I have scrutinized the relationship between various geographic features and the counts of bigfoot or Sasquatch sightings observed not only at both state and county-level but also at the decade-level. The data comprises the reported number of bigfoots sighted in the Pacific Northwestern from $1968-2017$, specifically in the states of Washington, and Oregon.

In the entire dataset, the time-invariant variables are the five variables measuring distance characteristics including the spatial location via latitude and longitude variables. Moreover, the names of the states and counties are time-invariant as they are constant over time. Alternatively, the land characteristics and the number of bigfoot sightings are time-dependent covariates.

From the bigfoot dataset with measures of bigfoot sightings in Oregon and Washington, the summary statistics of the land characteristics variables show that the proportions very minutely change over time, as also evident from the very low standard deviations. Due to the high volatility of the variable population ($sd = 203,648.5$), I transformed it by taking its log to reduce the deviation from its mean, creating a variable called *log_pop*.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Bigfoot	450	1.436	2.791	0	0	2	21
Population	450	99,051.330	203,648.500	1,400	13,825.8	78,420.2	1,931,249
Water	450	0.014	0.020	0.0005	0.005	0.016	0.171
Developed	450	0.063	0.065	0.006	0.023	0.074	0.289
Forest	450	0.419	0.244	0.0001	0.223	0.614	0.783
Agriculture	450	0.144	0.165	0.002	0.025	0.259	0.715
Wetlands	450	0.023	0.022	0.001	0.008	0.032	0.090
Snow.Barren	450	0.009	0.015	0.00000	0.0004	0.010	0.080
Grass.Shrub	450	0.327	0.218	0.033	0.146	0.481	0.936

Below, I have summarized the number of bigfoots sighted in each decade. Over time, the average reported sightings have risen, reaching the peak to 3.547 in $1998-2007$ before falling to 1.547, albeit the maximum of bigfoots were reportedly sighted in $2008-2017$.

Summary Statistics of the Reported Number of Bigfoots Sighted by Decade

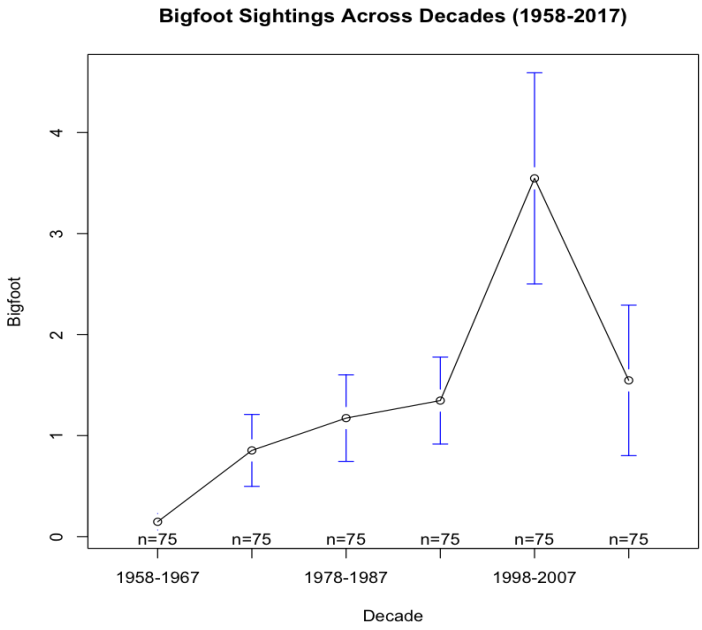
	Decade	min	q1	mean	q3	max
1	1958-1967	0	0	0.147	0	1
2	1968-1977	0	0	0.853	1	7
3	1978-1987	0	0	1.173	1.500	11
4	1988-1997	0	0	1.347	2	8
5	1998-2007	0	0	3.547	4.500	19
6	2008-2017	0	0	1.547	2	21

Likewise, a summary statistics by state showcases that people have observed on average 2 more bigfoots in Washington than in Oregon.

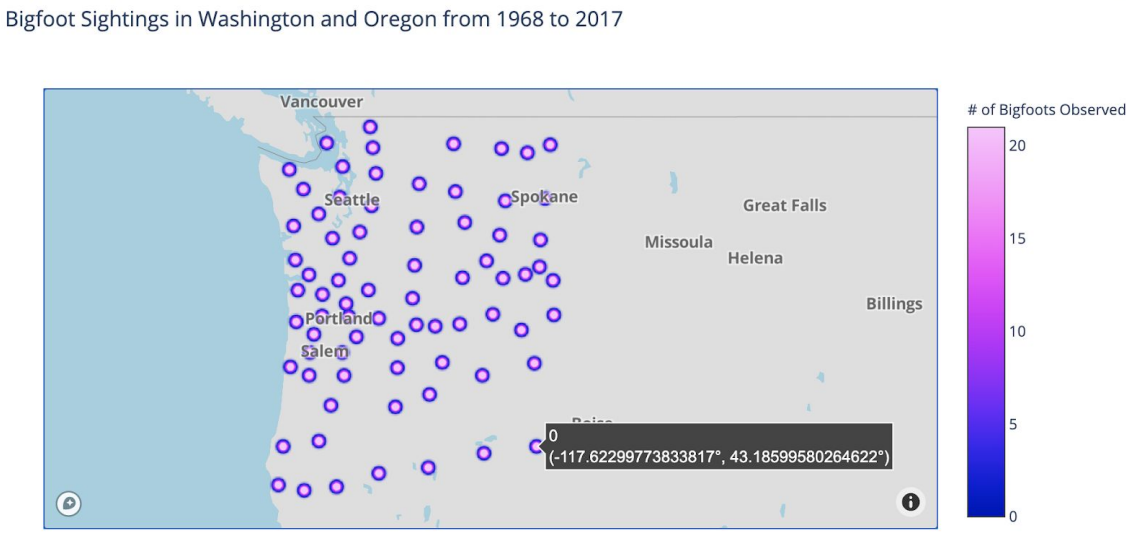
Summary Statistics of the Reported Number of Bigfoots Sighted by State

	State	min	q1	mean	q3	max
1	Oregon	0	0	0.806	1	8
2	Washington	0	0	2.017	2.750	21

A graphical illustration of the mean sightings over the decades corroborates the finding that many people saw bigfoots in the previous decade (1998 − 2007), and in general, a rising trend of sightings from approximately 0 in 1958-1967 to slightly more than 3 in 1998-2007. The vertical lines around the means or the bandwidths represent the 95 percent confidence intervals around the means.



The following spatial location plot of the map shows the regions' latitude, longitude.



Next, I have fitted a few regression models, whose estimates are shown on page $4-5$.

OLS with stepwise AIC approach with interactions

First, I fit 10 different OLS models by performing 10-fold cross-validation on the bigfoot dataset. I included separate intercepts for each decade and state. Thus, the state-specific, and decade-specific model via dummy variables controls for differences across states and decades. This allows each estimate of the factor variable (state and decade) to absorb effects pertaining to each state, and decade, respectively. After inserting all the explanatory variables, I chose a stepwise regression approach (a combination of forward and backward approaches) and chose the model with the lowest AIC. This automated process considers not only the combinations of linear terms but also their interactions. However, this approach yielded the best model as a very convoluted model best model consisting of all the variables and interactions with one another, most of which are statistically significant. However, the residuals vs fitted plot are heteroskedastic as the residuals are clustered on one side of the plot. Moreover, they are not normally distributed as they tail-off on both ends, implying they are skewed. The Appendix shows the model and the two diagnostic plots.

Fixed Effects

Due to the panel data structure, I fitted generalized linear mixed models, such as the fixed effects (FE), random effects (RE), and pooled OLS model, and then used hypothesis testing to compare which model is most appropriate. Unlike in OLS models fitted via maximizing the log-likelihood, in this case, we cannot use AIC as a metric to compare models.

Since we have to model the relationship between the land cover characteristics (time-variant variables) and count data $-$ the number of bigfoot sightings and the model does not have a time-invariant variable (except for the dummy variable for state), I first tried the FE model. It only analyzes the net effect of variables that change across the decades by eliminating the impact of time-invariant characteristics. Furthermore, it presupposes that the time-invariant variables are uncorrelated with the other county-specific characteristics. The FE model is:

$$Y_{it} = \beta_1 X_{it} + ... + \beta_k X_k + \alpha_i + u_{it}, \text{ where } i = 1, ..., n;\ t = 1, ..., T$$

α_i is the county-specific intercept that captures heterogeneity across counties. These variables don't change over time.

X_{it} is a matrix of all the time-variant variables such as the land characteristics described before. These variables are specific to county *i,* and time *t,* where *t* refers to the time measured in decades. Likewise, Y_{it} is the counts of bigfoots sighted at a particular county and decade. The fixed-effects model assumes that the residuals u_{it} are exogenous or have a conditional mean of 0 i.e. $E[u_{it} / X_{i1},\ X_{i2},\ ...,\ X_{it}] = 0$.

From the fixed effects model, all variables except for Wetlands are insignificant. The positive slopes of variables population, forest, *Snow.Barren*, and a dummy variable that captures the state of Washington implies that increasing the percent of county land covered with forest, wetlands, and snow or barren land raises the reported sightings of bigfoots. Furthermore, a person will see 0.988 (or estimated 1) more bigfoot in a county of Washington than that in Oregon.

From the model, if the proportion of snow increases by 1 percent, the number of bigfoots reported would rise by 0.234. So, a 5 percent increase in the barren land will raise bigfoot sightings by 1.17, and a 20 percent raise will raise the counts by 4.68. Likewise, growing the forested region by 5 percent increases the bigfoot sightings by 0.12335. However, if the forested region expands by about 40 percent, then we expect that the sightings of bigfoot will rise by 1.

Alternatively, if the proportion of county that is developed increases by 1 percent, then we expect that the reported sightings will diminish by 0.004751 bigfoots. So, if the county is developed by 5 percent more, then it will substantially diminish the bigfoots sighted by 0.02375.

Random Effects

Unlike the assumption in the FE model, we assume that the variation across counties is random and serially uncorrelated with the counts of bigfoots. In contrast to the FE model where the intercept absorbs the time-invariant characteristics, the RE model can incorporate time-invariant variables such as distance and county characteristics given in the dataset.

The random-effects model is $X_{it} = \beta X_{it} + \alpha + u_{it} + \varepsilon_{it}$,

where ε_{it} is the fixed-effects or within-county error, and u_{it} is the between-county error.

Having noticed from the summary statistics that Sasquatch enthusiasts have sighted more bigfoots in Washington than in Oregon, I added a distance variable called *Seattle*, which measures the distance from county centroid to Seattle in kilometers. Additionally, I incorporated a variable called $SqMi$, which measures a county's area in square miles. Both variables are statistically significant. Other statistically significant variables are log_pop, $Forest$, and $Wetlands$. The coefficient of log_pop is 0.474, so if the population increases by 1 percent, then the number of bigfoots sighted will rise very minutely by 0.00474.

Pooled OLS

Next, I fitted a pooled ordinary least squares (POLS) model and expanded the RE model by interacting wetlands with agriculture and adding a polynomial term for $Snow.Barren$, resulting in statistically significant estimates. Whilst I tried other polynomials and interactions manually, they were not statistically significant and reduced the adjusted R^2 further.

Poisson model with random intercepts

Finally, I fit a type of generalized linear mixed model, called the Poisson model. Without any offset, I fit a random intercept only Poisson model wherein there are six different intercepts corresponding to each decade. The model is:

$$Y_{it} \sim Pois(\lambda_i) : log(\lambda_i) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \eta_{it} + \eta_{it} t,$$

where i refers to the state; $(\beta_0, \beta_1, ..., \beta_k)$ are fixed-effects, and (η_{i1}, η_{i2}) are the random-effects. The reference or base decade is from 1958 to 1967.

To control for non-constant variance in FE, I calculated heteroskedasticity consistent coefficients. For each slope coefficient, I computed the clustered standard error for all panel models, which allows for autocorrelated and heteroskedastic errors within a county but not across counties.

Panel Regression Results with Robust Standard Errors

	Dependent variable: Bigfoot			
	Fixed Effects	Random Effects	Pooled OLS	Poisson
	(1)	(2)	(3)	(4)
Population	0.00001***			
	(0.00000)			

Constant		-2.115***	-6.853***	-4.959***
		(0.455)	(1.459)	(0.439)
log pop		0.474***	0.722***	0.435***
		(0.060)	(0.145)	(0.040)
Water	-15.657***	-9.141***	-11.539***	-6.239*
	(2.698)	(3.033)	(3.167)	(3.558)
Developed	-4.751**	-0.769	-1.155	-4.064***
	(1.876)	(0.842)	(1.832)	(0.931)
Forest	2.467***	2.149*	1.257*	2.529***
	(0.923)	(1.206)	(0.707)	(0.248)
Agriculture		-1.505***	0.141	
		(0.554)	(0.363)	
poly(Snow.Barren, 2)2			-8.989**	
			(3.902)	
Wetlands:Agriculture			-169.887***	
			(55.037)	
Wetlands	0.230	-19.349**	2.258	
	(3.719)	(8.285)	(4.249)	
Snow.Barren	23.409***	0.678	21.001**	-1.037
	(8.747)	(9.716)	(9.225)	(2.257)
factor(State)Washington	0.988***		1.087***	
	(0.380)		(0.388)	
Seattle		-0.007***		-0.002***
		(0.003)		(0.0005)
SqMi		0.0003**		
		(0.0001)		
Observations	450	450	450	450
R2	0.348	0.308	0.293	
Adjusted R2	0.330	0.294	0.277	
Log Likelihood				-588.678
Akaike Inf. Crit.				1,233.357
Bayesian Inf. Crit.				1,348.416
F Statistic	33.258*** (df = 7; 437)	195.718***	18.217*** (df = 10; 439)	
Note:				*p<0.1; **p<0.05; ***p<0.01

To choose between the FE and RE models, I performed the Hausman Test, which measures the hypothesis that:
$H_o : cov(\alpha_i, X_{it}) = 0$ => use a random-effects model,

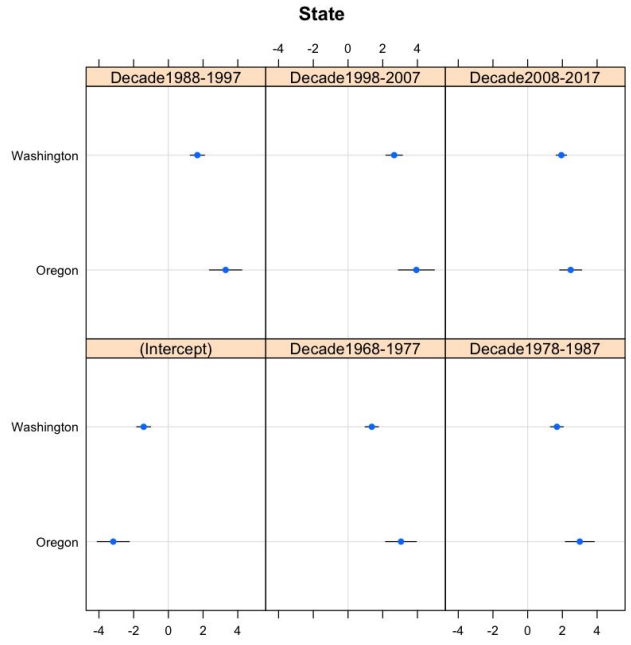
$H_A : cov(\alpha_i, X_{it}) \neq 0$ => use a fixed-effects model

Since the p-value of 2.2×10^{-16} is much less than the 0.05 significance level, we reject the null hypothesis and select the alternative i.e. FE model. However, in general, we have to be cautious of bias originating (if any) due to omitted factors that vary across counties, but not overtime, and other omitted factors that may also vary across time, biasing the estimates.

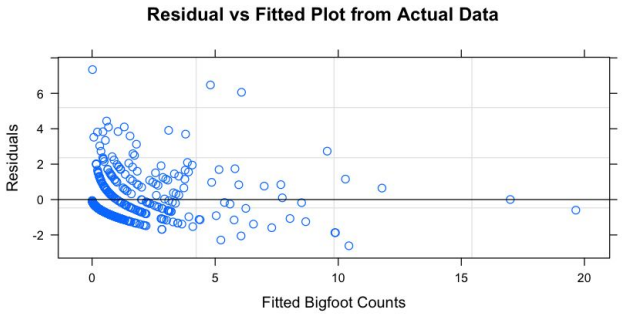
Moreover, I conducted the Breusch-Pagan Lagrange Multiplier test which helps select between the RE and the POLS model. The null hypothesis is that variances across counties are 0, signifying no panel effects. In this case, the POLS model would be sufficient. The results of the Breusch-Pagan LM test yield a very small p-value, so we reject the null hypothesis and conclude that there is a significant difference across counties, making the RE model more appropriate. This model does not distinguish between estimates of the same counties across different decades and is typically well-suited for cross-sectional data.

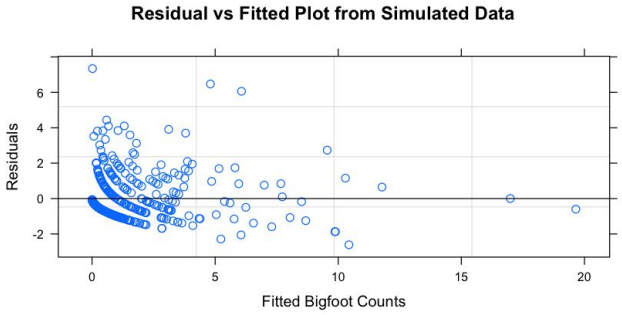
Additionally, I conducted the Breusch-Pagan test to check for the presence of heteroscedastic errors in the FE model. The null hypothesis is that the errors are homoskedastic. Again, a very small p-value statistically significant at 5 percent rejects the null hypothesis, so we conclude that the errors are heteroskedastic. To control for non-constant variance in FE, I calculated heteroskedasticity consistent coefficients. For each slope coefficient, I computed the clustered standard error for all panel models, which allows for autocorrelated and heteroskedastic errors within a county but not across counties.

In the Poisson model, through a visual inspection of a caterpillar plot below, for each state, the random-effects model shows different intercepts for each decade.



Below I have simulated a Poisson random variable where the mean is equal to the third model's fitted mean. Then, I plotted the residuals from the simulated data. The residual vs fitted plot from the simulated looks qualitatively very similar to that of the model fit from the true data.



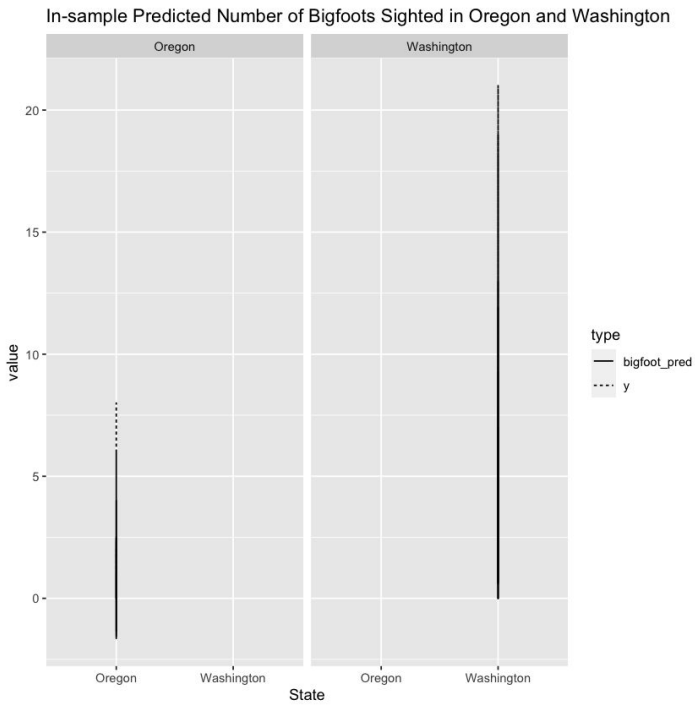


Predicted counts of bigfoot from the best model

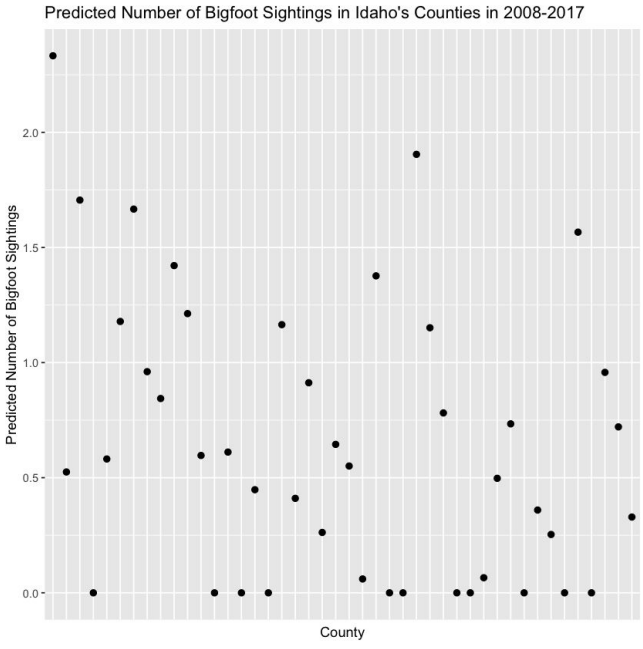
From the best model (FE) chosen from the Hausman test, I first predicted the count of bigfoot sightings for each county in Washington and Oregon and found that the results were very similar to the actual counts in the dataset. The summary statistics of the number of predicted sightings in Washington and Oregon, and the graphs of in-sample predictions show higher counts for Washington as expected. After obtaining the fitted values from the model, the root mean square error in the dataset was 2.0824.

Summary Statistics of the Predicted Number of Bigfoots Sighted by Decade in Washington and Oregon

	Decade	min	q1	mean	q3	max
1	1958-1967	-1.810	-0.738	0.157	0.705	5.507
2	1968-1977	-1.182	-0.036	0.869	1.360	7.502
3	1978-1987	-0.967	0.229	1.182	1.616	8.446
4	1988-1997	-0.875	0.348	1.337	1.795	10.006
5	1998-2007	1.200	2.510	3.548	4.051	13.435
6	2008-2017	-0.902	0.461	1.540	2.126	12.523



Finally, using the appropriate model, I predicted the number of bigfoot sightings for each county in the state of Idaho in $2008-2017$. After observing a few negative values of predicted counts, I set a floor value for all fitted values. That is, I set the value of 0, if the predicted count was negative, and then created a scatterplot for the predicted counts in each county. The scatterplot connotes a random spread of predicted counts across counties ranging from 0 to 3.



Conclusion

Thus, if the tourism board for Kittitas County in Washington would like to increase the number of Bigfoot sightings, it would be wise to expand the percent of barren and snow-clad land, forested region and shrink the percent of the developed region.

References

Christoph Hanck, M. (2020, September 15). Introduction to Econometrics with R. Retrieved from https://www.econometrics-with-r.org/10-rwpd.html

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. Retrieved from https://cran.r-project.org/web/packages/plm/vignettes/plmPackage.html

Fomite. (1961, June 01). Fitting a Poisson GLM mixed model with a random slope and intercept. Retrieved 2020, from https://stats.stackexchange.com/questions/27869/fitting-a-poisson-glm-mixed-model-with-a-random-slope-and-intercept

Reyna, O. T. (2010). Getting Started in Fixed/Random Effects Models using R. Retrieved from https://www.princeton.edu/~otorres/Panel101R.pdf

Appendix

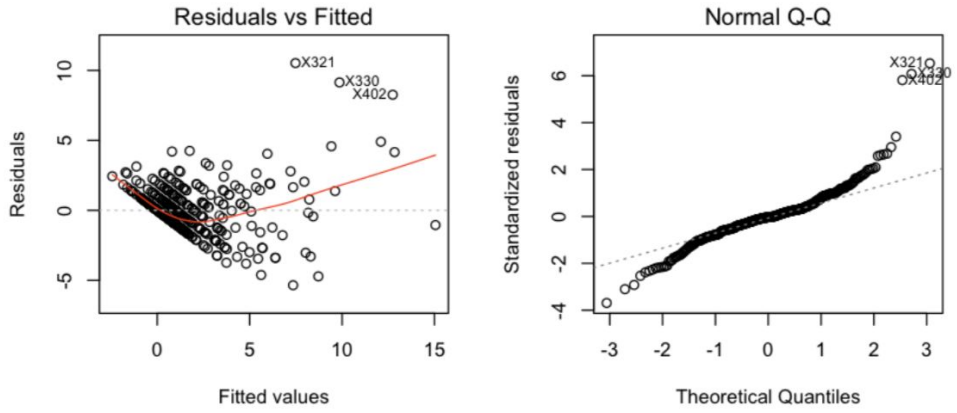
	Dependent variable:
	Bigfoot
Constant	-3,503.902***
	(539.000)
log_pop	956.418***
	(149.256)
Water	1,882.762***
	(663.256)
Developed	3,479.825***
	(546.176)
Forest	3,503.238***
	(538.498)
Agriculture	3,504.550***
	(538.275)
Wetlands	3,179.045***
	(546.610)
Grass.Shrub	3,510.452***
	(539.588)
`factor(Decade)1968-1977`	-3.737*
	(2.170)
`factor(Decade)1978-1987`	-4.598**
	(2.170)
`factor(Decade)1988-1997`	-3.679*
	(2.131)
`factor(Decade)1998-2007`	-4,877.224***
	(1,393.889)
`factor(Decade)2008-2017`	-5,438.553***
	(1,399.914)
`log_pop:Water`	-971.236***

	(149.775)
`log_pop:Developed`	-955.123***
	(149.727)
`log_pop:Forest`	-956.308***
	(149.170)
`log_pop:Agriculture`	-956.333***
	(149.183)
`log_pop:Wetlands`	-943.486***
	(149.768)
`log_pop:Snow.Barren`	-959.355***
	(149.468)
`log_pop:Grass.Shrub`	-957.200***
	(149.317)
`log_pop:factor(Decade)1968-1977`	0.412*
	(0.213)
`log_pop:factor(Decade)1978-1987`	0.496**
	(0.211)
`log_pop:factor(Decade)1988-1997`	0.457**
	(0.205)
`log_pop:factor(Decade)1998-2007`	1.045***
	(0.252)
`log_pop:factor(Decade)2008-2017`	0.765***
	(0.247)
`Water:Developed`	2,749.035***
	(644.984)
`Water:Forest`	1,582.380***
	(416.197)
`Water:Agriculture`	1,778.856***
	(417.501)
`Water:Wetlands`	4,156.858***
	(1,063.003)
`Water:Snow.Barren`	8,565.883***
	(1,479.180)
`Water:Grass.Shrub`	1,809.567***
	(442.074)
`Water:factor(Decade)1998-2007`	4,867.490***
	(1,395.654)
`Water:factor(Decade)2008-2017`	5,423.640***
	(1,401.662)
`Developed:Snow.Barren`	4,861.409***
	(569.254)
`Developed:factor(State)Washington`	-20.074***
	(6.217)
`Developed:factor(Decade)1998-2007`	4,858.778***
	(1,392.837)
`Developed:factor(Decade)2008-2017`	5,426.784***
	(1,398.878)

`Forest:Agriculture`	12.214***
	(4.044)
`Forest:Wetlands`	188.163***
	(64.990)
`Forest:Snow.Barren`	3,742.229***
	(560.973)
`Forest:factor(Decade)1998-2007`	4,871.989***
	(1,393.928)
`Forest:factor(Decade)2008-2017`	5,430.503***
	(1,399.950)
`Agriculture:Snow.Barren`	2,722.613***
	(545.336)
`Agriculture:factor(Decade)1998-2007`	4,863.872***
	(1,393.913)
`Agriculture:factor(Decade)2008-2017`	5,428.237***
	(1,399.932)
`Wetlands:Grass.Shrub`	328.658***
	(70.531)
`Wetlands:factor(State)Washington`	44.563***
	(10.025)
`Wetlands:factor(Decade)1998-2007`	4,837.672***
	(1,394.279)
`Wetlands:factor(Decade)2008-2017`	5,432.759***
	(1,400.294)
`Snow.Barren:Grass.Shrub`	3,468.761***
	(512.157)
`Snow.Barren:factor(State)Washington`	81.654***
	(25.229)
`Snow.Barren:factor(Decade)1968-1977`	21.666
	(15.986)
`Snow.Barren:factor(Decade)1978-1987`	42.077***
	(15.941)
`Snow.Barren:factor(Decade)1998-2007`	4,914.375***
	(1,400.599)
`Snow.Barren:factor(Decade)2008-2017`	5,506.642***
	(1,406.450)
`Grass.Shrub:factor(Decade)1998-2007`	4,867.628***
	(1,393.806)
`Grass.Shrub:factor(Decade)2008-2017`	5,431.785***
	(1,399.830)
`factor(State)Washington:factor(Decade)1998-2007`	2.861***
	(0.475)
`factor(State)Washington:factor(Decade)2008-2017`	1.582***
	(0.475)
Observations	450
R2	0.682

```
Adjusted R2                                                    0.634
Residual Std. Error                                      1.687 (df = 391)
F Statistic                                          14.433*** (df = 58; 391)
================================================================================
Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

The model's diagnostic plots that fail the normality and homoscedasticity assumptions are:



The full results of the poisson model with random intercepts:

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula: Bigfoot ~ log_pop + Water + Developed + Forest + Snow.Barren +
    Seattle + (1 + Decade | State)
   Data: bigfoot1

     AIC      BIC   logLik deviance df.resid
  1233.4   1348.4   -588.7   1177.4      422

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6110 -0.7115 -0.3465  0.3330  7.3405

Random effects:
 Groups Name             Variance Std.Dev. Corr
 State  (Intercept)       7.913   2.813
        Decade1968-1977   7.309   2.703    -1.00
        Decade1978-1987   7.659   2.768    -1.00  1.00
        Decade1988-1997   8.881   2.980    -1.00  1.00  1.00
        Decade1998-2007  14.377   3.792    -0.99  0.99  1.00  0.99
        Decade2008-2017   6.237   2.497    -0.97  0.97  0.99  0.98  1.00
Number of obs: 450, groups:  State, 2

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.9593243  0.4387237 -11.304  < 2e-16 ***
log_pop      0.4351563  0.0398343  10.924  < 2e-16 ***
Water       -6.2387837  3.5577281  -1.754   0.0795 .
Developed   -4.0638284  0.9309540  -4.365 1.27e-05 ***
Forest       2.5288438  0.2475158  10.217  < 2e-16 ***
Snow.Barren -1.0374103  2.2570834  -0.460   0.6458
```

```
Seattle     -0.0023364  0.0004624  -5.053 4.36e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) log_pp Water  Devlpd Forest Snw.Br
log_pop     -0.840
Water       -0.057 -0.039
Developed    0.312 -0.636 -0.162
Forest      -0.328 -0.048  0.053  0.122
Snow.Barren  0.092 -0.261 -0.034  0.341 -0.157
Seattle     -0.316 -0.083 -0.010  0.477  0.196  0.365
```