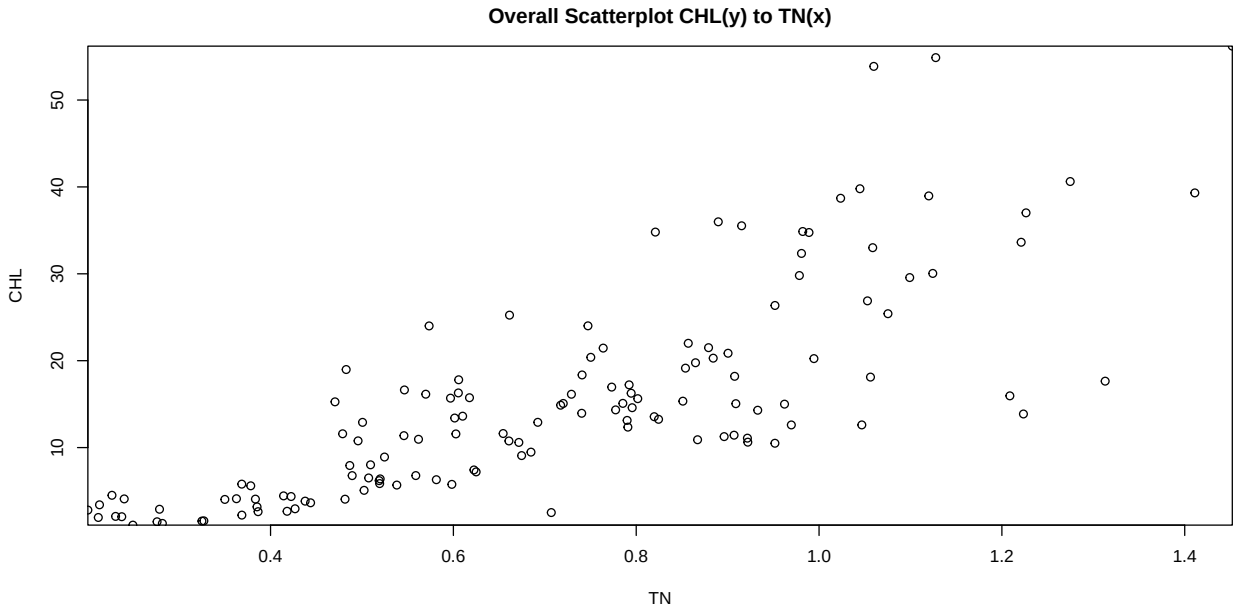
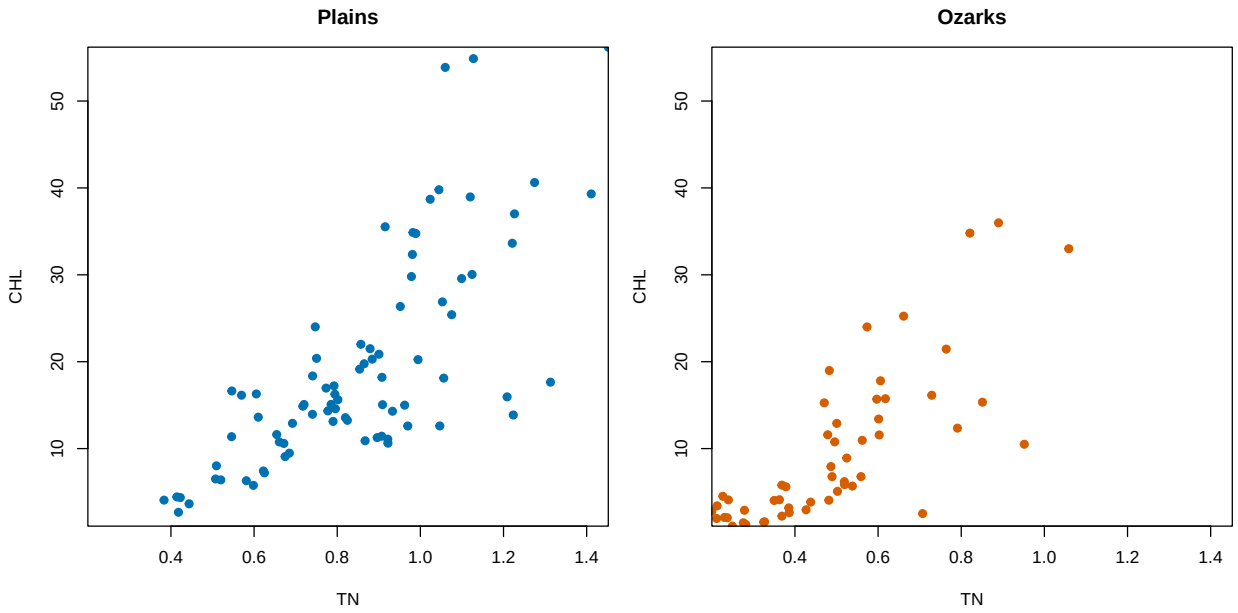
5200 Take-Home

Sam Olson

Q1: CHL & TN (Plains vs. Ozarks)

Overall Distribution & Approach





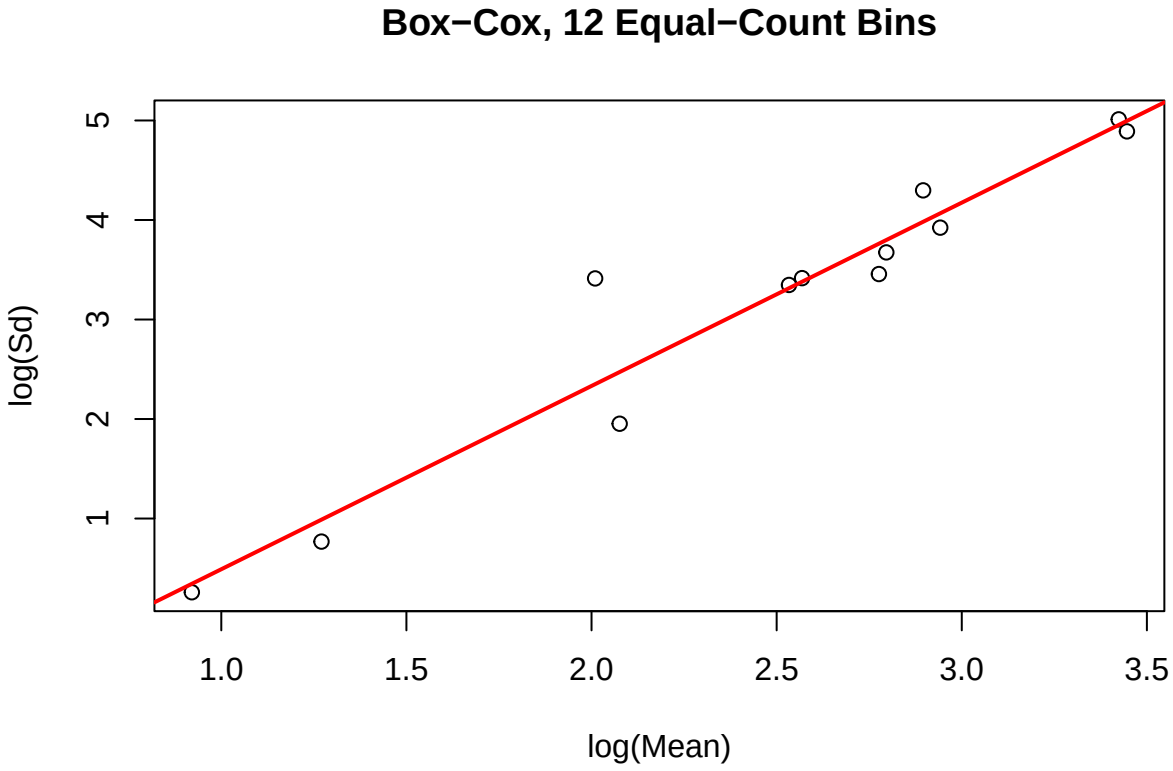
CHL is right-skewed and increases with TN, with variance also rising at higher TN values. The regional scatterplots (Plains vs. Ozarks) show the same general pattern but suggest possible differences in the strength of the TN–CHL relationship.

Let Y_i de a random variable denoting CHL for lake i ($i = 1, \cdots, 134$), and also denote the corresponding covariate TN value x_i. Each lake appears once, judging by the column `lake` in the dataset, so observations may be treated as independent.

Assuming a common underlying biological mechanism across Missouri reservoirs, which seems generally reasonable, we first fit an overall TN–CHL model to determine an appropriate mean–variance structure (in addition to potential link functions). Once a suitable overall model is identified, we then apply this same model structure separately to the Plains and Ozarks so that any differences in fitted curves or parameters may be assessed as indications of regional differences rather than artifacts of using different model families (a risk we would run if we instead assessed model fit by region).

We then proceed with potential candidate models, starting with GLMs.

Generalized Linear Models

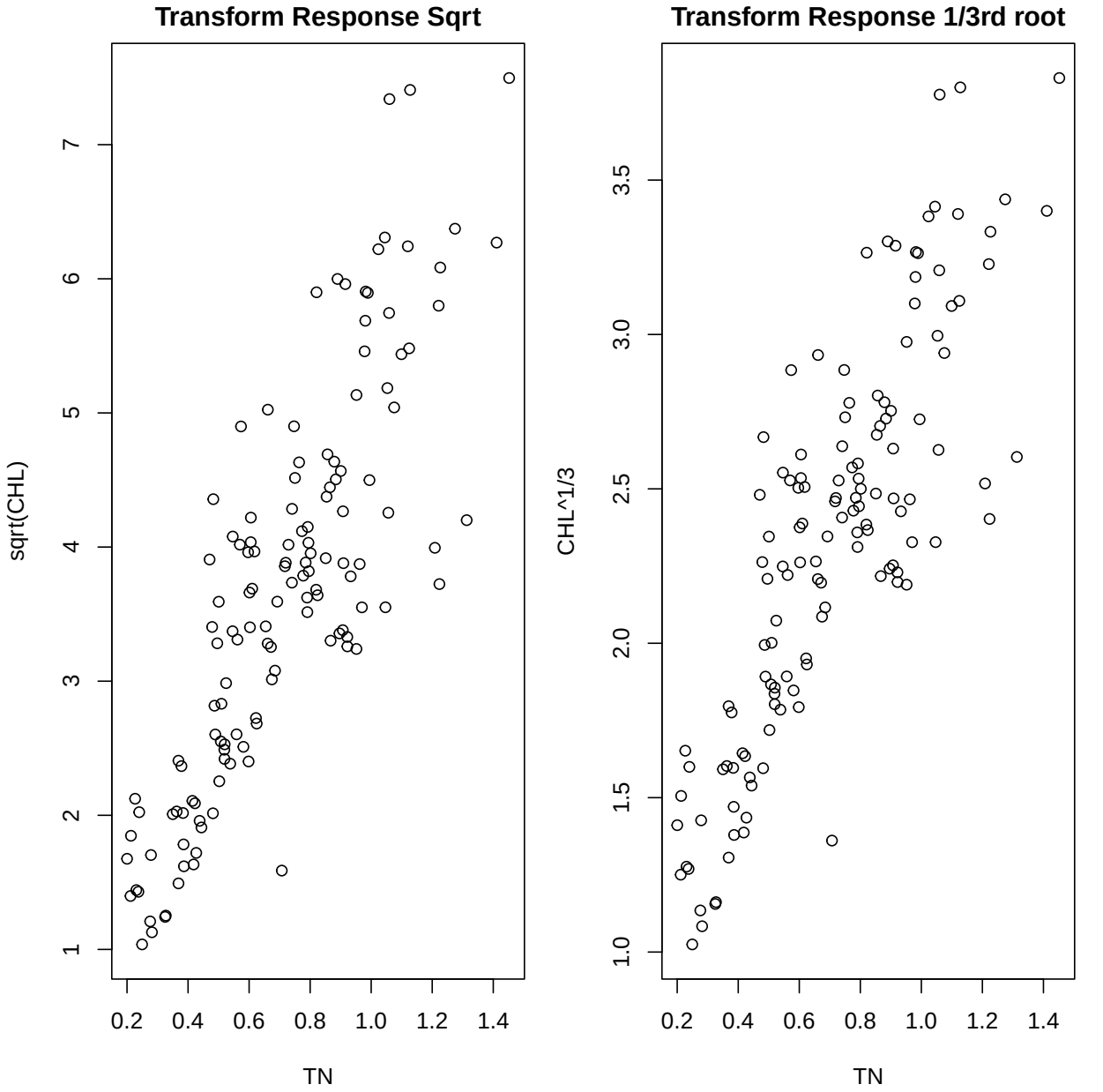


nbins	Equal.Count	Equal.Spaced
6	1.76	1.85
10	2.04	1.95
12	1.84	1.94
14	1.97	2.10
16	1.82	1.86
18	1.77	1.87

nbins	Equal.Count	Equal.Spaced
22	1.88	1.61

To select an appropriate GLM family, I used `boxcoxfctns` to estimate Box–Cox log(mean)–log(variance) relationships across various specifications. Across all binning schemes, the slopes lay between 1.6 and 2.1, indicating that Var(CHL) increases roughly by a factor of μ^2 to μ^3. Such variance patterns are consistent with Gamma or Inverse Gaussian random components, and as such were considered candidate random components.

With the random component identified, the next step is to determine a suitable link function for modeling the TN–CHL mean relationship.

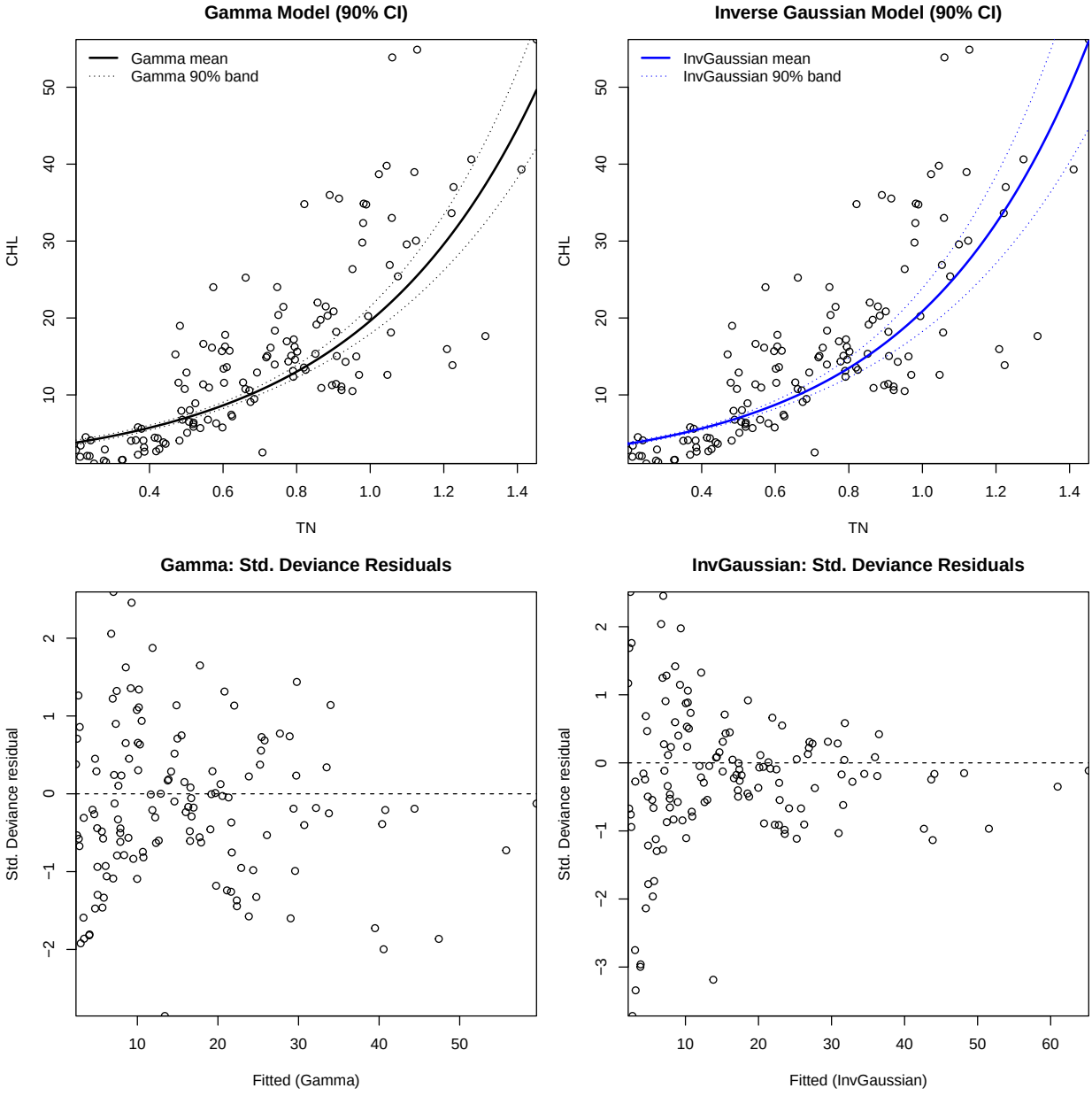


To select a link function, I examined transformations of CHL that approximately linearize its relationship with TN. Exploratory plots showed that the cube-root transformation yielded the most linear trend, with

the square-root transformation performing reasonably well (these two being shown, though multiple link functions were considered, including link functions not available in the `basic.glm` function). Notably, these transformations were used only to guide link choice and under no circumstances used to motivate a transformation of the response in the GLM itself.

Taken together, this motivated fitting Gamma and Inverse Gaussian GLMs paired with power links corresponding to the cube-root and square-root transformations.

Within each GLM family, the scaled deviance provides an appropriate criterion for comparing link functions. The cube-root power link consistently produced lower deviances (raw and scaled) than the square-root link within both the Gamma and Inverse Gaussian families; so the cube-root versions of these two models were treated as viable candidates in toto. However, the Gamma and IG models themselves are not compared by deviance, as they are not nested and thus do not allow LRT-like comparisons.



To compare the Gamma and Inverse Gaussian models, we then turn to a general assessment of deviance spread and fitted curves. Oddly, both yield nearly identical fitted curves, so scatterplots alone cannot distinguish

them. Residual diagnostics provide clearer evidence: The Gamma model exhibits more stable variance and better-behaved deviance residuals, whereas the IG model shows mild heteroscedasticity. Thus, the Gamma GLM is selected as the preferred model, despite evidence that there may be some outliers in both.

With the GLM choice established, additive error models are then considered for completeness before selecting a final model for the region-specific comparisons.

Other Models Considered – Additive Error Models

Transform Both Sides

A simple OLS fit shows a roughly linear TN–CHL trend but clear heteroscedasticity in the studentized residuals, motivating consideration of transform both sides (TBS) additive error models for the purposes of variance stabilization. Several variance-stabilizing transformations were considered, and ultimately the cube-root transformation performed best while maintaining a reasonable linearization of the relationship.

However, TBS models have some drawbacks, the least of which being that inference requires back-transformation, and only the estimates of the mean back-transforms cleanly. As we are interested in other quantities for assessing differences between regions (such as confidence intervals and differences), we start to envision quite a few appreciable drawbacks to a linear fit TBS. Combined with the remaining heteroscedasticity in the studentized residuals, and the TBS approach is not as appealing as the GLMs considered thus far.

Power of the Mean

Given the apparent nonlinearity and mean–variance relationship in the data, power of the mean (POM) additive error models were also considered. Several values of the variance-stabilizing power θ were tested, including models that allowed the mean curve itself to be nonlinear. Although $\theta = 0.5$ performed best in terms of consistency with the Box–Cox plots and visually when plotted against the observed data, its studentized (absolute and squared) residuals still indicated heteroscedasticity, and small fitted means made the weighting scheme unstable for larger θ.

Overall, even the best POM combined with TBS that was considered for modelling the CHL-TN relationship failed to produce generally satisfactory variance stabilization while also failing to predict higher observed CHL values to boot. So neither the POM, TBS, nor their combination was ultimately retained as a candidate model.

Comparing Different Models

Overall, the Gamma GLM with a cube-root link best matched the Box–Cox variance diagnostics, produced well-behaved residuals, and offered the most coherent and interpretable mean–variance structure. For such reasons, the Gamma GLM with cube-root link was used for region-specific comparisons.

Table 2: Region-specific optimized GLM coefficients with 95% Wald CIs

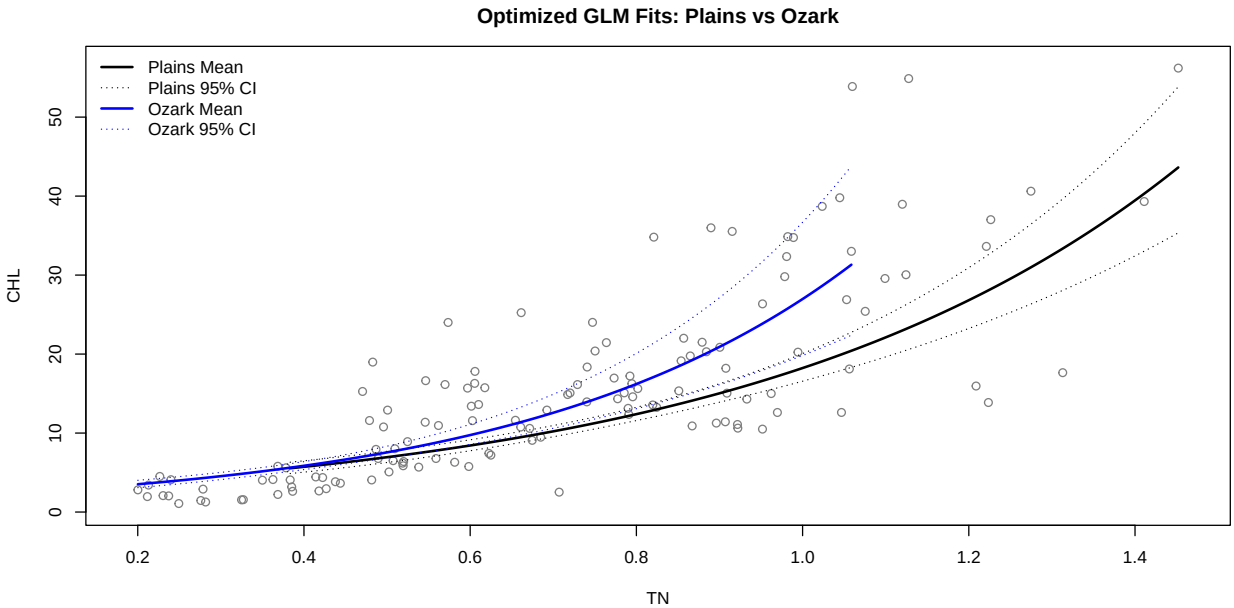
Region	Term	Estimate	SE	LCL	UCL
Plains	Intercept	0.9731	0.1180	0.7417	1.2044
Plains	TN	1.9301	0.1479	1.6402	2.2199
Ozark	Intercept	0.7500	0.1107	0.5331	0.9670
Ozark	TN	2.5447	0.2525	2.0499	3.0395

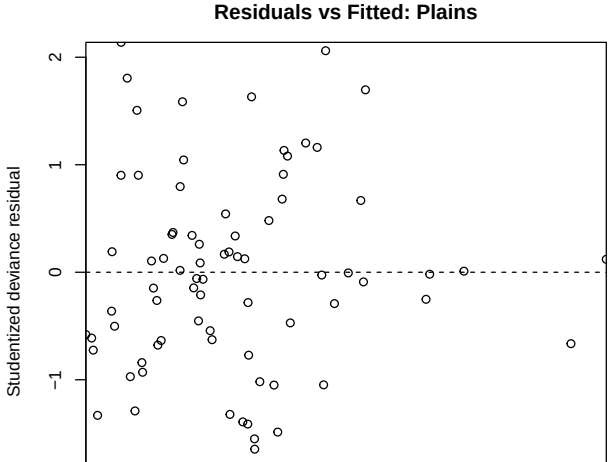
Table 3: Predicted CHL for TN = 0.4, 0.6, 0.8, 1.0 with 95% CIs by region

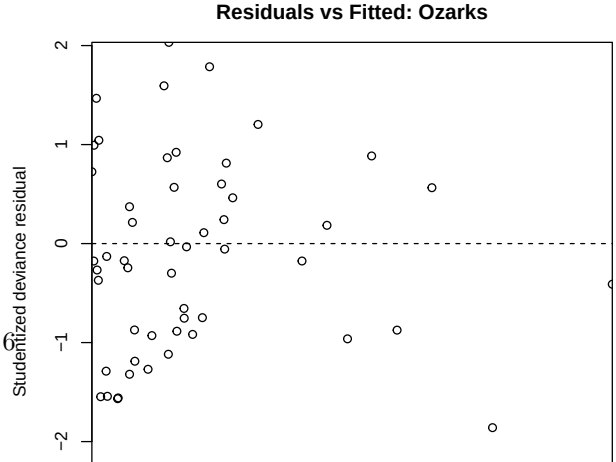
Region	TN	Fit	LCL	UCL
Plains	0.4	5.7264	5.0535	6.4889
Plains	0.6	8.4242	7.7594	9.1459
Plains	0.8	12.3928	11.5794	13.2633
Plains	1.0	18.2311	16.5652	20.0644
Ozark	0.4	5.8584	5.3762	6.3838
Ozark	0.6	9.7456	8.5571	11.0991
Ozark	0.8	16.2119	13.0813	20.0916
Ozark	1.0	26.9687	19.8236	36.6892

Extending to regions

Comparing Two Regions







Region	Scaled_Deviance
Plains	79.59165
Ozarks	54.28234

Using the selected Gamma GLM with a cube-root link, models were separately fit to the Plains and Ozarks regions. The fitted curves and pointwise 95% confidence bands show a fair amount of overlap at lower TN values ($\approx$ 0.4–0.6), with very similar predicted means. At higher TN ($\approx$ 0.8–1.0), the Ozarks curve becomes noticeably steeper, and the bands begin to separate, indicating higher CHL in the Ozarks at the upper end of the TN range.

The coefficient estimates reflect this pattern: The Ozarks model has a larger slope and a slightly smaller intercept, though the 95% Wald intervals overlap for both parameters. Model diagnostics support using the same GLM form in both regions. Studentized deviance plots show no systematic patterns and are generally well-behaved, albeit with some potential outlying points. In part due to not considering likelihoods at this stage of the exam, it bears noting the potential shortfall of Wald intervals, both in terms of sample size per region (53 in Ozarks, 81 in Plains), as well as possible issues with the presumed asymptotic normality of the estimator, even though there are no clear issues present (such as parameter intervals containing values outside the parameter space).

So overall, both regions share the same functional TN–CHL form, but the Ozarks appear to exhibit a stronger response at higher TN levels.

Probability Assessments

For a fixed TN value x_i, let Y_i denote the Plains CHL response and Z_i the Ozarks CHL response.

From the region-specific Gamma GLMs with cube-root link:

$$\mu_P(x_i) = E(Y_i \mid x_i), \qquad \mu_O(x_i) = E(Z_i \mid x_i),$$

with variance functions

$$\mathrm{Var}(Y_i \mid x_i) = \phi_P \mu_P(x_i)^2, \qquad \mathrm{Var}(Z_i \mid x_i) = \phi_O \mu_O(x_i)^2,$$

where ϕ_P and ϕ_O are the dispersion parameters for the Plains and Ozarks models.

The target quantity is:

$$\Pr(Y_i > Z_i \mid x_i)$$

Note: Although the data model is Gamma, the quantity $\Pr(Y_i > Z_i)$ depends on the distribution of a difference of two independent Gamma variables, which has no closed form; thus the plug-in Normal approximation below is used for tractability.

Using a plug-in Normal approximation to the Gamma distribution,

$$Y_i \mid x_i \approx N\big(\mu_P(x_i), \phi_P \mu_P(x_i)^2\big), \qquad Z_i \mid x_i \approx N\big(\mu_O(x_i), \phi_O \mu_O(x_i)^2\big),$$

and assuming conditional independence of Y_i and Z_i given x_i, a reasonable assumption given the setup, define:

$$D_i = Y_i - Z_i$$

Then

$$E(D_i \mid x_i) = \mu_P(x_i) - \mu_O(x_i),$$

and

$$\mathrm{Var}(D_i \mid x_i) = \mathrm{Var}(Y_i \mid x_i) + \mathrm{Var}(Z_i \mid x_i) = \phi_P \mu_P(x_i)^2 + \phi_O \mu_O(x_i)^2$$

Such that

$$D_i \mid x_i \approx N\Big(\mu_P(x_i) - \mu_O(x_i),\ \phi_P \mu_P(x_i)^2 + \phi_O \mu_O(x_i)^2\Big),$$

so that

$$\Pr(Y_i > Z_i \mid x_i) = \Pr(D_i > 0 \mid x_i) \approx \Phi\left(\frac{\mu_P(x_i) - \mu_O(x_i)}{\sqrt{\phi_P \mu_P(x_i)^2 + \phi_O \mu_O(x_i)^2}}\right),$$

where Φ is the standard Normal CDF.

Under the cube-root power link used in `basic.glm` (with `pwr = 1/3`),

$$\eta(x) = \mu(x)^{1/3}, \qquad \eta(x) = x^\top \hat{\beta},$$

so the fitted mean at covariate value x is

$$\hat{\mu}(x) = \left(x^\top \hat{\beta}\right)^{1/(1/3)} = \left(x^\top \hat{\beta}\right)^3$$

For each region and each TN value x_i,

$$\hat{\mu}_P(x_i) = \left((1, x_i)^\top \hat{\beta}_P\right)^3, \qquad \hat{\mu}_O(x_i) = \left((1, x_i)^\top \hat{\beta}_O\right)^3,$$

with corresponding dispersion estimates $\hat{\phi}_P$ and $\hat{\phi}_O$. The plug-in estimator of the desired probability is then

$$\widehat{\Pr}(Y_i > Z_i \mid x_i) = \Phi\left(\frac{\hat{\mu}_P(x_i) - \hat{\mu}_O(x_i)}{\sqrt{\hat{\phi}_P \hat{\mu}_P(x_i)^2 + \hat{\phi}_O \hat{\mu}_O(x_i)^2}}\right)$$

Note: The calculations that follow rely on the asymptotic Normal approximation to the Gamma GLM means and variances. Whether this approximation is appropriate for the sample sizes in each region should be considered.

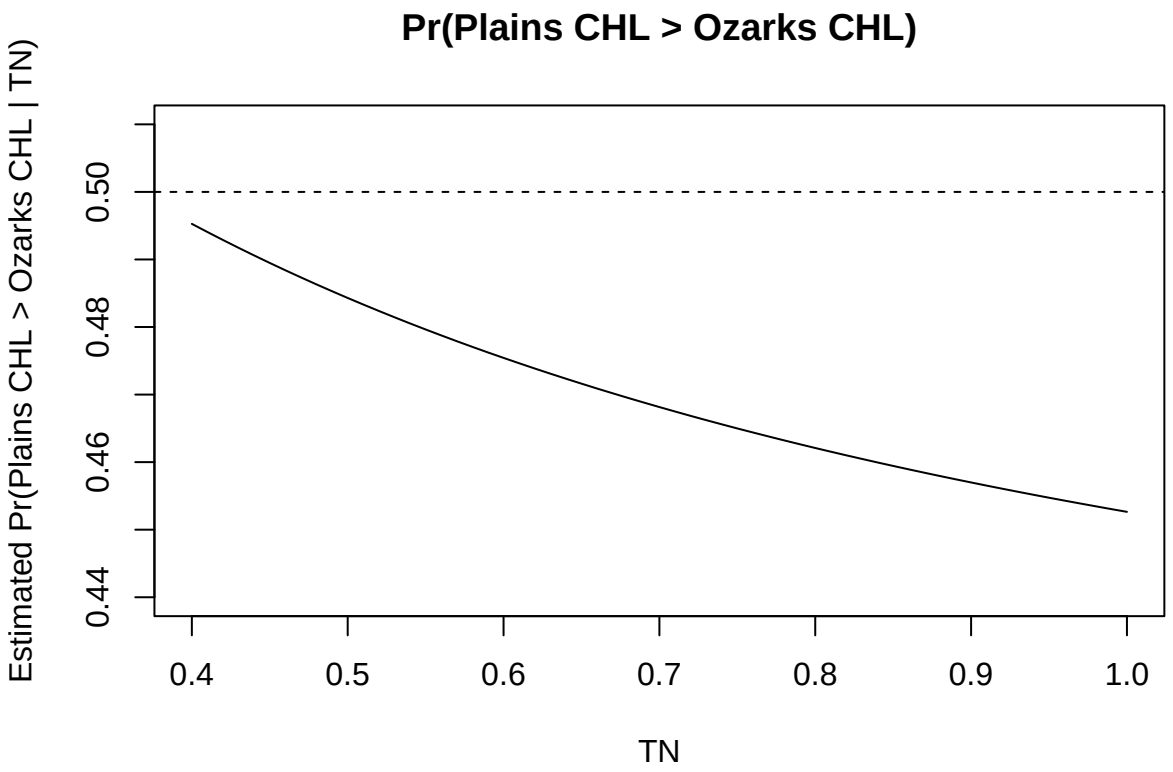
```
# Models fitted already
# mod_gamma_13_plains
# mod_gamma_13_ozark

# extract mean mu from basic.glm
predict_mu_power <- function(mod, x, pwr = 1/3) {
  # regression coefficients
  beta_hat <- mod$estb[, 1]
  # design matrix
  X <- cbind(1, x)
```

```r
  eta <- as.vector(X %*% beta_hat)
  # eta = mu^pwr
  mu <- eta^(1 / pwr)
  mu
}

# plug-in estimate
prob_YgtZ_normal <- function(x, mod_P, mod_O, pwr = 1/3) {
  # plains mean and dispersion
  mu_P <- predict_mu_power(mod_P, x, pwr = pwr)
  phi_P <- mod_P$ests$phi
  # ozarks mean and dispersion
  mu_O <- predict_mu_power(mod_O, x, pwr = pwr)
  phi_O <- mod_O$ests$phi
  # difference calculations
  mean_D <- mu_P - mu_O
  var_D <- phi_P * mu_P^2 + phi_O * mu_O^2
  sd_D <- sqrt(var_D)
  # probability
  pnorm(mean_D / sd_D)
}

# estimate x = 0.70
prob_0.70 <- prob_YgtZ_normal(0.70,
                              mod_gamma_13_plains,
                              mod_gamma_13_ozark)
cat("Estimated probability is:", round(prob_0.70, 3), "\n")
```

```
## Estimated probability is: 0.468
```

Our estimate is: $\Pr(Y_i > Z_i \mid x_i = 0.70) = 0.46815$.

Now, the sequence of values $x_i \in \{0.4, 0.41, 0.42, \cdots, 1.0\}$ is given by:



So, as TN concentration increases, the probability that Plains CHL exceeds Ozarks CHL decreases (from roughly 0.5 to 0.45).

Relation Between CHL and TN within Regions

In both regions, CHL increases with TN and the variance rises with the mean. The Gamma GLM with a cube-root link adequately captures this mean–variance relationship, with well-behaved residuals and no indication that different model forms are needed. Thus, the TN–CHL relationship has the same functional form in the Plains and Ozarks.

Fitting the model separately by region shows differences in magnitude. For TN $\approx$ 0.4–0.6, fitted curves and 95% bands nearly coincide, and predicted CHL is similar across regions. As TN increases toward 0.8–1.0, the Ozarks curve steepens and predicted CHL values in the Ozarks diverges upward; this also coincides with the confidence bands separating, and generally indicates a stronger TN response in the Ozarks at higher TN.

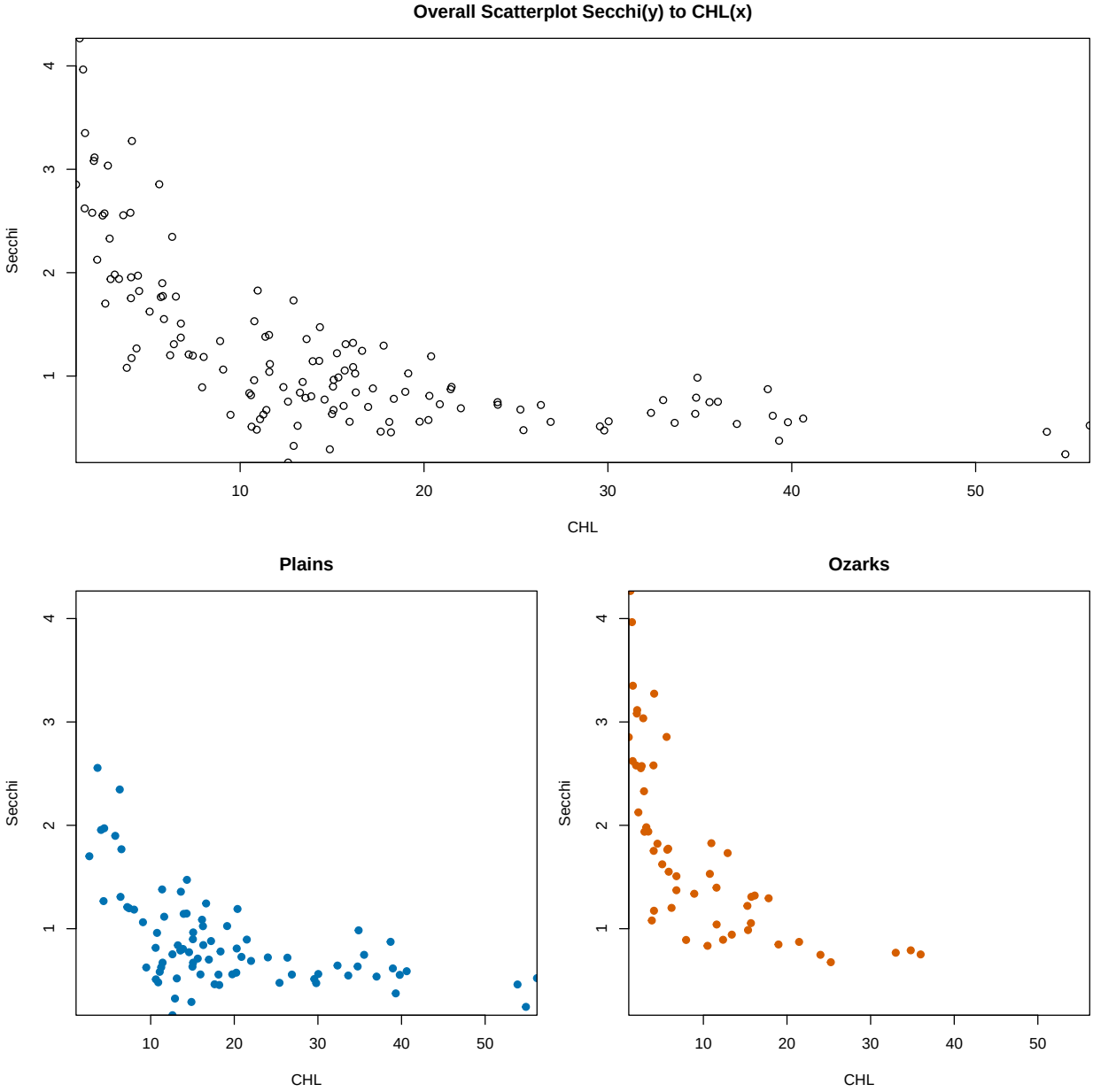
The coefficient estimates are consistent with this picture: both slopes are positive, with the Ozarks slope being larger. Although the 95% Wald intervals overlap and do not provide definitive parameter-wise separation, the combined evidence from curves, bands, and slopes suggests a stronger TN–CHL response in the Ozarks.

The probability assessment summarizes this difference: For moderate TN (0.4–0.7), Pr(Plains CHL $>$ Ozarks CHL | TN) $\approx$ 0.5, but as TN approaches 1.0 this probability falls below 0.5, indicating that Ozarks CHL is more likely to exceed Plains CHL at higher TN.

Overall, the regions share the same basic TN–CHL form, but the Ozarks show a stronger response to increasing TN, especially near the upper end of the observed TN range.

Q2: SECCHI & CHL (Plains vs. Ozarks)

Overall Distribution & Approach



We start by examining the scatterplot of Secchi depth versus CHL and also examine the scatterplot by region. Secchi depth is right-skewed, consistent with ecological expectations, and shows a clear negative, potentially nonlinear relationship with CHL, with variance possibly decreasing at larger CHL values. By region, the same basic pattern is also shown, while still suggesting possible differences in the strength of the CHL-Secchi relationship

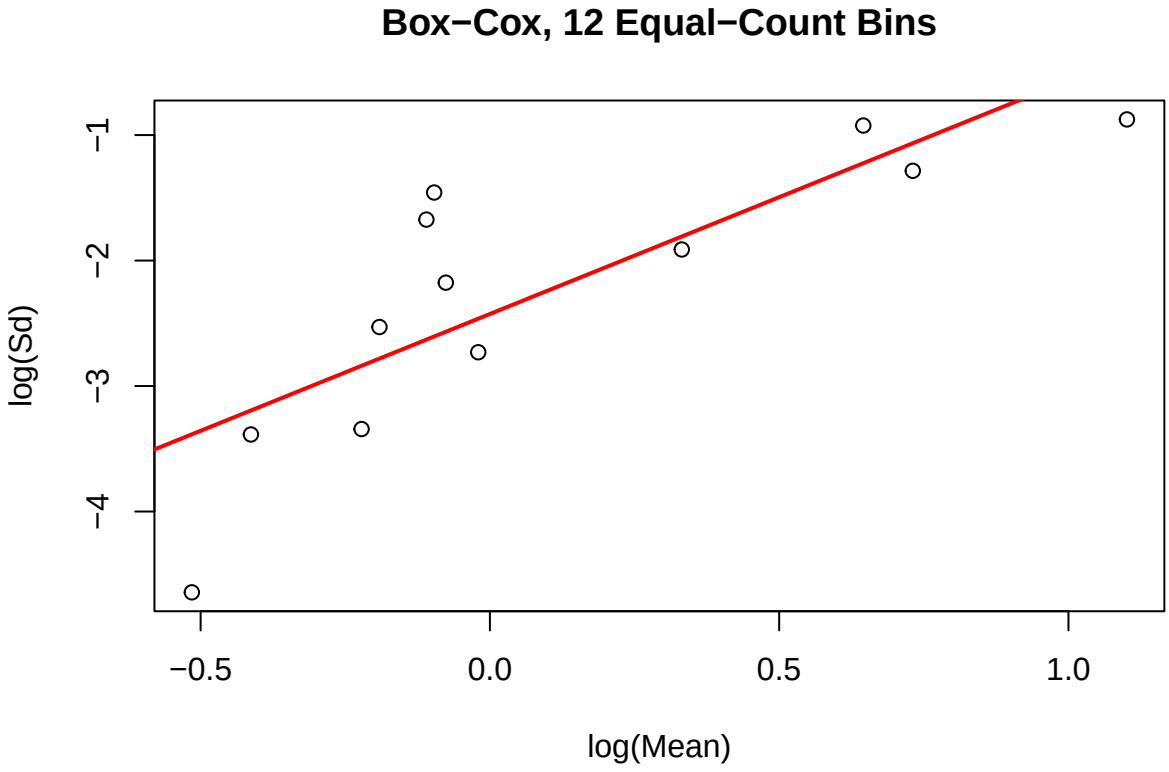
Let the random variable Y_i denote the Secchi depth for lake i ($i = 1, \cdots, 134$), with corresponding CHL covariate value x_i. Because each lake appears only once, it is reasonable to treat observations as independent.

Generally, the modeling strategy mirrors Part I: We first identify an appropriate overall model for Secchi as a function of CHL, then fit the same model structure separately to the Plains and Ozarks. Under this

approach then, differences in intercept, slope, curvature, or dispersion may be interpreted as potential regional differences in the CHL–Secchi relationship rather than artifacts of using different model families or transformations.

Henceforth, for brevity, I omit explicit "as in Part I" references, though the underlying modeling logic is the same.

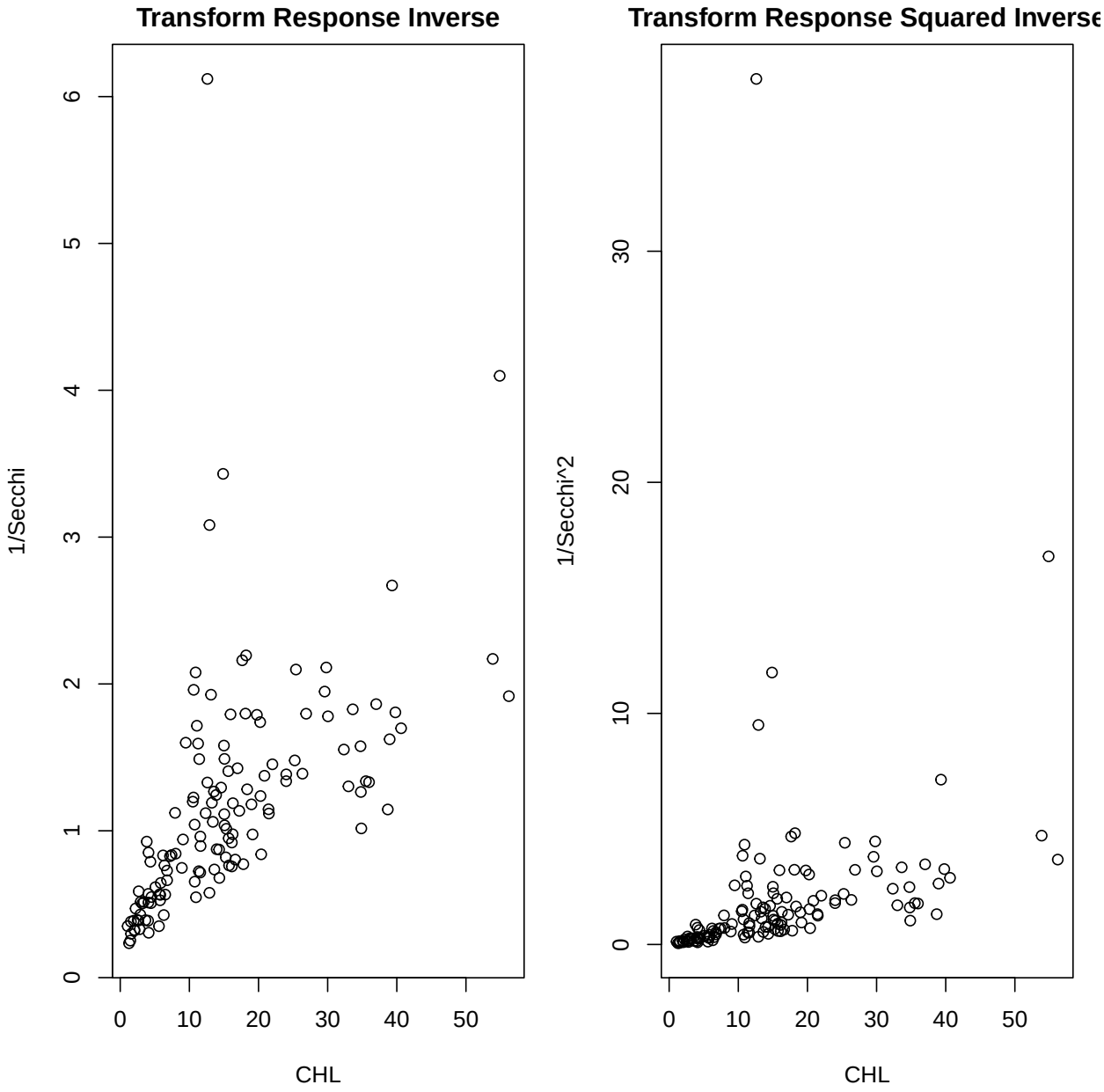
Generalized Linear Model



nbins	Equal.Count	Equal.Spaced
6	2.13	2.13
10	1.86	2.71
12	1.86	2.12
14	1.69	2.29
16	1.75	2.69
18	1.73	2.37
22	1.77	2.42

We first consider GLMs, and begin by identifying a suitable random component. Using `boxcoxfctns`, Box–Cox plots and slopes were calculated; across both equal-count and equal-width binning schemes, and across variable number of bins, the slopes ranged from 1.6 to 2.1. This indicates that Var($Secchi$) grows roughly like μ^2–μ^3, which is consistent with Gamma or Inverse Gaussian random components.

With the random component thus narrowed to Gamma or IG, the next step is to select an appropriate link function for modeling the CHL–Secchi mean relationship.



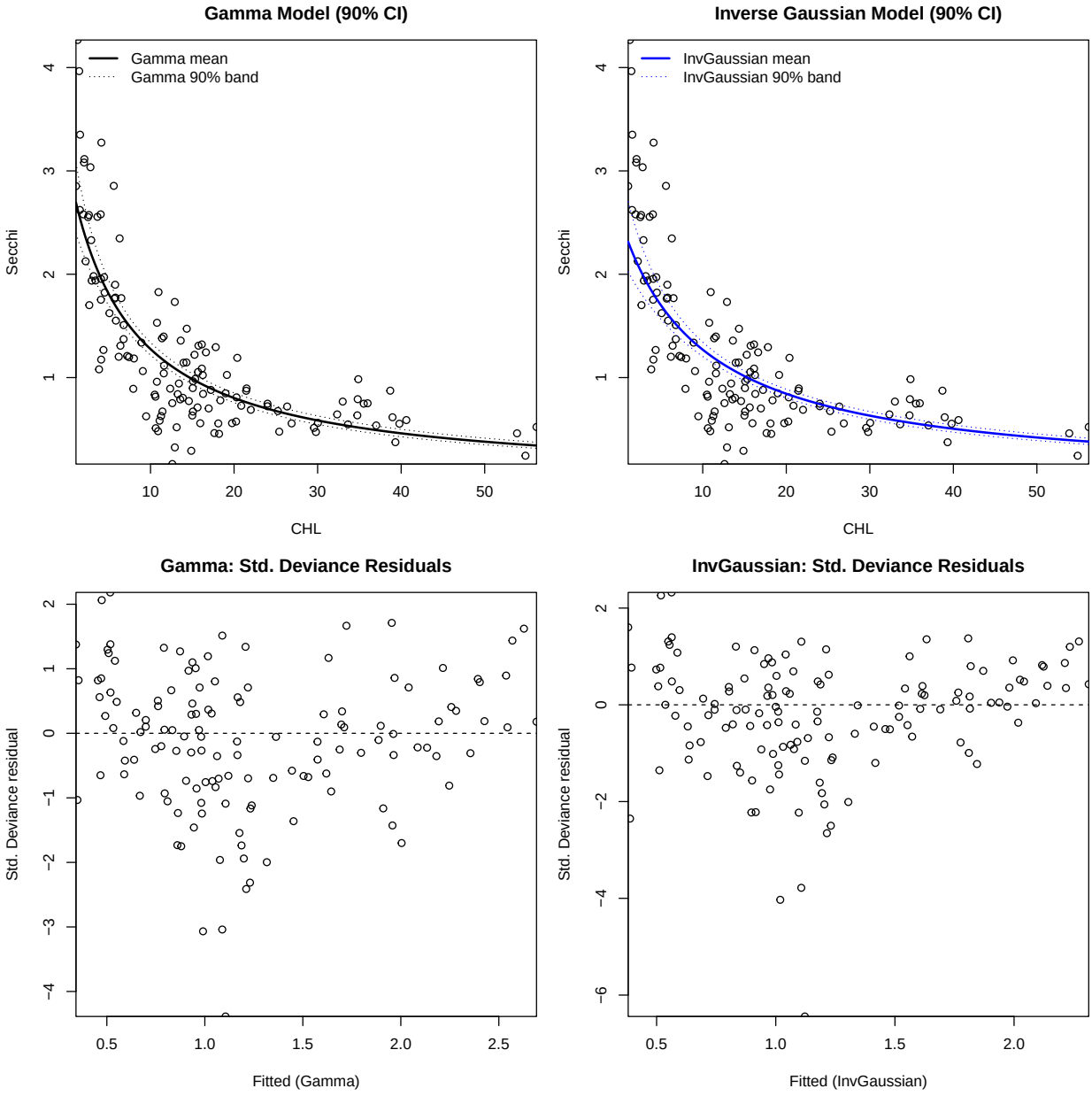
To identify a suitable link function, we seek a transformation of the mean that both linearizes the Secchi–CHL relationship and preserves the expected monotonically decreasing pattern. Exploratory transformation plots indicate that an inverse transformation of the mean yields an approximately linear trend, with the square-inverse transformation performing similarly well. Importantly, these transformations are used only to guide link selection, and under no circumstance should they been seen as justification for transforming the random variable for Secchi directly.

Combined with the earlier mean–variance assessment, this motivates fitting Gamma and Inverse Gaussian GLMs ($\theta = 2, 3$) paired with inverse and square-inverse links. These candidate models are then compared.

Several GLMs were considered, including the Inverse Gaussian model with its canonical inverse-squared link. Even after generating reasonable starting values from a preliminary `glm()` fit and restricting initial linear predictors to be positive, the `basic.glm` routine failed to converge because iterations quickly produced invalid η values. Due to this numerical instability, only a subset of models could be reliably fitted.

The Gamma GLM with inverse link and the Inverse Gaussian GLM with inverse link both converged and

exhibited stable behavior. These models were therefore carried forward for comparison and then evaluated them using fitted-curve overlays on the scatterplot of the observed in addition to deviance residual analysis.



The fitted curves from the Gamma and Inverse Gaussian GLMs are visually similar, and the scatterplot alone does not clearly favor one model. Although the Inverse Gaussian model appears to reflect the increased variability at low CHL values (CHL < 10), its fitted mean curve misses most observations in this range, indicating a poorer fit to the central trend.

Deviance residual diagnostics provide more discerning evidence to compare the two models. The Inverse Gaussian deviance residuals show more pronounced heteroscedasticity and noticeable asymmetry compared to the Gamma model. In contrast, the Gamma GLM with an inverse link produces fairly symmetric and homoscedastic deviance residuals. These patterns persist when using studentized residuals, which ultimately drove the model choice, though there are some potential outliers in both model fits.

On this basis, the Gamma GLM with inverse link is selected as the preferred model. With this GLM established, several additive error models were then assessed for completeness before identifying a single

working model for the region-specific comparisons.

Other Models Considered – Additive Error Models

Transform Both Sides

We begin by fitting a simple linear regression of Secchi on CHL. Although the linear trend is directionally fitting, the studentized residuals show heteroscedasticity, motivating consideration (but not necessarily adoption) of a TBS additive error model.

Several variance–stabilizing transformations were evaluated, with the inverse and cube-root transformations appearing most promising based on exploratory plots. The purpose of these transformations is not to model Secchi on a transformed scale per se, but to obtain an additive-error model with approximately constant variance.

The resulting TBS fits, however, indicate that this approach is not suitable for the Secchi–CHL relationship. Even under the best-performing transformations, the fitted curves fail to capture the pattern in the data, and the studentized residuals retain noticeable patterns and heteroscedasticity. Also, the cube-root transformation performs better than the inverse transformation while avoiding instability near very small Secchi values, though substantial funnel patterns remain.

Overall, the TBS additive error models provide a poorer fit than the GLMs and do not adequately account for the nonlinear mean–variance structure. For this reason, they are not competitive as working models, though they offer some guidance for selecting transformations within POM models considered next.

Power of the Mean

We next consider POM additive error models as an alternative to the GLM. The Box–Cox mean–standard-deviation patterns for Secchi versus CHL suggested

$$\mathrm{Var}(Y \mid x) \propto \mu(x)^{\theta}, \qquad \theta \approx 1.5\text{–}3,$$

which motivates variance weights of the form

$$w_i \propto \mu(x_i)^{-\theta},$$

So we examined POM models with

$$\theta = 1.5, 2, 2.5, 3$$

Across these specifications, the fitted curves and diagnostics revealed serious issues. For many choices of θ, the fitted Secchi–CHL curve becomes nearly flat—or even slightly increasing—across much of the CHL range, with an abrupt "upturn" for CHL > 40. This behavior contradicts the strongly decreasing relationship evident in the raw data and in the GLM fits.

Residual diagnostics reinforce this conclusion: Even after weighting by $\mu(x)^{-\theta}$, substantial heteroscedasticity remains. Large residuals (absolute and squared) persist at both low and moderate CHL, and the spread does not stabilize across fitted means, indicating that the assumed POM variance form does not capture the true mean–variance relationship.

Notably, the POM models above were all formulated directly with CHL as the predictor. Because Secchi depth reflects light penetration through a water column, it may be more fitting to express the model in terms of a simple geometric argument linking depth to the amount of material per unit volume. This motivates the "cylinder" formulation considered next.

A Possibly Better Approach? Disk & Volume

The motivation for an additive model based on 1/CHL follows from a geometric argument. If the visible water column above the Secchi disk is approximated as a cylinder of radius r and height h (the Secchi depth), then its volume is given by:

$$V = \pi r^2 h, \qquad h = \frac{V}{\pi r^2}$$

As CHL is measured as mass per unit volume (1 liter is equivalent to $1,000$ cm^3), if visibility is "lost" when a roughly fixed mass of CHL is present above the disk, then:

$$V \propto \frac{1}{\text{CHL}}$$

and

$$h \propto \frac{1}{\text{CHL}}$$

This suggests a reciprocal relationship between Secchi depth and CHL, motivating an additive model in the transformed predictor $\phi = 1/\text{CHL}$:

$$Y_i = \beta_0 + \beta_1 \phi_i + \varepsilon_i, \qquad E[\varepsilon_i \mid \phi_i] = 0$$

Expressed in the original CHL scale, the regression function is then:

$$m(x) = E[Y \mid x] = \beta_0 + \frac{\beta_1}{x},$$

After initial fit, the residual diagnostics (absolute and squared studentized residuals) for this model still show heteroscedasticity.

As this model seems fairly promising, further adjustments were considered. Following the cube-root approach used in the fish-weight (walleye) example, consider specifying a TBS model where:

$$Z = Y^{1/3}, \qquad \psi = \phi^{1/3} = x^{-1/3},$$

and fit

$$Z = \alpha_0 + \alpha_1 \psi + \eta, \qquad E[\eta \mid \psi] = 0$$

Back-transforming yields the approximate mean function

$$m(x) \approx \left(\alpha_0 + \alpha_1 x^{-1/3}\right)^3$$

which happens to be a smooth, strictly decreasing curve analogous to the cube-root linearization in the fish-weight model.

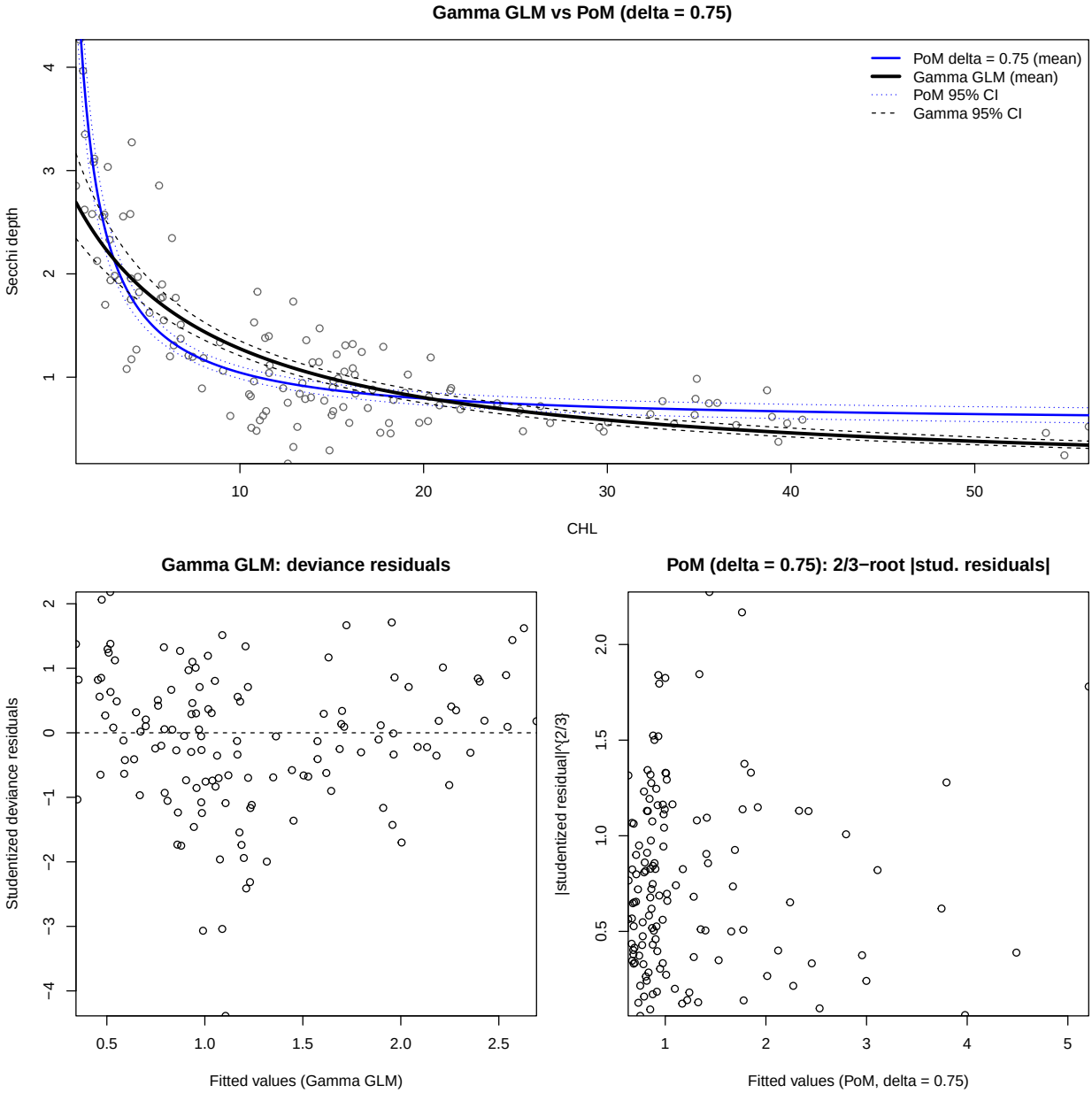
Although both the reciprocal additive model "as-is" and its extension via the cube-root TBS capture the general decreasing trend, neither provides satisfactory variance stabilization. Further, the POM extensions with $\delta = 0.5$ and $\delta = 0.75$ perform the best among the additive models considered, producing smoother curves while doing a better job at stabilizing the variance. However, even these weighted fits retain visible structure

in the residuals (absolute, squared, and now absolute $\frac{2}{3}$-root studentized) and fall short of the diagnostic quality achieved by the Gamma GLM with inverse link.

Between the two POM specifications, $\delta = 0.75$ provided slightly better performance at high CHL values: The fitted curve tracked the observed decline in Secchi depth more closely in the upper tail than the corresponding $\delta = 0.5$ model.

Note: The "best" additive model is not presented in full here, but its fitted curve and representative residual diagnostics appear in the figures that follow.

Comparing Models



The two models that performed adequately were the Gamma GLM with an inverse link and the POM additive model with $\delta = 0.75$ combined with a cube-root TBS transformation. Although both produce broadly similar decreasing curves, the Gamma GLM remains the preferred model.

First, the Gamma GLM better models the observed variance structure. The POM–TBS model, by contrast, imposes variance stabilization through estimated weights, making it more sensitive in the fitted means and more prone to irregularities at the extremes of CHL. Second, the fitted curve and confidence intervals from the Gamma GLM are smoother and more stable across the entire CHL range. The POM–TBS curve tends to flatten at higher CHL values, and its confidence bands widen noticeably at both ends due to back-transformation and weight variability.

A more solid nail in the coffin though is comparing respective residuals. The residual diagnostics generally favor the Gamma GLM: The deviance and studentized deviance residuals are more approximately symmetric and exhibit no major patterns, while the POM–TBS model retains curvature and uneven spread, indicating that its variance adjustments remain incomplete.

Taken together, these considerations make the Gamma GLM with inverse link the preferred modeling choice for the region-specific comparisons that follow.

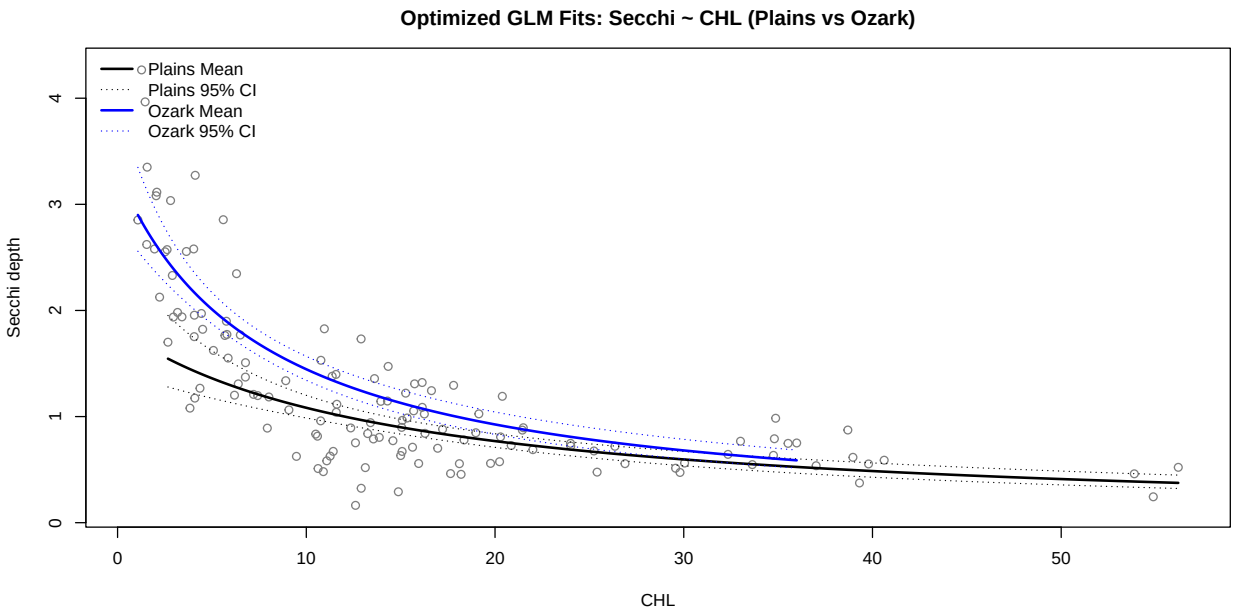
Applying to Regions

Table 6: Region-specific optimized GLM coefficients for Secchi CHL with 95% Wald CIs

Region	Term	Estimate	SE	LCL	UCL
Plains	Intercept	0.5464	0.0799	0.3898	0.7030
Plains	CHL	0.0377	0.0051	0.0277	0.0477
Ozark	Intercept	0.3030	0.0263	0.2514	0.3547
Ozark	CHL	0.0389	0.0039	0.0312	0.0465

Table 7: Predicted Secchi for CHL = 1, 5, 10, 20 with 95% CIs by region

Region	CHL	Fit	LCL	UCL
Plains	1	1.7119	1.3653	2.2944
Plains	5	1.3606	1.1728	1.6200
Plains	10	1.0829	0.9865	1.2001
Plains	20	0.7689	0.7101	0.8384
Ozark	1	2.9250	2.5747	3.3856
Ozark	5	2.0108	1.8740	2.1692
Ozark	10	1.4460	1.3418	1.5677
Ozark	20	0.9258	0.8322	1.0431



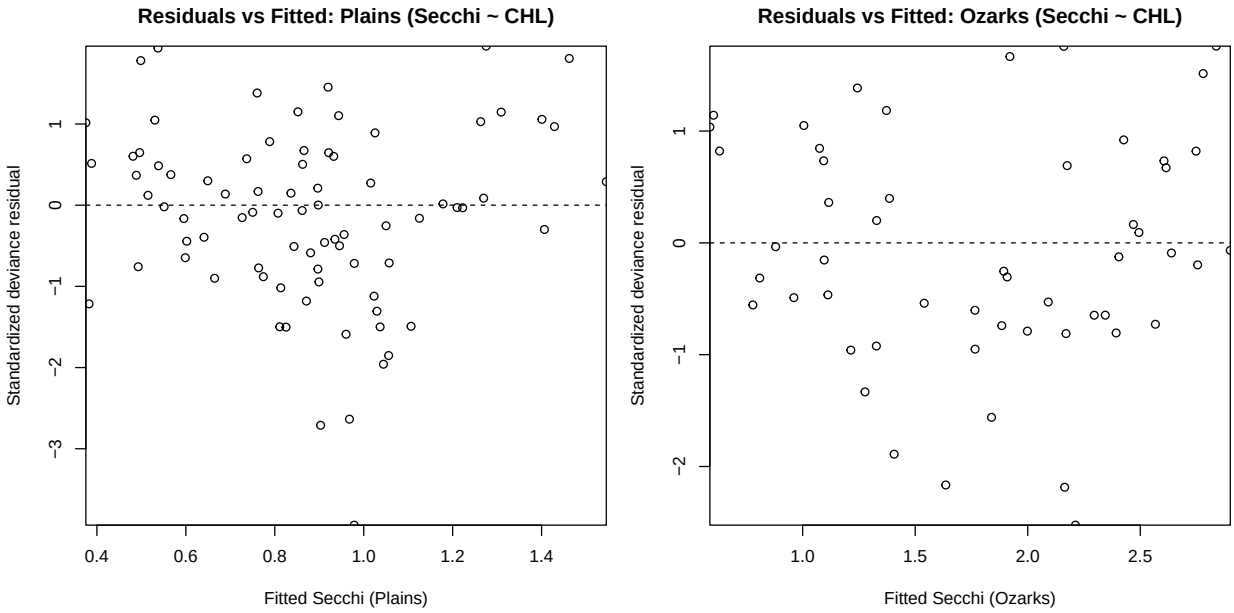


Table 8: Scaled deviance by region for Secchi ~ CHL Gamma GLMs

Region	Scaled_Deviance
Plains	93.27254
Ozarks	53.66591

Interpreting the Model & Results

Across both regions, Secchi depth shows the expected negative and nonlinear association with chlorophyll: Water clarity declines as CHL increases, and the Gamma GLM with an inverse link plausibly models the observed data. When the model is fit separately to the Plains and Ozarks using the same structure, the shapes of the fitted Secchi–CHL curves are nearly identical, indicating that the underlying ecological mechanism

appears consistent across Missouri lakes in both regions.

The magnitude of the response differs, however. For any fixed CHL value, Secchi depth is higher in the Ozarks than in the Plains. This vertical offset is evident even at low CHL (Ozarks roughly 2.9–3.2m vs. 1.7–2.5m in the Plains) and becomes more pronounced as CHL increases. By CHL values around 20–30, the two fitted curves are visibly distinct, with minimal overlap in their confidence bands.

The coefficient estimates reinforce this interpretation. The regions have virtually identical slopes on the inverse–link scale, implying similar rates at which Secchi declines with CHL, though the intercepts differ. The Ozarks exhibit a larger fitted intercept, indicating that, for a given chlorophyll concentration, Ozark lakes tend to have higher water clarity than Plains lakes.

Residual diagnostics further support these conclusions. Both regional fits display well-behaved studentized deviance residuals under the Gamma GLM, with no evidence of region-specific misspecification, differing functional forms, or dispersion anomalies. Thus, the observed separation reflects a genuine regional difference rather than a modeling artifact.

Overall, there is evidence to support that the Plains and Ozarks share the same functional Secchi–CHL relationship but differ in overall water clarity: Ozark lakes exhibit consistently higher Secchi depth than Plains lakes at matched CHL values.

Q3: LRT SECCHI & CHL (Plains vs. Ozarks)

Let Y_{ij} denote Secchi depth for observation i in region $j \in \{P, O\}$ (Plains, Ozarks), with covariate $x_{ij} = \text{CHL}$. Assume a Gamma model with common shape $\alpha > 0$ and mean $\mu_{ij} > 0$:

$$Y_{ij} \mid x_{ij} \sim \text{Gamma}(\alpha, \mu_{ij})$$

The density (shape–mean form) is

$$f(y_{ij} \mid \mu_{ij}, \alpha) = \frac{1}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_{ij}}\right)^{\alpha} y_{ij}^{\alpha-1} \exp\left(-\frac{\alpha y_{ij}}{\mu_{ij}}\right), \qquad y_{ij} > 0$$

The GLM chosen uses the inverse link

$$g(\mu) = \frac{1}{\mu}, \qquad \eta_{ij} = \frac{1}{\mu_{ij}}$$

The Reduced model is given by:

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij}, \qquad \mu_{ij} = \frac{1}{\beta_0 + \beta_1 x_{ij}}$$

The parameter vector is

$$\theta_0 = (\beta_0, \beta_1, \alpha)$$

With reduced log–likelihood

$$\ell_0(\beta_0, \beta_1, \alpha) = \sum_{i,j} \Big[(\alpha - 1)\log y_{ij} - \alpha y_{ij}(\beta_0 + \beta_1 x_{ij}) + \alpha \log(\beta_0 + \beta_1 x_{ij}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big]$$

The Full model is given by:

$$\eta_{iP} = \beta_{0P} + \beta_{1P} x_{iP}, \qquad \eta_{iO} = \beta_{0O} + \beta_{1O} x_{iO}, \qquad \text{(Regretting the subscripts right about now.)}$$

$$\mu_{iP} = \frac{1}{\beta_{0P} + \beta_{1P} x_{iP}}, \qquad \mu_{iO} = \frac{1}{\beta_{0O} + \beta_{1O} x_{iO}}$$

The parameter vector is

$$\theta_1 = (\beta_{0P}, \beta_{1P}, \beta_{0O}, \beta_{1O}, \alpha)$$

The full log–likelihood is

$$\begin{aligned}\ell_1(\theta_1) = &\sum_{i=1}^{n_P} \Big[(\alpha - 1)\log y_{iP} - \alpha y_{iP}(\beta_{0P} + \beta_{1P} x_{iP}) + \alpha \log(\beta_{0P} + \beta_{1P} x_{iP}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big] \\ &+ \sum_{i=1}^{n_O} \Big[(\alpha - 1)\log y_{iO} - \alpha y_{iO}(\beta_{0O} + \beta_{1O} x_{iO}) + \alpha \log(\beta_{0O} + \beta_{1O} x_{iO}) + \alpha \log \alpha - \log \Gamma(\alpha)\Big]\end{aligned}$$

The LRT compares (under regularity conditions):

Table 9: Model Fit, Log-Likelihoods, and LRT Results, Convergence using 'optim'

Quantity	Value
Convergence (Reduced)	0
Convergence (Full)	0
LogLik (Reduced MLE)	-58.30
LogLik (Full MLE)	-58.30
Lambda (raw)	-1e-05
LRT p-value	1.000

- H_0: reduced model is true (no region difference)
- H_1: full model is true (regions differ)

Let $\hat{\theta}_0$ be the MLE under H_0 and $\hat{\theta}_1$ the MLE under H_1. The likelihood ratio statistic is of the form:

$$-2\big[\ell_0(\hat{\theta}_0) - \ell_1(\hat{\theta}_1)\big]$$

Under H_0,

$$-2\big[\ell_0(\hat{\theta}_0) - \ell_1(\hat{\theta}_1)\big] \overset{a}{\sim} \chi^2_2$$

With degrees of freedom of the χ^2 given by the difference in parameters between full and reduced $(5 - 3 = 2)$. We reject H_0 at the α level when

$$-2\big[\ell_0(\hat{\theta}_0) - \ell_1(\hat{\theta}_1)\big] > \chi^2_{2,\,1-\alpha}$$

In short,

- Definitive Question: Do the regions differ in the relation between Chlorophyll and Secchi depth?
- Definitive Answer: No!