Homework 07 Solutions

10/30/2020

Sheather 8.3.2

First of all, note that the outcome values are not independent, because exactly 10 jurisdictions produce a top 10 finalist in a given year. Since the data ranges over 9 years, that means that the sum of the `Top10` values must be $9 \times 10 = 90$. That's actually a pretty strong constraint; it means, for example, that if we observe 10 jurisdictions with the value 9 for `Top10`, we automatically know the value for the other 41 jurisdictions (i.e., 0). Of course, we never expect regression assumptions to hold exactly in practice, but I'd take this model with a big grain of salt since the samples are so heavily dependent.

That said, let's go ahead and visualize the data (Figure 1).

Three predictors, `LogPopulation`, `LogContestants`, and `Latitude` have high marginal correlations with `Top10`. `LogTotalArea` and `Longitude` have pretty small correlations, but of course these are marginal correlations, so they could still be strongly related to the outcome.

There are some unsurprising correlations between some pairs of predictors. `LogPopulation` and `LogContestants` are positively correlated, which makes sense: larger populations produce more contestants. Longitude is positively correlated with `LogTotalArea`: western states are larger on average than eastern ones.

The predictors have reasonably symmetric and unimodal distributions, presumably thanks in part to the log transformations.

(a): Full model

Here's the full model, with a summary below. Standardized deviance residuals are plotted against each of the predictors in Figure 2. Marginal model plots are in Figure 3.

The residuals look essentially patternless, and the nonparametric curves for the fitted and the observed values in the marginal model plots look roughly the same. Since `Longitude` is not significant, and the problem requests a model in which all the predictors are significant, let's drop `Longitude` and refit (see below).

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.692	2.63	-2.925	0.003446
LogPopulation	0.6256	0.1845	3.391	0.0006959
LogContestants	1.417	0.4213	3.364	0.0007675
LogTotalArea	-0.3701	0.1393	-2.657	0.007892
Latitude	-0.06525	0.03028	-2.155	0.03115
Longitude	0.006509	0.009271	0.7021	0.4826

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	118.47 on 50 degrees of freedom
Residual deviance:	50.11 on 45 degrees of freedom

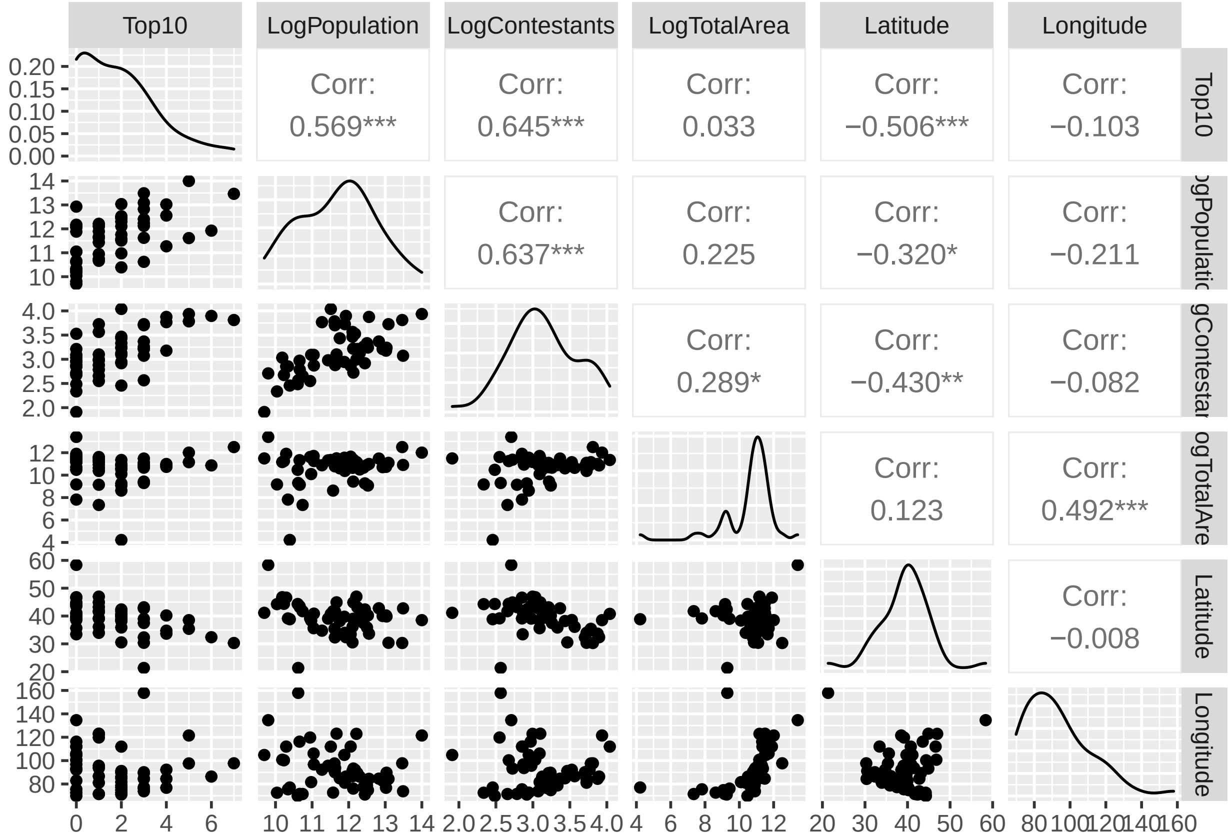

Figure 1: Pairs plot for Miss America data, excluding abbreviation column.

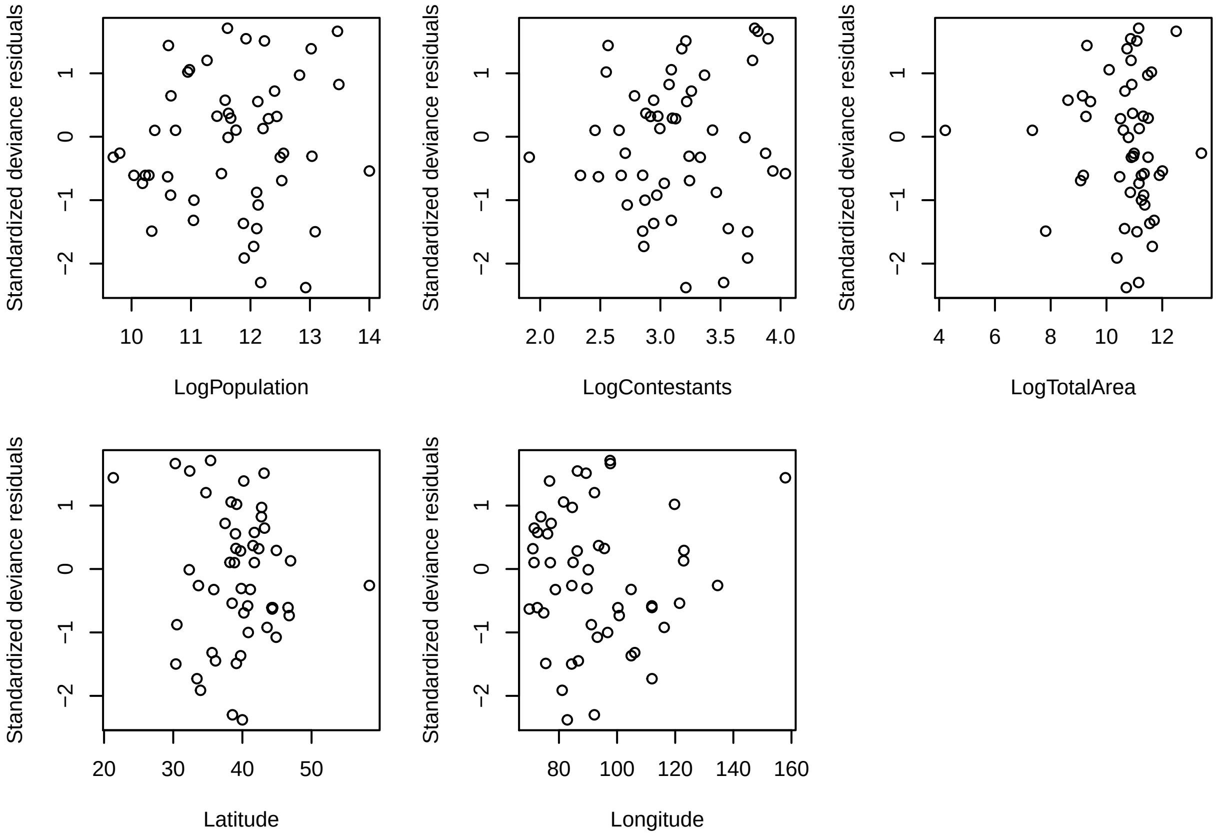

Figure 2: Standardized deviance residuals (left) against fitted values for the full model.

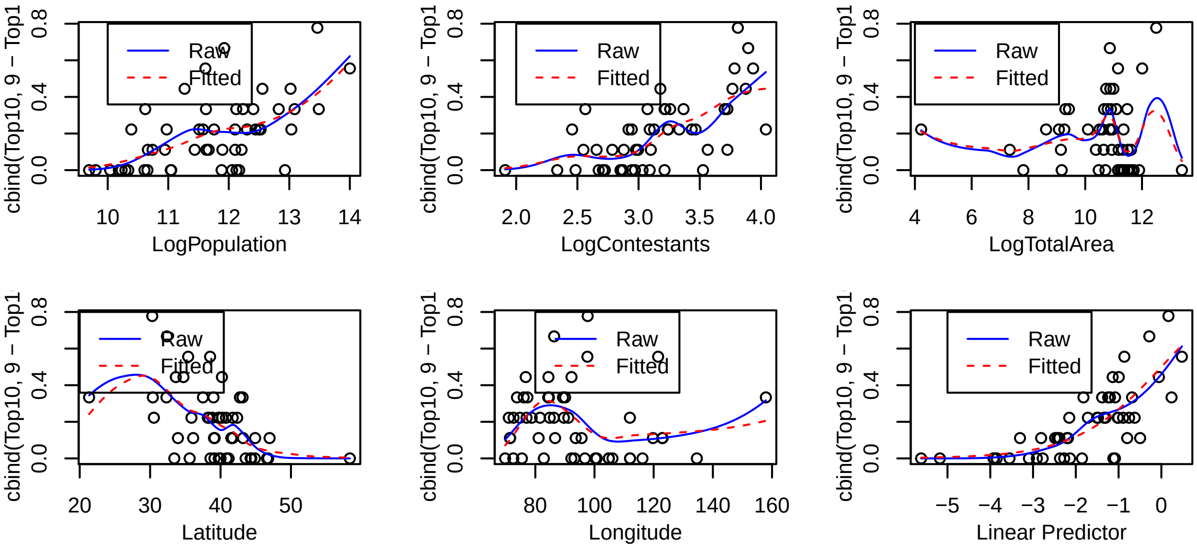

Figure 3: Marginal model plots for the full model

(a): Reduced model without Longitude

The model summary is below. Residuals are plotted in Figure 2 and the marginal model plots are in Figure 5.

Once again, the residuals look essentially patternless, and the curves in each marginal model plot follow each other closely. All the predictors are now statistically significant. The AIC for this model (142.79) is also lower than for the full model (144.3). This model seems reasonable.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.619	2.15	-3.079	0.002075
LogPopulation	0.5888	0.1758	3.35	0.0008074
LogContestants	1.337	0.4104	3.258	0.001123
LogTotalArea	-0.3198	0.1204	-2.656	0.007903
Latitude	-0.0733	0.029	-2.528	0.01148

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	118.47 on 50 degrees of freedom
Residual deviance:	50.59 on 46 degrees of freedom

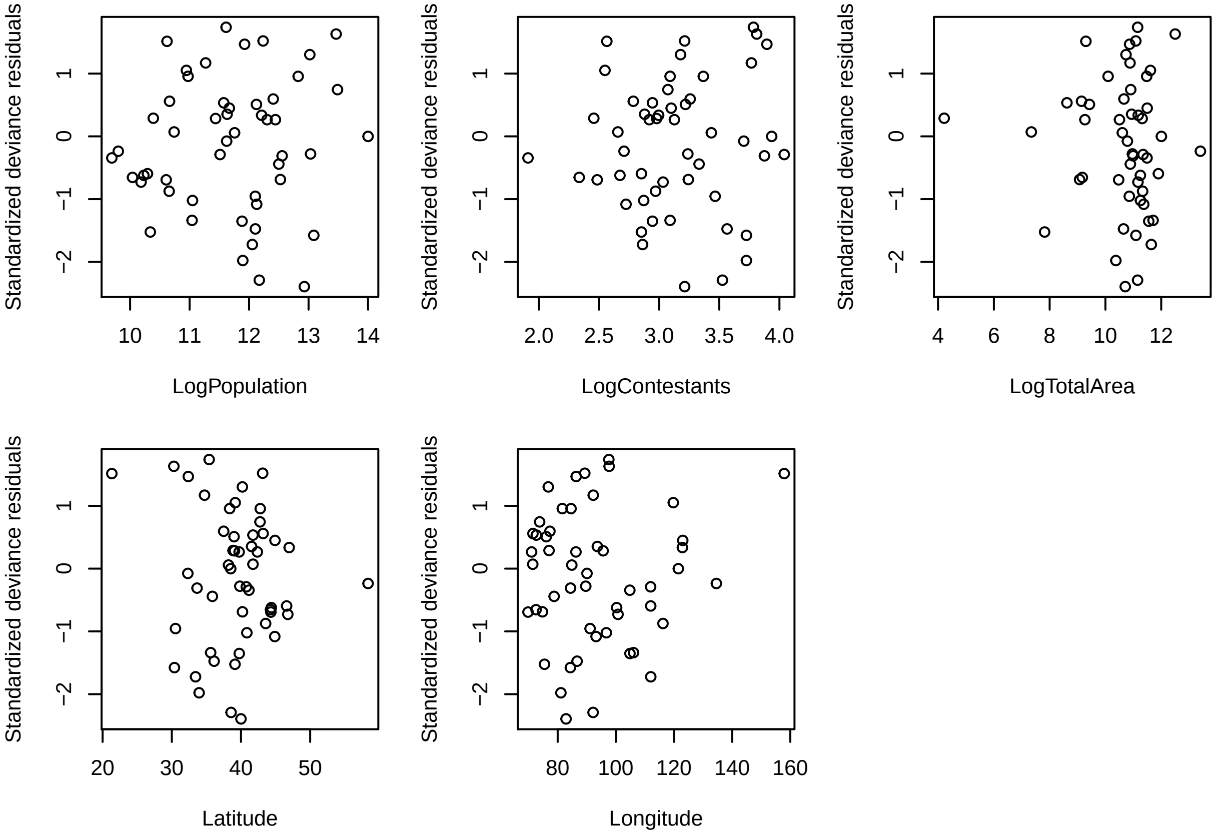

Figure 4: Standardized deviance residuals (left) against fitted values for the reduced model.

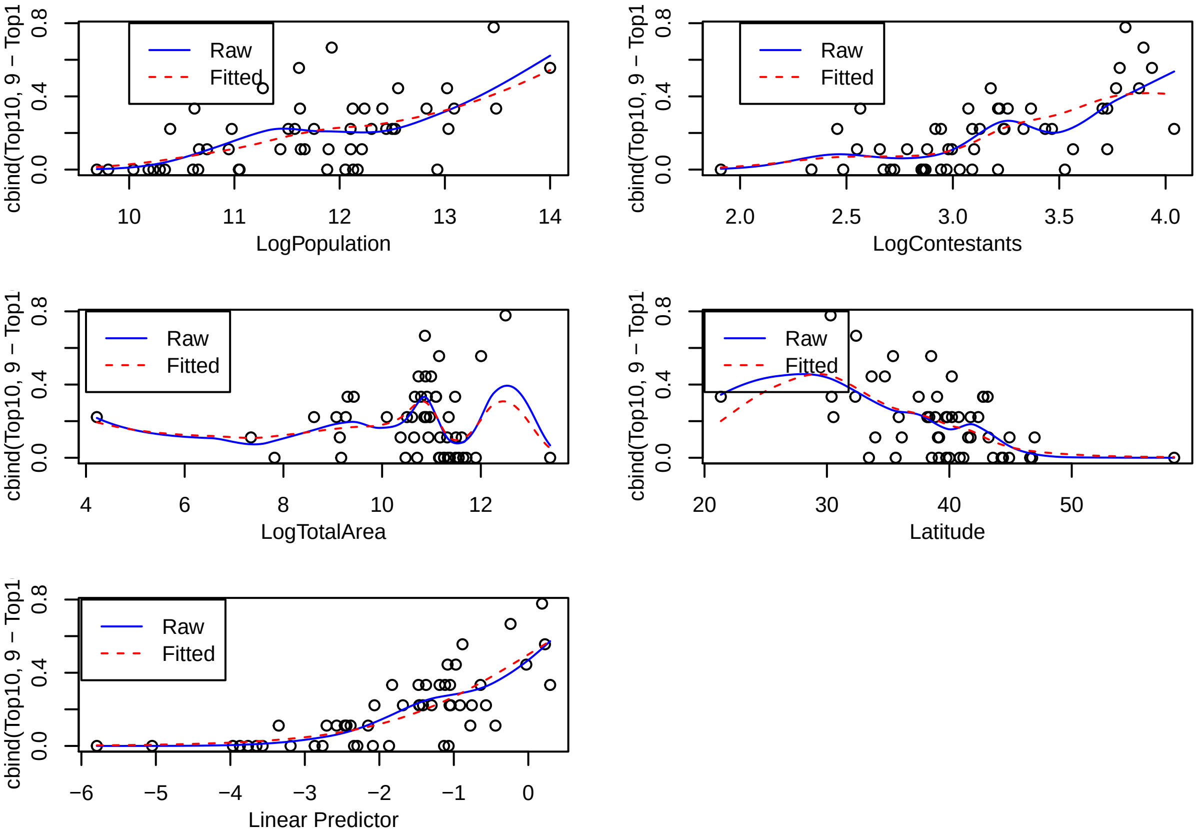

Figure 5: Marginal model plots for the reduced model

(b)

Recall that a "bad" leverage point was defined as a leverage point that is also an outlier. To detect this, let's examine the diagnostic plot that R generates with standardized Pearson residuals on the y-axis and leverage on the x-axis, in Figure 6.

Point 12 has both high leverage and a high standardized residual value, so it could be considered a bad leverage point. Let's refit the model without that point.

Once again, the residual plots (Figure 7) and marginal model plots (Figure 8) look good. From the summary below, we see that the coefficient for `Latitude` is no longer significant. Point 12 corresponded to Hawaii, which has a very low latitude, so it makes sense that that point was responsible for making `Latitude` statistically significant. The model without Hawaii is probably a better model of the continental US. (Alaska is still in the picture, but evidently it's not a bad leverage point, since it didn't show up as such on the plot.)

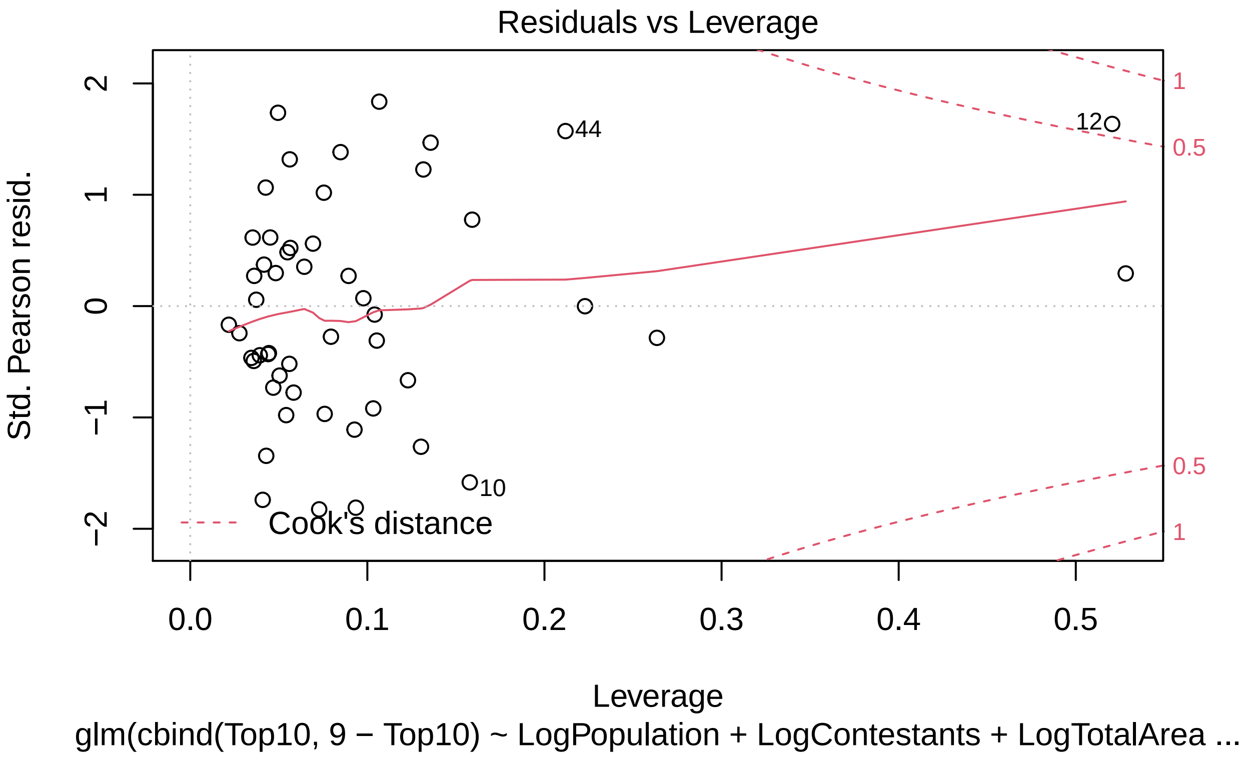

Figure 6: Residuals vs. leverage for the reduced model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.741	2.938	-3.315	0.0009168
LogPopulation	0.6171	0.1768	3.489	0.0004839
LogContestants	1.772	0.4994	3.549	0.0003872
LogTotalArea	-0.354	0.1286	-2.753	0.005905
Latitude	-0.02955	0.03938	-0.7504	0.453

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	117.51 on 49 degrees of freedom
Residual deviance:	48.07 on 45 degrees of freedom

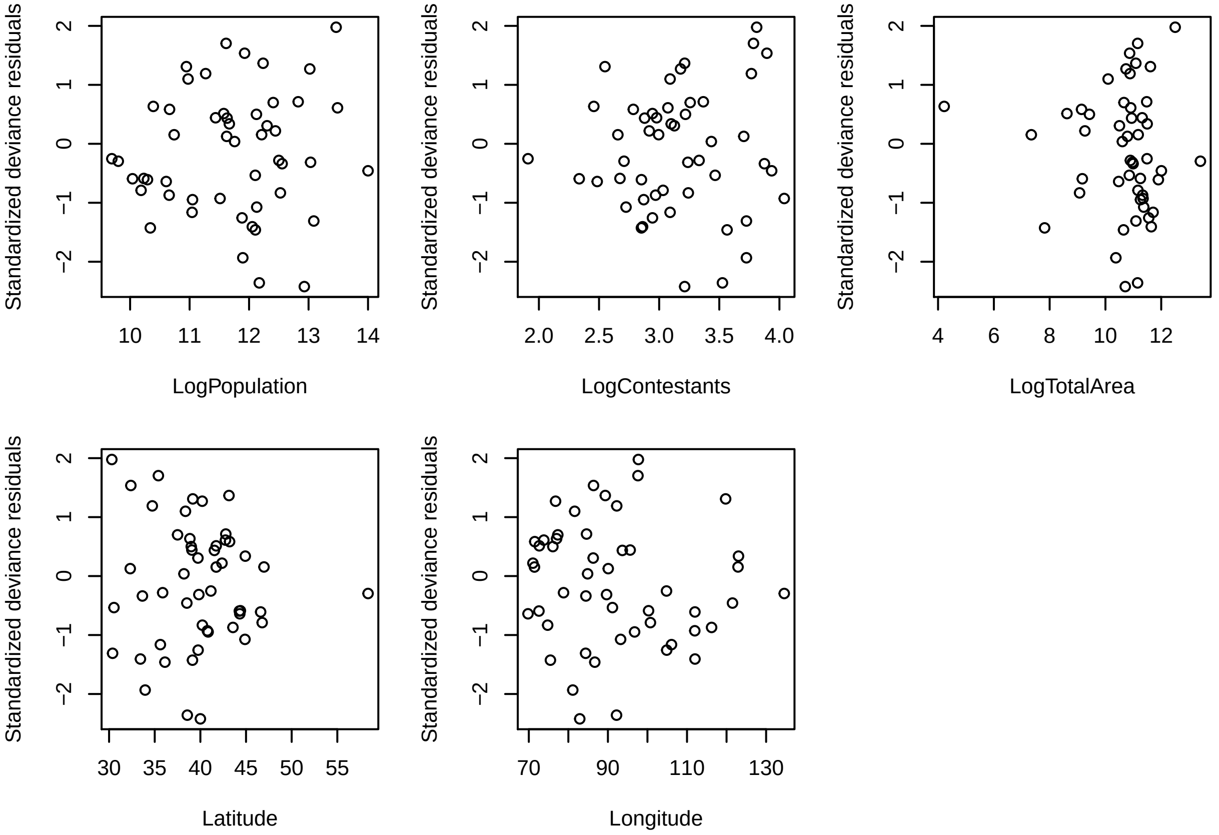

Figure 7: Standardized deviance residuals (left) against fitted values for the reduced model without point 12.

As described above, a “bad” point is related to high residuals and high leverage (making it influential). In order to assess this, we need the residuals to meet our expectations (leverage is only related to the dependent variables, not the model). In Figure 9 we can observe that, only once does a binned grouping have a average residual beyond the 2 standard deviations, this suggests that the relationship between residuals and fitted values isn’t too concerning. Moreover, we can’t observe any nonconstant pattern with the residual averages. Additionally, from the marginal plots we know that only 2 observations have expected values below -4, which might give us pause to overinterpret the plot. Under the model assumptions, we can simulate potential outcomes and compare these to the actual values. Figure 10 suggests that the residuals observed compared to those from the simulations do not suggest significant differences, specifically the ranks for the residuals vs what ranking would be expected seem to have similar distributions.

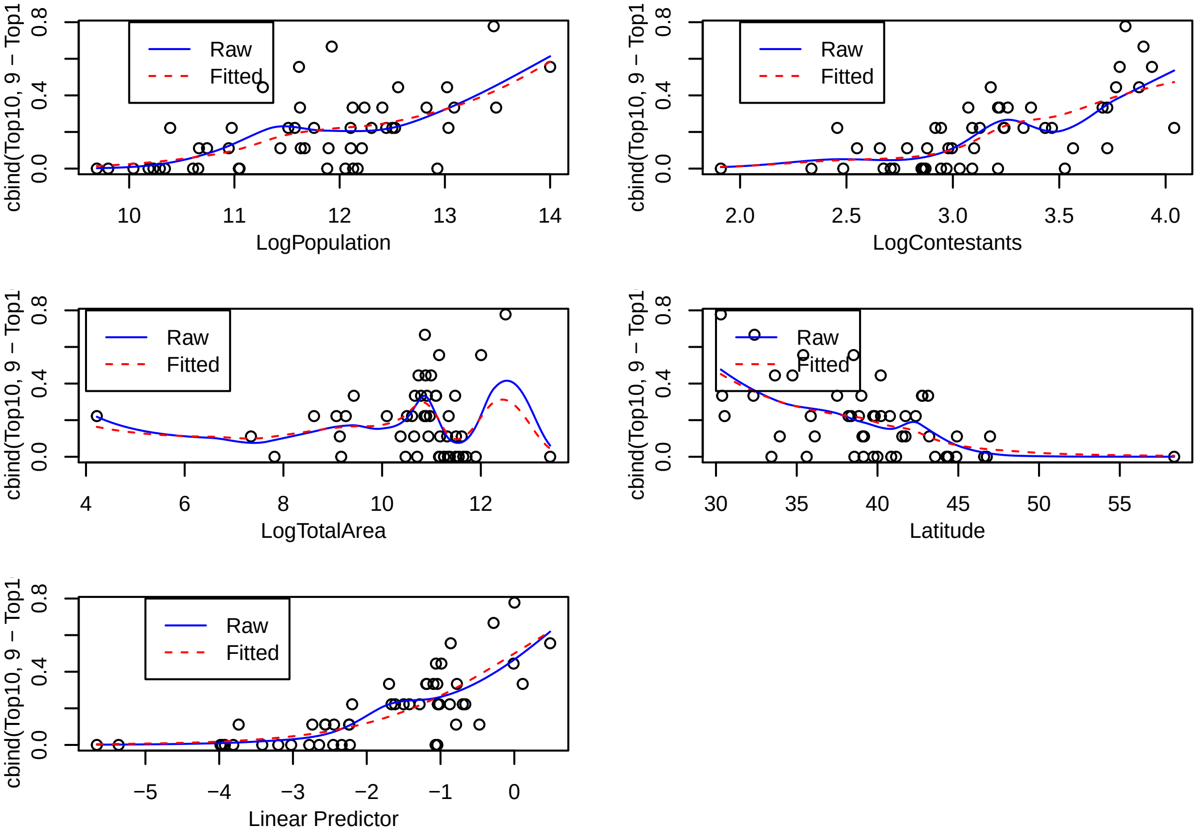

Figure 8: Marginal model plots for the reduced model without point 12.

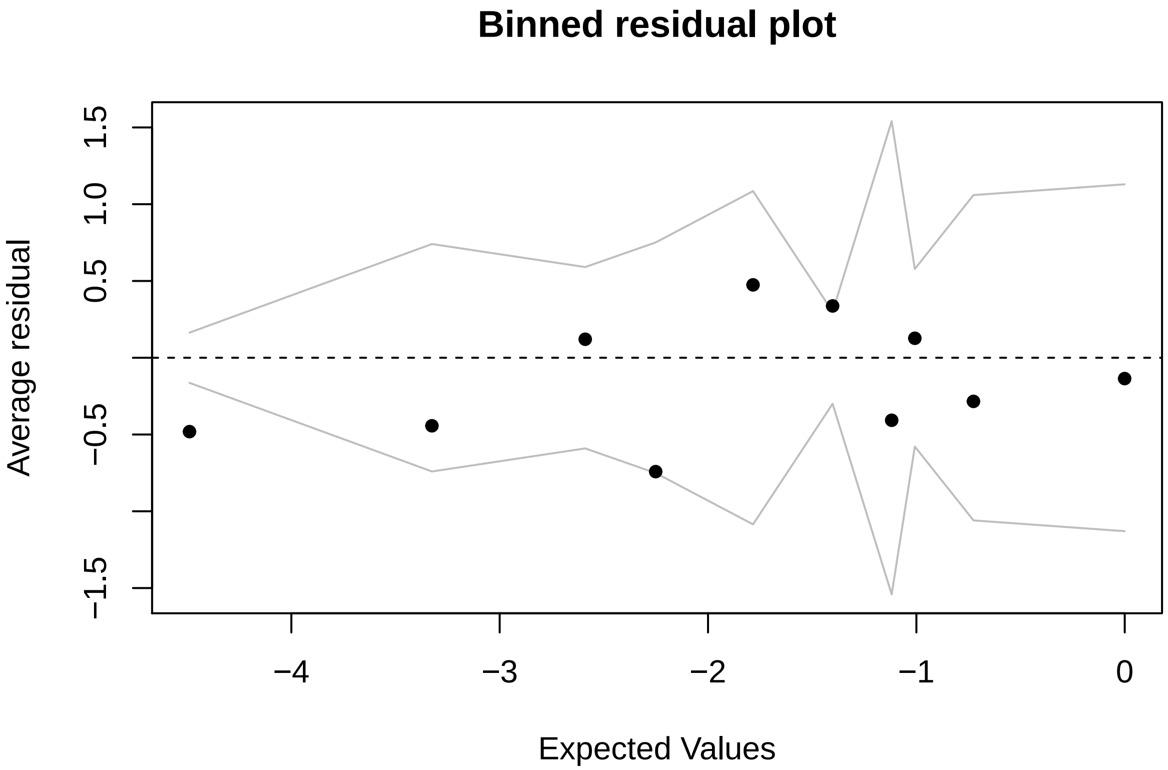

Figure 9: Binned residual plot for second model.

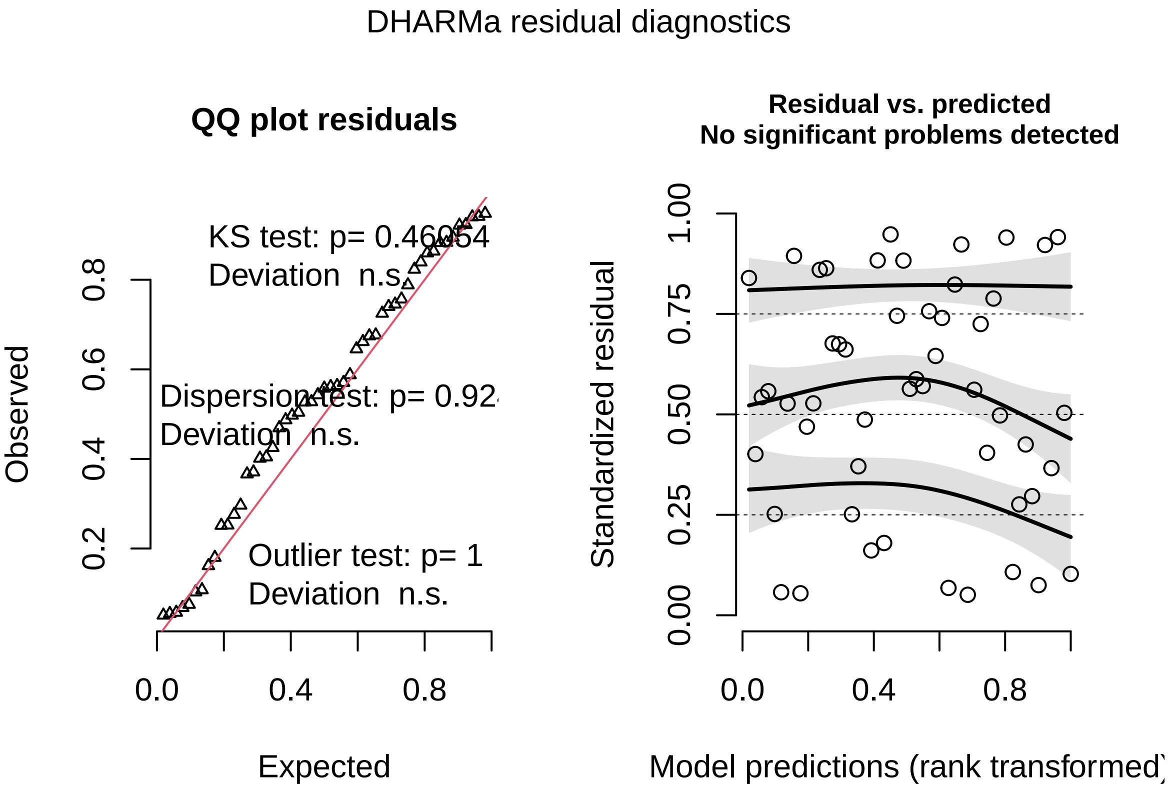

Figure 10: Simulation based assessment of residuals.

(c)

Let's interpet the coefficients from the model without Hawaii.

Mathematically, the intercept means that when the covariates are all 0, the estimated log odds of producing a top 10 finalist is -9.741, which corresponds to a probability of 0.000059. Of course, a jurisdiction with 0 for all covariate values is not realistic or of interest in the model.

Note that the odds are

$$\frac{\theta(X)}{1-\theta(X)}$$

where $\theta(X)$ represents the probability of producing a top 10 finalist (given covariates X) for a single Bernoulli trial. Since we're pretending that $Y_i \sim Bin(10, \theta(X_i))$, we have that $\theta(X)$ corresponds to the probability of producing a top 10 finalist in a single year. It does not represent the probability of producing at least one top 10 finalist over 10 ten years or anything like that.

The coefficient on `LogPopulation` indicates that, if we observe two jurisdictions that have the same covariate values except that they differ by 1 in `LogPopulation`, then the log odds of producing a top 10 finalist for the more populous jurisdiction are 0.6171 higher than for the less populous jurisdiction. Equivalently, the odds are `exp(0.671) = 1.9` higher.

The interpretation is analogous for the other coefficients.

Problem 2

(a)

It appears that the "bernoulli" dataset is just yearly for each state if their contestant was in the top 10. This means that the "binimial" dataset was a summary of the "bernoulli" for each state.

(b)

Using this Bernoulli data, we show the model without `Longitude` below. Because both this model and the second model in part 1a are trying to to model $\theta[X_i]$ - the probability in being in the top 10 contestants related to state information and they both use the same information we expect the model's parameters to be very similar.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.619	2.15	-3.079	0.002074
LogPopulation	0.5888	0.1757	3.35	0.0008067
LogContestants	1.337	0.4104	3.258	0.001123
LogTotalArea	-0.3198	0.1204	-2.656	0.007899
Latitude	-0.0733	0.029	-2.528	0.01148

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	454.3 on 458 degrees of freedom
Residual deviance:	386.5 on 454 degrees of freedom

(c)

We should not expect the AICs of the two models to be the same, nor are they (with the Bernoulli `m2` model having an AIC of 396.46 and the binomial `m2` model having an AIC of 142.79). This directly related to the definition of AIC, which is

$$-2 \cdot \text{Log likelihood} + 2 \cdot p \ ,$$

where p is the number of parameters of the model (5 in our models).

The first model has an assumption of the observations being Binomial$(9, \theta[X_i])$ and the second uses the assumption that the observations are Bernoulli$(\theta[X_i])$. The likelihoods for a state therefore look slightly different, specifically:

$$\binom{9}{x_i} \theta[X_i]^{y_i} (1 - \theta[X_i])^{9-y_i}$$

for the Binomial assumption and

$$\prod_{z=2000}^{2009} \theta[X_i]^{y_{iz}} (1 - \theta[X_i])^{1-y_{iz}}$$
$$\text{or } \theta[X_i]^{y_i} (1 - \theta[X_i])^{9-y_i} \ ,$$

for the Bernoulli assumption[1]. As the log likelihood is over all i, this means that the Log likelihoods for the model from part 1 (the Binomial assumption) has an additional

$$\sum_{i=1}^{51} \log \left(\binom{9}{y_i} \right)$$

term. THe difference between the two AICs should then be by $2 * \sum_{i=1}^{51} \log \left(\binom{9}{y_i} \right)$ (from the definition of the AIC). The final 2 lines of code below show that to be true.

```
binomial_m2_aic - bernoulli_m2_aic
```

```
## [1] -254
```

```
sapply(dat_original$Top10, function(x) choose(9, x)) %>%
  log() %>% sum() %>% "*"(.,-2)
```

```
## [1] -254
```

[1] This is because we can see $y_i = \sum_{z=2000}^{2009} y_{iz}$.