

STAT 420 - Final Project

Jonathan Bowen

2025-12-07

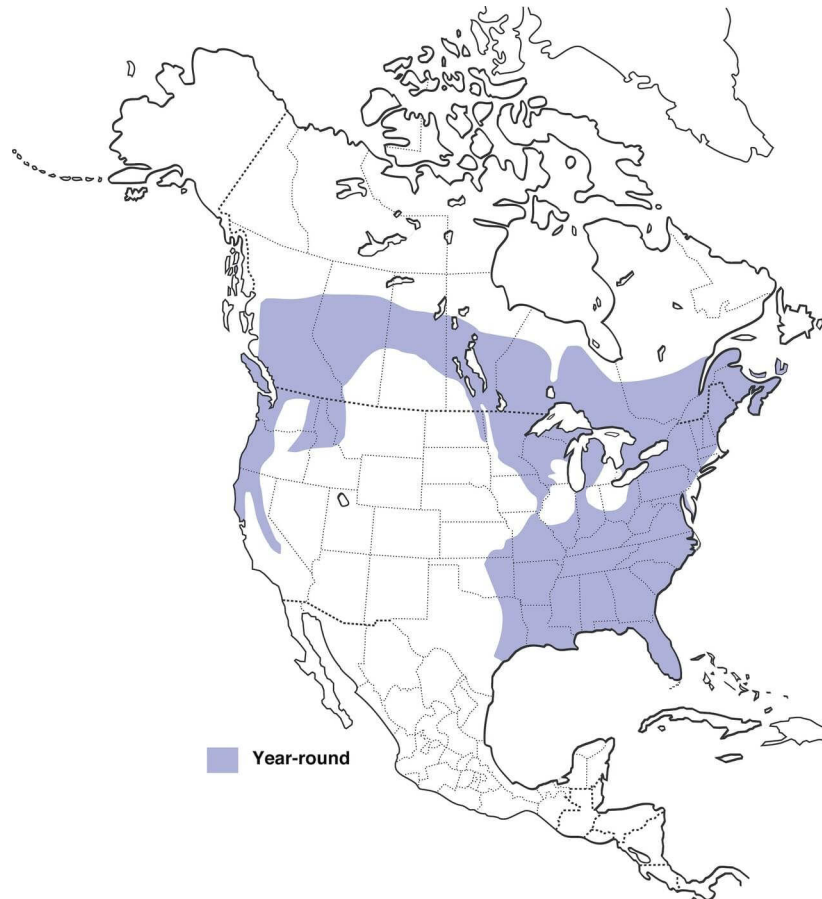
Contents

Introduction	2
Data Curation	3
Filtering eBird Data	3
Aggregation	4
Effort variables from eBird checklists	5
Habitat Covariates	5
Final Feature Engineering	6
Data Cleaning	8
Methods	9
Exploratory data analysis	9
Full/base model interpretation (no poly or interaction terms)	14
Diagnostics	16
Improvements	18
Conclusion	21
Appendix	22
Aggregation	22
EDA	22
Improvements and Model Selection	31
Citation for eBird data use:	37

Introduction

My goal was to gain insight into how environmental/habitat factors impact detection rates of a specific bird species, the pileated woodpecker.

The below map shows the pileated woodpecker's range in North America [source]:



I used a composite dataset for my project. The primary source of my data is the Cornell Lab of Ornithology's eBird Basic Dataset, sometimes called "EBD". As described on the eBird website, "eBird is among the world's largest biodiversity-related science projects" and "eBird data document bird distribution, abundance, habitat use, and trends through checklist data collected within a simple, scientific framework." eBird makes available the full dataset of observations upon request.

Each entry/row of the eBird dataset includes fields/columns for species of bird observed, number of individuals observed, date, time, location, and many more.

I then compiled habitat-related data from several other data sources:

- **Elevation** data using the `elevatr` R package from the US EPA
- **Distance to the nearest coast** from the NASA Ocean Biology Processing Group
- **Land-cover classifications** from the National Land Cover Database
- Annual **mean temperature** and **cumulative precipitation** using the `prism` package

Data Curation

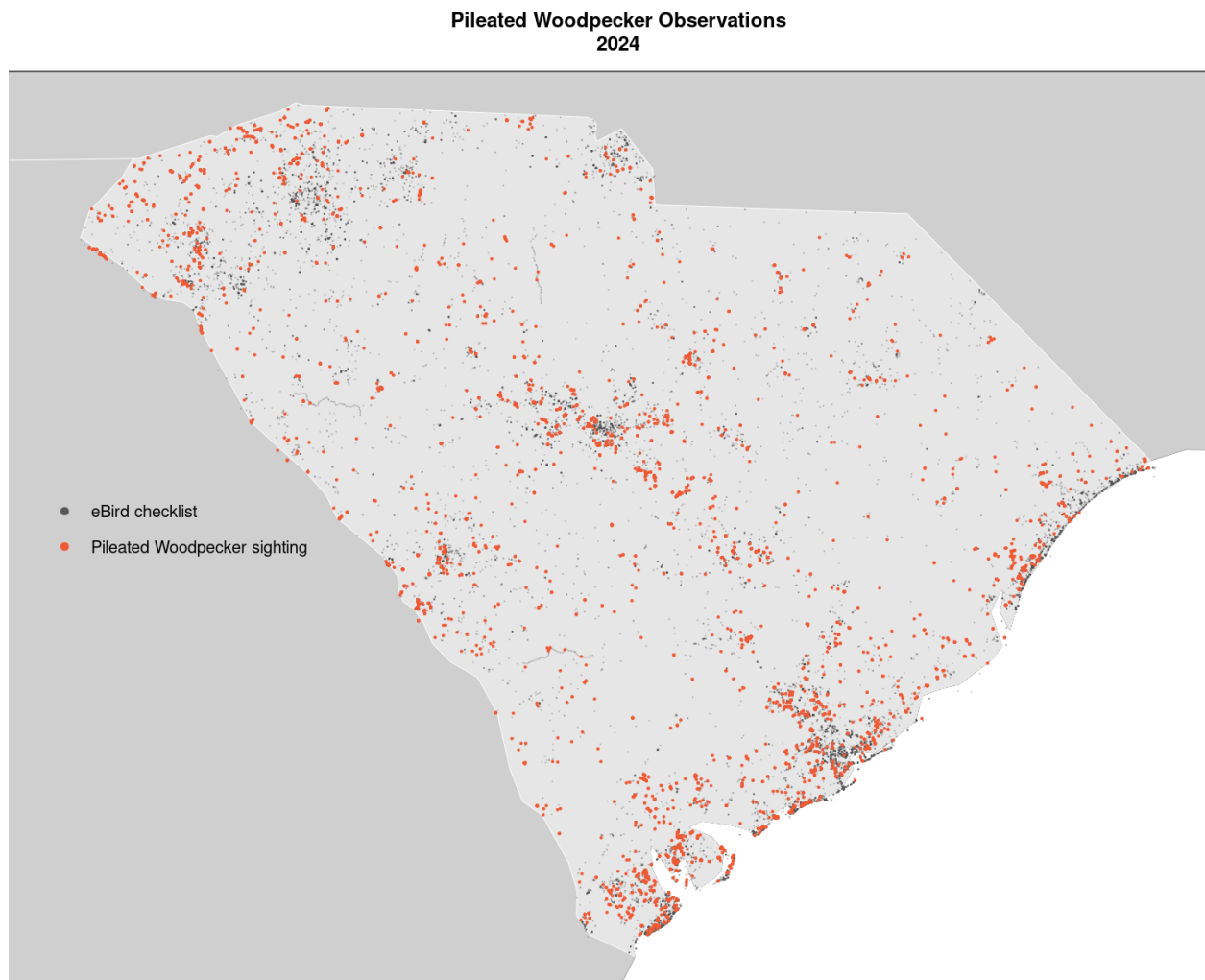
I adapted my approach from the eBird Best Practices Workshop notes and the Best Practices for Using eBird Data eBook, both written or contributed to by Matthew Strimas-Mackey, a lead contributor to the `auk` package I used to filter and pre-process eBird data. I followed the preprocessing guidance and read through the inference and modeling sections, gleaned what I could apply to a linear regression model. A unique property of eBird checklist data is the ability to infer null samples by allowing user to assert non-detections, discussed more below.

Filtering eBird Data

The current version of the eBird Basic Dataset contains 5,656,350 observations (any bird, all time, anywhere). I chose to limit the analysis to my home state of South Carolina and the year 2024 to reduce the dataset to a manageable size while hopefully including enough variation in habitat, weather, and topology. Filtered eBird checklist count (detections and non-detections): 56,168

I imposed some restrictions to checklists included in the analysis to those that were collected over a period of more than 15 minutes and less than 5 hours. I also limited checklists to those with 5 or fewer observers, 10 or fewer km traveled, and speed < 100 kph (I guess people in cars?). These types of filters were recommended in the eBird guide to reduce variability and remove outliers.

Here are detections (red) and non-detections (grey) locations plotted on a map of the state:

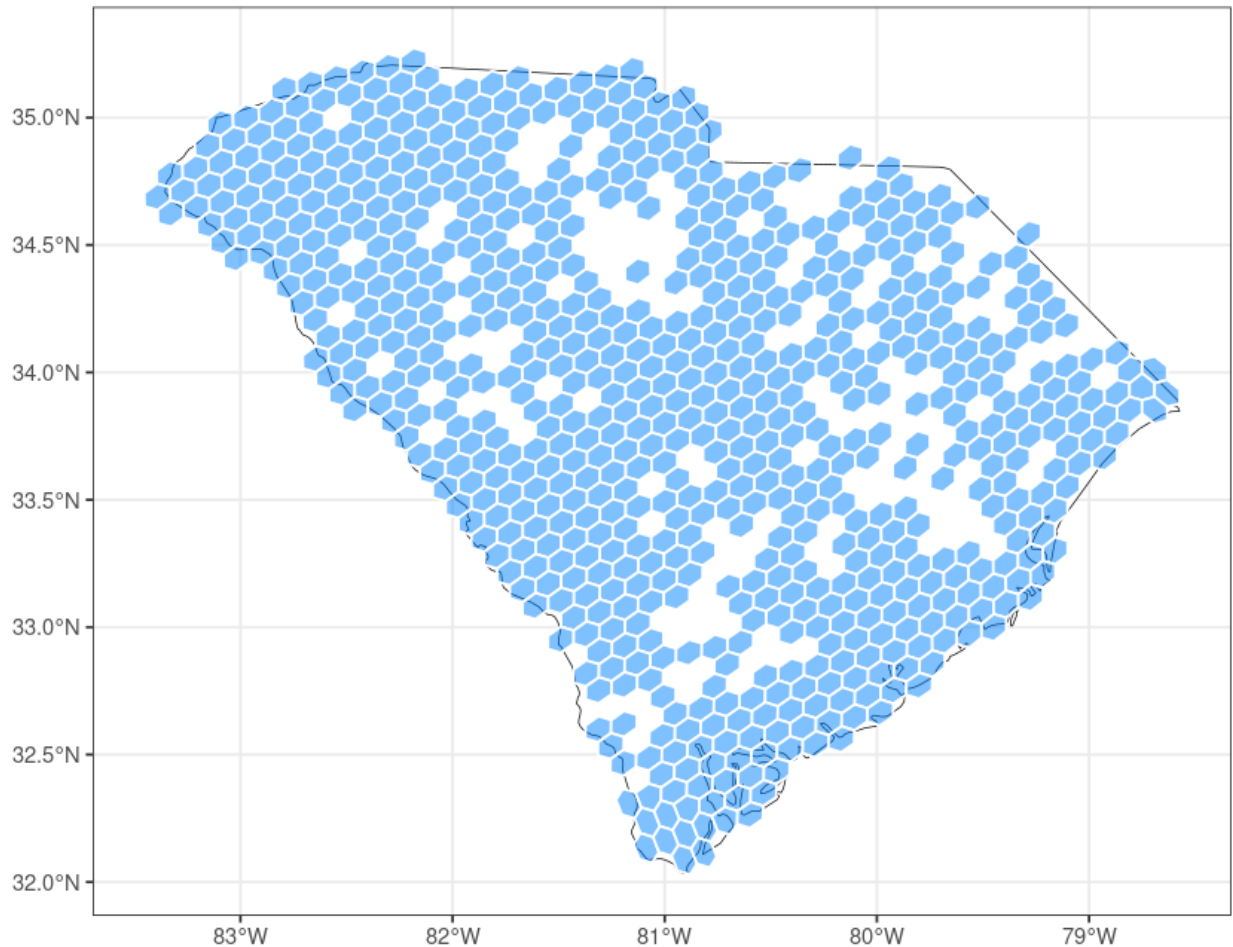


Aggregation

I can't use individual checklists as the observations because they don't have a useful (numerical) response variable (discussed more in the Appendix). I could build a logistic-regression classifier using success as >0 individuals observed, but I need a numeric response variable. Number of individuals detected on a checklist is “drawing the lines in the wrong place” - one person observing 10 birds will send a different signal to the model than 10 people each observing one bird. I want my model predictions to reflect species occupancy/abundance rather than just checklist “score”. Predicting checklist score would also not control for location popularity, making correlation with habitat variables less useful. More popular locations would tend to have better checklist scores than less popular places. To get the dataset into a form conducive to a regression analysis, I aggregated over space and time.

I segmented the state into 10-km hex cells using the `dggridR` package. Looking at the grid overlaid on the state, this seems like a good compromise to provide sufficient data reduction without completely smoothing out local variation.

There are “missing” cells where either no complete checklists were recorded in 2024, or I had to omit that cell due to lack of one or more predictor values available for it (discussed in the Data Cleaning section below). Total hex cell count: 732 hex cells



All the checklists from 2024 that fall into a given cell are collected together, and summary statistics are generated to use as that cell's contribution to the dataset. The hex cells become the “rows” in the dataset I'll perform regression on. Then the response becomes the “detection frequency” - how many complete checklists

in 2024 in this area (hex cell) included an observation (any number) of the species of interest.

Effort variables from eBird checklists

At the recommendation of the eBird Best Practices documentation, I included so-called “effort” variables to account for observers’ variation in probability of seeing a species that is present. Here are the effort variables I included, along with their description from the eBird documentation:

- **Number of observers:** “The total number of observers participating the sampling event”
I used the mean number of observers as the summary statistic for each hex cell.
- **Start time:** “The time observations were initiated” (hours after midnight)
I used the mean start time as the summary statistic for each hex cell.
- **Duration:** “The duration of the sampling event.” (hours)
I used the median duration as the summary statistic for each hex cell.
- **Distance traveled:** “The distance traveled during the sampling event” (km)
I used the mean distance traveled as the summary statistic for each hex cell.

These effort variables allow my model to account for locations where birders tend to “put in more effort”. I’d expect a hex cell in a state park to have a higher number of observers, a lower start time, longer duration, and longer distance traveled than a hex cell filled with strip malls.

Habitat Covariates

Instead of using raw GPS coordinates as predictors, I wanted to use more direct and specific locational attributes as predictors, ideally capturing habitat-relevant characteristics. Here are the covariates I chose:

- **Elevation**
I used the `elevatr` R package from the US EPA to translate the each hex cell’s center latitude and longitude coordinates into elevation above sea level, to serve as a more interesting numerical predictor. I expect this to be constant throughout 2024.
- **Distance from the coast** (obviously important for shorebirds, but maybe it matters for other species too?)
I used the NASA Ocean Biology Processing Group’s Distance to the Nearest Coast dataset to associate each hex cell’s location (lat/long) with distance from the nearest coast. This dataset is a text file NASA makes available for download.
I expect this to be constant throughout 2024.
- **Land cover classification**
This is something I wasn’t familiar with when I started this project. It is basically the landscape characterization (forest, wetland, developed, grassland, etc.). I downloaded the 2024 Land Cover (CONUS) data product from this page on the Multi-Resolution Land Characteristics (MRLC) Consortium site. These data are often called the National Land Cover Database (NLCD), and the USGS contributes to them as well.
Land-cover data are captured by both by field survey and analysis of remotely sensed imagery. [Wikipedia]
I used the downloaded land-cover raster to determine for each hex cell, what are the proportions of each land-cover type (so it became several variables). I did some further processing on these data, explained in the Feature Engineering section below.
I am making the assumption that this is constant throughout 2024, although for some hex cells, that won’t be true (land gets developed, ecology changes).
- **Annual mean temperature**
I used the `prism` package to download and employ rasters from the Northwest Alliance for Computational Science and Engineering (NACSE), which is part of Oregon State University.
According to their website, “To create a continuous weather map across the country, our automated

systems collect weather station observations from tens of thousands of reporting stations per day. These observations are fed into a sophisticated computer model called PRISM (Parameter-elevation Regressions on Independent Slopes Model)."

For each hex cell, I collected all associated pixels from the annual-mean-temp raster and calculated the average temp for that cell over 2024.

This variable is pretty coarse since it's averaging over the whole year (through seasons), but I still saw plenty of variability (See Exploratory Data Analysis section below).

- **Annual cumulative precipitation**

This is also acquired from the PRISM data.

For each hex cell, I collected all associated pixels from the annual-precipitation raster and averaged to create the total precipitation for that cell in 2024.

Final Feature Engineering

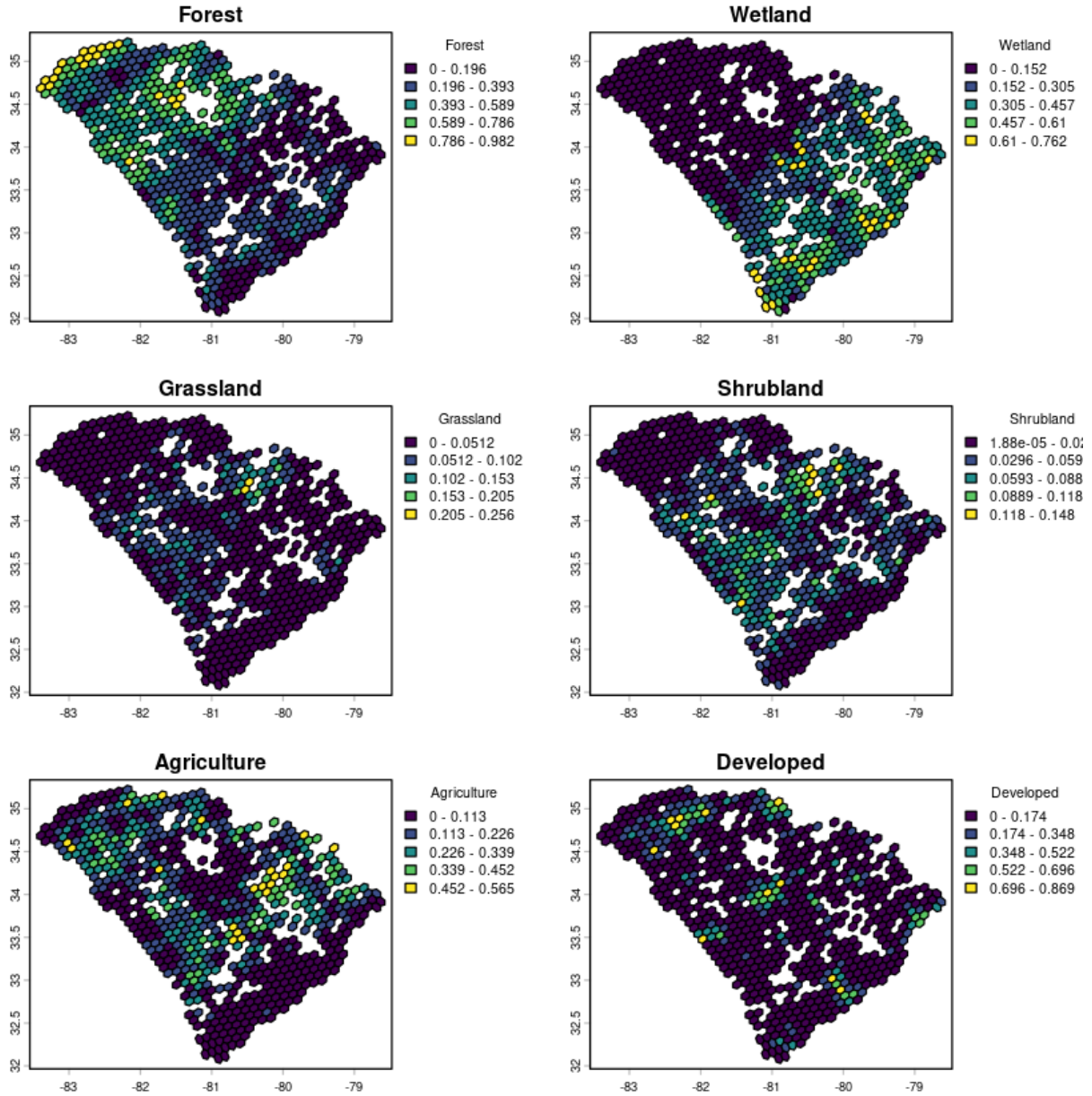
The NLCD dataset I used includes 15 different land-cover classes:

```
11 - "Open Water",
21 - "Developed, Open Space",
22 - "Developed, Low Intensity",
23 - "Developed, Medium Intensity",
24 - "Developed, High Intensity",
31 - "Barren Land",
41 - "Deciduous Forest",
42 - "Evergreen Forest",
43 - "Mixed Forest",
52 - "Shrub/Scrub",
71 - "Herbaceous",
81 - "Pasture/Hay",
82 - "Cultivated Crops",
90 - "Woody Wetlands",
95 - "Emergent Herbaceous Wetlands"
```

I grouped these to simplify a bit, for example mapping all “Developed” subcategories into one. Here are all the mappings:

NLCD categories	Simplified “ecological group” predictor
Deciduous Forest, Mixed Forest, Evergreen Forest	Forest
Pasture/Hay, Cultivated Crops	Agriculture/Open
Grassland/Herbaceous	Grassland
Woody Wetlands, Emergent Herbaceous Wetlands	Wetland
Open Water	Water
Developed, all intensities	Developed
Shrub/Scrub	Shrubland

This yields seven simplified land-cover categories I’m calling “ecological groups”. I calculated land-cover proportions for each hex cell and stored proportions of each. Here are the distributions of the different land-cover categories across the state:



I could have used the ecological-group proportions themselves as numerical predictors, but I wanted to convert them into a categorical variable for lower model complexity, and possibly a more interesting result. For example, including the proportion of “Forest” ecology in a regression model predicting woodpecker detections yielded a significant (low p-value) beta parameter, but that’s an obvious result. I thought it would be more interesting to see how a cell’s dominant ecology classification (largest proportion among its constituent ecological groups) influenced the prediction or added explanatory value.

Each hex cell’s “dominant ecological group” (categorical variable) was then translated into a dummy/one-hot encoding in the model.

This is the final list of predictors I included in the regression analysis:

- Checklist average number of observers
- Checklist average start time of day (hours past midnight)

- Checklist median duration
- Checklist average distance traveled
- Distance from coast
- Elevation
- Dominant ecological-group classification ('Forest', 'Agriculture', 'Grassland', 'Wetland', 'Water', 'Developed', 'Shrubland')
- Annual mean temperature
- Annual precipitation

All are continuous/numerical except for the ecology classification, which is categorical. Each row represents a cell in the hex grid across the state with its values aggregated over all of 2024 (e.g., precipitation summed for the year, checklist distance traveled averaged over the year). The response variable is detection frequency (number of checklists with species observed vs. total number of complete checklists in the hex cell).

Data Cleaning

I converted the dominant ecological-group character variable to a factor variable. I also removed 15 (out of 732) hex cells where, for whatever reason, temp and precip data was unavailable. This left a final observation/row count of $n=717$ for the model to fit.

Resulting data structure (with a few extra non-predictor variables shown for context):

```
cols <- c("cell", "cell_ctr_lat", "cell_ctr_lon", "n_checklists", "n_detected", "det_freq", "mean_num_obs", "mean_start_time_hr", "med_duration", "mean_dist_km", "dist_to_coast_mi", "elevation_ft", "dominant_eco_group", "tmean_2024_C", "ppt_2024_mm")
str(ebird_features_df_pwo[cols])
```

```
## tibble [717 x 15] (S3: tbl_df/tbl/data.frame)
## $ cell : num [1:717] 493475 493476 493477 493478 493479 ...
## $ cell_ctr_lat : num [1:717] 35.1 35.1 35.2 35.2 35.2 ...
## $ cell_ctr_lon : num [1:717] -82.8 -82.7 -82.5 -82.4 -82.3 ...
## $ n_checklists : int [1:717] 1 11 180 20 10 2 35 5 15 91 ...
## $ n_detected : int [1:717] 0 8 80 4 5 0 10 2 5 20 ...
## $ det_freq : num [1:717] 0 0.727 0.444 0.2 0.5 ...
## $ mean_num_observers: num [1:717] 2 1.55 1.13 1.2 2 ...
## $ mean_start_time_hr: num [1:717] 12.78 10.12 11.66 9.82 11.25 ...
## $ med_duration : num [1:717] 0.417 2.35 0.583 0.333 1.083 ...
## $ mean_dist_km : num [1:717] 1.043 2.238 0.392 0.408 2.41 ...
## $ dist_to_coast_mi : num [1:717] 213 211 208 206 205 ...
## $ elevation_ft : num [1:717] 2690 3010 2850 2277 2385 ...
## $ dominant_eco_group: Factor w/ 5 levels "Agriculture",... 3 3 3 3 3 3 3 3 3 3 ...
## $ tmean_2024_C : num [1:717] 13.5 13.3 14.3 14.6 14.9 ...
## $ ppt_2024_mm : num [1:717] 2448 2553 2312 2077 1949 ...
```

Methods

Exploratory data analysis

Response transformation

Detection frequency is my basic response variable - what proportion of checklists in a hex cell in 2024 included a positive detection of the species. This is a numerical prediction, but its [0-1] range (and “counting” basis) is unconventional for a standard regression model. In particular, the diagnostics indicated very poor adherence to linearity, normality or errors, and equivariance (see graphs in Appendix).

Here is the first model I started with that has no response transformation:

```
mod_add_pwo = lm(det_freq ~ mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
                  dist_to_coast_mi + elevation_ft + dominant_eco_group +
                  tmean_2024_C + ppt_2024_mm,
                  data = ebird_features_df_pwo, weight = n_checklists)
```

It took some experimenting to find an appropriate transformation. I tried a log transformation (consistent w/ the recommendation from a Box-Cox analysis), which improved things somewhat, but ultimately I settled on the logit transformation, ultimately adopting the `glm()` binomial framework to model my detection rate (but still as a proportion, not binary prediction).

```
form_pwo_glm_base = det_freq ~
                    (mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
                     dist_to_coast_mi + elevation_ft + dominant_eco_group +
                     tmean_2024_C + ppt_2024_mm)
mod_pwo_glm_base = glm(form_pwo_glm_base,
                       weights = n_checklists,
                       data = ebird_features_df_pwo,
                       family = binomial)
```

All subsequent discussion relates to models this logit-transformed response, unless otherwise specified. See Appendix for more output and graphs supporting this direction.

Collinearity

I did my best to avoid choosing predictors with very high correlation, but there still ended up being some. I used a pairs plot to visualize trends among predictors, and saw that distance to coast and elevation were very well correlated. This correlation is not uniform everywhere in the country. SC has a relatively steady slope from the ocean to the Appalachian mountain range, but a different state (like CA) could show a more interesting relationship.

I then used `cor()` to get an idea of individual predictors’ correlation magnitude with each other the response (see table in Appendix). I found the effort variables all have higher correlation w/ the response (detection frequency) than do the climate or ecology (at least when aggregated annually into hex cells as I did).

Distance to coast and elevation have a correlation coefficient of 84%. Elevation and mean temp have a -85% correlation, meaning that the yearly average temp tends to decrease with distance from the coast, as one would expect.

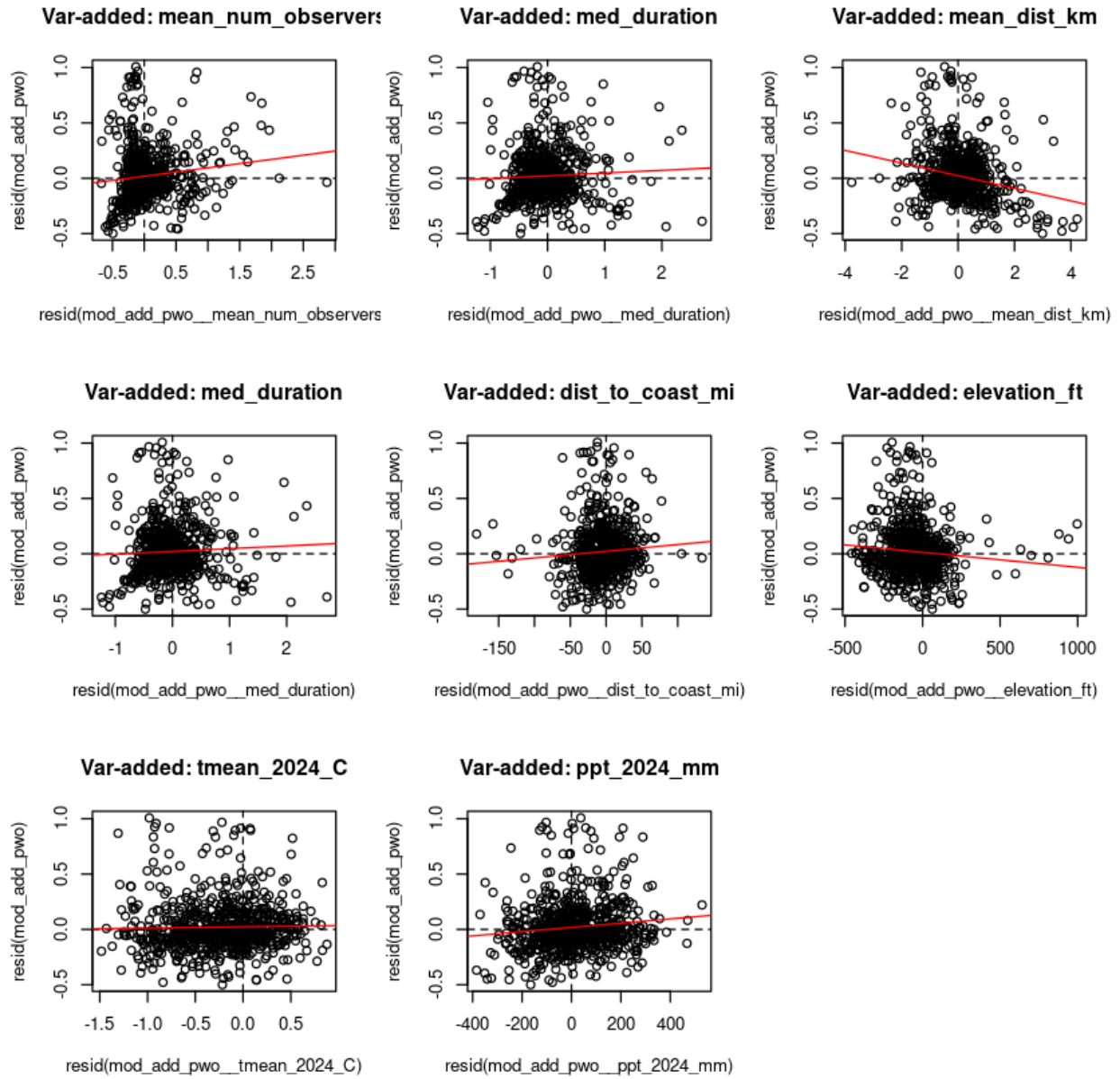
Elevation and precipitation have a correlation coefficient of 57%, indicating that higher elevations tend to have higher precipitation.

Next I evaluated the generalized VIF for the base model, and the highest values were for distance to coast, elevation, and mean temp. This makes sense based on the correlation coefficient among them and the roughly linear trend visible in the pairs plot. None are higher than the conventional “high VIF” cutoff of 5 though, so I don’t expect excessive parameter-estimate variance.

```
car::vif(mod_pwo_glm_base)
```

```
##              GVIF Df  GVIF^(1/(2*Df))
## mean_num_observers 1.348  1           1.161
## mean_start_time_hr 1.064  1           1.032
## med_duration       1.270  1           1.127
## mean_dist_km       1.555  1           1.247
## dist_to_coast_mi   8.461  1           2.909
## elevation_ft       9.645  1           3.106
## dominant_eco_group 3.225  4           1.158
## tmean_2024_C       9.882  1           3.144
## ppt_2024_mm        2.310  1           1.520
```

I also created variable-added plots for each of my numerical predictors, but most looked like noise. The ones with an appreciable slope to the trend line had high-leverage points that seemed responsible for that.



Finally, I made new models, omitting individual predictors to see how significance of other predictors changed. I noticed that some predictors actually get less significant when certain other predictors were omitted, e.g., when I removed median checklist duration as a predictor, dominant eco groups became less significant. I haven't been able to make sense of this, but it could be due to invalid model assumptions (linearity or normality of errors, etc.) undermining the hypothesis test's validity.

Removing `mean_dist_km` made `mean_num_observers`, `dist to coast`, `elevation`, `precip`, and `dominant eco group` all more significant, so `mean_dist_km` seems to have collinearity with these other predictors.

With `dominant eco group` removed, `temp` and `precip` go from insignificant to very significant. I believe this is due to the ecology changing with `elevation` (which `temp` and `precip` are correlated with).

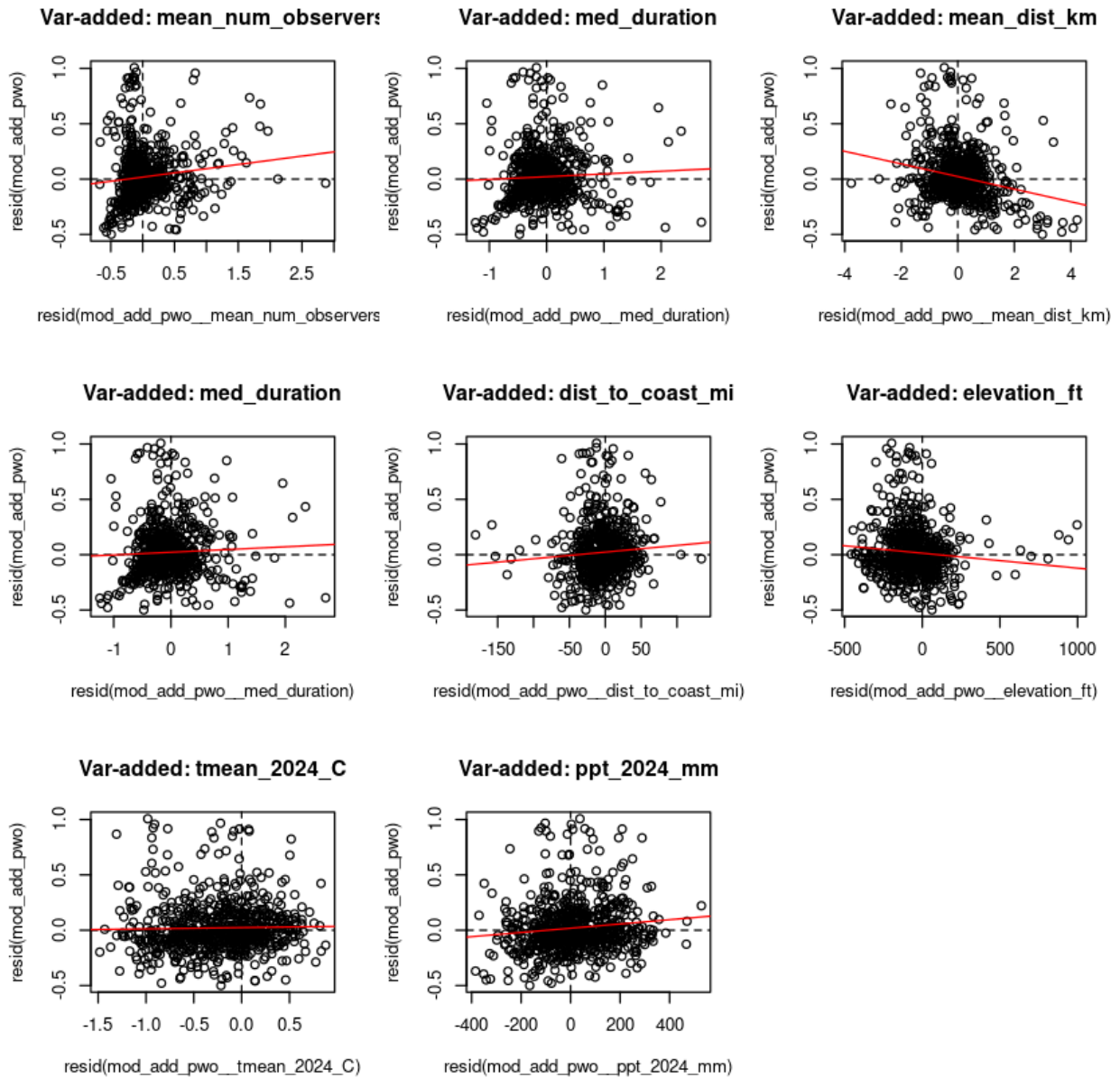
I thought removing `elevation` or `dist to coast` would effect each other more; they have second and third highest VIFs, and 84% correlation w/ each other. Since they're also correlated with `Tmean` though (each pair's correlation magnitude is >80%), its presence kept me from seeing much effect when only one was removed. When I removed two of any three of `elevation`, `temp`, and `distance to coast`, the significance level of the remaining predictor increased substantially. Each of them becomes highly significant when the others are

removed:

	Pr(> t)	Pr(> t)	Pr(> t)	Pr(> t)	Pr(> t)	Pr(> t)	Pr(> t)
dist_to_coast_mi	0.28981		0.22959	0.09093			0.00038
elevation_ft	0.82237	0.542068		0.45839		0.00131	
tmean_2024_C	0.40732	0.119541	0.27595		0.00044		

(These were all done w/ additive model and no response transformation, so the values don't reflect my chose base model, but the effect is still valid.)

Here are the variable-added plots, showing relatively low multicollinearity:



Scope Exploration

I spent some time trying different species of birds (source of response variable). Since I wasn't seeing very high correlations between most habitat predictors and the response, I built models for around a dozen other species in an attempt to see if any showed better correlation with my predictors. Models predicting other species detections performed a worse both as additive models and models w/ all 2-way interactions, so I stuck with the pileated woodpecker. In my proposal, I planned to use the common raven, but since I chose to analyze only data in SC, and its SC range is limited, I went with a species that ranges over the whole state.

I also experimented with different spatial resolutions to get a feel for the best aggregation unit. I saw higher correlation with values with larger cell sizes (20 km better and 5 km worse), but I didn't see any clear changes to the trends, for example the individual response-vs-predictor row in the pairs plot. I decided to leave it at the 10km level since I felt this was the most interesting for my analysis. I suspect the correlation just increased due to less noise and not a real change in predictive value; this could use some more exploration though.

Unusual Points

I explored high-leverage points and outliers and found that ~13% of the data are high-leverage points (using rule of thumb of $> 2h$). A more concerning value is that 27% of (standardized) residuals are > 2 , as opposed to the 5% we'd expect for a normal distribution.

This high number of high residuals matches the heavy tails I saw in the normal Q-Q plot.

Using the rule of thumb of Cook's distance $> 4/n$, I determined ~21% of the data points are "influential". This tells me the model could use some work before being used for real-world inference.

```
get_percent_high_leverage(ebird_features_df_pwo, mod_pwo_glm_base)
```

```
## [1] 0.1283
```

```
get_percent_outliers(ebird_features_df_pwo, mod_pwo_glm_base)
```

```
## [1] 0.2706
```

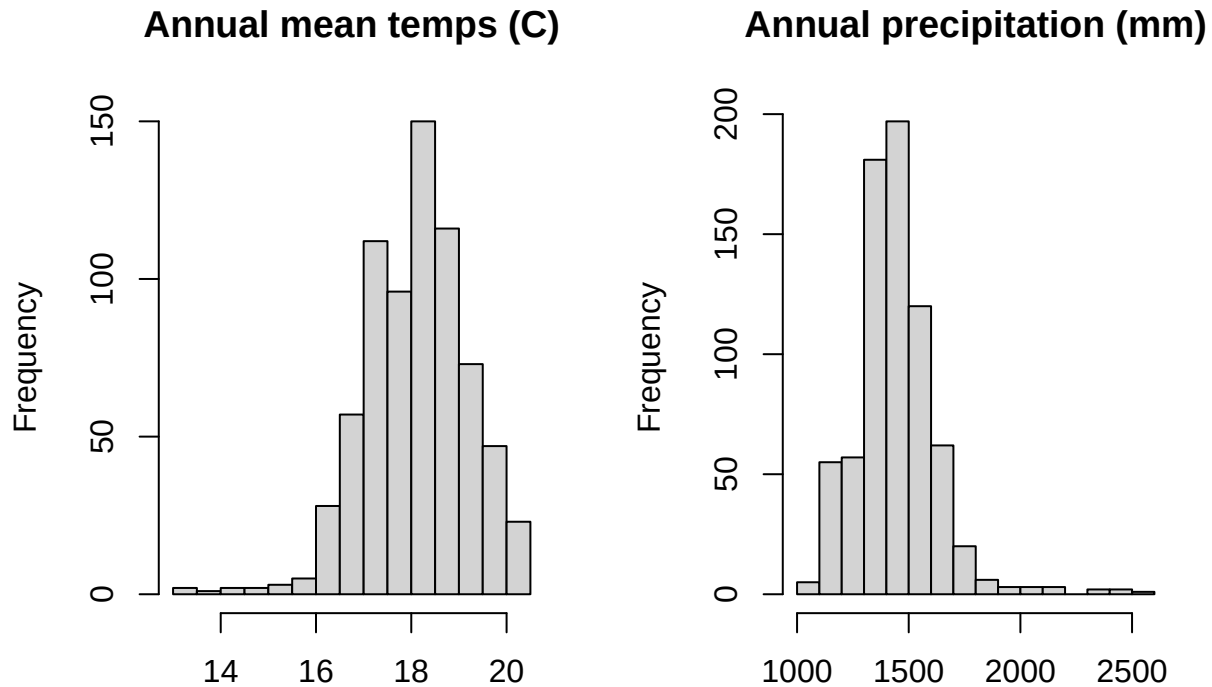
```
get_percent_high_cooks_distance(ebird_features_df_pwo, mod_pwo_glm_base)
```

```
## [1] 0.205
```

Various Plotting

I looked at some distributions of different predictors to ensure there was enough variation for the model to "fit". Here's an example of the spread of annual average temp and cumulative precipitation:

```
par(mfrow = c(1, 2))
hist(ebird_features_df_pwo$tmean_2024_C, main = "Annual mean temps (C)")
hist(ebird_features_df_pwo$ppt_2024_mm, main = "Annual precipitation (mm)")
```



`ebird_features_df_pwo$tmean_2024_C`

`ebird_features_df_pwo$ppt_2024_mm`

I also analyzed certain individual trends, like the effect of the eBird-checklist “effort” variables on the detection rate. See the Appendix for separate plots of these.

I used the `car::crPlots()` to create component + residual plots and evaluate nonlinear trends in the data. This gave me some indication of what variables could benefit from a transformation. See the plots in the Appendix.

Full/base model interpretation (no poly or interaction terms)

```
form_pwo_base = det_freq ~
  (mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
   dist_to_coast_mi + elevation_ft + dominant_eco_group +
   tmean_2024_C + ppt_2024_mm)
mod_pwo_glm_base = glm(form_pwo_base,
  weights = n_checklists,
  data = ebird_features_df_pwo,
  family = binomial)
summary(mod_pwo_glm_base)
```

```
##
## Call:
## glm(formula = form_pwo_base, family = binomial, data = ebird_features_df_pwo,
##      weights = n_checklists)
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.7208718   0.6199635   -2.78   0.0055 **
## mean_num_observers   -0.1089660   0.0752673   -1.45   0.1477
## mean_start_time_hr   -0.0987868   0.0123688   -7.99  1.4e-15 ***
## med_duration         0.6542476   0.0379704   17.23  < 2e-16 ***
```

```
## mean_dist_km          0.5429238  0.0230561  23.55 < 2e-16 ***
## dist_to_coast_mi      0.0022312  0.0005059   4.41  1.0e-05 ***
## elevation_ft          0.0001309  0.0000914   1.43  0.1518
## dominant_eco_groupDeveloped -0.5827603  0.0798888  -7.29  3.0e-13 ***
## dominant_eco_groupForest  -0.0832003  0.0784516  -1.06  0.2889
## dominant_eco_groupWater   -0.6625551  0.0912765  -7.26  3.9e-13 ***
## dominant_eco_groupWetland  0.3451259  0.0826658   4.17  3.0e-05 ***
## tmean_2024_C           0.0064977  0.0310714   0.21  0.8344
## ppt_2024_mm            -0.0000427  0.0001060  -0.40  0.6872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 7478.3 on 716 degrees of freedom
## Residual deviance: 4218.1 on 704 degrees of freedom
## AIC: 5752
##
## Number of Fisher Scoring iterations: 5
```

In the base model, mean_num_observers, elevation, annual avg temp, and precip are each showing as individually not significant per the reported p-values.

These are the “significant” predictors and my interpretation:

Predictor	Coefficient Sign	Interpretation
Start time	Negative	Earlier start time associated w/ higher detection rate
Duration	Positive	Longer checklist duration associated w/ higher detection rate
Distance	Positive	Farther distance traveled associated w/ higher detection rate
“Developed” eco group	negative	associated w/ lower detection rate
“Water” eco group	negative	associated w/ lower detection rate
“Wetland” eco group	positive	associated w/ higher detection rate Woody wetlands (class 90) is included in this.

The sign of each coefficient indicates the relationship “direction”, and the magnitude tells us how the response changes from a one-unit change in this predictor (all other predictors held constant).

For example, the mean start time increasing by 1 hour will correspond to a decrease in predicted detection frequency of -0.098, or ~10% lower. And a one-mile increase in distance to the coast corresponds to an increased predicted detection frequency of 0.0022, or 0.22%.

10-fold CV RMSE: **0.2324**

I used the `caret` package’s `trainControl` and `train` functions to perform 10-fold cross-validation. It splits the rows randomly into 10 groups (folds) and trains the provided model with 9 of the folds then evaluates the model’s prediction accuracy using the tenth (held out) fold. Each fold gets a turn as the test fold. The resulting RMSE values are averaged to yield the 10-fold avg RMSE.

My binomial model doesn’t produce an R^2 , so I’m using cross-validated RMSE as my judge of model performance. (So that you know I know what R^2 means, if this model had an multiple R^2 of 0.63, that would mean that the predictors account for 63% of the variance in the response. And adjusted R^2 includes a penalty to account for model complexity).

Diagnostics

Linear-regression modeling relies on assumptions that the data being fit will conform to certain structural patterns - that the relationship between predictors and response will fundamentally be linear (as opposed to quadratic, exponential, etc.), that the errors will be independent/random, that the prediction errors will approximate a normal distribution, and that the variance of the errors will be constant across the range of fitted/predicted values. If any of these assumptions don't hold true, inferences can be invalidated. The regressor may still provide predictive value, but confidence intervals and significance levels for the coefficient estimates are not to be trusted.

To evaluate linearity and equivariance, I used a fitted-vs-residuals plot. I also used the Breusch-Pagan test to assess equivariance. To evaluate normality of errors, I used both a normal Q-Q plot and the Shapiro-Wilk test.

Initially, all diagnostics indicated the model lacked linearity, equivariance, and normality of errors. The model w/ raw detection frequency as response shows a very heavy right tail on the normal Q-Q plot and a diagonal trend on the fitted-vs-residuals plot. The variance/spread of the residuals is not centered at 0 or constant as the fitted value varies.

I was able to get a good-looking Q-Q plot and "pass" the Shapiro-Wilk test by transforming my response variable to be the log of raw detection frequency. Unfortunately, it still failed to exhibit homoscedasticity or linearity on the fitted-vs-residuals plot. I also tried a reciprocal transformation to hopefully stabilize the variance but with no luck.

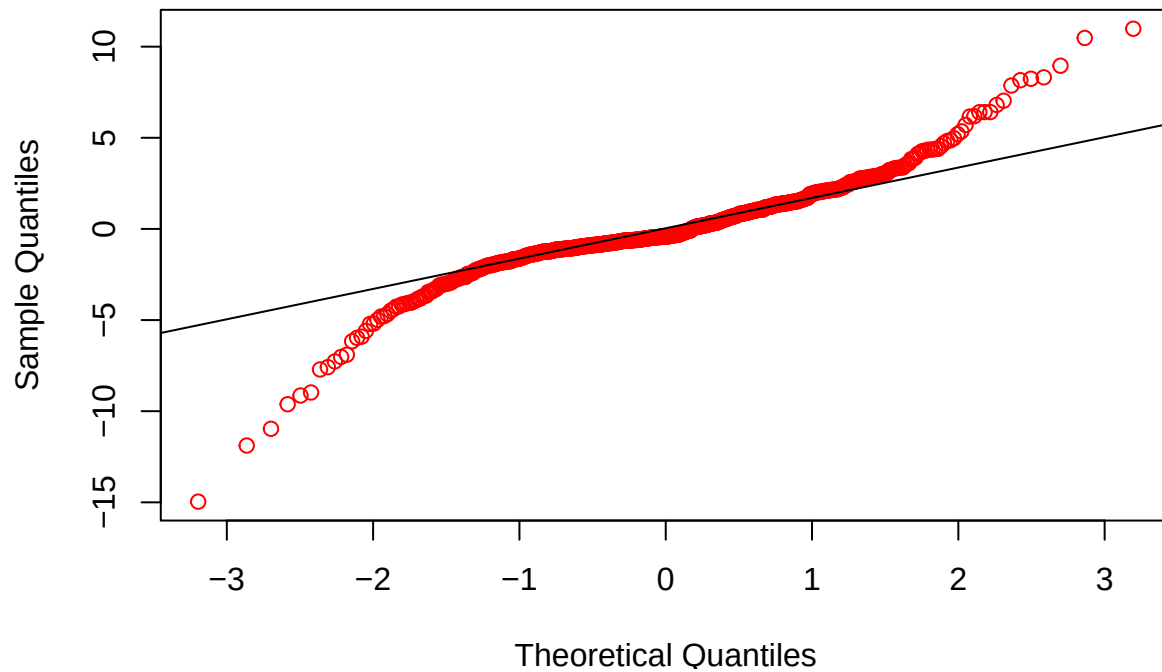
See the graphs of the non-transformed and log-transformed models in the Appendix.

After the professor's recommendation, I opted to use the logit transformation on my response variable based on its theoretical underpinning matching my proportion/counting-based variable. This transformation provides a broader quantitative range for the response variable instead of the $[0, 1]$ range detection frequency has. Despite rejecting the null hypothesis for both Shapiro-Wilk and Breusch-Pagan tests, the fitted-vs-residuals plot is now centered at ~ 0 for all fitted values; the spread just decreases with fitted value magnitude. This gives me confidence in the linearity assumption at least, although the variance is still not constant. The normal Q-Q plot also looks more symmetrical than the original one, although it still shows the error distribution has heavy tails.

These are the resulting diagnostic plots and test results for the `glm()`/logit-transformed/binomial model:

```
qqnorm(resid(mod_pwo_glm_base), col = "red")  
qqline(resid(mod_pwo_glm_base))
```

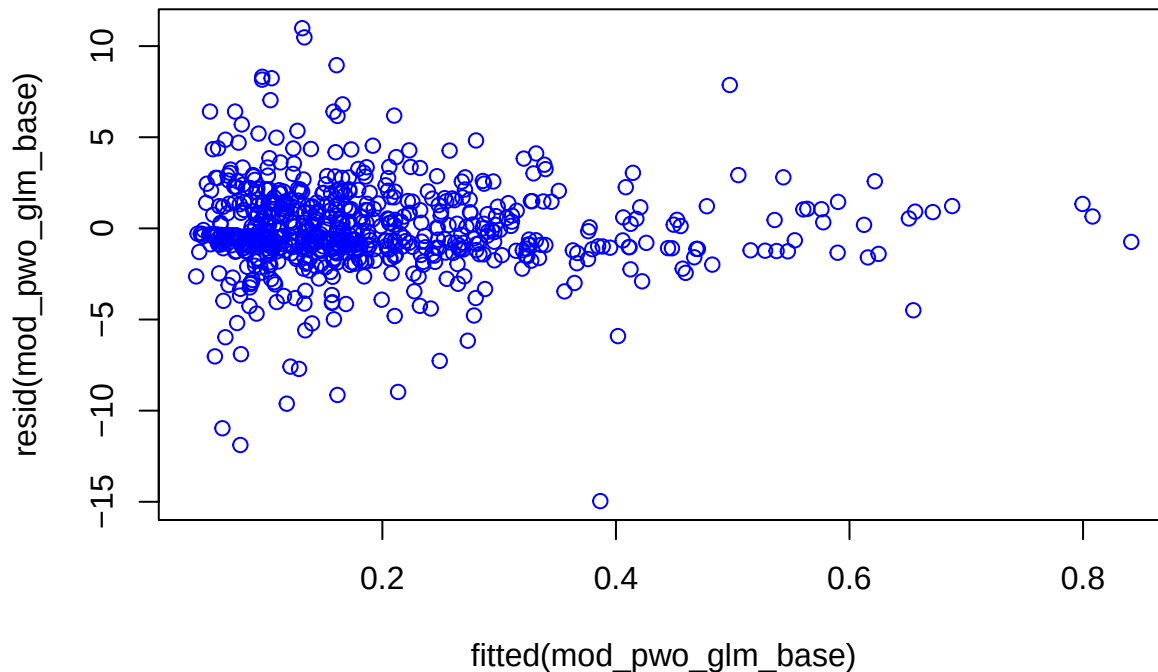
Normal Q-Q Plot



```
shapiro.test(resid(mod_pwo_glm_base))
```

```
##  
##  Shapiro-Wilk normality test  
##  
## data:  resid(mod_pwo_glm_base)  
## W = 0.91, p-value <2e-16
```

```
plot(resid(mod_pwo_glm_base) ~ fitted(mod_pwo_glm_base), col = "blue")
```



```
lmtest::bptest(mod_pwo_glm_base)
```

```
##
## studentized Breusch-Pagan test
##
## data: mod_pwo_glm_base
## BP = 55010, df = 12, p-value <2e-16
```

Since I was unable to transform the model to one with normally-distributed or equal-variance errors, some of my future inferences will be invalid. I'll mostly ignore this from now on for the sake of the assignment and comment on what would be a useful inference, if the LINE assumptions held.

Improvements

Response transformations were discussed in a previous section. I settled on using the logit transformation, and accomplished this by using the `glm()` function with family set to “binomial”. I set up the parameters to output `logit(detection frequency)` as the response variable.

Predictor transformations

I tried second and (second+)third-order polynomial terms for every numerical predictor and kept ones whose inclusion yielded a significant β estimator in the resulting model. Some polynomial terms even improved the significance level of the main-effect predictor's parameter estimate, indicating lack of flexibility was causing the base predictor to underfit.

Next I used 10-fold cross-validation to evaluate whether each additional polynomial term's inclusion helped or hurt the prediction accuracy. This was necessary because additional parameters can only ever help the error/deviance when evaluated against “seen” data, but it could introduce significant variance on unseen data. So I only kept higher-order terms when the cross-validated RMSE was consistently and appreciably below the baseline model (having main-effect predictor variables only).

For almost every predictor, adding at least a second-order component seemed to improve significance level(s), but after running each through cross-validation, I only kept polynomial terms for checklist distance traveled,

distance from coast, and elevation. I don't want to go higher than third-order because the data is so noisy, the additional does not appear to truly be fitting any signal. I'm also trying to avoid introducing too much complexity because it will increase prediction variance.

I think in many cases, interpretation of "significance" based on the hypothesis test is dubious because some LINE assumptions don't hold, so that makes it even more important to let cross-validation guide the component selection.

Here is the model with all chosen polynomial terms included:

```
form_pwo_glm_full_poly = det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration + mean_dist.  
                                     I(mean_dist_km^2)+I(mean_dist_km^3) +  
                                     dist_to_coast_mi +  
                                     I(dist_to_coast_mi^2)+I(dist_to_coast_mi^3) +  
                                     elevation_ft +  
                                     I(elevation_ft^2) +  
                                     dominant_eco_group + tmean_2024_C + ppt_2024_mm)  
mod_pwo_glm_full_poly = glm(form_pwo_glm_full_poly,  
                             weights = n_checklists,  
                             data = ebird_features_df_pwo,  
                             family = binomial)
```

(Keep in mind that what's actually being predicted is $\text{logit}(\text{det_freq})$)

An ANOVA (compared to model w/ no poly terms) shows that the addition of polynomial terms is significant.
10-fold CV RMSE: **0.2276** (compared to 0.2324 for base model w/o poly terms)

This model has lower RMSE than the one without polynomial terms.

Model selection and interaction terms

I chose to use a backward selection procedure with BIC to prune predictors from my model that weren't adding enough value. I chose BIC instead of AIC because I'm more interested in interpretability and gaining insight from the model (what predictors have what effect), rather than pure predictive power, and BIC penalizes more than AIC when n is relatively large (like in my case).

I started with the model that includes polynomial terms but no interactions then did backward BIC. The selection procedure got rid of number of observers, temp, and precip, but it kept all polynomial terms.

10-fold CV RMSE: **0.2276** (better)

An ANOVA on the model with all poly terms vs. the model selected using BIC shows that the predictors removed by the stepwise process are likely insignificant. This smaller model has basically the same RMSE as before, so it is preferred.

Next I took the poly-term + full-interaction model and ran backward BIC on that:

10-fold CV RMSE: **0.3173** (worse)

The RMSE is worse than what I saw on the model without any interaction terms, so I don't want to use this model. The complexity is causing too much prediction variance.

Next I took a more measured approach to adding interaction terms. I restricted interactions to only be allowed among certain groups - among the effort variables, and among the habitat variables, but not across the groups.

The resulting model was significant according to the an ANOVA.

10-fold CV RMSE: **0.2496** (worse)

Then I ran backward selection procedure with BIC on this. The resulting RMSE was better than before pruning, but worse than before adding the interactions.

10-fold CV RMSE: **0.2436** (worse)

Next I tried a few different models with hand-picked interaction terms, based on what combinations appeared significant in the fully-interactive model (summary output), and what variables tended to survive BIC-based selection procedures.

I tried a few of these until I got a model that, after backward-BIC selection, performed better in cross-validated RMSE than the model without interactions:

```
form_pwo_glm_poly_cint3 = det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
  mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) +
  mean_num_observers:mean_start_time_hr +
  mean_num_observers:mean_dist_km +
  mean_start_time_hr:mean_dist_km +
  mean_start_time_hr:I(mean_dist_km^2) +
  mean_start_time_hr:I(mean_dist_km^3) +
  med_duration:mean_dist_km +
  med_duration:I(mean_dist_km^2) +
  med_duration:I(mean_dist_km^3) +
  dominant_eco_group +
  dist_to_coast_mi + I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) +
  elevation_ft + I(elevation_ft^2) +
  tmean_2024_C +
  ppt_2024_mm +
  dist_to_coast_mi:ppt_2024_mm +
  I(dist_to_coast_mi^2):ppt_2024_mm +
  I(dist_to_coast_mi^3):ppt_2024_mm
)
mod_pwo_glm_poly_cint3 = glm(form_pwo_glm_poly_cint3,
  weights = n_checklists,
  data = ebird_features_df_pwo,
  family = binomial)
```

10-fold CV RMSE: **0.2318** (worse)

After running backward selection with BIC:

10-fold CV RMSE: **0.2292** (similar to original no-interaction model)

I also tried using LASSO and Ridge regression with poor results. I used `cv.glm()` to find the optimal lambda penalty coefficient. I used the one-std-dev-larger lambda value and ran cross-validation with this model. RMSE with either LASSO or Ridge is worse than before (~0.69 instead of ~0.24), so it doesn't seem to be a good choice in this case.

Conclusion

I found a no-interaction model and one with some interactions that score roughly the same at cross-validated RMSE. All other things being equal, we should prefer the simpler model then.

1. Basic polynomial model after backward selection w/ BIC:

```
det_freq ~ (mean_start_time_hr + med_duration +
  mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) +
  dist_to_coast_mi + I(dist_to_coast_mi^2)+I(dist_to_coast_mi^3) +
  elevation_ft + I(elevation_ft^2) +
  dominant_eco_group)
```

Backward BIC selection removed Tmean, Precipitation, and observer count.

2. “Curated” interactive model after backward selection w/ BIC

```
det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
  mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) +
  mean_num_observers:mean_start_time_hr +
  mean_start_time_hr:mean_dist_km +
  med_duration:mean_dist_km +
  med_duration:I(mean_dist_km^2) + med_duration:I(mean_dist_km^3) +
  dist_to_coast_mi + I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) +
  elevation_ft + I(elevation_ft^2) +
  dominant_eco_group +
  ppt_2024_mm +
  dist_to_coast_mi:ppt_2024_mm + I(dist_to_coast_mi^2):ppt_2024_mm +
  I(dist_to_coast_mi^3):ppt_2024_mm)
```

Backward BIC selection removed Tmean and some effort interaction variables.

Both of these models achieve a 10-fold cross-validated RMSE of < 0.230 . This is pretty unimpressive, considering the full scale of the response variable is 1 (so the model has ~23% error on average).

Note that the two best models are not nested, so I can't do a proper ANOVA w/o hierarchy in place.

I don't have confidence that this model will provide useful predictive power for detections or bird populations. I think it will take a different approach, probably with different variables, to create something of any explanatory or predictive power.

The effort variables provided the most information, but those effects are mostly common across species. I was hoping to learn something more about woodpecker habitat preferences, like that they prefer higher/lower elevations or dislike living near the ocean. I don't see strong enough correlations to make any such inferences.

A significant part of my time investment was learning how to extract and work with the eBird data extraction, then habitat-data curation and learning how to use all those datasets in R, and then feature engineering to formulate a dataset for regression. Then I spent significant time slicing the hex grids different ways, trying different species, and trying different predictor and response transformations in an attempt to get the predictor-response relationship to conform to the linear-regression assumptions (linearity, independence, normality of errors, and equivariance of errors).

The more exploration I did, the more direction I saw that the analysis could take - more granular timescale, less granular spatial scale, season-to-season variation, etc. There is certainly more to explore here.

Appendix

Aggregation

My original plan was to build a regression model to predict observation count based on conditions and environment - hour of day, weather variables like temp and precip, windspeed, and geographic variables like elevation and distance to coast. Challenges with this approach:

- Observation count is how many individual birds a given observer counted. If I use that as the response, a data point where 60 birds were seen by 1 observer will be treated by model as much different than 60 birds seen by 60 observers (60 data points each with 1 bird). What I'm really interested in is number of birds seen in a given area. Observation count doesn't do a great job of that.
- Observer "prevalence" needs to be controlled for somehow. Observation count will be high in an area where birders frequent, even if the species is rarely present, whereas another area having rare observers but high bird "occupancy" will look sparse. Detection frequency is how I chose to address this - dividing the number of "positive" checklists (where any number of individuals of the species were observed) by the number of "negative" checklists (checklists w/ no observations of the species). This essentially normalizes the observation count and maps observations to the range [0, 1]. To enable this, I filtered the eBird checklists by those asserted to be "complete" by the user. This mechanism is used to indicate that during the checklist collection, the user recorded every single individual bird they were able to positively identify, allowing non-detections to be inferred. Of course, there's still the issue of failure to detect (a bird being quiet and inconspicuous, a birder being inexperienced, distraction, etc.), but it at least takes out the possibility of the user just not recording the detection, because the species is common/boring for instance.
- Practical limitation - I can't afford to pay for all the API queries it would take to associate weather data to the tens of thousands of individual checklist locations and times. I could restrict the scope of my analysis to a small geographic area and/or time, but then the range of each predictor variable will shrink (temp, elevation, etc. ranges will narrow), which I expect will inflate the parameter variance and just make the result less interesting.

EDA

Here is the model I started with, without a response transformation:

```
mod_add_pwo = lm(det_freq ~ mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
                  dist_to_coast_mi + elevation_ft + dominant_eco_group +
                  tmean_2024_C + ppt_2024_mm,
                  data = ebird_features_df_pwo, weight = n_checklists)
summary(mod_add_pwo)
```

```
##
## Call:
## lm(formula = det_freq ~ mean_num_observers + mean_start_time_hr +
##     med_duration + mean_dist_km + dist_to_coast_mi + elevation_ft +
##     dominant_eco_group + tmean_2024_C + ppt_2024_mm, data = ebird_features_df_pwo,
##     weights = n_checklists)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -5.873  -0.337  -0.061   0.460   4.936
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.98e-01   1.91e-01   1.56  0.11996
## mean_num_observers -2.65e-02   2.48e-02  -1.07  0.28616
```

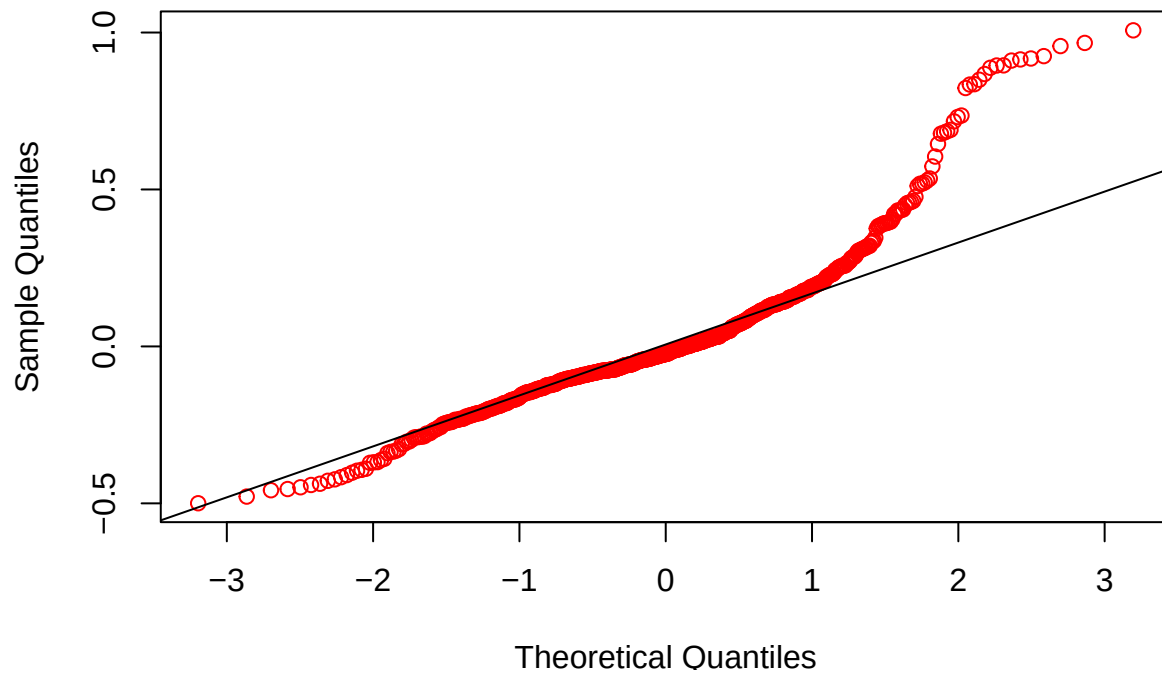
```
## mean_start_time_hr      -1.26e-02  3.76e-03  -3.35  0.00085 ***
## med_duration            1.13e-01  1.36e-02   8.34  3.8e-16 ***
## mean_dist_km            8.22e-02  7.84e-03  10.49 < 2e-16 ***
## dist_to_coast_mi        1.67e-04  1.57e-04   1.06  0.28980
## elevation_ft            6.65e-06  2.96e-05   0.22  0.82237
## dominant_eco_groupDeveloped -7.32e-02  2.62e-02  -2.80  0.00531 **
## dominant_eco_groupForest  -2.00e-02  2.64e-02  -0.76  0.45018
## dominant_eco_groupWater   -8.11e-02  2.84e-02  -2.85  0.00447 **
## dominant_eco_groupWetland  3.97e-02  2.75e-02   1.44  0.14947
## tmean_2024_C            -7.97e-03  9.62e-03  -0.83  0.40732
## ppt_2024_mm             1.58e-05  3.30e-05   0.48  0.63224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.895 on 704 degrees of freedom
## Multiple R-squared:  0.45, Adjusted R-squared:  0.441
## F-statistic: 48 on 12 and 704 DF, p-value: <2e-16
get_k10_cv_rmse(formula(mod_add_pwo), ebird_features_df_pwo)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:httr':
##
##   progress
##
## The following object is masked from 'package:purrr':
##
##   lift
## [1] 0.2346
```

Here are diagnostic plots and test results that signaled that a response transformation was needed:

```
qqnorm(resid(mod_add_pwo), col = "red")
qqline(resid(mod_add_pwo))
```

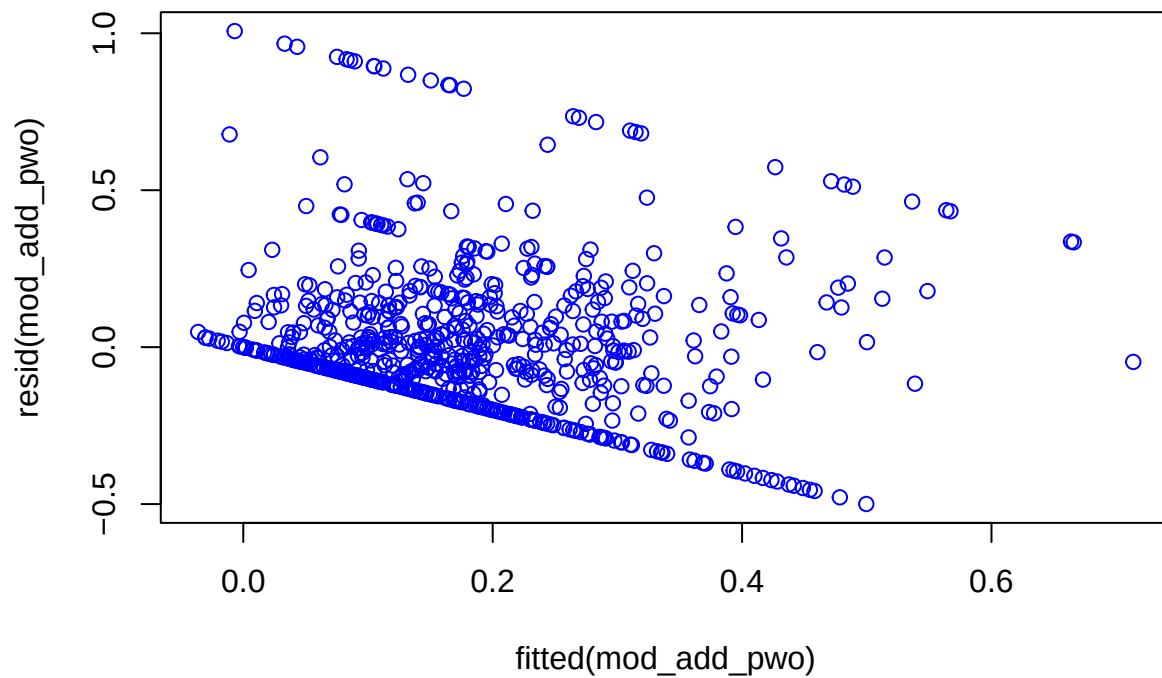
Normal Q-Q Plot



```
shapiro.test(resid(mod_add_pwo))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data:  resid(mod_add_pwo)  
## W = 0.9, p-value <2e-16
```

```
plot(resid(mod_add_pwo) ~ fitted(mod_add_pwo), col = "blue")
```



```
lmtest::bptest(mod_add_pwo)
```

```
##
## studentized Breusch-Pagan test
##
## data: mod_add_pwo
## BP = 55010, df = 12, p-value <2e-16
```

Here is the log(y) model I then tried:

```
# Using log(y+0.1) transformation. Used offset because there are some 0-detection cells.
```

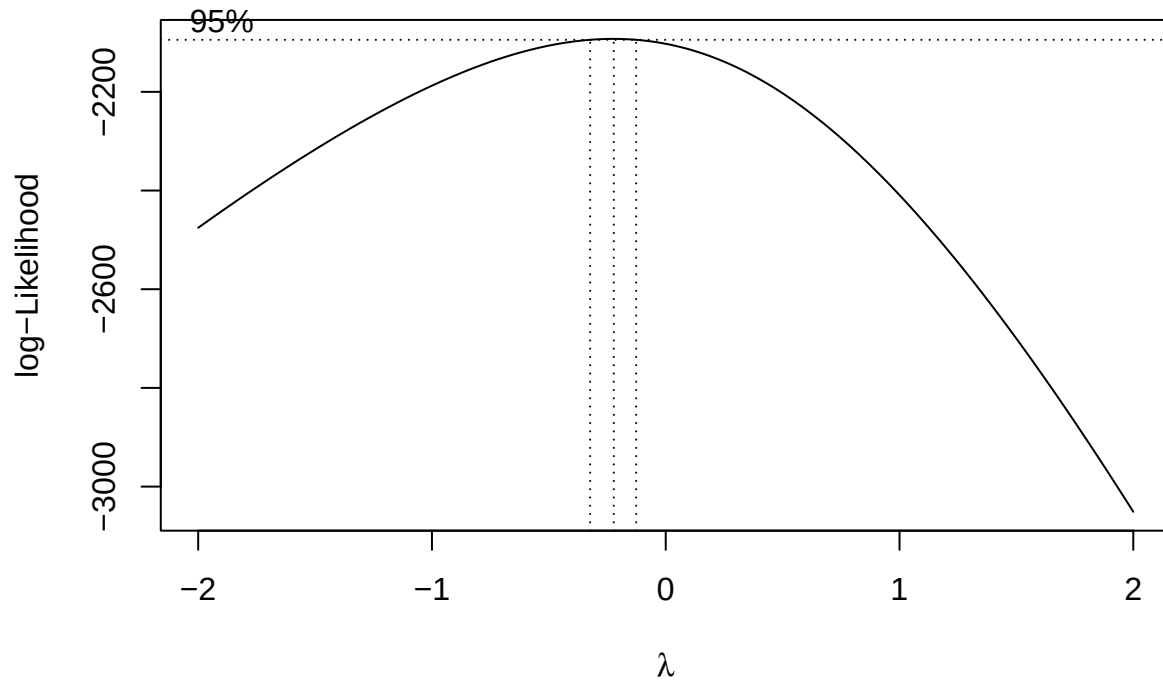
```
mod_add_pwo_logy = lm(log(det_freq + 0.1) ~ (mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
      dist_to_coast_mi + elevation_ft + dominant_eco_group +
      tmean_2024_C + ppt_2024_mm),
      data = ebird_features_df_pwo, weight = n_checklists)
summary(mod_add_pwo_logy)
```

```
##
## Call:
## lm(formula = log(det_freq + 0.1) ~ (mean_num_observers + mean_start_time_hr +
##   med_duration + mean_dist_km + dist_to_coast_mi + elevation_ft +
##   dominant_eco_group + tmean_2024_C + ppt_2024_mm), data = ebird_features_df_pwo,
##   weights = n_checklists)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -21.134  -1.332  -0.262   1.555  14.238
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.907406    0.693650  -4.19  3.1e-05 ***
## mean_num_observers    -0.070105    0.090022  -0.78  0.43639
## mean_start_time_hr    -0.046740    0.013651  -3.42  0.00065 ***
## med_duration         0.302184    0.049203   6.14  1.4e-09 ***
## mean_dist_km         0.299308    0.028431  10.53 < 2e-16 ***
## dist_to_coast_mi      0.001335    0.000571   2.34  0.01962 *
## elevation_ft         0.000240    0.000107   2.24  0.02564 *
## dominant_eco_groupDeveloped -0.223936    0.094963  -2.36  0.01864 *
## dominant_eco_groupForest    0.018880    0.095855   0.20  0.84391
## dominant_eco_groupWater    -0.366179    0.103194  -3.55  0.00041 ***
## dominant_eco_groupWetland   0.164270    0.099751   1.65  0.10004
## tmean_2024_C         0.090176    0.034885   2.58  0.00994 **
## ppt_2024_mm          -0.000203    0.000120  -1.70  0.09008 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.25 on 704 degrees of freedom
## Multiple R-squared:  0.418, Adjusted R-squared:  0.408
## F-statistic: 42.2 on 12 and 704 DF, p-value: <2e-16
```

This was informed by the result of a Box-Cox transformation, where the optimum λ was around 0:

```
mod_add_pwo_shift = lm((det_freq + 0.1) ~ mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
      dist_to_coast_mi + elevation_ft + dominant_eco_group +
      tmean_2024_C + ppt_2024_mm,
      data = ebird_features_df_pwo, weight = n_checklists)
```

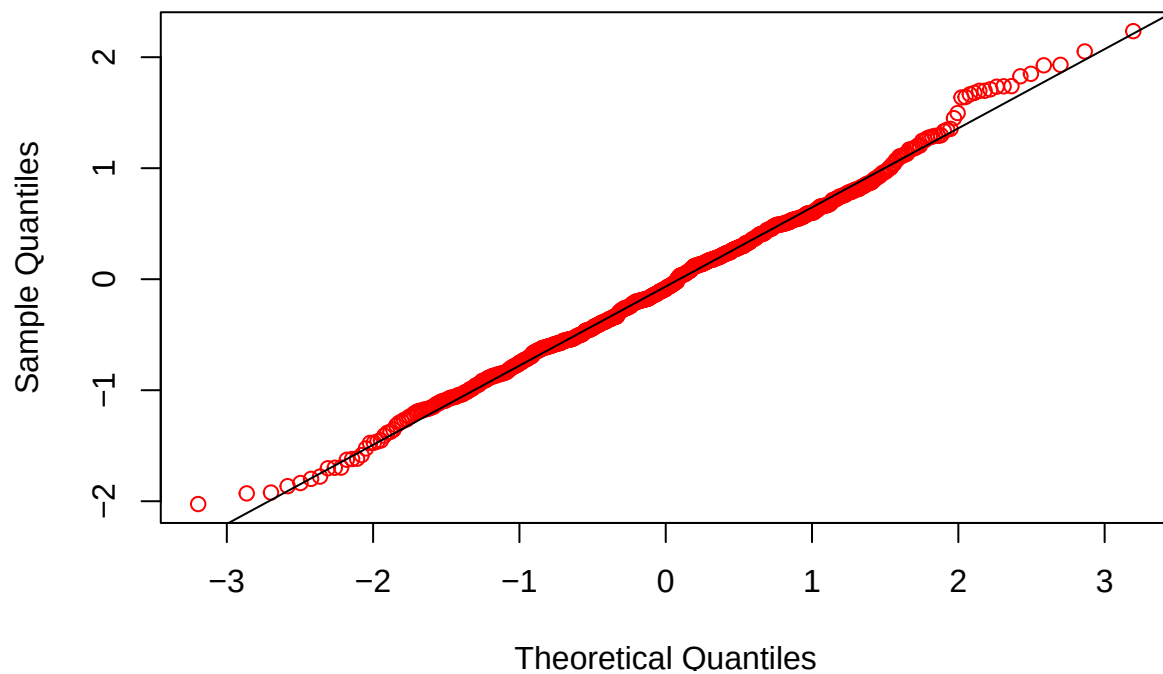
```
MASS::boxcox(mod_add_pwo_shift)
```



The normality diagnostics looked pretty good for the $\log(y)$ transformation. The fitted-vs-residuals plot still looked poor though, especially the nonlinearity indicated by spread not being centered at 0 as fitted value varied:

```
qqnorm(resid(mod_add_pwo_logy), col = "red")  
qqline(resid(mod_add_pwo_logy))
```

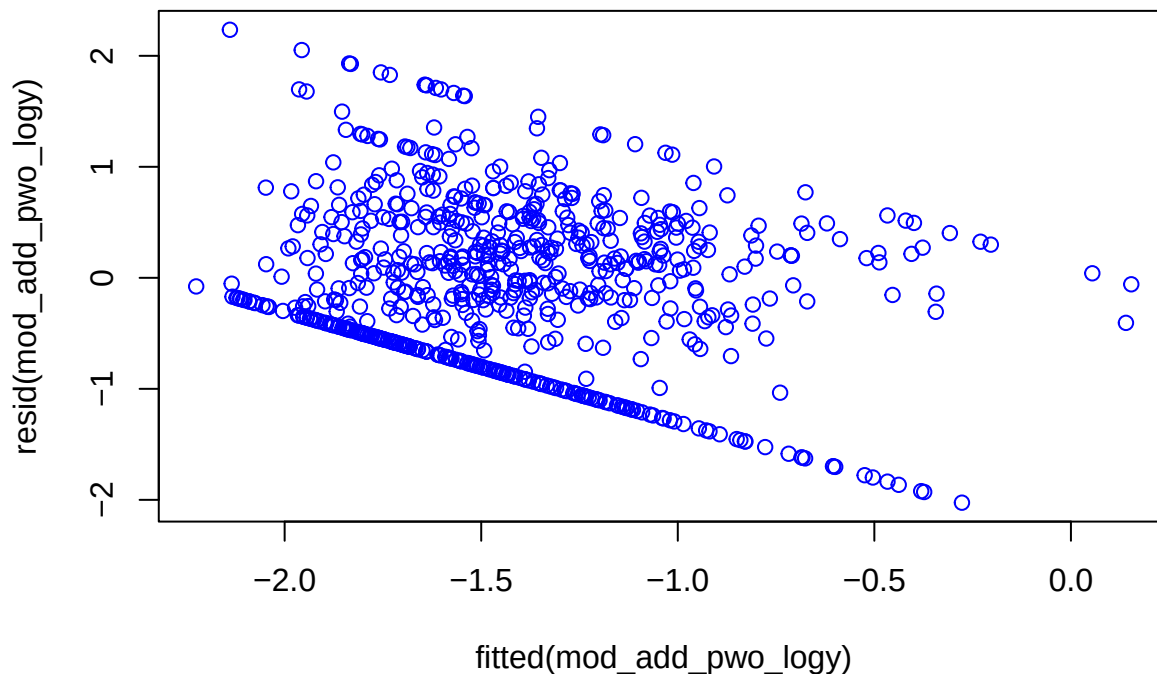
Normal Q-Q Plot



```
shapiro.test(resid(mod_add_pwo_logy))
```

```
##
## Shapiro-Wilk normality test
##
## data: resid(mod_add_pwo_logy)
## W = 1, p-value = 0.09
```

```
plot(resid(mod_add_pwo_logy) ~ fitted(mod_add_pwo_logy), col = "blue")
```



```
lmtest::bptest(mod_add_pwo_logy)
```

```
##
## studentized Breusch-Pagan test
##
## data: mod_add_pwo_logy
## BP = 54099, df = 12, p-value <2e-16
```

These are all the correlations among the predictors:

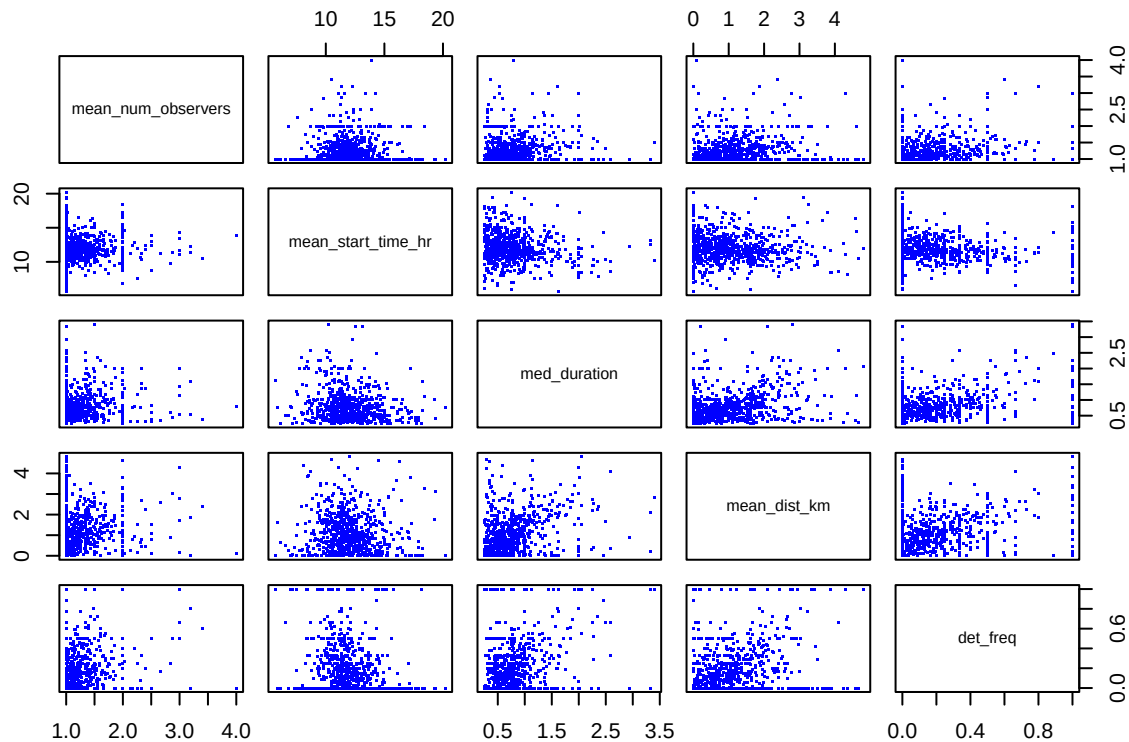
```
cor(select(ebird_features_df_pwo, mean_num_observers, mean_start_time_hr, med_duration, mean_dist_km,
          dist_to_coast_mi, elevation_ft,
          tmean_2024_C, ppt_2024_mm, det_freq))
```

```
##
## mean_num_observers mean_start_time_hr med_duration
## mean_num_observers      1.000000      0.001156      0.094743
## mean_start_time_hr      0.001156      1.000000     -0.141349
## med_duration            0.094743     -0.141349      1.000000
## mean_dist_km            0.197314     -0.054657      0.292165
## dist_to_coast_mi       -0.113967      0.031555     -0.049021
## elevation_ft           -0.051169     -0.006946     -0.001695
## tmean_2024_C            0.092338     -0.028316      0.017955
## ppt_2024_mm             0.033556     -0.091877      0.038975
## det_freq               0.124954     -0.200040      0.303986
```

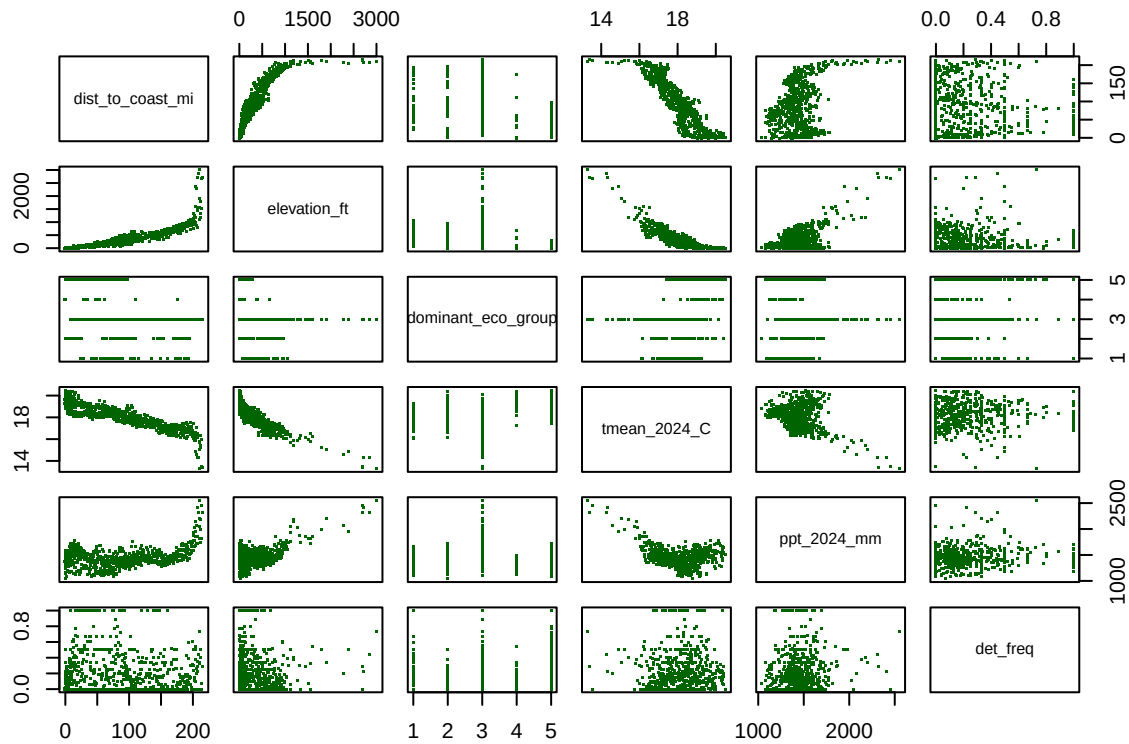
```
##               mean_dist_km dist_to_coast_mi elevation_ft tmean_2024_C
## mean_num_observers      0.19731      -0.11397      -0.051169      0.09234
## mean_start_time_hr     -0.05466       0.03156      -0.006946     -0.02832
## med_duration            0.29216      -0.04902      -0.001695      0.01795
## mean_dist_km            1.00000      -0.14243      -0.113575      0.10063
## dist_to_coast_mi       -0.14243       1.00000       0.841250     -0.88868
## elevation_ft           -0.11357       0.84125       1.000000     -0.84782
## tmean_2024_C           0.10063      -0.88868      -0.847816      1.00000
## ppt_2024_mm            -0.01565       0.32617       0.567524     -0.32850
## det_freq               0.21209      -0.11460      -0.064983      0.08308
##               ppt_2024_mm det_freq
## mean_num_observers      0.03356  0.12495
## mean_start_time_hr     -0.09188 -0.20004
## med_duration            0.03898  0.30399
## mean_dist_km           -0.01565  0.21209
## dist_to_coast_mi        0.32617 -0.11460
## elevation_ft            0.56752 -0.06498
## tmean_2024_C           -0.32850  0.08308
## ppt_2024_mm             1.00000  0.08319
## det_freq                0.08319  1.00000
```

I used these values and pairs plots as part of my assessment of interaction among predictors:

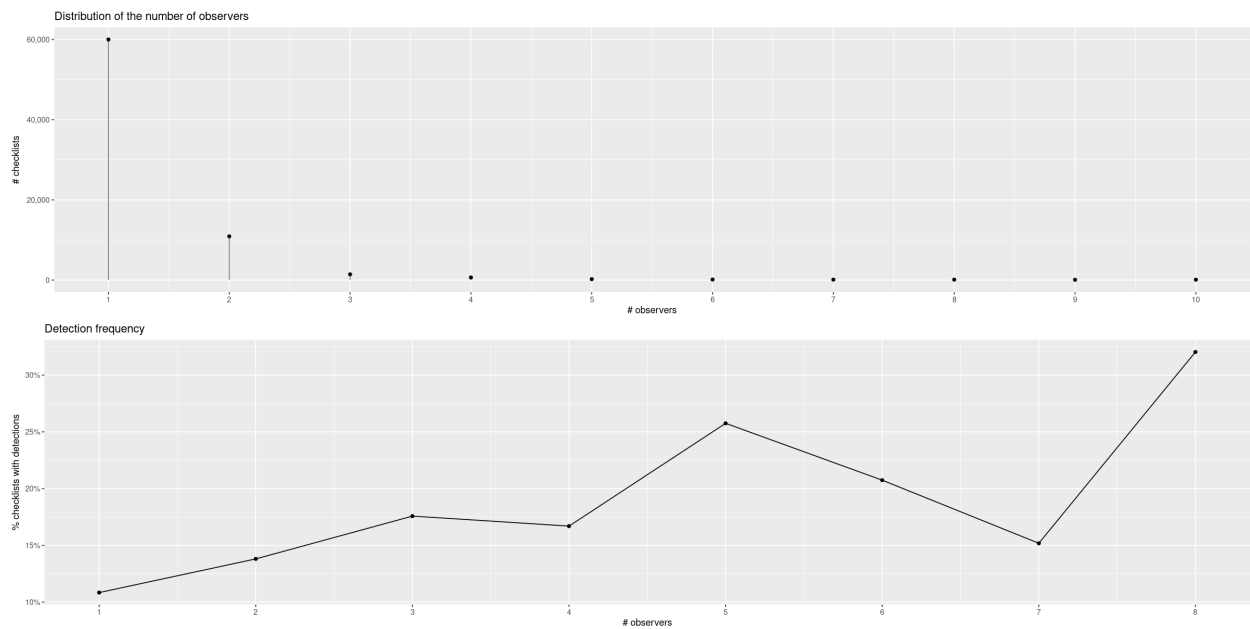
```
pairs(ebird_features_df_pwo[c(
  "mean_num_observers", "mean_start_time_hr", "med_duration", "mean_dist_km", "det_freq"
)], pch = ".", col = "blue")
```

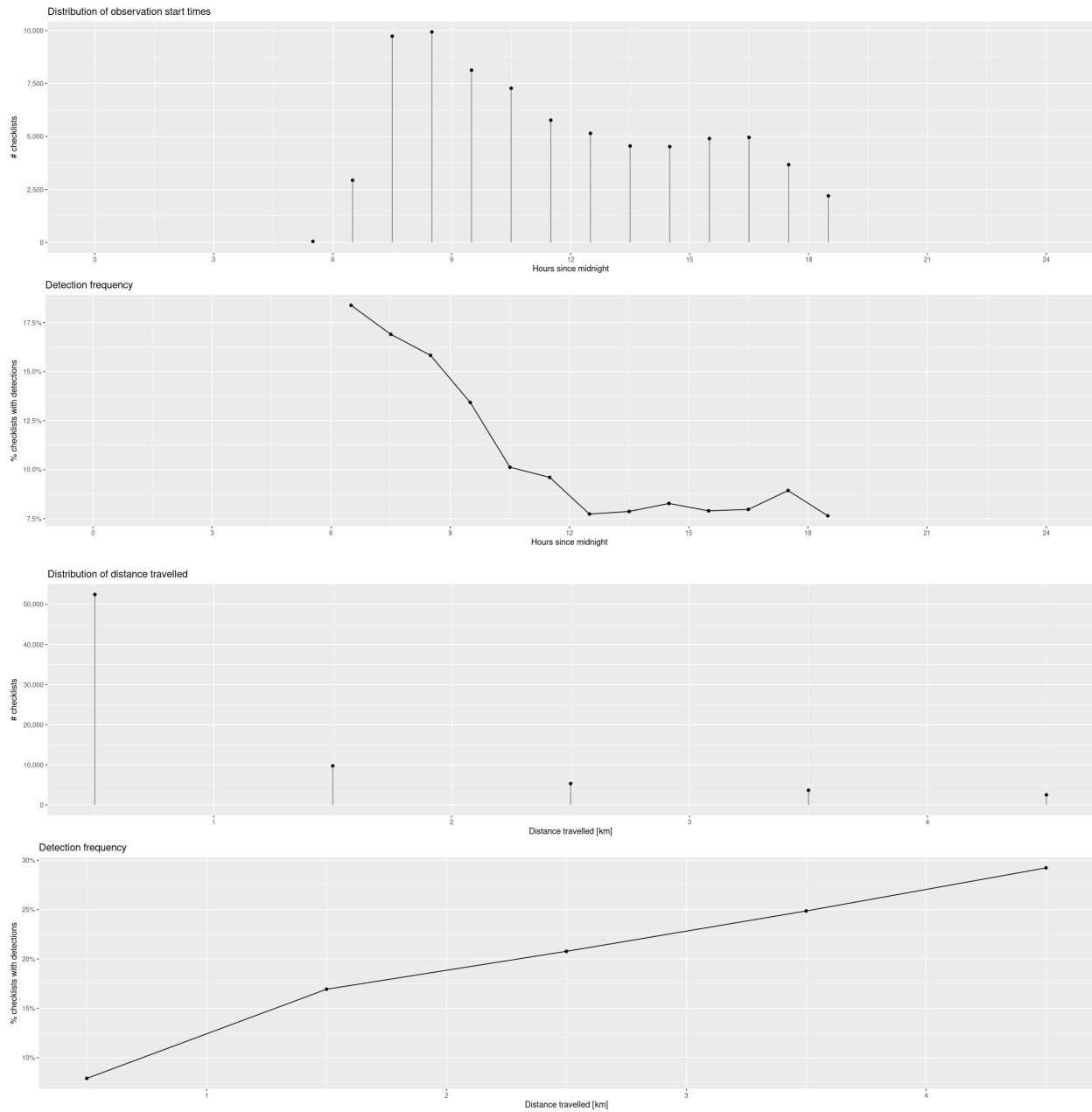


```
pairs(ebird_features_df_pwo[c(
  "dist_to_coast_mi", "elevation_ft", "dominant_eco_group",
  "tmean_2024_C", "ppt_2024_mm", "det_freq")], pch = ".", col = "darkgreen")
```



Here are some individual eBird effort-predictor trends I found interesting:

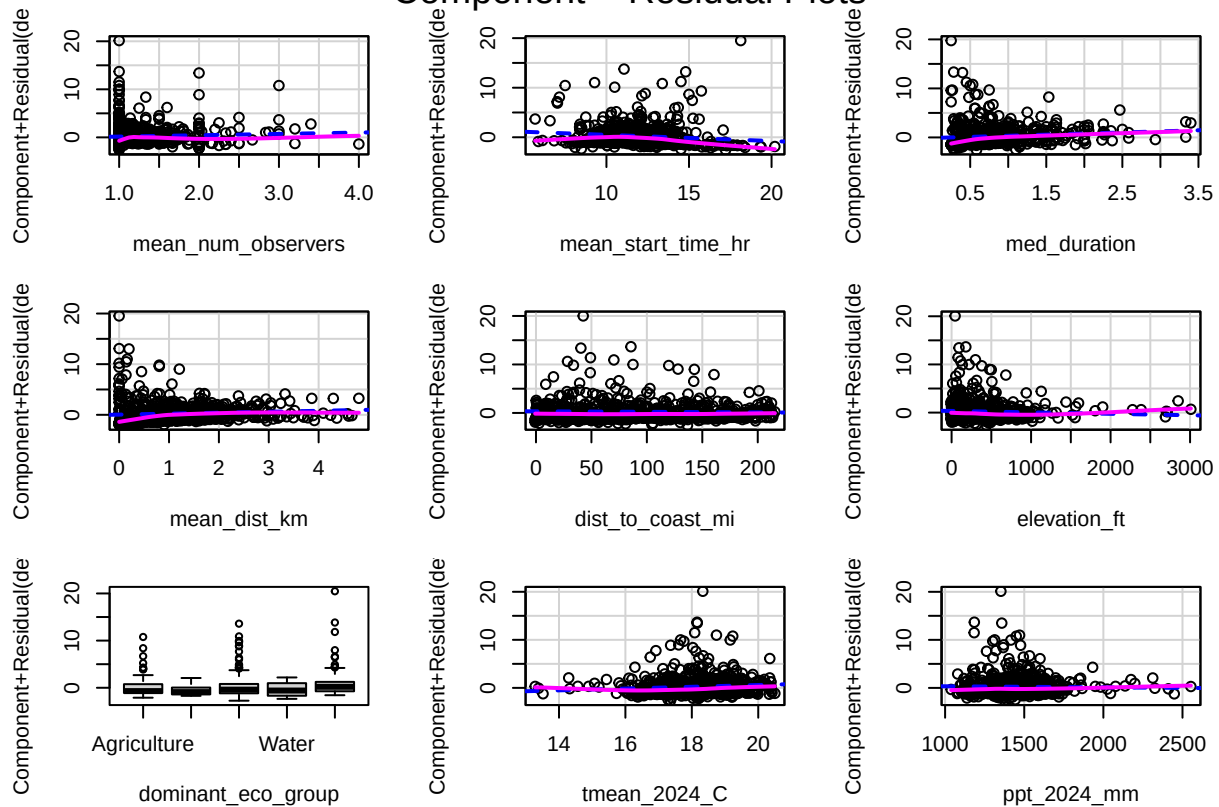




Here are the component + residual plots I evaluated to help me determine which predictor transformations to attempt. I experimented with polynomial terms for any variable with a significant deviation between the line and curve.

```
car::crPlots(mod_pwo_glm_base)
```

Component + Residual Plots



Improvements and Model Selection

Base model:

```
# base model
form_pwo_glm_base = det_freq ~
  (mean_num_observers + mean_start_time_hr + med_duration + mean_dist_km +
   dist_to_coast_mi + elevation_ft + dominant_eco_group +
   tmean_2024_C + ppt_2024_mm)
mod_pwo_glm_base = glm(form_pwo_glm_base,
  weights = n_checklists,
  data = ebird_features_df_pwo,
  family = binomial)
summary(mod_pwo_glm_base)
```

```
##
## Call:
## glm(formula = form_pwo_glm_base, family = binomial, data = ebird_features_df_pwo,
##      weights = n_checklists)
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.7208718   0.6199635   -2.78   0.0055 **
## mean_num_observers -0.1089660   0.0752673   -1.45   0.1477
## mean_start_time_hr -0.0987868   0.0123688   -7.99  1.4e-15 ***
## med_duration      0.6542476   0.0379704   17.23 < 2e-16 ***
## mean_dist_km      0.5429238   0.0230561   23.55 < 2e-16 ***
## dist_to_coast_mi   0.0022312   0.0005059    4.41  1.0e-05 ***
```

```
## elevation_ft          0.0001309  0.0000914    1.43  0.1518
## dominant_eco_groupDeveloped -0.5827603  0.0798888   -7.29  3.0e-13 ***
## dominant_eco_groupForest   -0.0832003  0.0784516   -1.06  0.2889
## dominant_eco_groupWater    -0.6625551  0.0912765   -7.26  3.9e-13 ***
## dominant_eco_groupWetland   0.3451259  0.0826658    4.17  3.0e-05 ***
## tmean_2024_C              0.0064977  0.0310714    0.21  0.8344
## ppt_2024_mm               -0.0000427  0.0001060   -0.40  0.6872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 7478.3  on 716  degrees of freedom
## Residual deviance: 4218.1  on 704  degrees of freedom
## AIC: 5752
##
## Number of Fisher Scoring iterations: 5
```

```
set.seed(42)
get_k10_cv_rmse_glm(form_pwo_glm_base, ebird_features_df_pwo)
```

```
## [1] 0.2324
```

Full poly model w/o interaction terms:

```
form_pwo_glm_full_poly = det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration + mean_dist.
I(mean_dist_km^2)+I(mean_dist_km^3) +
dist_to_coast_mi +
I(dist_to_coast_mi^2)+I(dist_to_coast_mi^3) +
elevation_ft +
I(elevation_ft^2) +
dominant_eco_group + tmean_2024_C + ppt_2024_mm)
mod_pwo_glm_full_poly = glm(form_pwo_glm_full_poly,
weights = n_checklists,
data = ebird_features_df_pwo,
family = binomial)
summary(mod_pwo_glm_full_poly)
```

```
##
## Call:
## glm(formula = form_pwo_glm_full_poly, family = binomial, data = ebird_features_df_pwo,
## weights = n_checklists)
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.49e+00  6.45e-01  -5.41  6.4e-08 ***
## mean_num_observers -1.62e-02  7.60e-02  -0.21  0.83089
## mean_start_time_hr -1.42e-01  1.30e-02 -10.92 < 2e-16 ***
## med_duration 7.87e-01  4.13e-02  19.06 < 2e-16 ***
## mean_dist_km 1.51e+00  1.39e-01  10.86 < 2e-16 ***
## I(mean_dist_km^2) -6.35e-01  9.71e-02  -6.54  6.0e-11 ***
## I(mean_dist_km^3) 9.50e-02  1.90e-02   4.99  5.9e-07 ***
## dist_to_coast_mi 4.70e-02  2.11e-03  22.25 < 2e-16 ***
## I(dist_to_coast_mi^2) -5.10e-04  2.89e-05 -17.65 < 2e-16 ***
## I(dist_to_coast_mi^3) 1.60e-06  9.83e-08  16.31 < 2e-16 ***
## elevation_ft -1.24e-03  2.33e-04  -5.33  1.0e-07 ***
```

```
## I(elevation_ft^2)          4.34e-07    6.52e-08    6.66    2.7e-11 ***
## dominant_eco_groupDeveloped -5.28e-01    8.19e-02   -6.45    1.1e-10 ***
## dominant_eco_groupForest    1.76e-02    8.04e-02    0.22    0.82638
## dominant_eco_groupWater     -5.71e-01    9.28e-02   -6.16    7.3e-10 ***
## dominant_eco_groupWetland    2.78e-01    8.28e-02    3.36    0.00079 ***
## tmean_2024_C                6.27e-02    3.19e-02    1.97    0.04940 *
## ppt_2024_mm                 1.51e-04    1.09e-04    1.39    0.16437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 7478.3  on 716  degrees of freedom
## Residual deviance: 3549.9  on 699  degrees of freedom
## AIC: 5094
##
## Number of Fisher Scoring iterations: 5
```

```
set.seed(42)
get_k10_cv_rmse_glm(form_pwo_glm_full_poly, ebird_features_df_pwo)
```

```
## [1] 0.2276
```

```
anova(mod_pwo_glm_base, mod_pwo_glm_full_poly)
```

```
## Analysis of Deviance Table
##
## Model 1: det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
##   mean_dist_km + dist_to_coast_mi + elevation_ft + dominant_eco_group +
##   tmean_2024_C + ppt_2024_mm)
## Model 2: det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
##   mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) + dist_to_coast_mi +
##   I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) + elevation_ft +
##   I(elevation_ft^2) + dominant_eco_group + tmean_2024_C + ppt_2024_mm)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         704         4218
## 2         699         3550  5         668    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Addition of polynomial terms is significant.

Next use backward selection w/ BIC:

```
n = nrow(ebird_features_df_pwo)
best_bic_mod_pwo_glm_poly = step(mod_pwo_glm_full_poly, direction = "backward", k = log(n), trace = FALSE)
```

Predictors removed:

```
setdiff(names(coef(mod_pwo_glm_full_poly)), names(coef(best_bic_mod_pwo_glm_poly)))
```

```
## [1] "mean_num_observers" "tmean_2024_C"          "ppt_2024_mm"
```

Get error and see if removed predictors were significant:

```
set.seed(42)
get_k10_cv_rmse_glm(formula(best_bic_mod_pwo_glm_poly), ebird_features_df_pwo)
```

```
## [1] 0.2276
```

```
anova(best_bic_mod_pwo_glm_poly, mod_pwo_glm_full_poly)
```

```
## Analysis of Deviance Table
##
## Model 1: det_freq ~ mean_start_time_hr + med_duration + mean_dist_km +
##      I(mean_dist_km^2) + I(mean_dist_km^3) + dist_to_coast_mi +
##      I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) + elevation_ft +
##      I(elevation_ft^2) + dominant_eco_group
## Model 2: det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
##      mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) + dist_to_coast_mi +
##      I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) + elevation_ft +
##      I(elevation_ft^2) + dominant_eco_group + tmean_2024_C + ppt_2024_mm)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         702        3558
## 2         699        3550  3      7.83    0.05 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Depending on significance level, removed predictors may or may not be significant.

This is my “curated” interaction model:

```
# Based on what interaction terms seem the most significant in first curated-interaction model
form_pwo_glm_poly_cint3 = det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
      mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) +
      mean_num_observers:mean_start_time_hr +
      mean_num_observers:mean_dist_km +
      mean_start_time_hr:mean_dist_km +
      mean_start_time_hr:I(mean_dist_km^2) +
      mean_start_time_hr:I(mean_dist_km^3) +
      med_duration:mean_dist_km +
      med_duration:I(mean_dist_km^2) +
      med_duration:I(mean_dist_km^3) +
      dominant_eco_group +
      dist_to_coast_mi + I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) +
      elevation_ft + I(elevation_ft^2) +
      tmean_2024_C +
      ppt_2024_mm +
      dist_to_coast_mi:ppt_2024_mm +
      I(dist_to_coast_mi^2):ppt_2024_mm +
      I(dist_to_coast_mi^3):ppt_2024_mm
      )
mod_pwo_glm_poly_cint3 = glm(form_pwo_glm_poly_cint3,
      weights = n_checklists,
      data = ebird_features_df_pwo,
      family = binomial)
summary(mod_pwo_glm_poly_cint3)
```

```
##
## Call:
## glm(formula = form_pwo_glm_poly_cint3, family = binomial, data = ebird_features_df_pwo,
##      weights = n_checklists)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) 1.88e+00 1.10e+00 1.70 0.08828 .
## mean_num_observers -4.82e+00 6.95e-01 -6.93 4.1e-12 ***
## mean_start_time_hr -4.48e-01 7.19e-02 -6.24 4.4e-10 ***
## med_duration -7.34e-02 1.33e-01 -0.55 0.57979
## mean_dist_km 4.24e+00 1.19e+00 3.58 0.00035 ***
## I(mean_dist_km^2) -2.36e+00 8.80e-01 -2.68 0.00744 **
## I(mean_dist_km^3) 3.86e-01 1.74e-01 2.22 0.02630 *
## dominant_eco_groupDeveloped -5.99e-01 8.55e-02 -7.00 2.5e-12 ***
## dominant_eco_groupForest -8.08e-02 8.50e-02 -0.95 0.34179
## dominant_eco_groupWater -6.01e-01 9.70e-02 -6.20 5.7e-10 ***
## dominant_eco_groupWetland 1.80e-01 8.66e-02 2.08 0.03731 *
## dist_to_coast_mi -5.23e-03 1.82e-02 -0.29 0.77389
## I(dist_to_coast_mi^2) 3.66e-04 2.33e-04 1.57 0.11550
## I(dist_to_coast_mi^3) -1.31e-06 7.66e-07 -1.71 0.08718 .
## elevation_ft -1.30e-03 2.44e-04 -5.31 1.1e-07 ***
## I(elevation_ft^2) 4.48e-07 7.05e-08 6.36 2.0e-10 ***
## tmean_2024_C 1.75e-02 3.57e-02 0.49 0.62371
## ppt_2024_mm 1.27e-04 2.48e-04 0.51 0.61003
## mean_num_observers:mean_start_time_hr 3.91e-01 5.57e-02 7.01 2.3e-12 ***
## mean_num_observers:mean_dist_km 1.70e-01 1.07e-01 1.59 0.11135
## mean_start_time_hr:mean_dist_km -3.34e-01 1.00e-01 -3.33 0.00085 ***
## mean_start_time_hr:I(mean_dist_km^2) 1.79e-01 7.34e-02 2.43 0.01498 *
## mean_start_time_hr:I(mean_dist_km^3) -3.21e-02 1.46e-02 -2.20 0.02773 *
## med_duration:mean_dist_km 1.77e+00 3.78e-01 4.69 2.8e-06 ***
## med_duration:I(mean_dist_km^2) -8.60e-01 2.57e-01 -3.34 0.00082 ***
## med_duration:I(mean_dist_km^3) 1.52e-01 4.97e-02 3.06 0.00218 **
## dist_to_coast_mi:ppt_2024_mm 3.95e-05 1.30e-05 3.03 0.00246 **
## I(dist_to_coast_mi^2):ppt_2024_mm -6.54e-07 1.63e-07 -4.00 6.3e-05 ***
## I(dist_to_coast_mi^3):ppt_2024_mm 2.18e-09 5.30e-10 4.12 3.8e-05 ***
```

```
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## (Dispersion parameter for binomial family taken to be 1)
```

```
## Null deviance: 7478.3 on 716 degrees of freedom
```

```
## Residual deviance: 3348.0 on 688 degrees of freedom
```

```
## AIC: 4914
```

```
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
set.seed(42)
```

```
get_k10_cv_rmse_glm(form_pwo_glm_poly_cint3, ebird_features_df_pwo)
```

```
## [1] 0.2318
```

```
anova(mod_pwo_glm_base, mod_pwo_glm_poly_cint3)
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
```

```
## mean_dist_km + dist_to_coast_mi + elevation_ft + dominant_eco_group +
```

```
## tmean_2024_C + ppt_2024_mm)
```

```
## Model 2: det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
```

```
## mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) + mean_num_observers:mean_start_time_hr +
```

```
## mean_num_observers:mean_dist_km + mean_start_time_hr:mean_dist_km +
```

```
## mean_start_time_hr:I(mean_dist_km^2) + mean_start_time_hr:I(mean_dist_km^3) +
## med_duration:mean_dist_km + med_duration:I(mean_dist_km^2) +
## med_duration:I(mean_dist_km^3) + dominant_eco_group + dist_to_coast_mi +
## I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) + elevation_ft +
## I(elevation_ft^2) + tmean_2024_C + ppt_2024_mm + dist_to_coast_mi:ppt_2024_mm +
## I(dist_to_coast_mi^2):ppt_2024_mm + I(dist_to_coast_mi^3):ppt_2024_mm
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1 704 4218
## 2 688 3348 16 870 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Addition of interaction terms is significant.

Next use backward selection w/ BIC:

```
n = nrow(ebird_features_df_pwo)
best_bic_mod_pwo_glm_poly_cint3 = step(mod_pwo_glm_poly_cint3, direction = "backward", k = log(n), trace = TRUE)
summary(best_bic_mod_pwo_glm_poly_cint3)
```

```
##
## Call:
## glm(formula = det_freq ~ mean_num_observers + mean_start_time_hr +
## med_duration + mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) +
## dominant_eco_group + dist_to_coast_mi + I(dist_to_coast_mi^2) +
## I(dist_to_coast_mi^3) + elevation_ft + I(elevation_ft^2) +
## ppt_2024_mm + mean_num_observers:mean_start_time_hr + mean_start_time_hr:mean_dist_km +
## med_duration:mean_dist_km + med_duration:I(mean_dist_km^2) +
## med_duration:I(mean_dist_km^3) + dist_to_coast_mi:ppt_2024_mm +
## I(dist_to_coast_mi^2):ppt_2024_mm + I(dist_to_coast_mi^3):ppt_2024_mm,
## family = binomial, data = ebird_features_df_pwo, weights = n_checklists)
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.65e+00 8.43e-01 3.14 0.00166 **
## mean_num_observers -4.48e+00 6.45e-01 -6.96 3.5e-12 ***
## mean_start_time_hr -5.08e-01 6.47e-02 -7.85 4.1e-15 ***
## med_duration -4.98e-02 1.32e-01 -0.38 0.70475
## mean_dist_km 1.50e+00 3.80e-01 3.96 7.5e-05 ***
## I(mean_dist_km^2) -2.61e-01 2.46e-01 -1.06 0.28829
## I(mean_dist_km^3) 1.10e-02 4.91e-02 0.22 0.82213
## dominant_eco_groupDeveloped -5.90e-01 8.42e-02 -7.01 2.4e-12 ***
## dominant_eco_groupForest -5.89e-02 8.38e-02 -0.70 0.48218
## dominant_eco_groupWater -5.85e-01 9.42e-02 -6.21 5.3e-10 ***
## dominant_eco_groupWetland 2.04e-01 8.54e-02 2.38 0.01709 *
## dist_to_coast_mi -6.16e-03 1.78e-02 -0.35 0.72961
## I(dist_to_coast_mi^2) 3.49e-04 2.30e-04 1.52 0.12821
## I(dist_to_coast_mi^3) -1.19e-06 7.57e-07 -1.58 0.11448
## elevation_ft -1.35e-03 2.40e-04 -5.62 2.0e-08 ***
## I(elevation_ft^2) 4.63e-07 7.01e-08 6.61 3.8e-11 ***
## ppt_2024_mm 1.23e-04 2.21e-04 0.55 0.57954
## mean_num_observers:mean_start_time_hr 3.80e-01 5.43e-02 7.00 2.6e-12 ***
## mean_start_time_hr:mean_dist_km -8.21e-02 1.96e-02 -4.19 2.7e-05 ***
## med_duration:mean_dist_km 1.74e+00 3.75e-01 4.65 3.3e-06 ***
## med_duration:I(mean_dist_km^2) -8.53e-01 2.55e-01 -3.35 0.00081 ***
## med_duration:I(mean_dist_km^3) 1.50e-01 4.91e-02 3.06 0.00222 **
```

```
## dist_to_coast_mi:ppt_2024_mm          4.06e-05  1.27e-05   3.19  0.00141 **
## I(dist_to_coast_mi^2):ppt_2024_mm     -6.49e-07  1.62e-07  -4.02  5.8e-05 ***
## I(dist_to_coast_mi^3):ppt_2024_mm      2.13e-09  5.25e-10   4.06  5.0e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 7478.3  on 716  degrees of freedom
## Residual deviance: 3356.8  on 692  degrees of freedom
## AIC: 4915
##
## Number of Fisher Scoring iterations: 5
```

Predictors removed:

```
setdiff(names(coef(mod_pwo_glm_full_poly)), names(coef(best_bic_mod_pwo_glm_poly)))
```

```
## [1] "mean_num_observers" "tmean_2024_C"          "ppt_2024_mm"
```

Get error and see if removed predictors were significant:

```
set.seed(42)
get_k10_cv_rmse_glm(formula(best_bic_mod_pwo_glm_poly_cint3), ebird_features_df_pwo)
```

```
## [1] 0.2293
```

```
anova(best_bic_mod_pwo_glm_poly, mod_pwo_glm_poly_cint3)
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: det_freq ~ mean_start_time_hr + med_duration + mean_dist_km +
```

```
## I(mean_dist_km^2) + I(mean_dist_km^3) + dist_to_coast_mi +
```

```
## I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) + elevation_ft +
```

```
## I(elevation_ft^2) + dominant_eco_group
```

```
## Model 2: det_freq ~ (mean_num_observers + mean_start_time_hr + med_duration +
```

```
## mean_dist_km + I(mean_dist_km^2) + I(mean_dist_km^3) + mean_num_observers:mean_start_time_hr +
```

```
## mean_num_observers:mean_dist_km + mean_start_time_hr:mean_dist_km +
```

```
## mean_start_time_hr:I(mean_dist_km^2) + mean_start_time_hr:I(mean_dist_km^3) +
```

```
## med_duration:mean_dist_km + med_duration:I(mean_dist_km^2) +
```

```
## med_duration:I(mean_dist_km^3) + dominant_eco_group + dist_to_coast_mi +
```

```
## I(dist_to_coast_mi^2) + I(dist_to_coast_mi^3) + elevation_ft +
```

```
## I(elevation_ft^2) + tmean_2024_C + ppt_2024_mm + dist_to_coast_mi:ppt_2024_mm +
```

```
## I(dist_to_coast_mi^2):ppt_2024_mm + I(dist_to_coast_mi^3):ppt_2024_mm)
```

```
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1      702      3558
```

```
## 2      688      3348 14      210  <2e-16 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Citation for eBird data use:

eBird Basic Dataset. Version: EBD_relSep-2025. Cornell Lab of Ornithology, Ithaca, New York. Sep 2025.
(<https://science.ebird.org/en/use-ebird-data/citation>)