LCWMD Chloride Data Analysis

Curtis C. Bohlen

3/2/2021

# Load Libraries

(code omitted)

#> Loading required package: MASS  
#> Loading required package: survival  
#> -- Attaching packages --------------------------------------- tidyverse 1.3.0 --  
#> v ggplot2 3.3.3 v purrr 0.3.4  
#> v tibble 3.0.5 v dplyr 1.0.3  
#> v tidyr 1.1.2 v stringr 1.4.0  
#> v readr 1.4.0 v forcats 0.5.0  
#> -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
#> x dplyr::filter() masks stats::filter()  
#> x dplyr::lag() masks stats::lag()  
#> x dplyr::select() masks MASS::select()

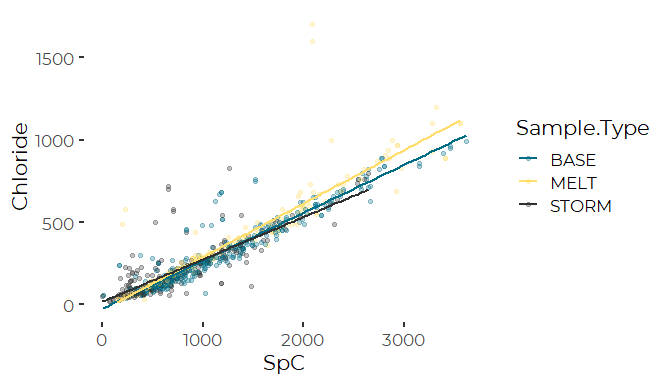
# Load Data

#> New names:  
#> \* `Sample Type` -> Sample.Type  
#> \* `Sample ID` -> Sample.ID  
#> \* `Site ID` -> Site.ID  
#> \* `Sample Date and Time` -> Sample.Date.and.Time  
#> \* `Sample Date` -> Sample.Date

# Flag Outliers

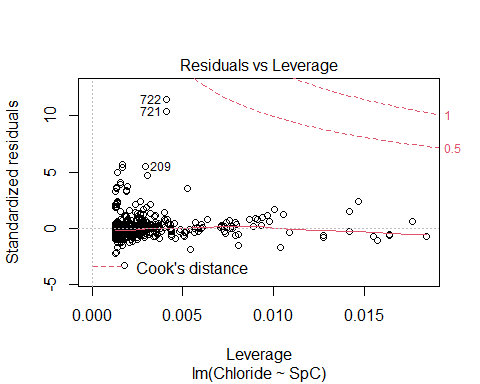
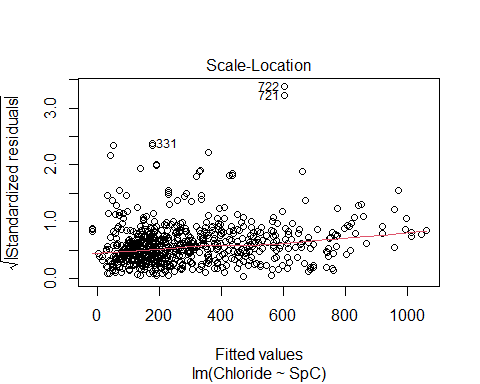
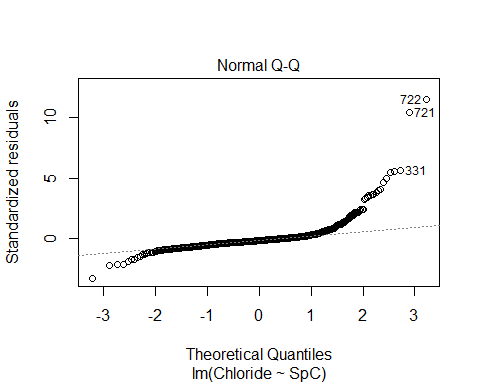
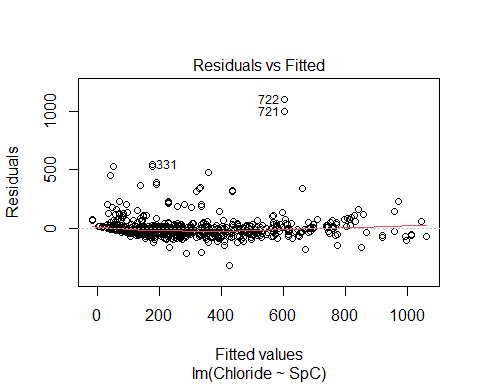
# Plot Data

#> `geom\_smooth()` using formula 'y ~ x'



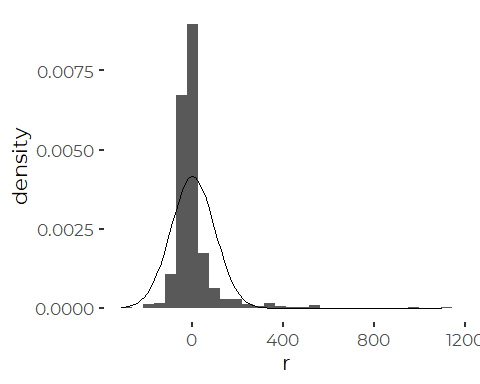
# Simple Linear Regression

#>   
#> Call:  
#> lm(formula = Chloride ~ SpC, data = chl\_data)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -314.12 -36.88 -15.29 9.05 1098.13   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) -16.986429 6.100374 -2.784 0.00549 \*\*   
#> SpC 0.297239 0.004863 61.123 < 2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 96.06 on 775 degrees of freedom  
#> Multiple R-squared: 0.8282, Adjusted R-squared: 0.828   
#> F-statistic: 3736 on 1 and 775 DF, p-value: < 2.2e-16

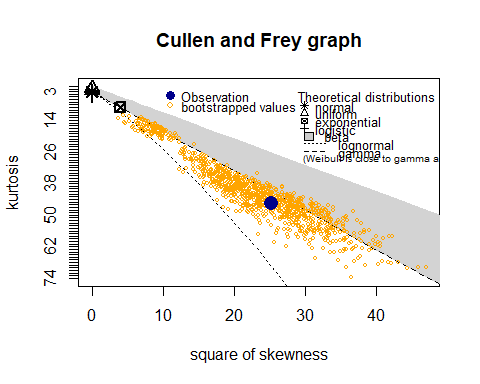


#Reviewing Residuals So the regression diagnostics clearly show the residuals deviate substantially from a normal distribution. The strongest deviations are from our two outliers (# 721 and 722), but even in their absence, the residuals are kurtotic and slightly skewed. There is also a small scale - location relationship, which is probably to be expected.

#> `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 It suggests a different model form may be more useful

# Cullen and Frey Moments Graph

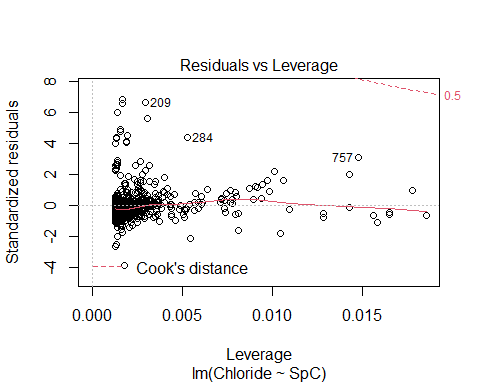
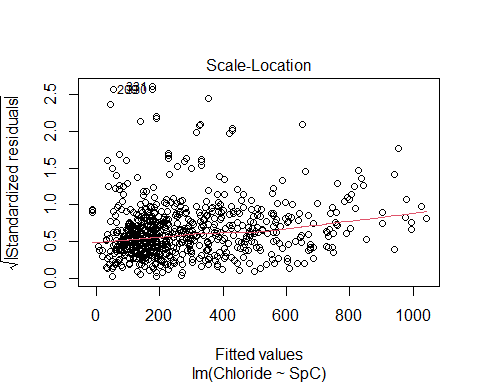
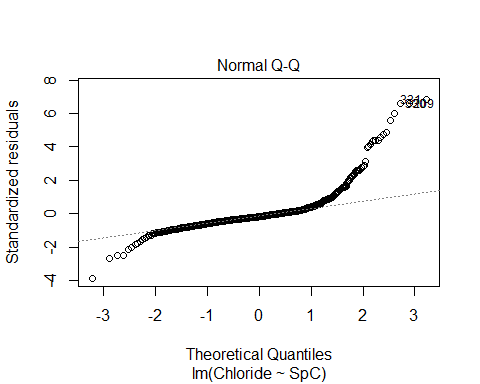
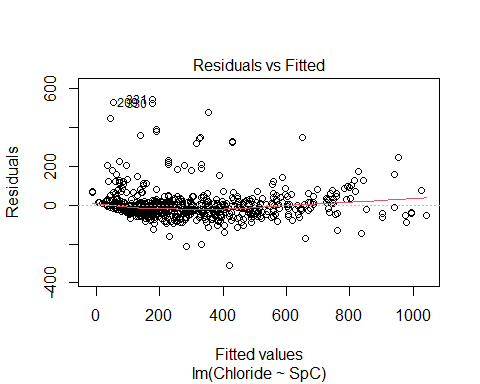


#> summary statistics  
#> ------  
#> min: -314.1169 max: 1098.134   
#> median: -15.28513   
#> mean: 6.400348e-15   
#> estimated sd: 95.99855   
#> estimated skewness: 5.016615   
#> estimated kurtosis: 45.57593

Which is little help. That suggest the distributional properties are not too far from a gamma distribution, but the gamma distribution is defined over the positive real numbers, and our residuals (by definition) have mean zero.

# Omitting Outliser

#>   
#> Call:  
#> lm(formula = Chloride ~ SpC, data = chl\_data, subset = !Outlier)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -308.85 -35.53 -14.14 11.28 544.03   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) -13.836887 5.074979 -2.726 0.00655 \*\*   
#> SpC 0.291569 0.004055 71.908 < 2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 79.87 on 773 degrees of freedom  
#> Multiple R-squared: 0.8699, Adjusted R-squared: 0.8698   
#> F-statistic: 5171 on 1 and 773 DF, p-value: < 2.2e-16



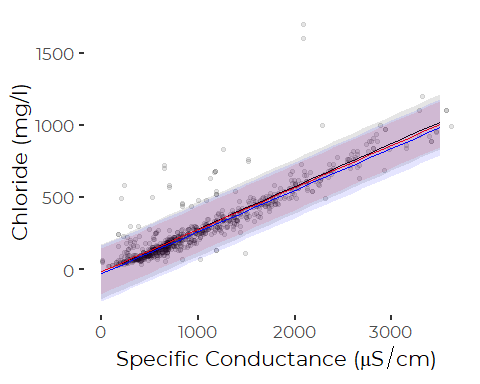
It’s still not a great model, give the structure of the errors. The kurtotic pattern of the residuals remains. It suggests a model that has left out some important predictor variables

# Robust Regression

We are using a Modified “Thiel-Sen Estimator” which is highly resistant to outliers.

#>   
#> Call:  
#> mblm(formula = Chloride ~ SpC, dataframe = chl\_data)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -291.14 -18.77 2.06 27.03 1125.94   
#>   
#> Coefficients:  
#> Estimate MAD V value Pr(>|V|)   
#> (Intercept) -27.96672 21.78284 46087 <2e-16 \*\*\*  
#> SpC 0.28916 0.02525 300622 <2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 98.16 on 775 degrees of freedom

# Compare Simple Linear Models

 Note that the robust regression (blue)line is parallel to, but somewhat below the other two regression lines. Error bands (95% confidence intervals of prediction) overlap. Many observations fall outside of those 95% confidence intervals, especially on the high side, suggesting some process may be generating unusually high chloride values. The NUMBER of high values is not extreme, but they are further from the 95% confidence band than we might expect. (As also shown bu the histogram of deviations, above).

# Polynomial Regressions

In reviewing preliminary models, I observed that the relationship between conductivity and chlorides is not quite linear. This actually corresponds to theory, which suggests we should see a shallow convex up relationship.

We fit nested polynomial models to test if that’s reasonable.

#> Analysis of Variance Table  
#>   
#> Response: Chloride  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> SpC 1 32985148 32985148 5309.4995 < 2.2e-16 \*\*\*  
#> I(SpC^2) 1 84200 84200 13.5534 0.0002481 \*\*\*  
#> I(SpC^3) 1 62017 62017 9.9826 0.0016420 \*\*   
#> I(SpC^4) 1 1260 1260 0.2028 0.6525743   
#> Residuals 770 4783608 6212   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#>   
#> Call:  
#> lm(formula = Chloride ~ SpC + I(SpC^2) + I(SpC^3) + I(SpC^4),   
#> data = chl\_data, subset = !Outlier)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -299.23 -35.68 -13.48 13.45 550.13   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 4.712e+01 1.825e+01 2.582 0.0100 \*\*  
#> SpC 1.206e-01 6.747e-02 1.788 0.0742 .   
#> I(SpC^2) 1.231e-04 7.862e-05 1.565 0.1179   
#> I(SpC^3) -3.031e-08 3.480e-08 -0.871 0.3841   
#> I(SpC^4) 2.288e-12 5.081e-12 0.450 0.6526   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 78.82 on 770 degrees of freedom  
#> Multiple R-squared: 0.8738, Adjusted R-squared: 0.8732   
#> F-statistic: 1333 on 4 and 770 DF, p-value: < 2.2e-16

Note that although the polynomial terms are significant by F test, they are not individually significant by T test, probably because of co-linearity. We can fit orthogonal polynomials using the poly() function.

Although a third order polynomial fits better than a second order polynomial, we stick with a second order polynomial for its simplicity. a Very lightly smoothed GAM might also work.

#> Analysis of Variance Table  
#>   
#> Response: Chloride  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> poly(SpC, 2) 2 33069348 16534674 2633.6 < 2.2e-16 \*\*\*  
#> Residuals 772 4846885 6278   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Call:  
#> lm(formula = Chloride ~ poly(SpC, 2), data = chl\_data, subset = !Outlier)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -297.52 -35.18 -13.94 12.62 544.28   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 287.958 2.846 101.171 < 2e-16 \*\*\*  
#> poly(SpC, 2)1 5758.940 79.460 72.476 < 2e-16 \*\*\*  
#> poly(SpC, 2)2 290.272 79.263 3.662 0.000267 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 79.24 on 772 degrees of freedom  
#> Multiple R-squared: 0.8722, Adjusted R-squared: 0.8718   
#> F-statistic: 2634 on 2 and 772 DF, p-value: < 2.2e-16

That fits better than the linear model

#> Analysis of Variance Table  
#>   
#> Model 1: Chloride ~ SpC  
#> Model 2: Chloride ~ poly(SpC, 2)  
#> Res.Df RSS Df Sum of Sq F Pr(>F)   
#> 1 773 4931085   
#> 2 772 4846885 1 84200 13.411 0.0002672 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

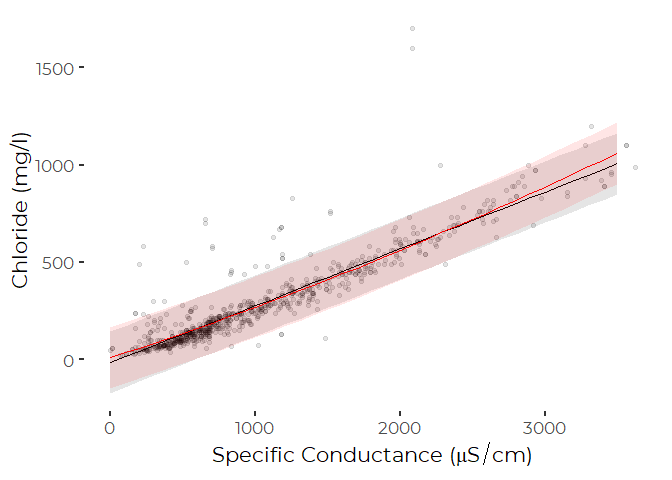
Theory suggests concentration of solutes should combine a linear term with a term raised to the three half power.

fitting a square root term, and this model actually fits far better than the straight polynomial model.

#> Analysis of Variance Table  
#>   
#> Model 1: Chloride ~ SpC  
#> Model 2: Chloride ~ SpC + sqrt(SpC)  
#> Model 3: Chloride ~ SpC + I(SpC^1.5)  
#> Model 4: Chloride ~ poly(SpC, 2)  
#> Res.Df RSS Df Sum of Sq F Pr(>F)   
#> 1 773 4931085   
#> 2 772 4792535 1 138550 22.318 2.744e-06 \*\*\*  
#> 3 772 4828908 0 -36373   
#> 4 772 4846885 0 -17977   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The square root model fits the best. Since all three models have the same degrees of freedom, there is no formal statistical test of what “better” means here.

# Compare Polynomial Models



# Multiple Regression Models

The wide outliers suggest our models are incomplete, omitting some predictor that could help explain extreme values. We can try to look at that by adding the few predictors we have available.

We chose not to include the Site term in the interactions, as we have few observations from some sites in several years, leading to risk of bias. We might be better off fitting a hierarchical model, with random intercepts and slopes by Site. We do not pursue that in this quick and dirty analysis.

## Maximum Model

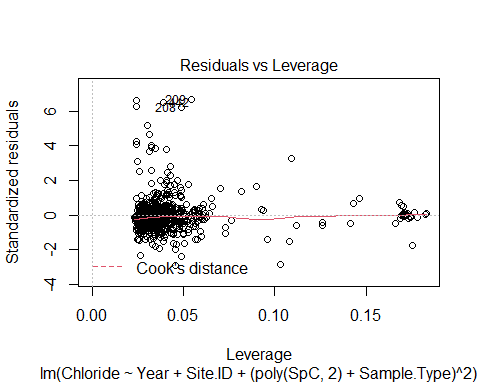
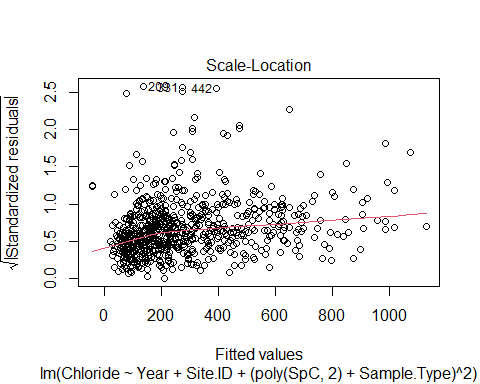
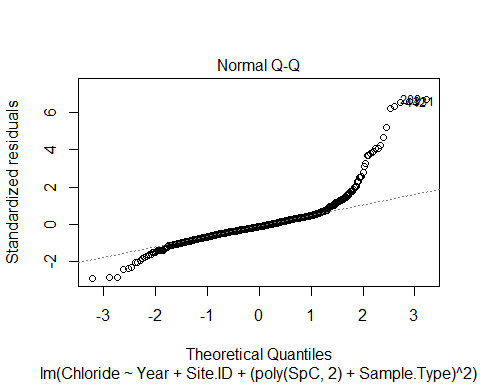
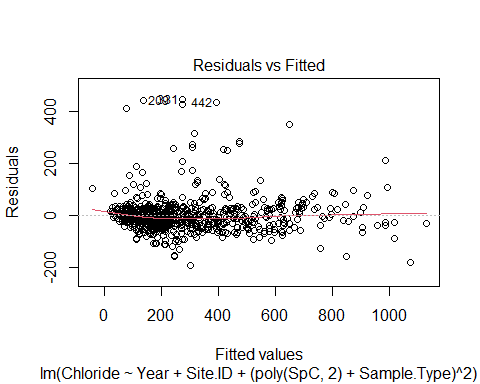
#> Analysis of Variance Table  
#>   
#> Response: Chloride  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> Year 10 3247889 324789 70.0548 < 2.2e-16 \*\*\*  
#> Site.ID 14 18343104 1310222 282.6060 < 2.2e-16 \*\*\*  
#> poly(SpC, 2) 2 12725066 6362533 1372.3554 < 2.2e-16 \*\*\*  
#> Sample.Type 2 40559 20280 4.3742 0.01293 \*   
#> poly(SpC, 2):Sample.Type 4 119544 29886 6.4462 4.217e-05 \*\*\*  
#> Residuals 742 3440071 4636   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Step Model Selection

All terms are “significant”, but it’s worth wondering if they are informative. We test with step(), which compares models based on AIC.

#> Start: AIC=6574.56  
#> Chloride ~ Year + Site.ID + (poly(SpC, 2) + Sample.Type)^2  
#>   
#> Df Sum of Sq RSS AIC  
#> <none> 3440071 6574.6  
#> - poly(SpC, 2):Sample.Type 4 119544 3559614 6593.0  
#> - Year 10 609416 4049486 6681.0  
#> - Site.ID 14 857965 4298036 6719.1  
#> Analysis of Variance Table  
#>   
#> Response: Chloride  
#> Df Sum Sq Mean Sq F value Pr(>F)   
#> Year 10 3247889 324789 70.0548 < 2.2e-16 \*\*\*  
#> Site.ID 14 18343104 1310222 282.6060 < 2.2e-16 \*\*\*  
#> poly(SpC, 2) 2 12725066 6362533 1372.3554 < 2.2e-16 \*\*\*  
#> Sample.Type 2 40559 20280 4.3742 0.01293 \*   
#> poly(SpC, 2):Sample.Type 4 119544 29886 6.4462 4.217e-05 \*\*\*  
#> Residuals 742 3440071 4636   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

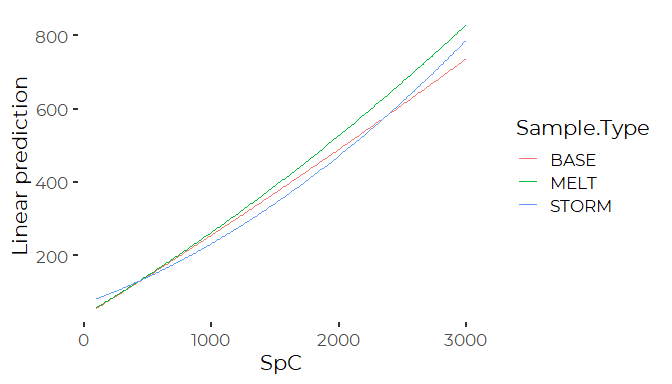
The step() function retains all terms from the full model

 So while the “optimal” model includes site and sample type terms, as well a interactions with SpC, the model weaknesses are little improved. The model still does not do a very good job of predicting outliers.

#>   
#> Call:  
#> lm(formula = Chloride ~ Year + Site.ID + (poly(SpC, 2) + Sample.Type)^2,   
#> data = chl\_data, subset = !Outlier)  
#>   
#> Residuals:  
#> Min 1Q Median 3Q Max   
#> -192.44 -31.02 -8.60 19.45 444.49   
#>   
#> Coefficients:  
#> Estimate Std. Error t value Pr(>|t|)   
#> (Intercept) 194.033 32.397 5.989 3.28e-09 \*\*\*  
#> Year2011 13.564 14.978 0.906 0.365436   
#> Year2012 10.671 15.066 0.708 0.478980   
#> Year2013 27.752 15.164 1.830 0.067627 .   
#> Year2014 73.986 14.817 4.993 7.40e-07 \*\*\*  
#> Year2015 98.227 16.153 6.081 1.91e-09 \*\*\*  
#> Year2016 73.621 17.209 4.278 2.13e-05 \*\*\*  
#> Year2017 64.671 15.107 4.281 2.11e-05 \*\*\*  
#> Year2018 94.884 15.657 6.060 2.16e-09 \*\*\*  
#> Year2019 78.662 16.645 4.726 2.74e-06 \*\*\*  
#> Year2020 100.812 16.496 6.111 1.59e-09 \*\*\*  
#> Site.IDS15 -33.278 39.423 -0.844 0.398875   
#> Site.IDS14 13.219 41.025 0.322 0.747368   
#> Site.IDS13 13.319 41.028 0.325 0.745556   
#> Site.IDS06 10.544 33.402 0.316 0.752334   
#> Site.IDS12 -23.798 31.790 -0.749 0.454337   
#> Site.IDS05 -6.804 30.175 -0.225 0.821655   
#> Site.IDS02 23.046 30.932 0.745 0.456480   
#> Site.IDS06B -16.817 30.481 -0.552 0.581292   
#> Site.IDS11 -28.232 31.917 -0.885 0.376692   
#> Site.IDS07 26.230 30.249 0.867 0.386151   
#> Site.IDS17 1.701 31.348 0.054 0.956735   
#> Site.IDS03 43.389 30.688 1.414 0.157819   
#> Site.IDS01 107.603 30.651 3.511 0.000474 \*\*\*  
#> Site.IDS04 97.036 31.333 3.097 0.002029 \*\*   
#> poly(SpC, 2)1 4618.834 139.202 33.181 < 2e-16 \*\*\*  
#> poly(SpC, 2)2 116.288 118.521 0.981 0.326836   
#> Sample.TypeMELT 14.587 6.383 2.285 0.022567 \*   
#> Sample.TypeSTORM -8.247 6.982 -1.181 0.237881   
#> poly(SpC, 2)1:Sample.TypeMELT 578.681 163.392 3.542 0.000422 \*\*\*  
#> poly(SpC, 2)2:Sample.TypeMELT 229.001 161.248 1.420 0.155976   
#> poly(SpC, 2)1:Sample.TypeSTORM 74.330 263.927 0.282 0.778305   
#> poly(SpC, 2)2:Sample.TypeSTORM 567.127 245.286 2.312 0.021045 \*   
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
#>   
#> Residual standard error: 68.09 on 742 degrees of freedom  
#> Multiple R-squared: 0.9093, Adjusted R-squared: 0.9054   
#> F-statistic: 232.4 on 32 and 742 DF, p-value: < 2.2e-16

Note that in this formulation, the second order polynomial term is involved with a large and significant interaction. The polynomial may only be needed to model storm event samples. This deserves further exploration, by plotting data and predictions, by Sample Type, but we do not bother. At a guess, the curvature in part reflects the fact that both chlorides and conductivity tend to be slightly lower during storm events.

## Marginal Means and Trends

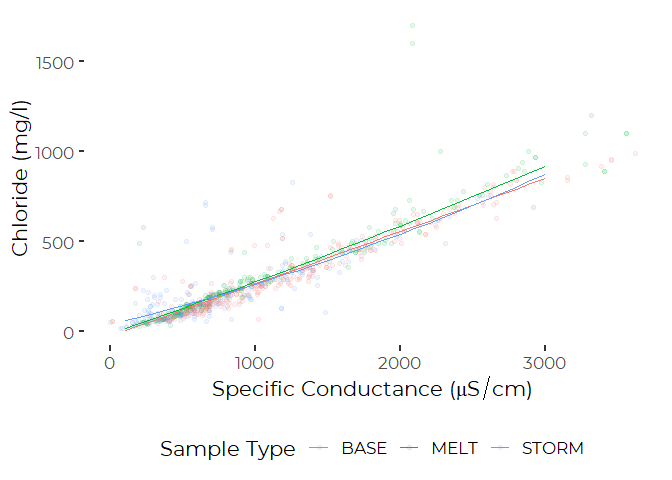
 That’s some indications that the Conductivity:Chloride relationship differs by Sample Type. note, however, that we actually have relatively little data for storm samples from the upper ranges, suggesting the main effect is that the conductivity - chloride slope may be less steep for the Storm samples.

I use a simplified model to generate a nice interaction plot to show the relationships.

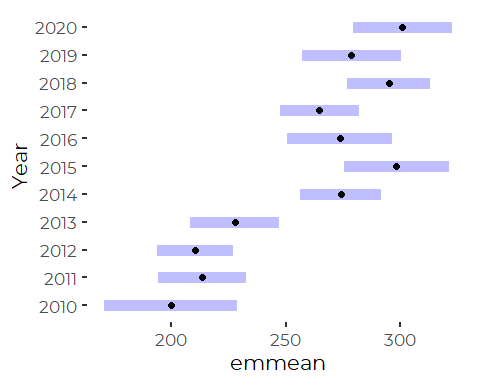
#> Analysis of Variance Table  
#>   
#> Model 1: Chloride ~ poly(SpC, 2) \* Sample.Type  
#> Model 2: Chloride ~ Year + Site.ID + (poly(SpC, 2) + Sample.Type)^2  
#> Res.Df RSS Df Sum of Sq F Pr(>F)   
#> 1 766 4700296   
#> 2 742 3440071 24 1260225 11.326 < 2.2e-16 \*\*\*  
#> ---  
#> Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

So the simplified model is definitely worse than the full model. The difference is from omitting the Site and Year terms.

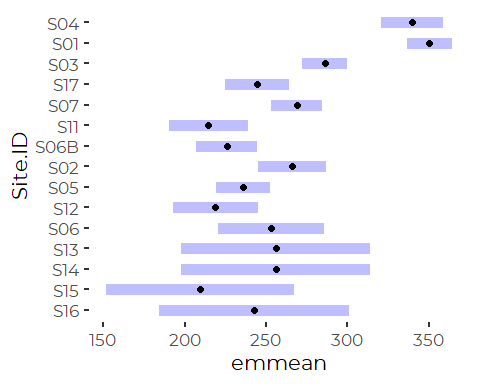
# Combined Graphic



# Additional Observations

WE can regain the high recent salinity conclusions we found for State of the Bay. 

And can examine differences among sites. These have been sorted by median chlorides. I fear there’s some bias here because of different sampling histories that can’t be fully resolved. But note only the two (or maybe three) highest salinity sites are significantly different from the others.



# Conclusions

1. As suspected, the use of a “resistant” regression methods has little effect on the Chloride concentration predictions.

2.We have something on the order of 25 or 30 observations that fall outside – and often well outside – of those 95% prediction intervals. That’s about the number we expect, as we have a few under 800 observations, so should expect about 800 \* 0.05 = 40 observations outside the 95% confidence intervals. But the deviations from prediction for those points are large. The distribution of errors is both slightly skewed (expected, since chloride < 0 is impossible), and heavy-tailed (a bit of a surprise). The best interpretation of these data is that for some unknown reason, we sometimes (rarely – maybe a couple percent of observations) get chloride values well above what we expect based on conductivity.

1. The multivariate models show “significant” differences in the conductivity to chloride regression relationships among sites and between melt, baseflow, and storm samples. This is not a surprise. We’ve seen that in prior years, but with more data, the differences are becoming “more significant”. They’re still small effects.
2. I noticed an odd pattern in the model results. When fitting multivariate models, several sites with high average chloride ALSO tend to show chlorides slightly higher at a given conductivity (or at least that’s what the MODEL shows). That got me wondering why that might be. One possibility is that the simple linear models are not appropriate. If there’s “really” some curvature to the conductivity - chloride regression, we’d expect exactly this result.
3. Digging a little into the theory of conductivity suggests a SMALL non-linearity in the conductivity-chloride relationship is to be expected. For “ideal solutions” of ions that dissociate completely in solution, conductivity increases slightly more slowly per unit increase in concentration, as the concentration increases. That means predicted concentrations based on conductivity should go up slightly faster at higher conductivities. The effect is small (much smaller than our 95% prediction intervals), but real. Fitting polynomial models does significantly improve model fit.
4. We now have evidence that there are subtle differences in the chloride-conductivity relationship between base flow, storm flow, and melt events. The differences are small. While they are “significant”, they reflect the fact that most of our high conductivity samples are associated with melt events and most of our low conductivity samples are from storm events. Given the small (expected) curvature in the conductivity-chloride relationship, the analysis is fitting curves to data pulled from different ranges. Observed slope is lower for lower conductivities – associated here with the storm events – and higher for the higher conductivities – associated with melt events.