

Lecture 4

Logistic Regression + Residual Analysis

Colin Rundel

1/29/2018

HW 1

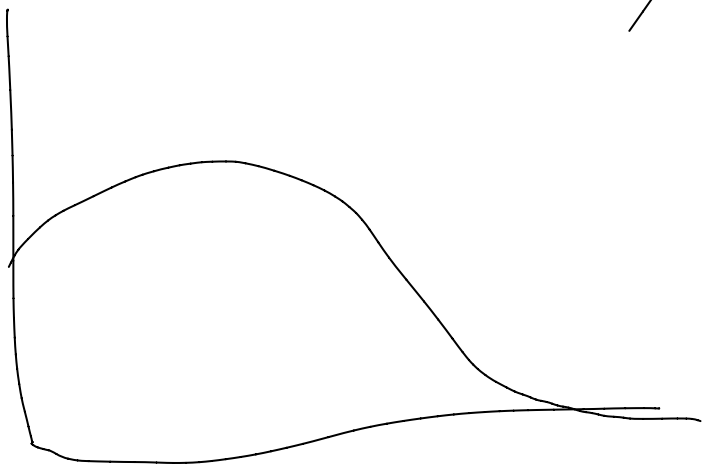
Q3

$$y^* \sim N(x\beta, \tau)$$

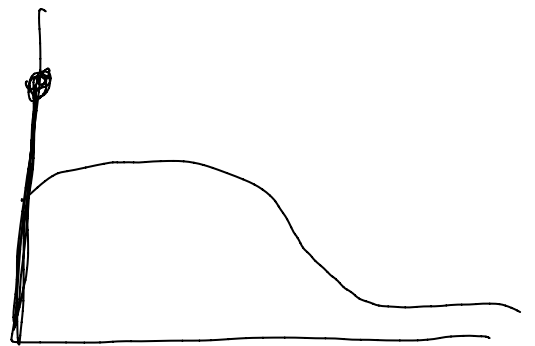
$$y = \max(y^*, 0)$$

Fails Jass

~~$y \sim \text{dnorm}(x^*, \tau) \cdot \mathbb{1}(0,)$~~



vs



Lecture 4

Logistic Regression + Residual Analysis

Colin Rundel

~~11/20/2018~~

9/12/2018

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

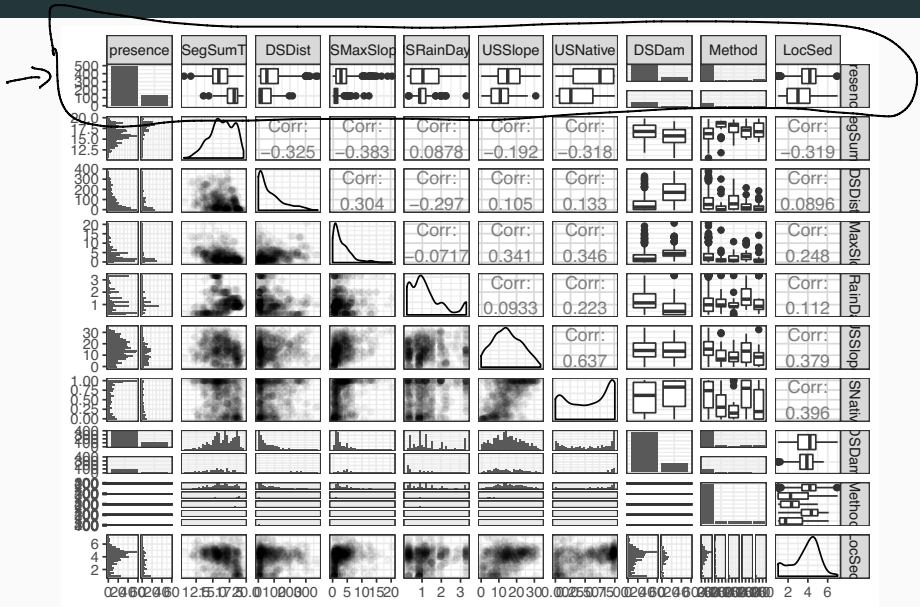


Codebook:

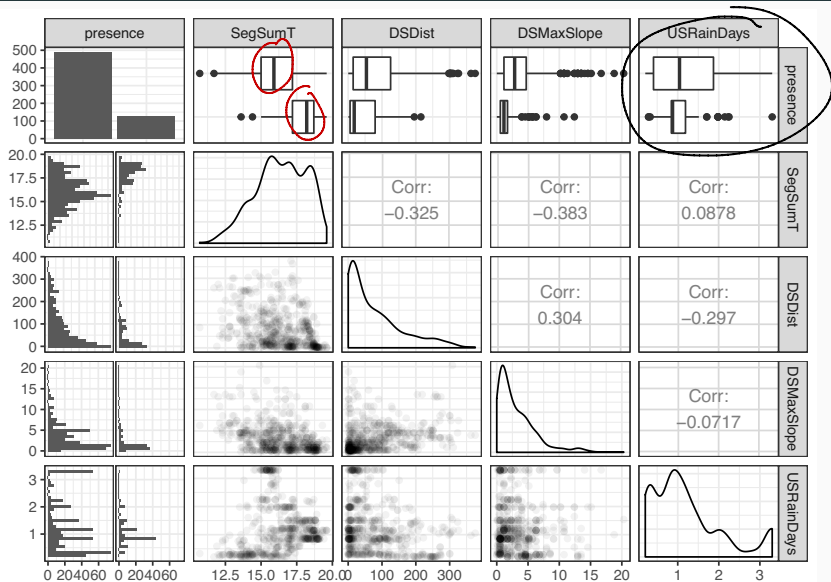
- **presence** - presence (1) or absence (0) of *Anguilla australis* at the sampling location
- **SegSumT** - Summer air temperature (degrees C)
- **DSDist** - Distance to coast (km)
- **DSMaxSlope** - Maximum downstream slope (degrees)
- **USRainDays** - days per month with rain greater than 25 mm
- **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

```
anguilla
## # A tibble: 617 x 10
##   presence SegSumT DSDist DSMaxSlope USRainDays USSlope USNative DSDam
## *   <int>   <dbl> <dbl>     <dbl>     <dbl>     <dbl> <dbl> <dbl> <int>
## 1     1    18.7  133.     1.15     1.15     8.3   0.34    0
## 2     0    18.3  107.     0.570    0.847    0.4   0       0
## 3     0    16.7  167.     1.72     0.21     0.4   0.22    1
## 4     0    15.1  11.2     1.72     3.3     25.7   1       0
## 5     0    12.7  42.4     2.86     0.43     9.6   0.09    0
## 6     1    18.2  94.4     3.43     0.847   20.5   0.92    0
## 7     1    18.3  91.9     1.72     0.861    6.7   0.580   1
## 8     1    17.1   6.8     0.52     0.62     0.7   0       0
## 9     0    13.4  190.     3.43     0.77    20.1   0.99    0
## 10    0    13.1  224.     6.84     0.290    9.8   0.98    0
## # ... with 607 more rows, and 2 more variables: Method <fct>, LocSed <dbl>
```

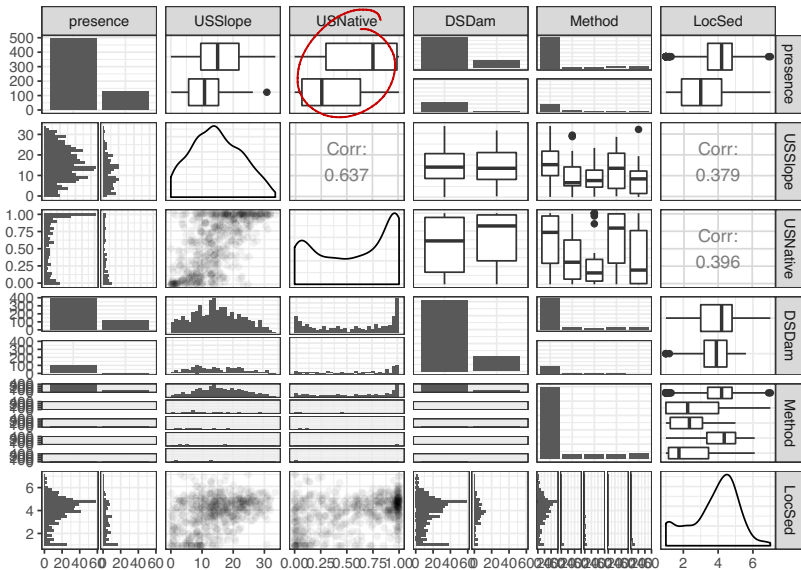
EDA



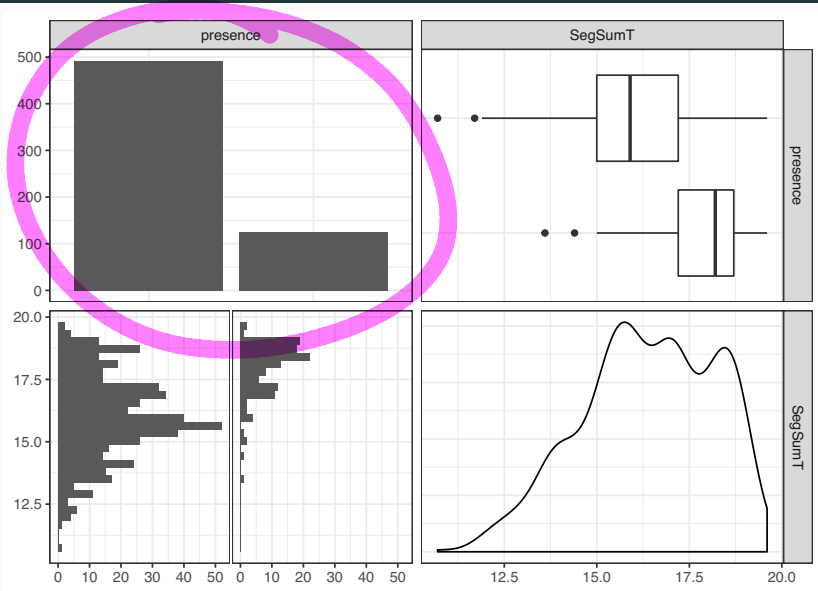
EDA (part 1)



EDA (part 2)



EDA (part 3)



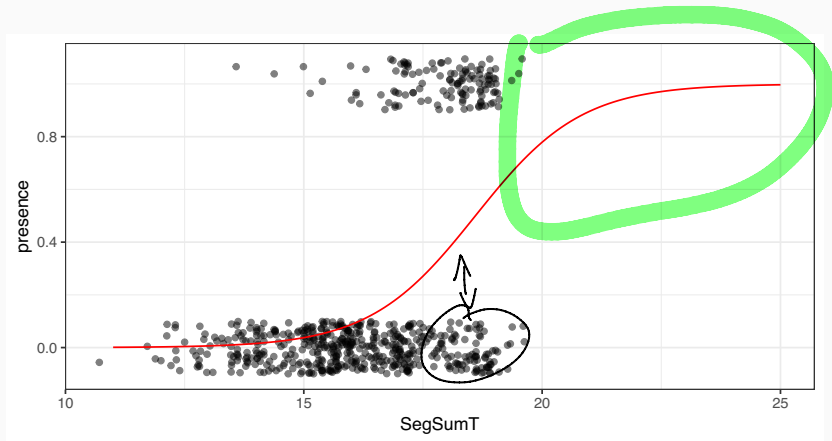
Simple Model

Model

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

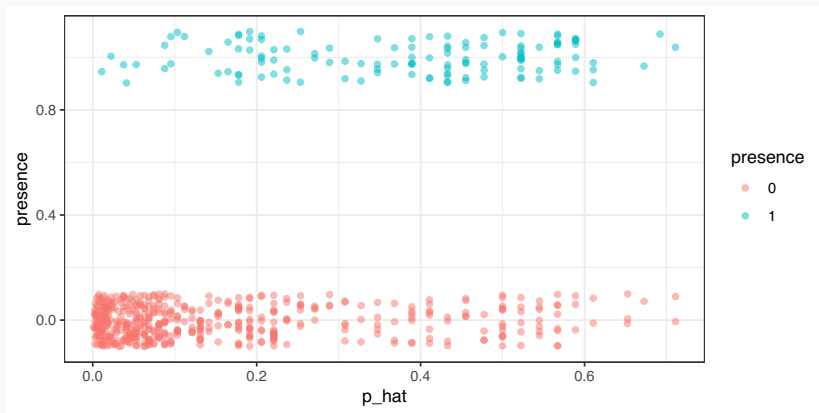
g = glm(presence~SegSumT, family=binomial, data=anguilla)
summary(g)
##
## Call:
## glm(formula = presence ~ SegSumT, family = binomial, data = anguilla)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.5755  -0.6260  -0.3452  -0.1299   3.0039
##
## Coefficients:
##              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
```

```
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

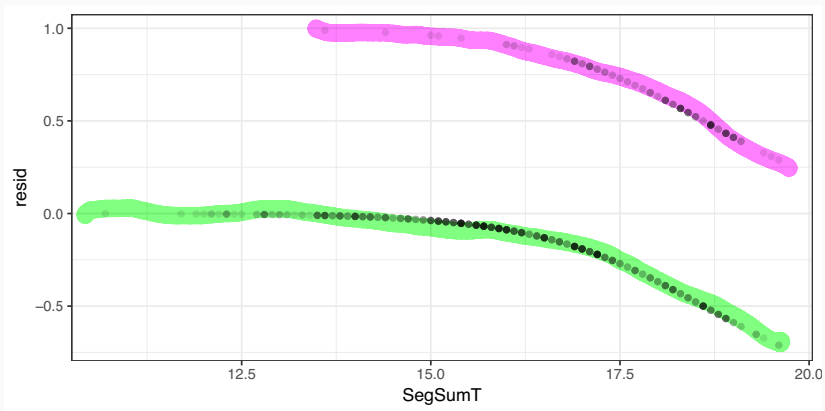
```
ggplot(d_g, aes(x=p_hat, y=presence, color=as.factor(presence))) +  
  geom_jitter(height=0.1, alpha=0.5) +  
  labs(color="presence")
```



Residuals

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

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

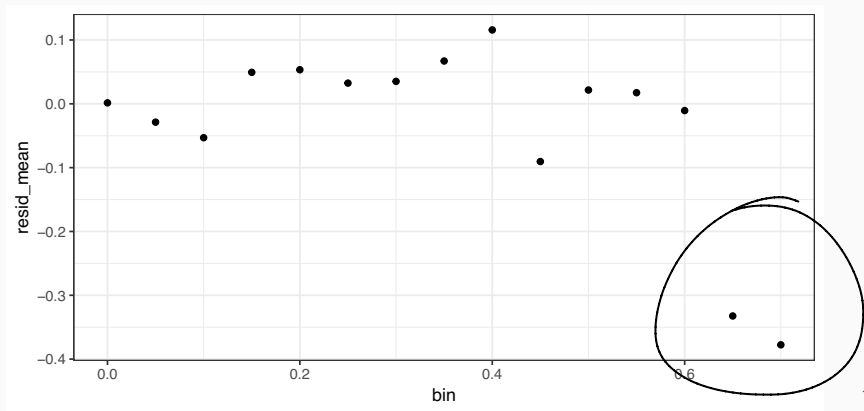


Binned Residuals

```
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()
```

$$E(Y_i) = p_i$$

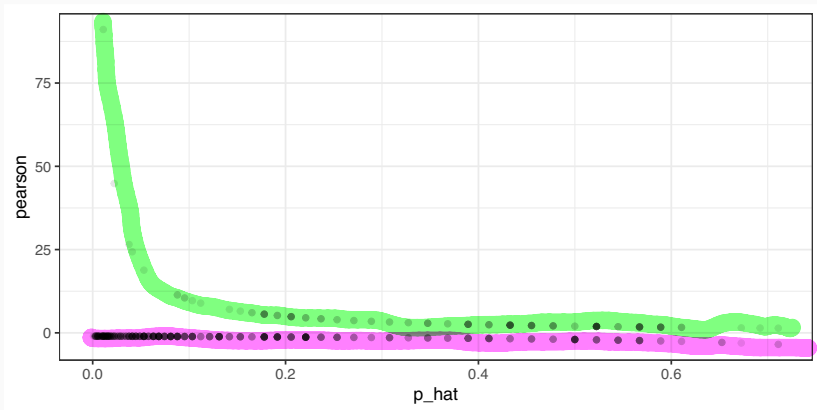
$$\text{Var}(Y_i) = p_i(1-p_i)$$



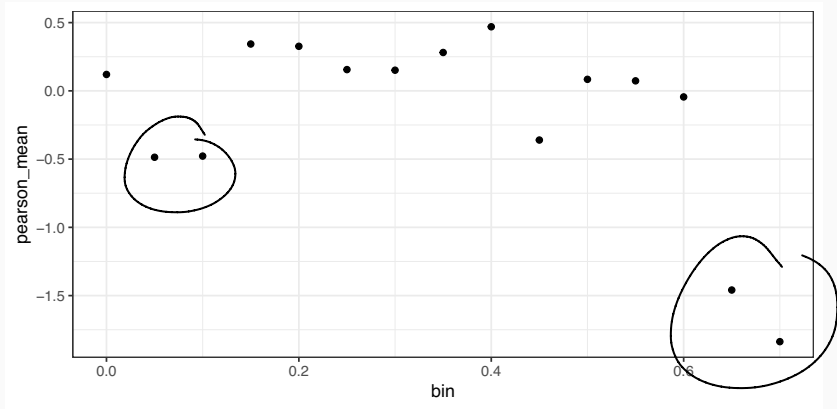
Pearson Residuals

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

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



Binned Pearson Residuals

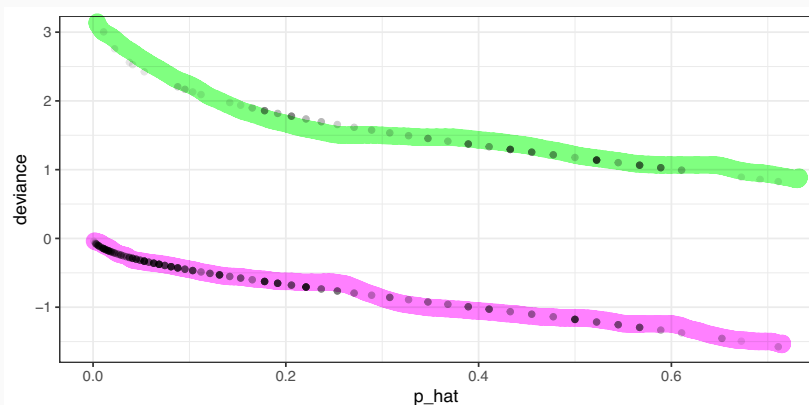


Deviance Residuals

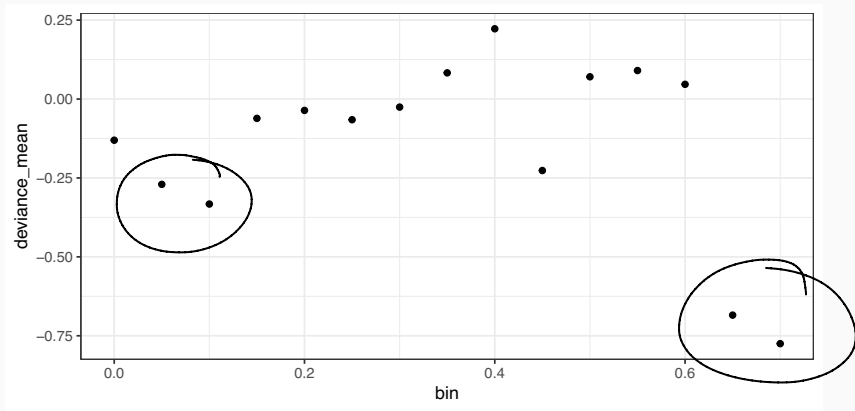
$$d_i = \text{sign}(Y_i - \hat{p}_i) \sqrt{-2 (Y_i \log \hat{p}_i + (1 - Y_i) \log(1 - \hat{p}_i))}$$

```
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 = anguilla)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)      SegSumT
```

```
##    -16.7418         0.9001
```

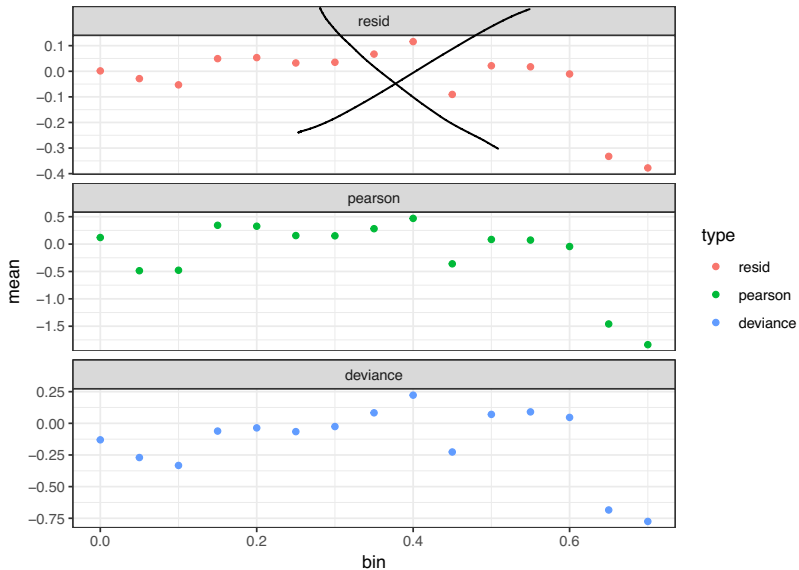
```
##
```

```
## Degrees of Freedom: 616 Total (i.e. Null); 615 Residual
```

```
## Null Deviance:      621.9
```

```
## Residual Deviance: 479.4    AIC: 483.4
```

All together



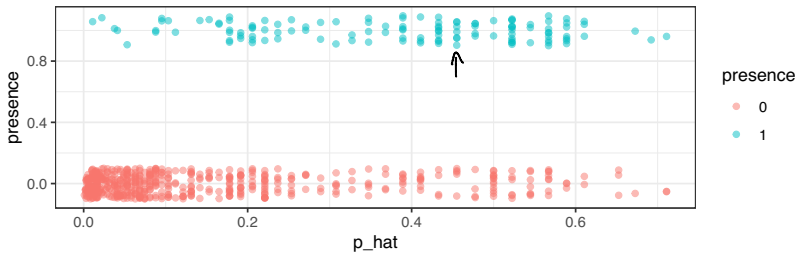
Full Model

Model

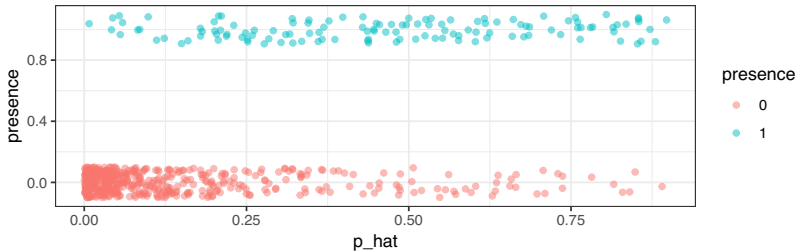
```
f = glm(presence ~ ., family=binomial, data=anguilla)
summary(f)
##
## Call:
## glm(formula = presence ~ ., family = binomial, data = anguilla)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10254  -0.53092  -0.27156  -0.08821   3.12463
##
## Coefficients:
##              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

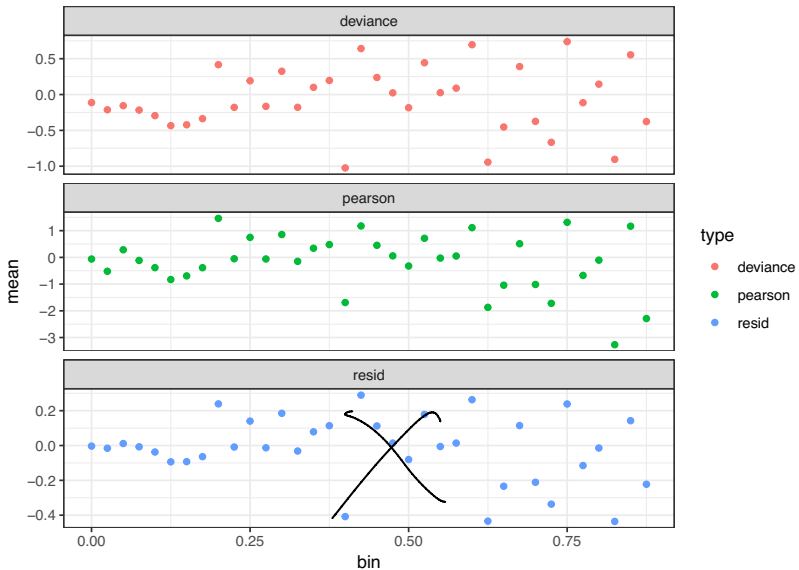
SegSumT Model



Full Model

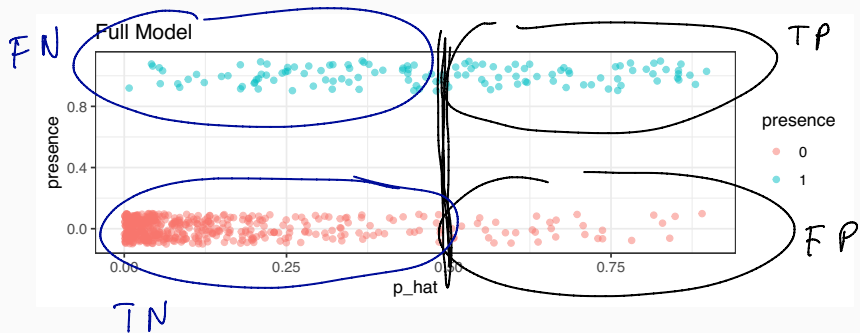


Residuals vs fitted



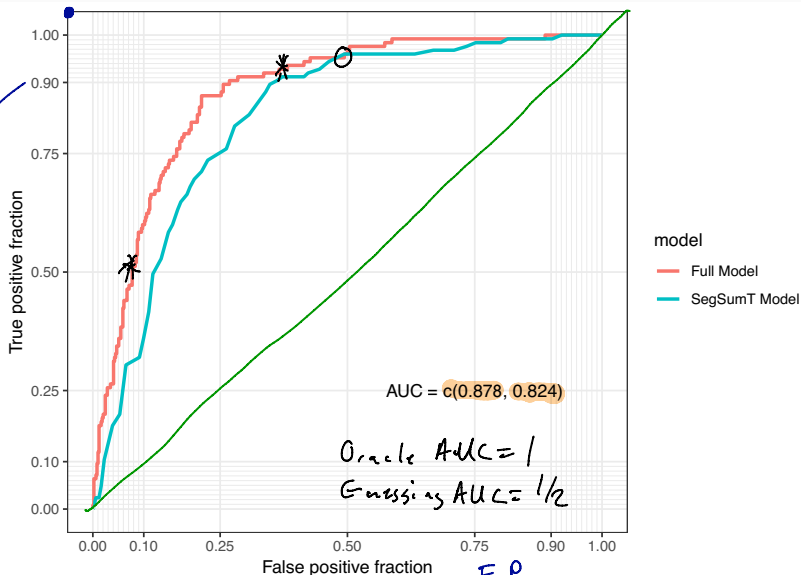
Model Performance

Confusion Tables



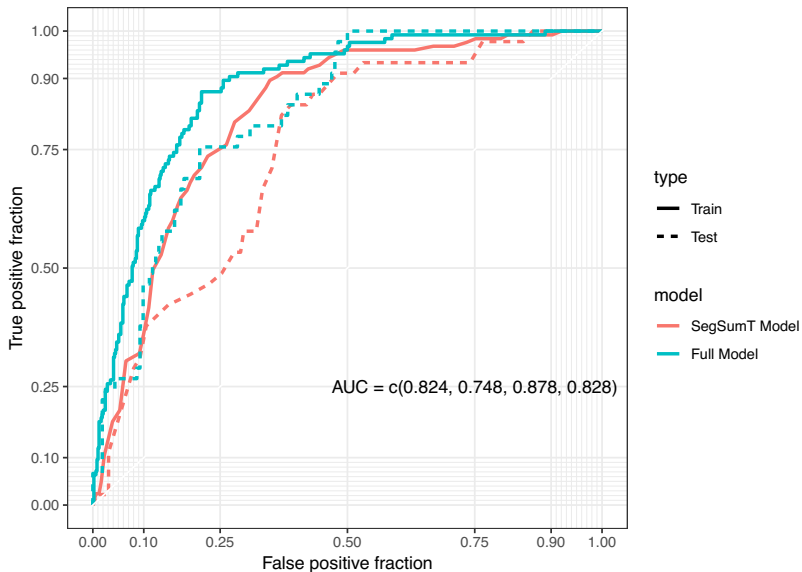
Predictive Performance (ROC / AUC)

TP



FP

Out of sample predictive performance

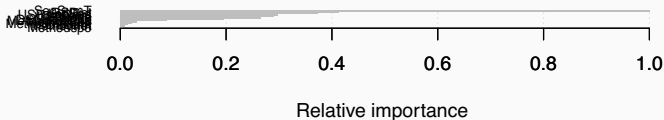


What about something
non-parametric?

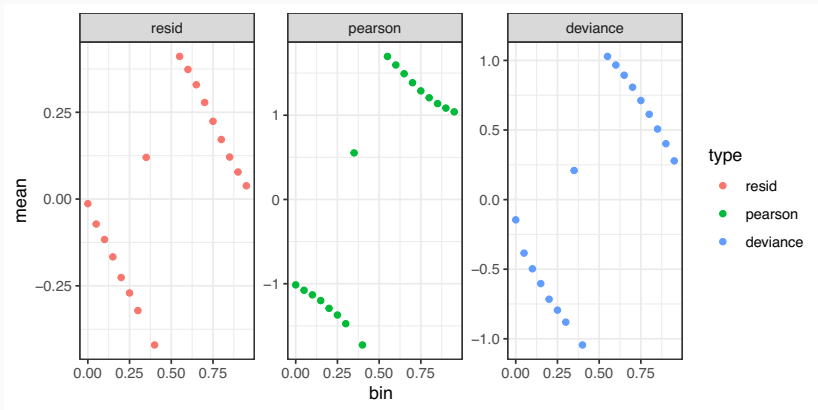
Gradient Boosting Model

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

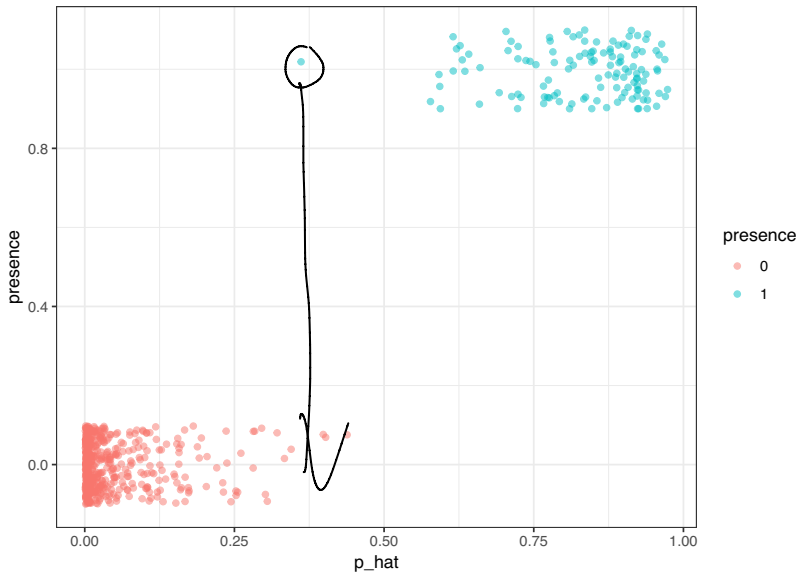
xg = xgboost::xgboost(data=x, label=y, nthread=4, nround=25,
                     objective="binary:logistic", verbose = FALSE)
```



Residuals?

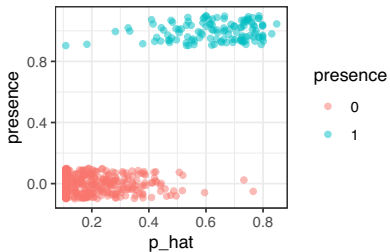


Separation?

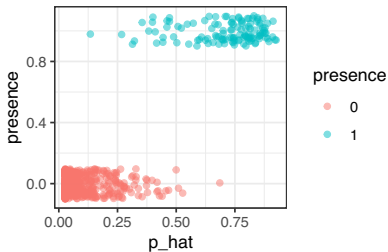


Effect of nround - Training Data

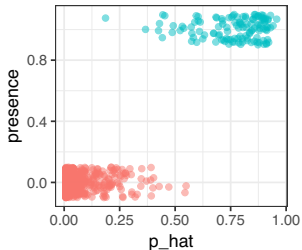
XGBoost - 5 rounds - Training Data



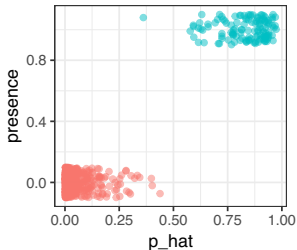
XGBoost - 10 rounds - Training Data



XGBoost - 15 rounds - Training Data

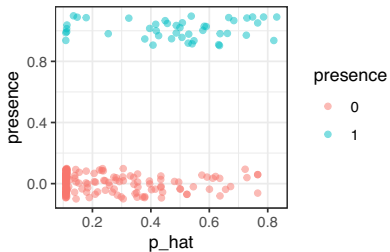


XGBoost - 25 rounds - Training Data

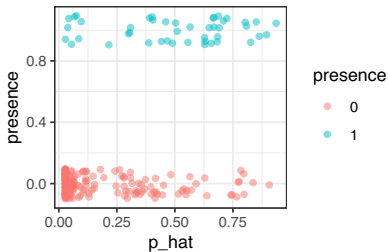


Effect of nround - Test Data

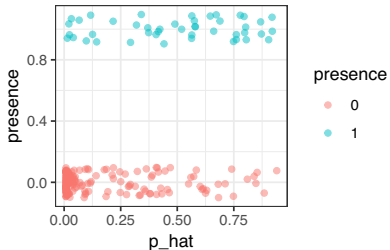
XGBoost – 5 rounds – Test Data



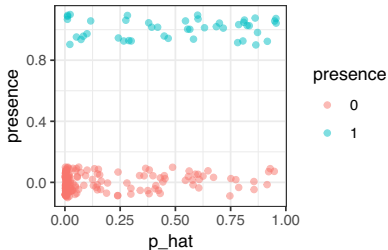
XGBoost – 10 rounds – Test Data



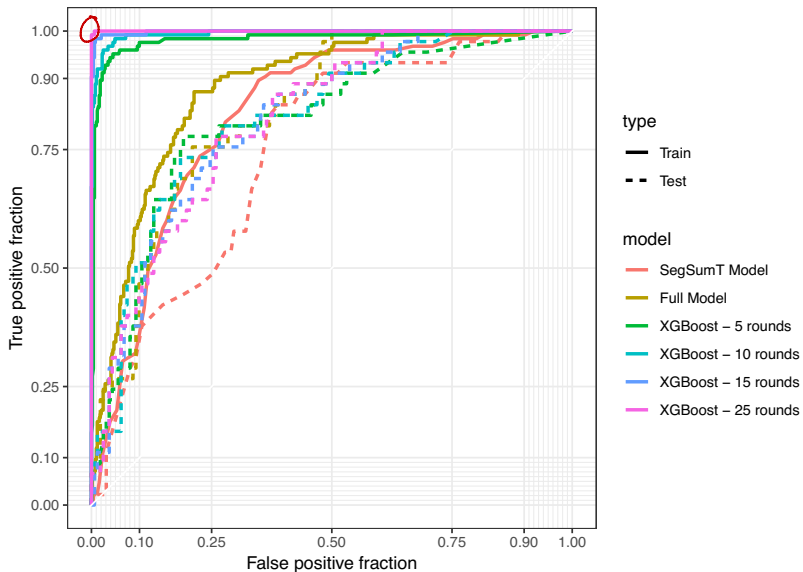
XGBoost – 15 rounds – Test Data



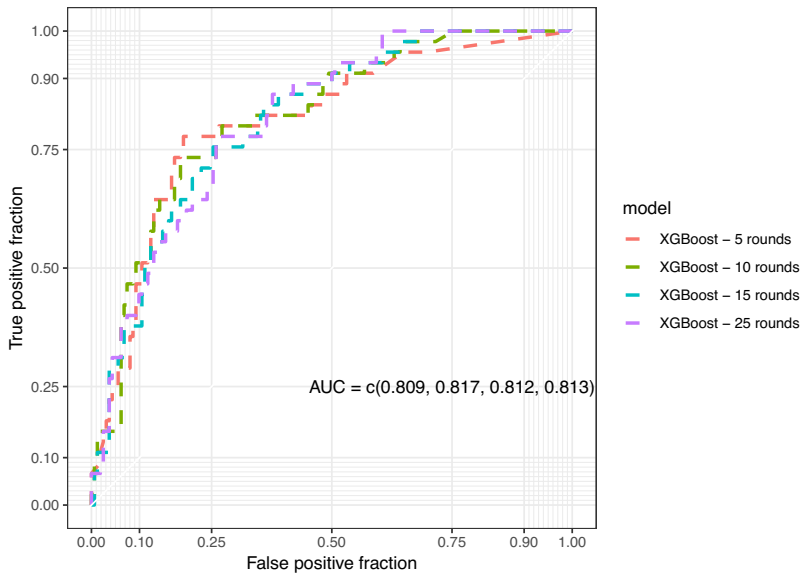
XGBoost – 25 rounds – Test Data



ROC Curves



ROC Curves (test only)



Aside: Species Distribution Modeling

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

$$y_i \sim \text{Bern}(p_i)$$
$$\text{logit}(p_i) = \mathbf{X}_i \cdot \boldsymbol{\beta}$$

Interpretation of y_i and p_i ?

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

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

$$y_i \sim \text{Bern}(q_i z_i)$$

$$z_i \sim \text{Bern}(p_i)$$

$$\text{logit}(q_i) = \mathbf{X}_i \cdot \boldsymbol{\gamma}$$

$$\text{logit}(p_i) = \mathbf{X}_i \cdot \boldsymbol{\beta}$$

Interpretation of y_i , z_i , p_i , and q_i ?