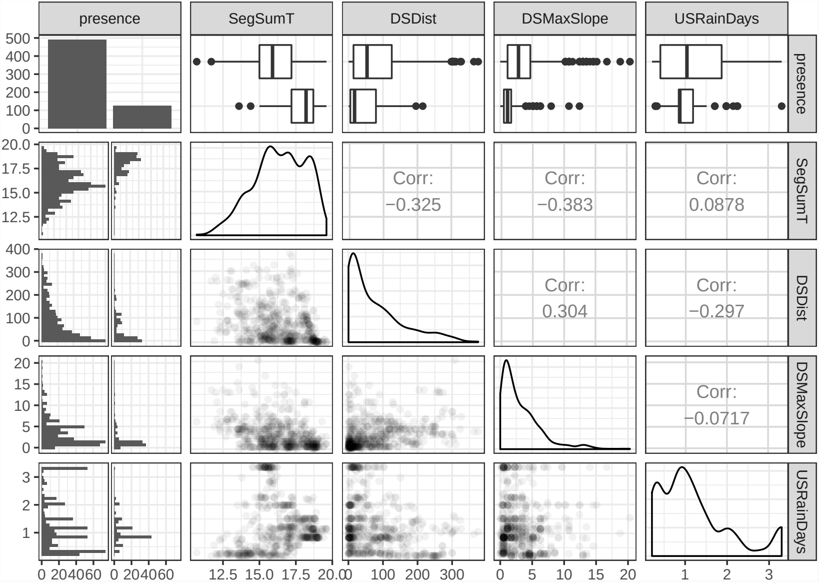
EDA (part 1)


presence
SegSumT
DSDist
DSMaxSlope
USRainDays
Corr:
−0.325
Corr:
−0.383
Corr:
0.0878
Corr:
0.304
Corr:
−0.297
Corr:
−0.0717
500
400
300
200
100
0
20.0
17.5
15.0
12.5
400
300
200
100
0
20
15
10
5
0
3
2
1
0 204060
0 204060
12.5 15.0 17.5 20.0
0 100 200 300
0 5 10 15 20
1 2 3

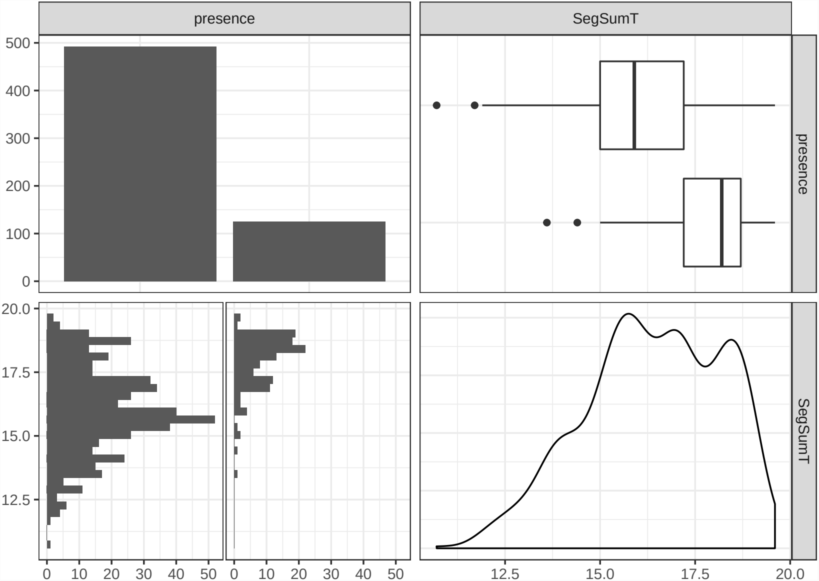
EDA (part 2)


presence
USSlope
USNative
DSDam
Method
LocSed
Corr:
0.637
Corr:
0.379
Corr:
0.396

EDA (part 1)


presence
SegSumT
DSDist
DSMaxSlope
USRainDays
Corr: −0.325
Corr: −0.383
Corr: 0.0878
Corr: 0.304
Corr: −0.297
Corr: −0.0717

EDA (part 3)



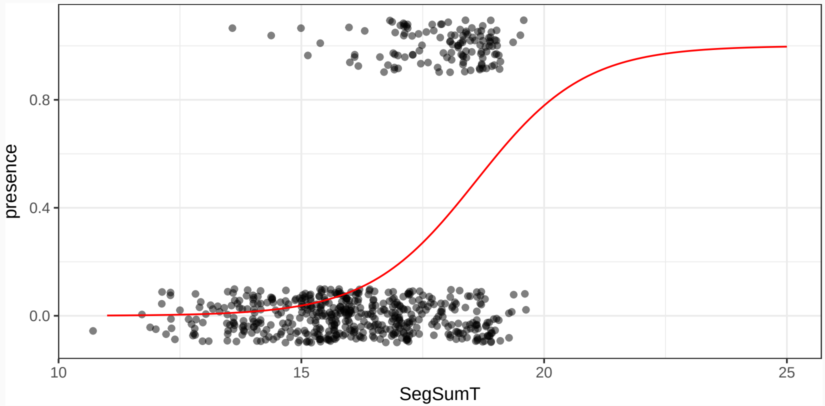
presence
SegSumT
500
400
300
200
100
0
20.0
17.5
15.0
12.5
0
10
20
30
40
50
12.5
15.0
17.5
20.0
presence
SegSumT

Simple Model

```
inv_logit = function(x) 1/(1+exp(-x))

g = glm(presence~SegSumT, family=binomial, data=anguilla)
summary(g)
##
## Call:
## glm(formula = presence ~ SegSumT, family = binomial, data = anguilla)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5755  -0.6260  -0.3452  -0.1299   3.0039
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.74184    1.65897 -10.092   <2e-16 ***
## SegSumT       0.90009    0.09413   9.562   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 621.91  on 616  degrees of freedom
## Residual deviance: 479.39  on 615  degrees of freedom
## AIC: 483.39
```

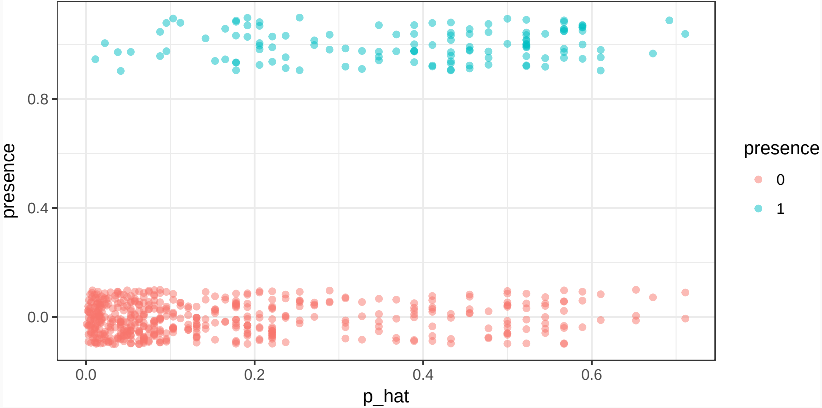
Fit

```r
d_g = anguilla %>%
  mutate(p_hat = predict(g, anguilla, type="response"))

d_g_pred = data.frame(SegSumT = seq(11,25,by=0.1)) %>%
  modelr::add_predictions(g,"p_hat") %>%
  mutate(p_hat = inv_logit(p_hat))
```



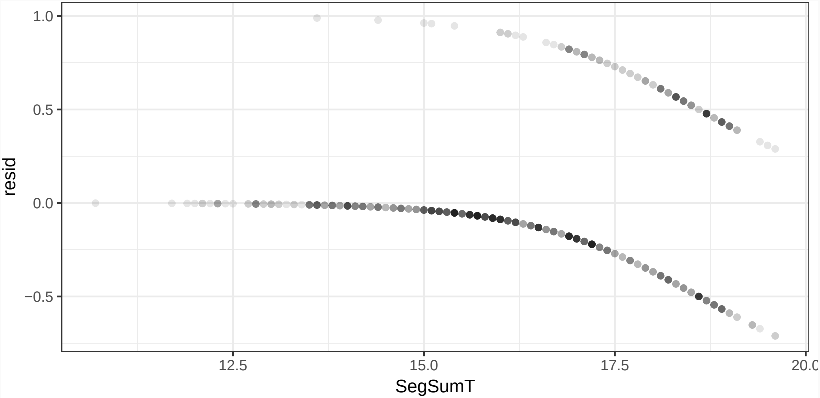
Separation

```
ggplot(d_g, aes(x=p_hat, y=presence, color=as.factor(presence))) +
  geom_jitter(height=0.1, alpha=0.5) +
  labs(color="presence")
```



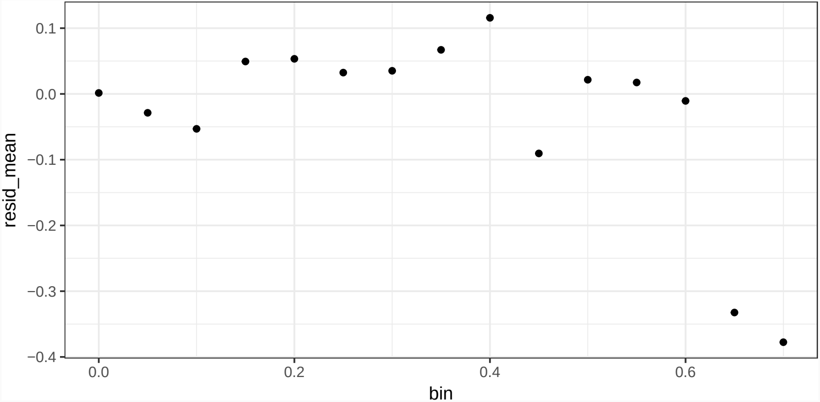
Residuals

```
d_g = d_g %>% mutate(resid = presence - p_hat)

ggplot(d_g, aes(x=SegSumT, y=resid)) +
  geom_point(alpha=0.1)
```



Binned Residuals

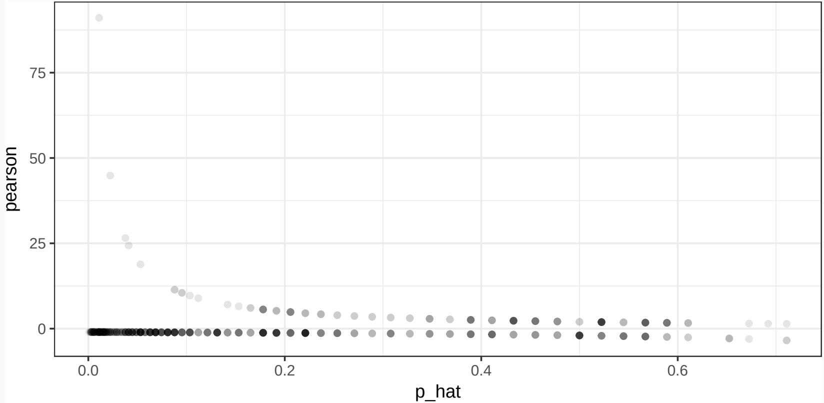
```
d_g %>%
  mutate(bin = p_hat - (p_hat %% 0.05)) %>%
  group_by(bin) %>%
  summarize(resid_mean = mean(resid)) %>%
  ggplot(aes(y=resid_mean, x=bin)) +
    geom_point()
```



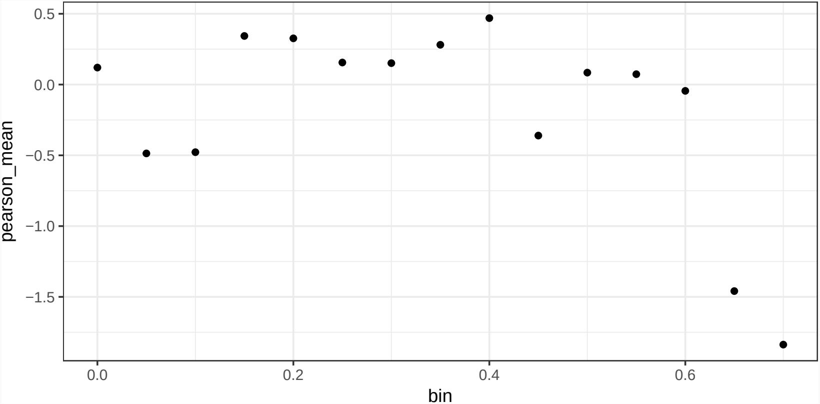
Pearson Residuals

$$r_i = \frac{Y_i - E(Y_i)}{Var(Y_i)} = \frac{Y_i - \hat{p}_i}{\hat{p}_i(1 - \hat{p}_i)}$$

```
d_g = d_g %>% mutate(pearson = (presence - p_hat) / (p_hat * (1-p_hat)))
```



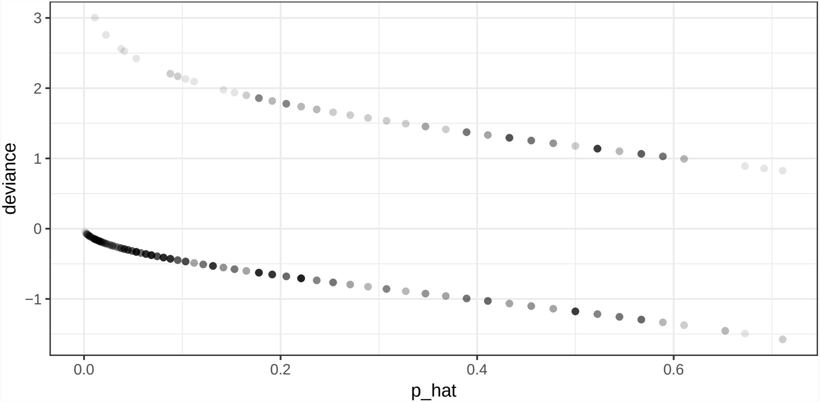
Binned Pearson Residuals



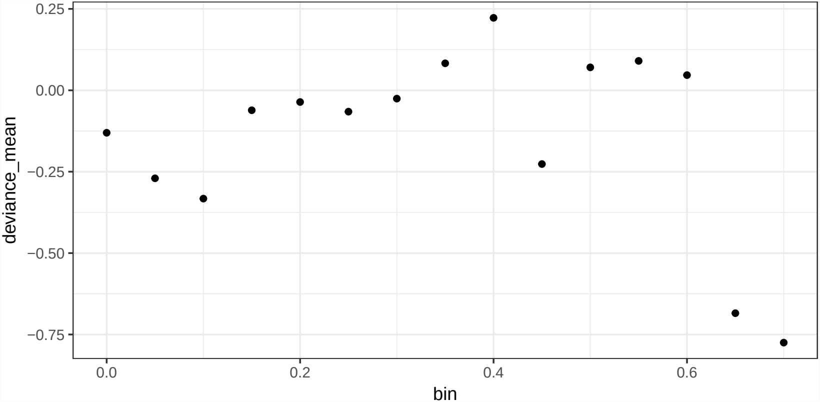
Deviance Residuals

$$d_i = \text{sign}(Y_i - \hat{p}_i)\sqrt{-2\left(Y_i \log \hat{p}_i + (1 - Y_i)\log(1 - \hat{p}_i)\right)}$$

```
d_g = d_g %>%
  mutate(deviance = sign(presence - p_hat) *sqrt(-2 * (presence*log(p_hat) +
```



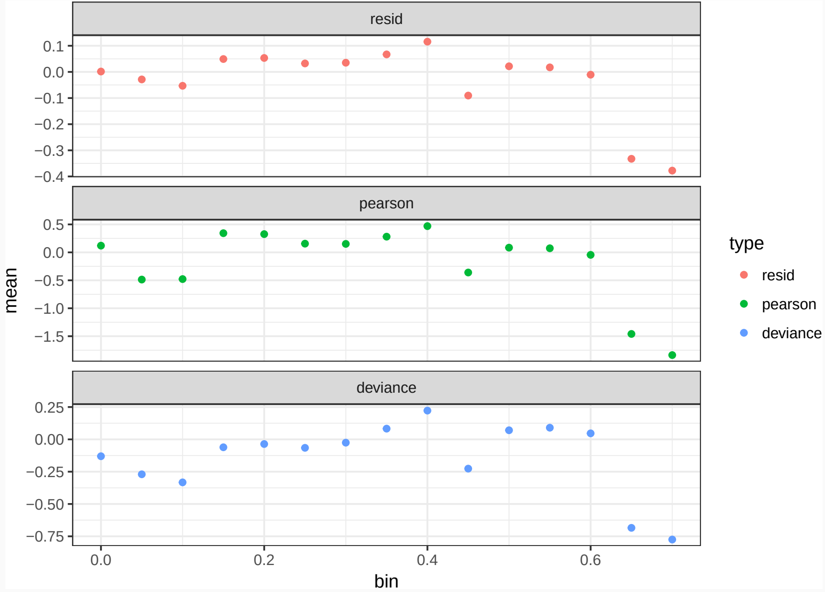
Binned Deviance Residuals



Checking Deviance

```
sum(d_g$deviance^2)
## [1] 479.3914

glm(presence~SegSumT, family=binomial, data=anguilla)
##
## Call:  glm(formula = presence ~ SegSumT, family = binomial, data = anguil
##
## Coefficients:
## (Intercept)      SegSumT
##    -16.7418       0.9001
##
## Degrees of Freedom: 616 Total (i.e. Null);  615 Residual
## Null Deviance:       621.9
## Residual Deviance: 479.4     AIC: 483.4
```

All together

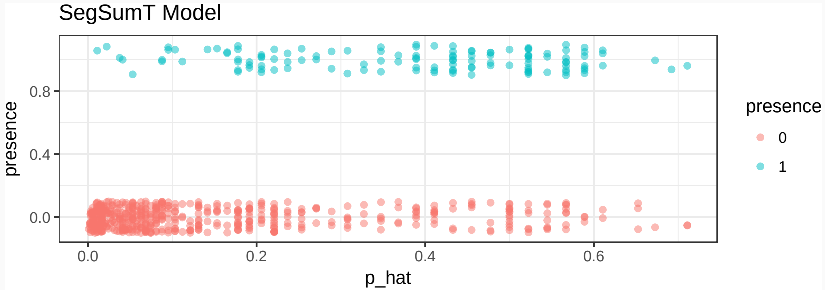


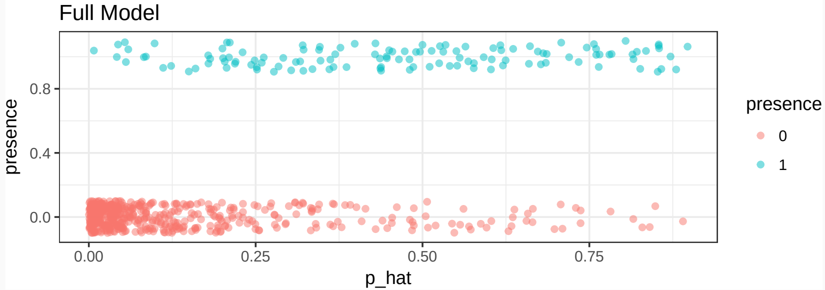
Full Model

Model

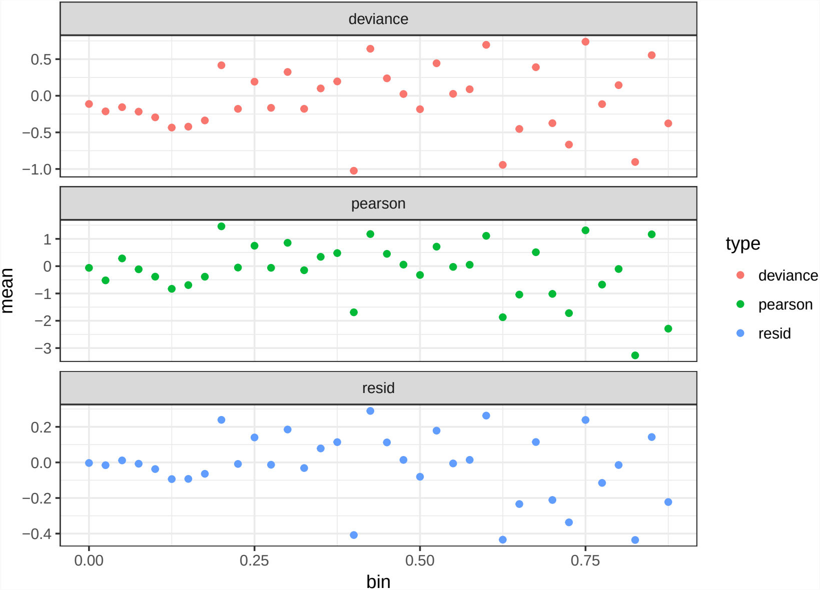
```
f = glm(presence~., family=binomial, data=anguilla)
summary(f)
##
## Call:
## glm(formula = presence ~ ., family = binomial, data = anguilla)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.10254  -0.53092  -0.27156  -0.08821   3.12463
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -11.554287   1.872102  -6.172 6.75e-10 ***
## SegSumT         0.765864   0.103173   7.423 1.14e-13 ***
## DSDist         -0.002551   0.002103  -1.213  0.22523
## DSMaxSlope     -0.062525   0.063093  -0.991  0.32169
## USRainDays     -0.619025   0.227316  -2.723  0.00647 **
## USSlope        -0.041399   0.024657  -1.679  0.09315 .
## USNative       -0.607045   0.475456  -1.277  0.20169
## DSDam          -0.922073   0.483492  -1.907  0.05651 .
## Methodmixture  -0.231175   0.498189  -0.464  0.64263
## Methodnet      -1.229762   0.534845  -2.299  0.02149 *
## Methodspo      -1.493876   0.733468  -2.037  0.04168 *
## Methodtrap     -2.476408   0.628486  -3.940 8.14e-05 ***
## LocSed         -0.175944   0.098204  -1.792  0.07319 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 621.91  on 616  degrees of freedom
## Residual deviance: 420.18  on 604  degrees of freedom
## AIC: 446.18
##
## Number of Fisher Scoring iterations: 6
```

Separation

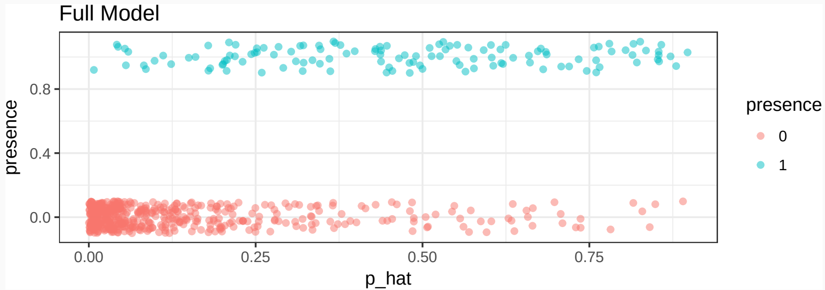

SegSumT Model
presence
0.8
0.4
0.0
0.0
0.2
0.4
0.6
p_hat
presence
0
1


Full Model
presence
0.8
0.4
0.0
0.00
0.25
0.50
0.75
p_hat
presence
0
1

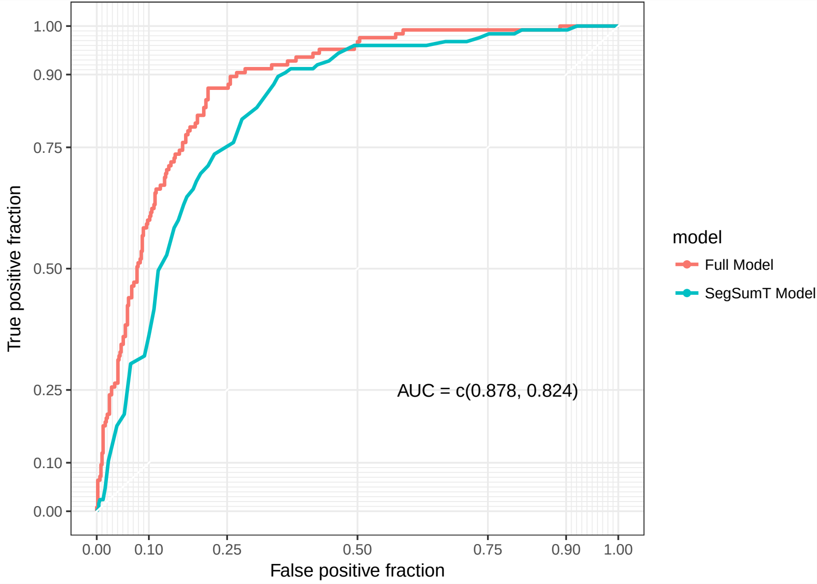
Residuals vs fitted



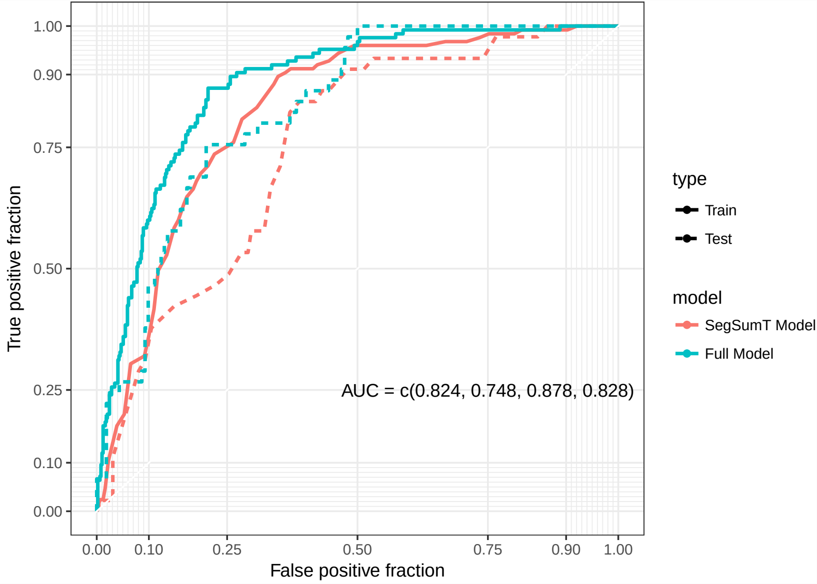
Model Performance


Full Model
presence
0.8
0.4
0.0
0.00
0.25
0.50
0.75
p_hat
presence
0
1

Predictive Performance (ROC / AUC)



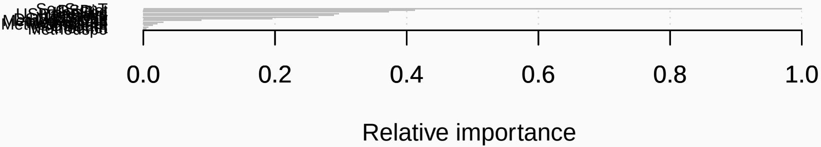
Out of sample predictive performance



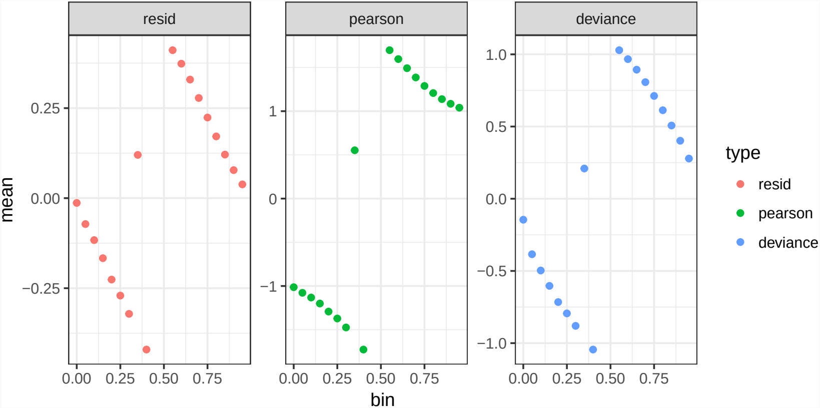
What about something non-parametric?

Gradient Boosting Model

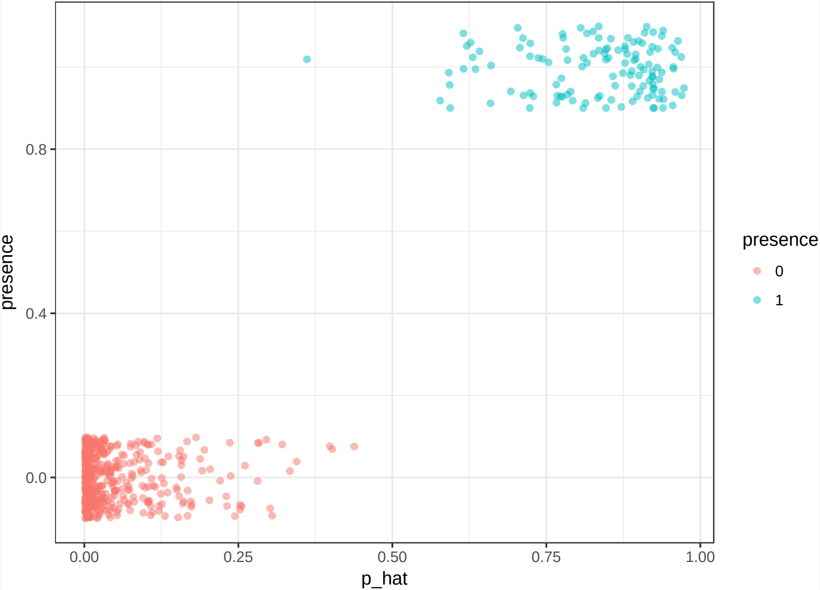
```r
y = anguilla$presence %>% as.integer()
x = model.matrix(presence~.-1, data=anguilla)
x_test = model.matrix(presence~.-1, data=anguilla_test)

xg = xgboost::xgboost(data=x, label=y, nthead=4, nround=25,
                      objective="binary:logistic", verbose = FALSE)
```



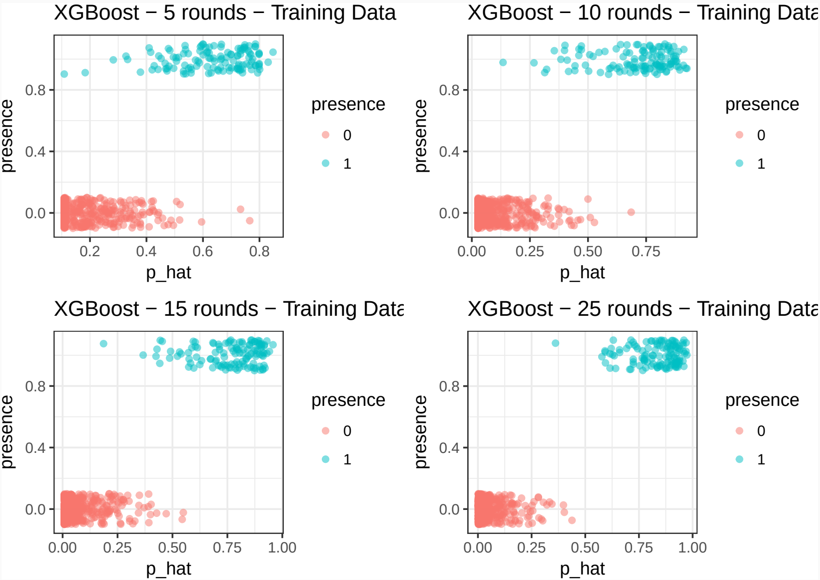
Residuals?


resid
pearson
deviance
mean
bin
0.25
0.00
−0.25
1
0
−1
1.0
0.5
0.0
−0.5
−1.0
0.00
0.25
0.50
0.75
type
resid
pearson
deviance

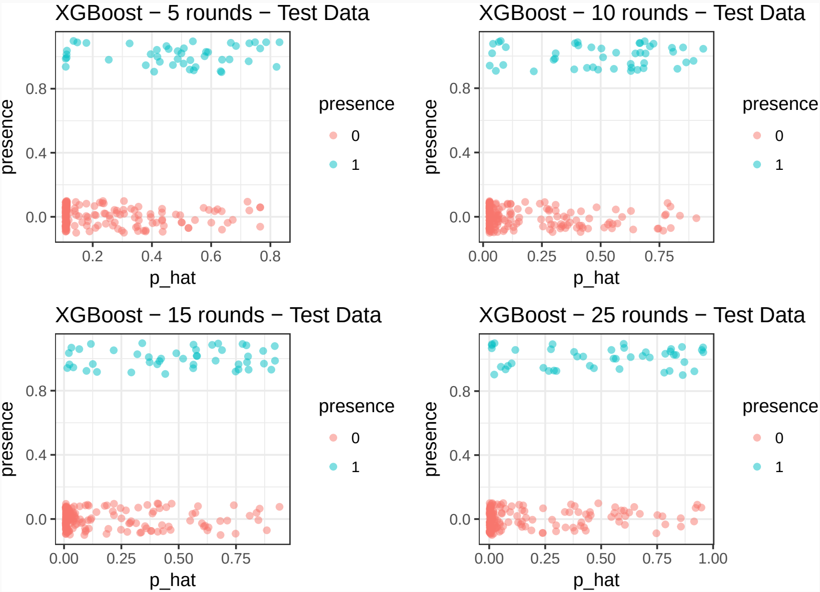
Separation?



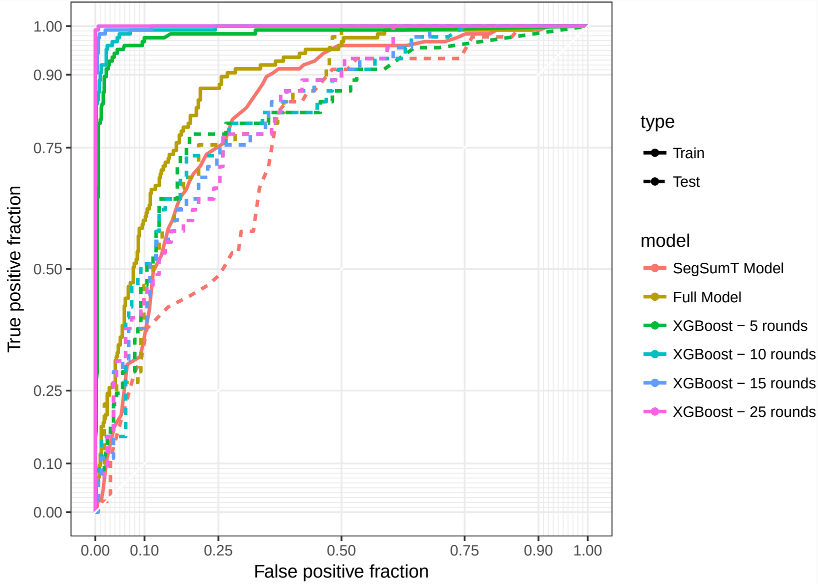
Effect of `nround` - Training Data



Effect of `nround` - Test Data



ROC Curves



True positive fraction
False positive fraction
0.00
0.10
0.25
0.50
0.75
0.90
1.00
type
Train
Test
model
SegSumT Model
Full Model
XGBoost − 5 rounds
XGBoost − 10 rounds
XGBoost − 15 rounds
XGBoost − 25 rounds

Aside: Species Distribution Modeling

Model Choice

We have been fitting a model that looks like the following,

$$
\begin{aligned}
y_i &\sim \text{Bern}(p_i) \\
\text{logit}(p_i) &= \mathbf{X}_{i\cdot}\boldsymbol{\beta}
\end{aligned}
$$

Interpretation of y_i and p_i?

Absence of evidence ...

If we observe a species at a particular location what does that tell us?

If we *don't* observe a species at a particular location what does that tell us?

If we allow for crypsis, then

$$
\begin{aligned}
y_i &\sim \text{Bern}(q_i\, z_i) \\
z_i &\sim \text{Bern}(p_i) \\
\text{logit}(q_i) &= \mathbf{X}_{i\cdot}\boldsymbol{\gamma} \\
\text{logit}(p_i) &= \mathbf{X}_{i\cdot}\boldsymbol{\beta}
\end{aligned}
$$

Interpretation of y_i, z_i, p_i, and q_i?