# Analysis of data collected under the Berry Protocol

2017-03-27

# 1. Summary of protocol:

## 1.1 Basic protocol

Areas rich in soapberries is located for permanent monitoring. In each area, 10 robust bushes are chosen, and two stems on each plant are chosen for sampling. The stems and bushes are marked with permanent tags so that they can be revisited each year.

There are three measures taken in this protocol.

* **Berry count**. The number of berries produced on the stem is recorded as an index of soapberry production.
* **Stem diameter.** The diameter (in millimeters) of the stem near its base is measured.
* **Mean berry weight**. A collection of 25-50 ripe red berries is obtained in August and weighed so the average wet weight of a single berry from each area is obtained.

If the tagged stem has died (or is damaged or the tag on the stem has “disappeared”), a new stem is chosen. This may be from a new bush or the same bush. If the stem has been browsed, then no count is conducted on this stem this year.

If the tagged bush has died (or the tags on all of the stems have “disappeared”), a new bush is selected for subsequent monitoring.

## 1.2 Cautions about the protocol.

#### 1.2.1 Don’t use 0 to indicate a missing value.

If a branch is present but the berries cannot be counted (e.g. browsed), a standardized codes should be entered into the data base. The berry count should be entered as MISSING rather than as zero.

### 1.2.2.Codes for stem and bushes.

The Current field uses a *plantxx-stemxx* notation (e.g. plant1-stem1). If a bush dies and is replaced by a new bush, a different “plant” number should be used. Similarly if a stem is replaced on the same plant, use a different stem number (but linked to the same bush). If a new stem on new bush is used, both the plant (bush) and stem number should be new. Do NOT reuse bush numbers on different bushed; do not reuse stem numbers on the same bush.

## 2. Database structure

The database for this protocol is a series of Excel workbooks with multiple sheets in each workbook. The *General Survey* sheet contains the information collected. There is one line per stem.

The relevant fields on the worksheet are:

* *Study Area Name*. The name of the study area.
* *Sample Station Label*. The bush/stem label.
* *Date*. The date the data was collected. The *Year* is extracted from this date.
* *Berry Count*. The number of soap berries on this stem. If the stem is browed (or damaged) a missing value should be entered here and not the value of 0.
* *Stem Diameter*. The diameter (mm) of the stem.
* *Average Weight.* The average weight of a sample of berries is collected. Notice that there is only ONE mean weight found so this value is replicated on every stem line of the sheet. The sample size used to determine the weight is in a separate column.

# 3. Sample Analyses – Single Sites.

A sample analysis is presented on the *Eskers Park* study area. Data is available from 2013 to 2015.

This design has multiple transects that are repeated measured over time with multiple plots measured on each transect that are also repeated measured over time. Please refer to the *Fitting Trends with Complex Study Designs* document in the *CommonFile* directory for information on fitting trends with complex study designs.

All analyses were done using the *R* (R Core Team, 2016) analysis system. An HTML document showing the results of the analysis is available. All plots are also saved as separate \*png files for inclusion into reports.

## 3.1 Mean berry weight.

This measurement is taken at the site level and so there is one measurement available per site/year. Notice that this value is replicated multiple times in the database for each individual stem, These are NOT real replicated readings but only an artifact of the database so some care is needed to extract only a single value per individual stem on a site/year..

A simple linear regression can be used to look for changes over time using the model



where *MBW* is the mean berry weight and *Year* is the calendar year over time. This model can be fit using the *lm()* function in *R.* Figure 1 shows asummary plot, along with estimates of the slope, its standard error, and the p-value of the hypothesis of no trend. With only 3 year of data, there is no evidence (p=0.66) of a trend with an estimated slope of 0.02 (SE 0.03) g/year.

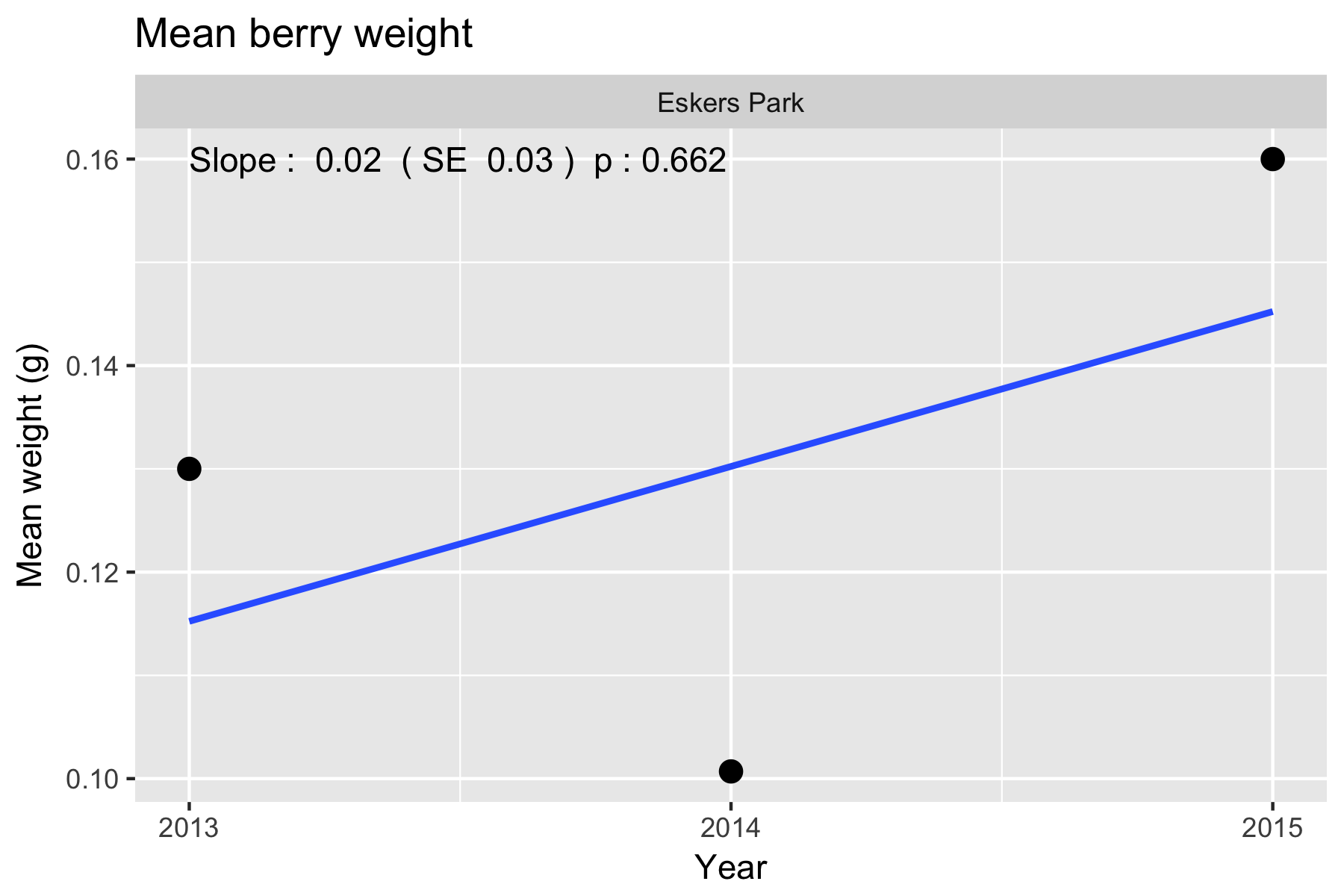


Figure 1. Summary plot of the trend in mean berry weight at Eskers Park.

Following the fit, the diagnostic plots should be examined. An illustration of such a plot is shown in Figure 2.

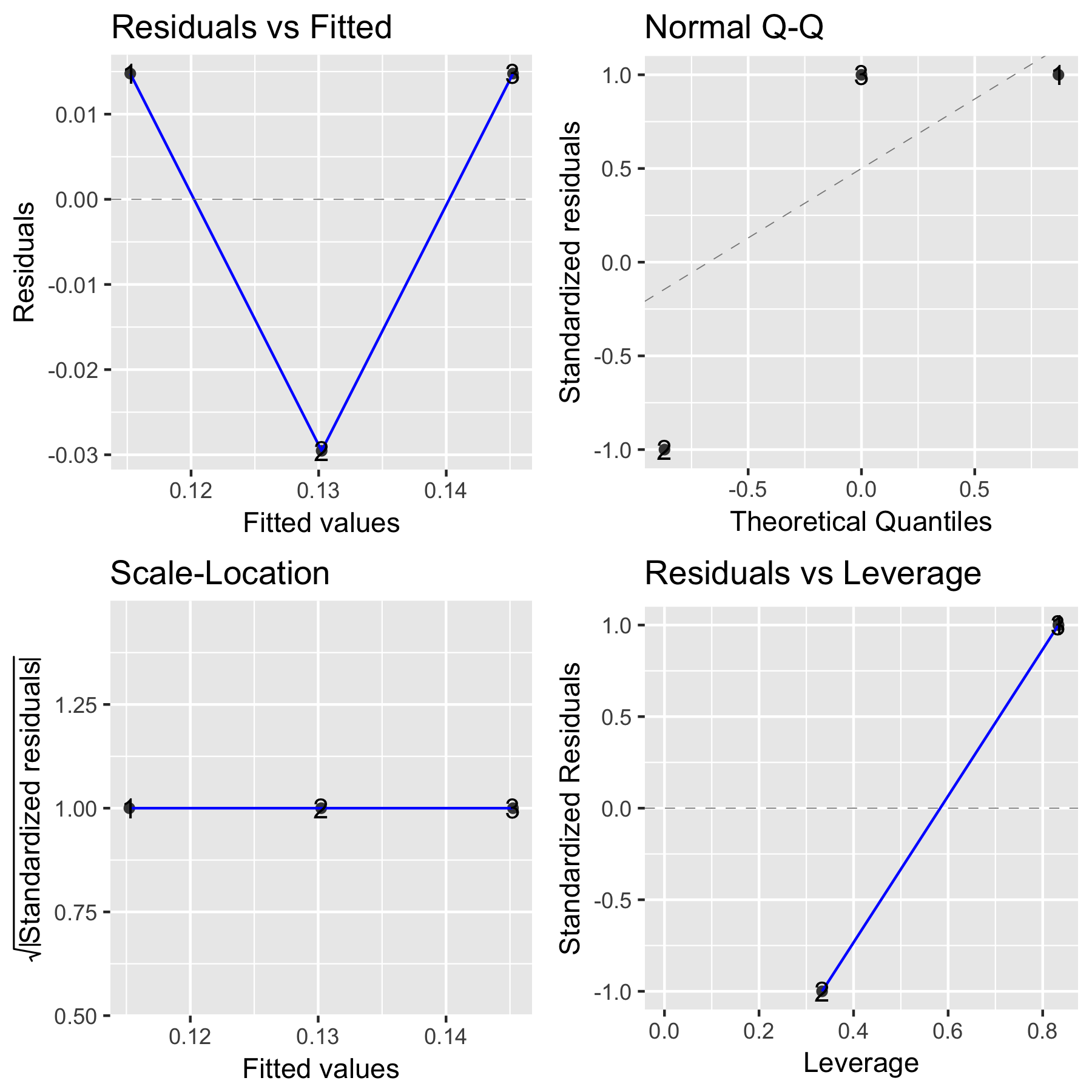


Figure 2. A sample diagnostic plot for the analysis of mean berry weight at Eskers Parks

With only 3 years of data, the plots are not very informative. In the upper left corner is a plot of residuals vs. the fitted values. A good plot will show a random scatter around 0. Any large deviations from 0 should be investigated as potential outliers. In the upper right is a normal probability plot. Points should be close to the dashed reference line. Fortunately, the analysis is fairly robust against non-normality so only extreme departures are worrisome. The bottom left plot examine the assumption that the variation about the line is constant over the line. You would expect to see a constant band of points. Finally the bottom right plot is a leverage plot – this is not useful for this simple model and can be ignored.

It will also be possible to covariates such as mean winter temperature or degree days in the year to try and explain some of the variation over time using a multiple regression. With only three years of data available, this not sensible.

Whenever an analysis of a trend over time is conducted, the analysis should test and adjust for autocorrelation. Autocorrelation usually isn’t a problem (and likely cannot be detected) unless you have 10+ years of data. The test for autocorrelation commonly used is the Durbin-Watson test.

If the number of berries used to compute the average is quite different over years, a weighted analysis (number of berries used in computing the mean) may be needed.

## 3.2 Stem Diameter.

This measurement is taken at the stem level and so there is one number per stem/bush/year. The same stem is repeatedly measured over time, but stems may leave the protocol (damaged or dead) or be added to the protocol (replacement stem) over time. All of the models below automatically will account for stems that are removed or added as long as each stem has a unique label within a site.

A linear mixed model will be used to look for changes over time to account for the repeated measurements over time of each branch on the same plant:



where *StemD* is the measured stem diameter, Yearis the trend; *YearF(R), BushF(R)* and *StemF(R)* are the random effects of year-specific factors, bushes and stems respectively. These random effects are needed to account for the repeated measurement of the same stem (the *StemF(R))* term; multiple stems measured from the same plant (*BushF(R))* term; and year-specific factors (also known as process error, *YearF(R)*). The process error term is distinguished from the simple trend term *Year*.

For example, all of the stems on the same bush may have related diameters because they are all similar aged. Similarly, the repeated measurements on the same stem over time will be related.

This model is fit using the *lmer()* function in the *lmerTest* package (Kuznetsova, et al. 2016) and a summary is shown in Figure 3.

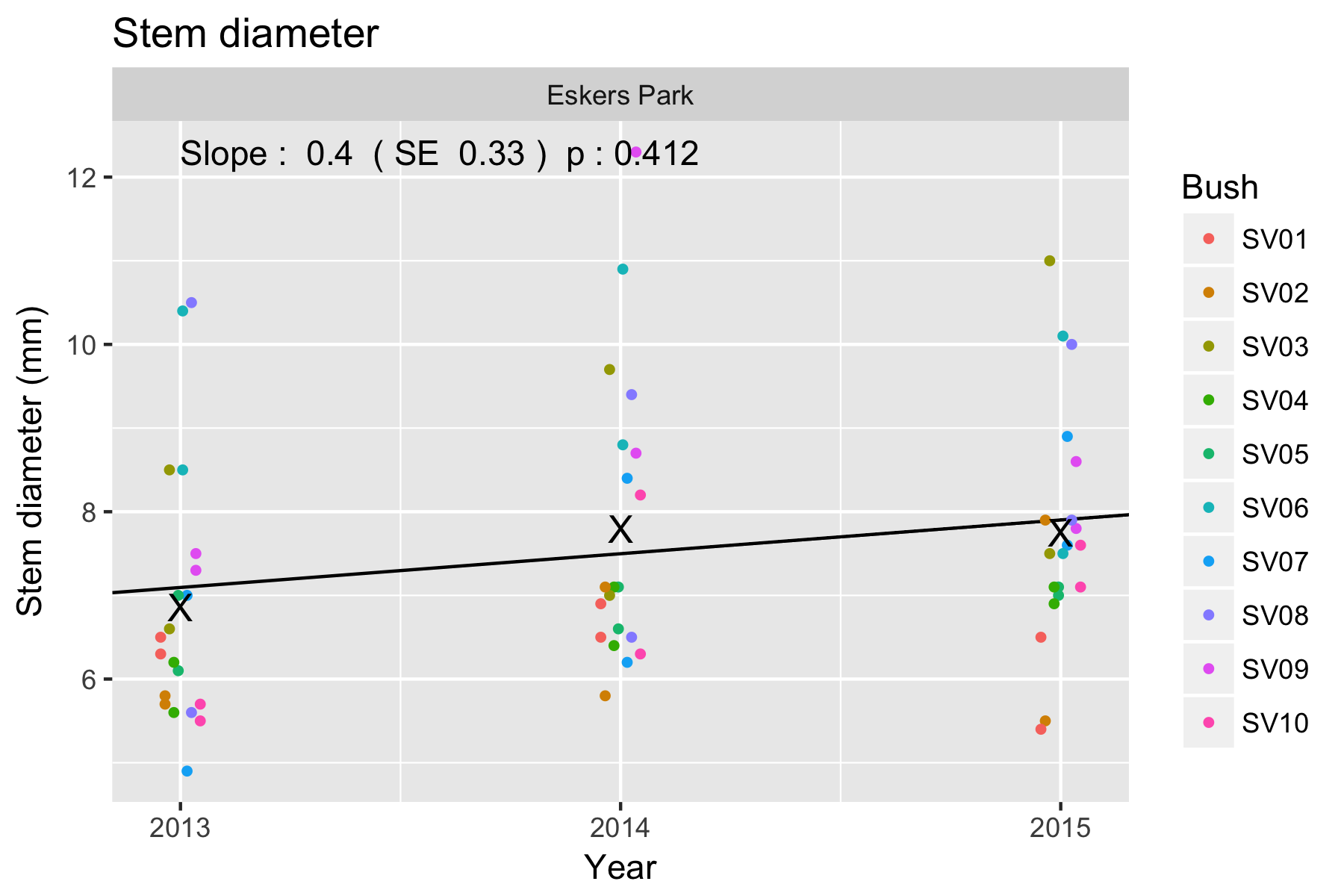


Figure 4. Summary plot from analysis of stem diameter data from Eskers Park. Points are jittered to reduce overplotting. The X indicate the observed mean of the data points in each year and deviations of the X from the trend line would represent (approximately) process error.

There was no evidence of a trend (p=0.42) with an estimated slope of 0.40 (SE 0.33). Note that because of the presence of process error, the effective sample for testing a trend is the number of YEARS and not the total number of observations, i.e. the three X’s essentially define the trend, while the other data points provide information about bush-to-bush variation and stem-to-stem variation, they provide little information on trend.

One of the outputs from this analysis is the relative size of the standard deviations in the points due to year-specific factor, bushes, stems within bushes, and residual (unknown) sources. For this example the estimated variance components are:

Source Std.Dev.

Stem 0.96

Bush 0.88

Year 0.39

Residual 0.84

The stem-to-stem variation (within a bush) is comparable to the bush-to-bush variation and residual variation and all are much larger than year-specific effect (process error). This is not too surprising because it is hard to imagine that stem diameter could be readily influenced by year-specific factors (unlike, for example, berry counts).

Diagnostic plots can again be produced (Figure 5)

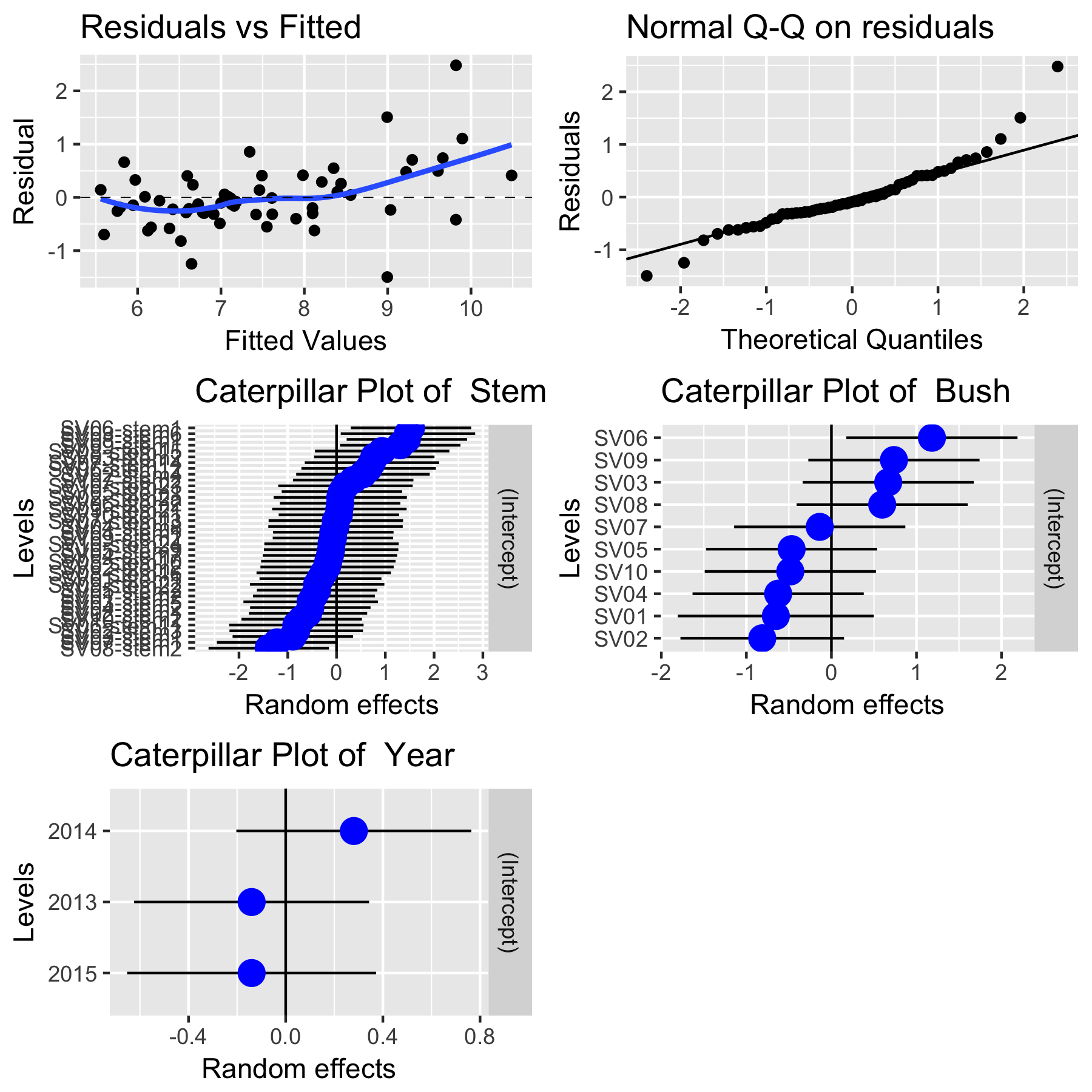


Figure 5.Diagnostic plots from a trend analysis on stem diameter.

The upper two plots are interpreted in the same way as noted previously. There is some (very weak) evidence of a lack of fit for stems with larger diameters, but it is not serious. Caterpiller plots attempt to show the distribution of the random effects. If the model fits well, you would expect the blue dots (the estimated random effect) to lie mostly in the +/- 2 std bands with no obvious outliers. There are no obvious outliers in the stem or bush random effects and there are too few years to say much.

As with the analysis of mean berry weight, covariates can also be added to the model to explain some of the year-specific effects.

Whenever an analysis of a trend over time is conducted, the analysis will have to test and adjust for autocorrelation in the year-specific effect. This usually isn’t a problem unless there are 10+ years of data.

## 2.3 Berry Count.

This measurement is taken at the stem level and so there is one number per stem/plant/year. The same stem/plant is repeated measured over time. All of the models below automatically will account for branches that are removed or added as long as each stem has a unique label within a site.

The models for the berry count are similar to those from the stem diameter except that the counts may be somewhat smallish. The average count is about 5 or less, then a Poisson regression can, in theory, be used. However, if there are several random effects, then a Poisson mixed effects model is extremely difficult to fit. However, for larger counts, a linear mixed model on the log(counts+0.5) will work well and avoid many of the problems in dealing with generalized linear mixed models.

A linear mixed model will be used to look for changes over time in the mean berry count to account for the repeated measurements over time of each branch on the same plant:



where *BerryCount* is the measured berry count on the stem, Yearis the trend; *YearF(R), Bush(R)* and *StemF(R)* are the random effects of year-specific factors, plants and stems respectively. These random effects are needed to account for the repeated measurement of the same stem (the *StemF(R))* term; multiple stems measured from the same plant (*BushF(R))* term; and year-specific factors (also known as process error, *YearF(R)*). Refer to the previous section for a discussion of process error.

This model is fit using the *lmer()* function in the *lmerTest* package (Kuznetsova, et al. 2016) and a summary is shown in Figure 6.

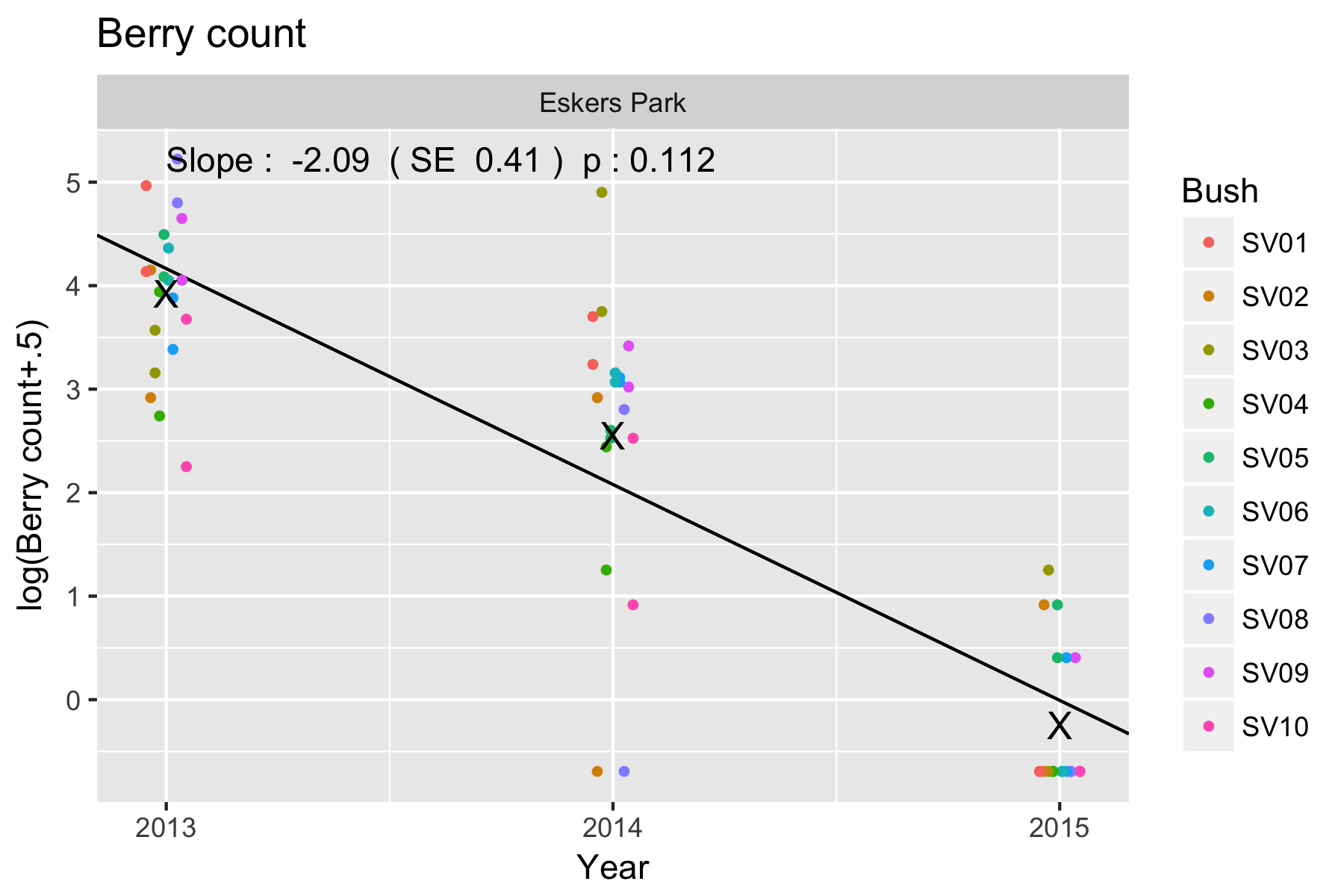


Figure 6. Summary plot of the trend analysis on the mean log(berry count). The X indicates the mean of the counts for that year showing a small year-specific effect because they are not centered on the trend line.

Despite the apparent step drop in mean berry production, there is no evidence of a trend (p=0.11) with the estimated slope of -2.09 (SE 0.41). Even though the SE of the estimated slope is quite small, there are only 3 years of data which leaves only a single degree of freedom in computing the confidence interval and the multiplier is much larger than the approximate of 2.

Diagnostic plots of the fit can be produced and interpreted as in the analysis of stem diameter (not shown here)

The variance component of the random effects can also be examined (not shown here, but see the output) and they show that now year-specific factors are on the same order of magnitude as bush-to-bush variation or stem-to-stem variation.

Additional covariates (e.g. temperature or rainfall) can be used to explain some of the year-specific factors (not shown).

# 4. Sample Analyses – Multiple Sites.

Now we are interested in comparing trends across two or more study areas. A sample analysis is presented using the *Eskers Park* and the *Schoen Lake* study areas. Data is available for *Eskers Park* from 2013 to 2015; we simulated another year of data for 2016. Data is available for *Schoen Lake* from 2013 to 2014. We simulated another two years of data for this study area.

This design has multiple transects that are repeated measured over time with multiple plots measured on each transect that are also repeated measured over time. Please refer to the *Fitting Trends with Complex Study Designs* document in the *CommonFile* directory for information on fitting trends with complex study designs.

All analyses were done using the *R* (R Core Team, 2016) analysis system. An HTML document showing the results of the analysis is available. All plots are also saved as separate \*png files for inclusion into reports.

## 3.1 Mean berry weight.

This measurement is taken at the site level and so there is one measurement available per site/year. Notice that this value is replicated multiple times (once for each stem in the data base) so some care is needed to extract only a single value per site/year.

A preliminary plot is shown in Figure 1

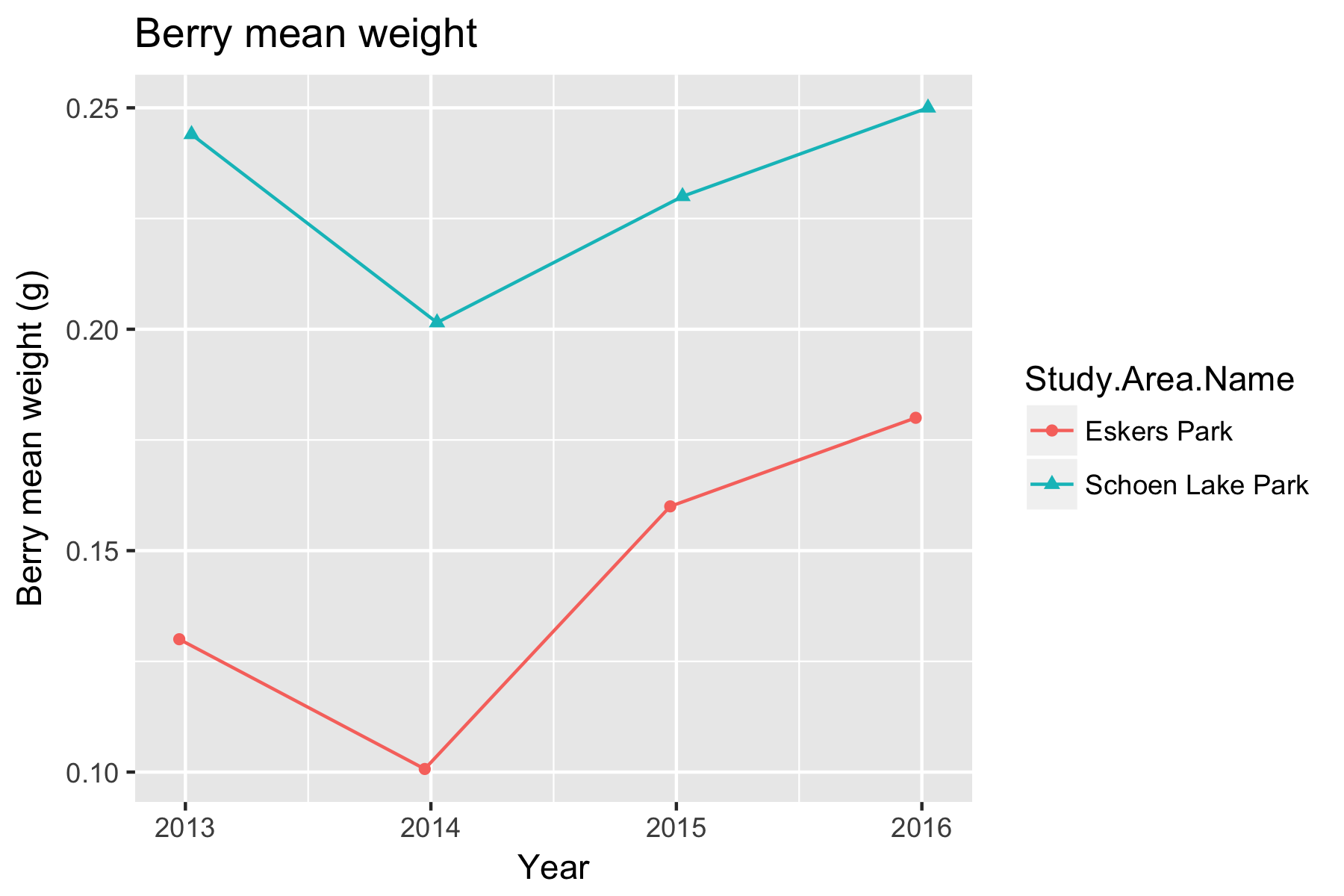


Figure 1. A preliminary plot of the trends over time in berry weight for two study areas.

The trends look similar and there is evidcnce of a year-specific factor because both weights are depressed in 2015.

A linear mixed model Analysis of Covariance (LMM ANCOVA) can be used to examine if the trends are the same in both study areas. There are two models that must be considered. First is the non-parallel slope model where each study area has its own trend line.



where *MeanWeight* is the mean weight of berries in the study area in each year; *YearF* represents the year specific factors (process error); *Year* represents the (common) calendar year trend over time; *StudyArea* represents the different intercepts for the trend. Because there is only one measurement per study area (unlike other protocols where multiple transects are measured in each study area), it is not necessary to include the *Year:StudyArea* term representing the differential trend in the two study areas. The *YearF* term allows for year-specific effects (process error) that affect counts in all study areas simultaneously.

Second is the parallel slope model where the trend is the same in both study areas, with a different intercept:



The model is identical to the previous model except that the *StudyArea:Year* term has been dropped which forces the slopes to be the same in all study areas.

Both models can be fit using the *lmer()* function in *R.* Following the fit, the diagnostic plots should be examined. An illustration of such a plot is shown in Figure 2 for the non-parallel slope model

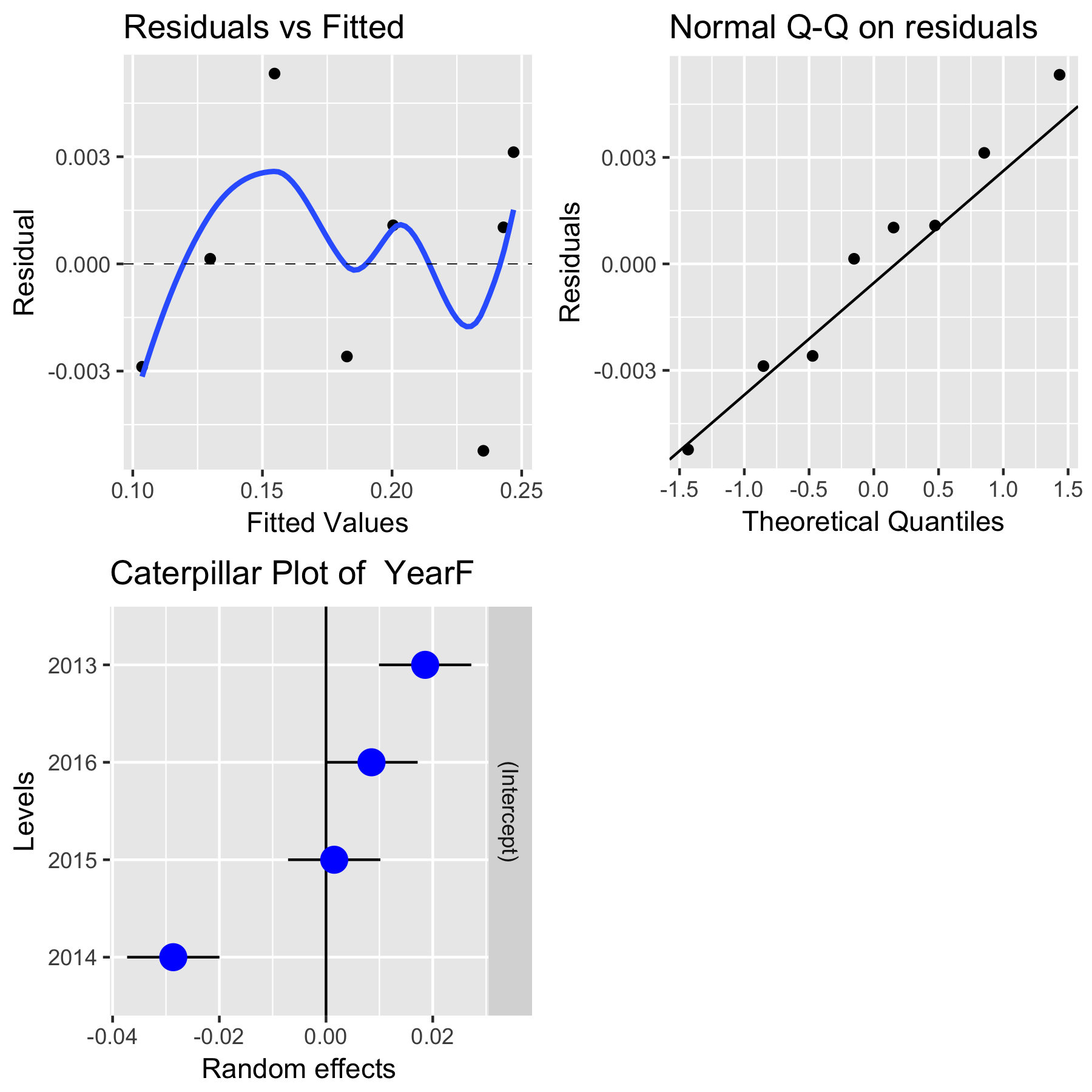


Figure 2. A sample diagnostic plot for the analysis of mean berry weight at *Eskers Park* and *Schoen Lake* for the non-parallel slope model.

With only a few years of data, the plots are not very informative. In the upper left corner is a plot of residuals vs. the fitted values. A good plot will show a random scatter around 0. Any large deviations from 0 should be investigated as potential outliers. In the upper right is a normal probability plot. Points should be close to the dashed reference line. Fortunately, the analysis is fairly robust against non-normality so only extreme departures are worrisome. The bottom left plot examines the year-specific effects. In this case, the year-specific effects for 2014 is quite pronounced.

Figures 3 and 4 shows asummary plot, along with estimates of the slope, its standard error for each of the models.

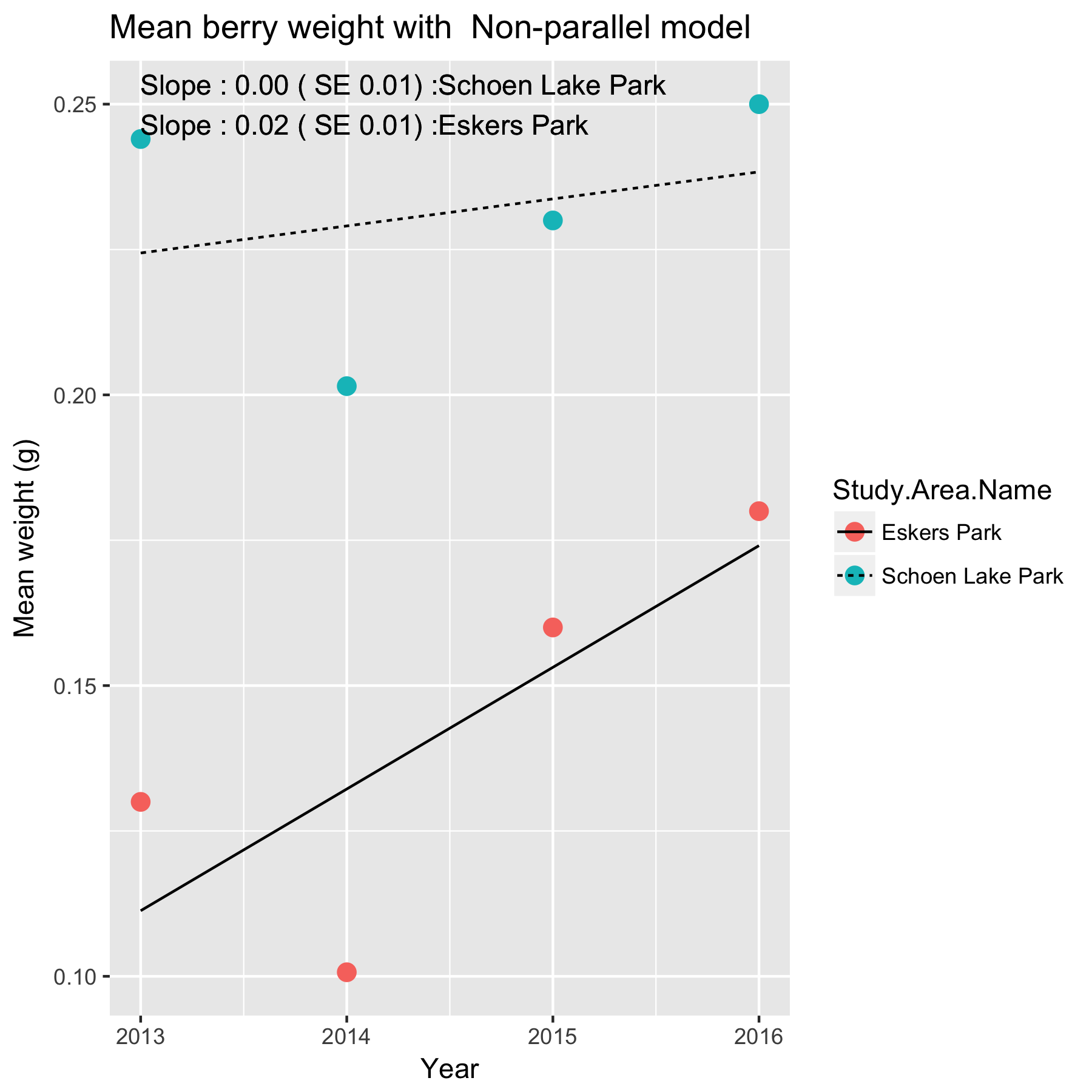


Figure 3. Summary plot of the trend in mean berry weight at *Eskers Park* and *Schoen Lake* under the non-parallel slope model.

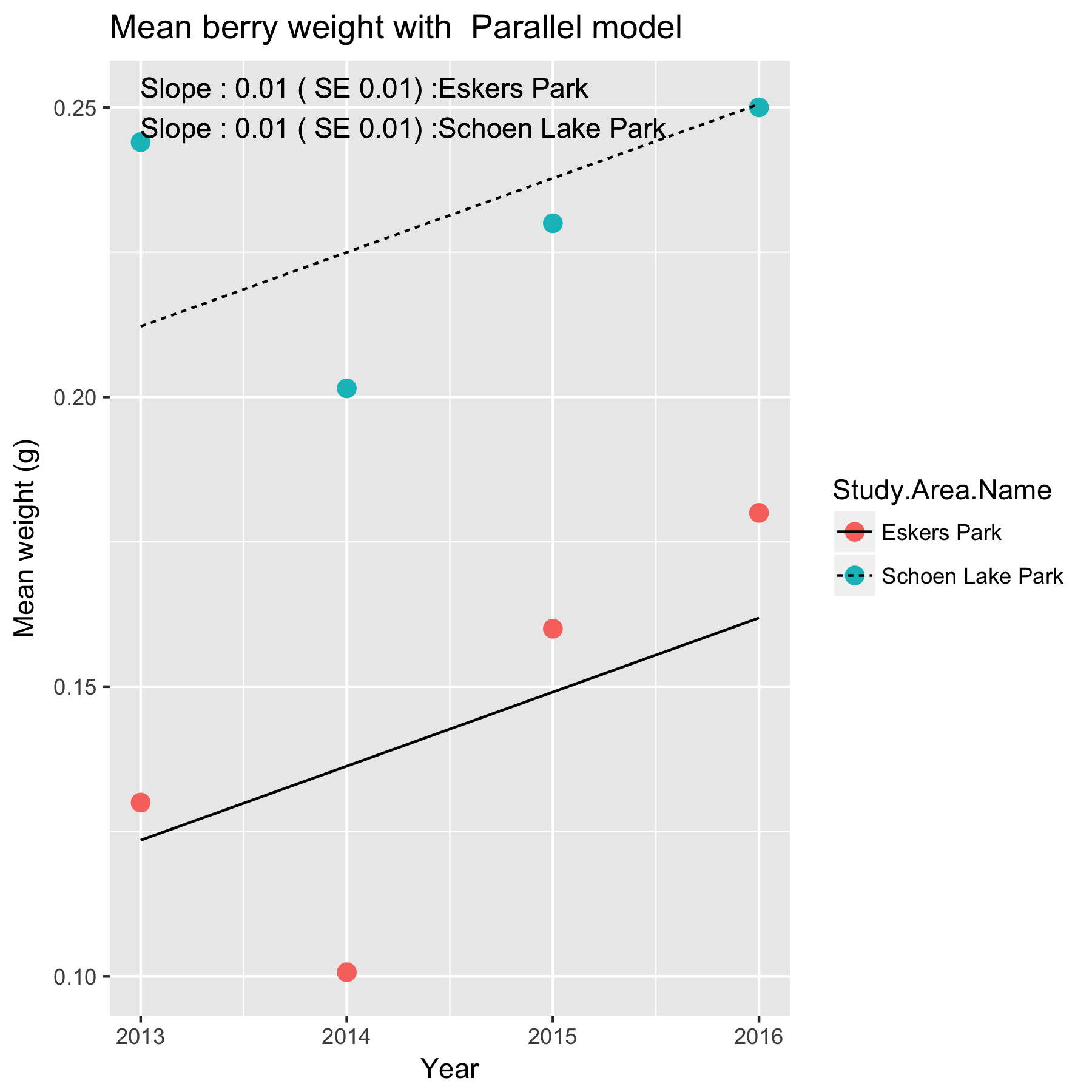


Figure 4. Summary plot of the trend in mean berry weight at *Eskers Park* and *Schoen Lake* under the parallel slope model.

The slopes are interpreted in the same way as in the single site case. For example, the estimated slope under the parallel slope model is of 0.01 (SE 0.01) /year for both study areas, i.e. the mean weight is trending upwards a .01 g/year.

Which model is preferable? A comparison between the two models (the parallel vs. the non-parallel slope model in the sites) is constructed using the *anova()* function in *R.* There was strong evidence against the parallel slope model (p=.002), implying that the non-parallel slope model is a better fit to the data. This indicates that there is evidence that the trend lines in the two sites have different slopes, i.e. the trends in the sites appear to be different.

In the event that you have more than two study sites and the non-parallel slope model was preferred, it is possible to do the equivalent of a Tukey multiple comparison procedure o the slope to see where the slopes may differ among the study areas. The typically output from the *lsmeans0* package looks like:

Study.Area Slope SE 95% LCL 95% UCL .group model

Schoen Lake 0.0046 0.011 -0.03296 0.0422 1 Non-parallel

Eskers Park 0.0209 0.011 -0.01668 0.0585 2 Non-parallel

Confidence level used: 0.95

The individual slopes along with standard errors and a 95% confidence interval for the slope are presented. The column labeled “group” indicates which study areas were found to have a different slope from other study areas. There is evidence that the slopes differ if a pair of study areas with different values of the grouping variable. In this case, all of the study areas slopes belong to the different “groups”, so there is no evidence of a difference in their slopes, i.e. the non-parallel slope model is to be preferred.

It will also be possible to covariates such as mean winter temperature or degree days in the year to try and explain some of the variation over time using a multiple regression. With only limited number of years of data available, this not sensible.

Whenever an analysis of a trend over time is conducted, the analysis should test and adjust for autocorrelation. Autocorrelation usually isn’t a problem (and likely cannot be detected) unless you have 10+ years of data. The test for autocorrelation commonly used is the Durbin-Watson test. There was no evidence of autocorrelation over time.

As in the single site case, an analysis can also be done by averaging over all transects in each years if the design is balanced. Please refer to the *Fitting Trends with Complex Study Designs* document in the *CommonFile* directory for information on fitting trends with complex study designs for more details.

If the number of berries used to compute the average is quite different over years, a weighted analysis (number of berries used in computing the mean) may be needed.

## 4.2 Stem Diameter.

This measurement is taken at the stem level and so there is one number per stem/bush/year. The same stem is repeatedly measured over time, but stems may leave the protocol (damaged or dead) or be added to the protocol (replacement stem) over time. All of the models below automatically will account for stems that are removed or added as long as each stem has a unique label within a site.

A preliminary plot is shown in Figure 5

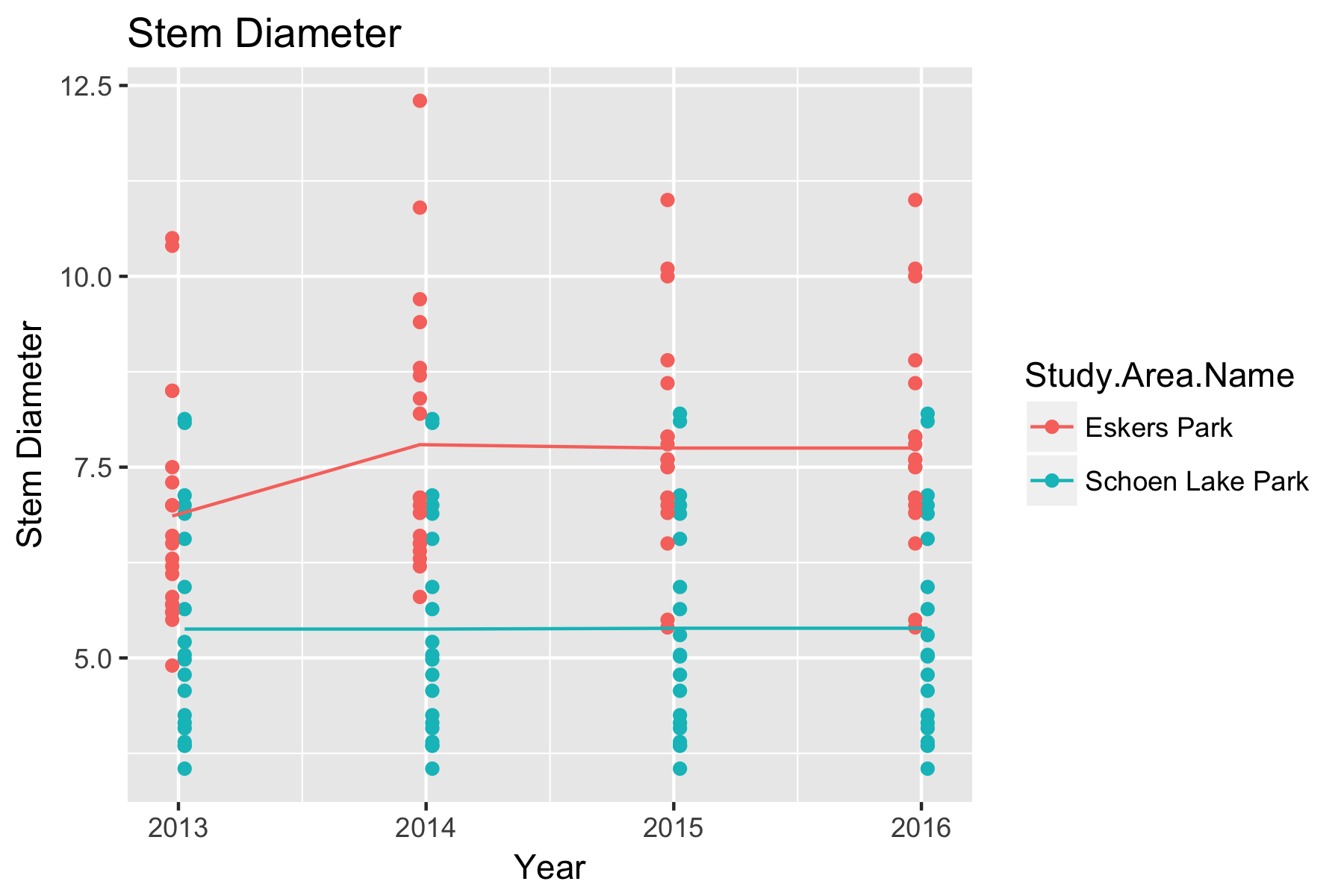


Figure 5. Preliminary plot of the stem diameter at *Eskers Lake* and *Schoen Lake.*

While not evident from the above plot, there is evidence from the single site analyses of plant effects (i.e. some plants are more robust than others) and stem effects (it is difficult for a stem to decrease in size over time).

Two linear mixed models will be used to look for differences in the trend among study area o accounting for the repeated measurements over time of each branch on the same plant. The first model is the non-parallel slope model:



and the second model is the parallel slope model:



where *StemDiameter* is the measured stem diameter, Yearis the trend; *YearF(R), StemF(R)* are the random effects of year-specific factors, bushes and stems respectively. These random effects are needed to account for the repeated measurement of the same stem (the *StemF(R))* term; and year-specific factors (also known as process error, *YearF(R)*). The process error term is distinguished from the simple trend term *Year*.

This model is fit using the *lmer()* function in the *lmerTest* package (Kuznetsova, et al. 2016). Following the fits, the diagnotsitc plots should be examined. An illustration of such a plot is shown in Figure 6 for the non-parallel slope model

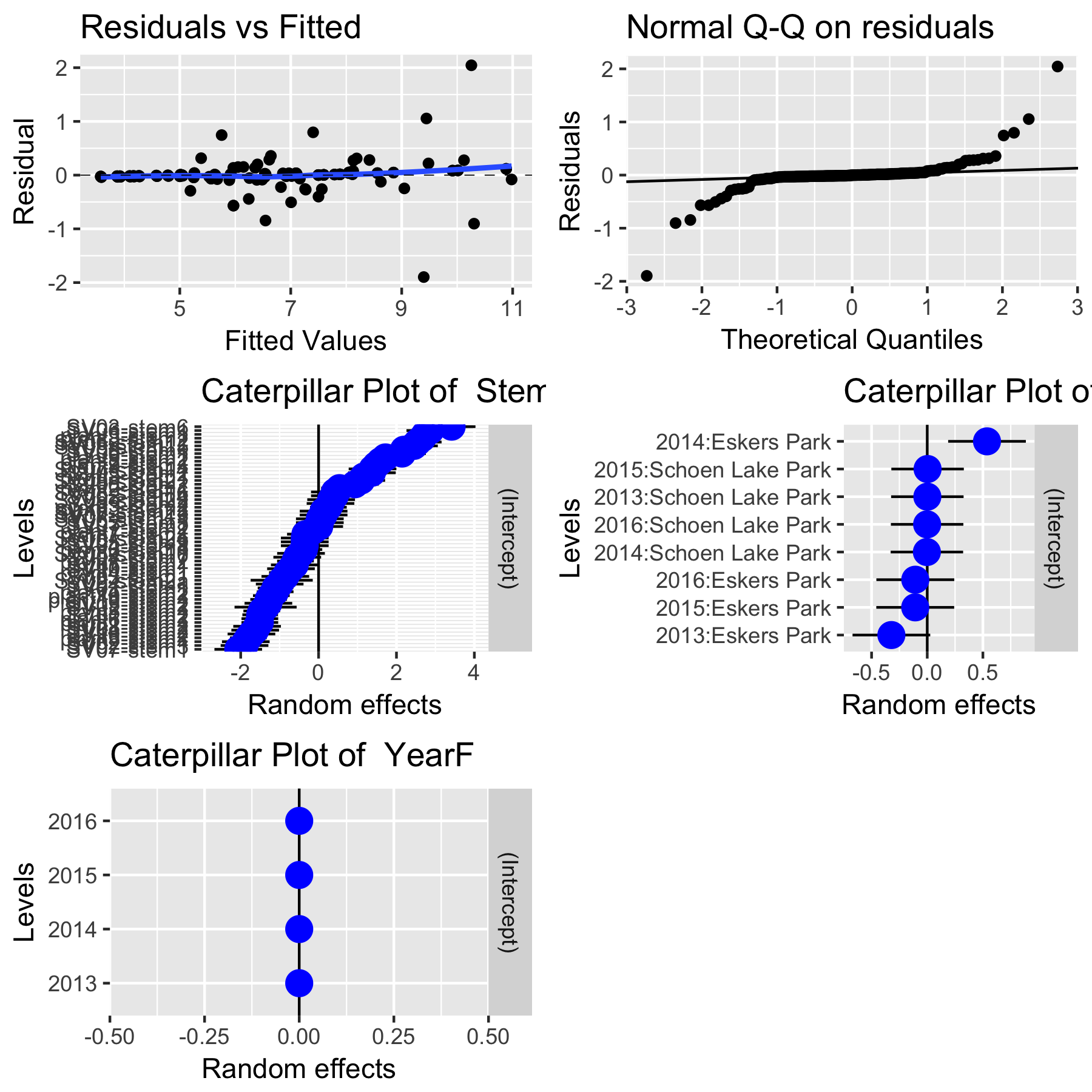


Figure 6. A sample diagnostic plot for the analysis of mean stem diameter at *Eskers Park* and *Schoen Lake* for the non-parallel slope model

With only a few years of data, the plots are not very informative. In the upper left corner is a plot of residuals vs. the fitted values. A good plot will show a random scatter around 0. Any large deviations from 0 should be investigated as potential outliers. In the upper right is a normal probability plot. Points should be close to the dashed reference line. Fortunately, the analysis is fairly robust against non-normality so only extreme departures are worrisome. The second and last row examine the distribution of the random effects. In this case, there year process error is very small (all of the dots are around 0).

A summary of the two fits is shown in Figures 7 and 8.



Figure 7. Summary plot of the trend in mean stem diameter at *Eskers Park* and *Schoen Lake* under the non-parallel slope model.

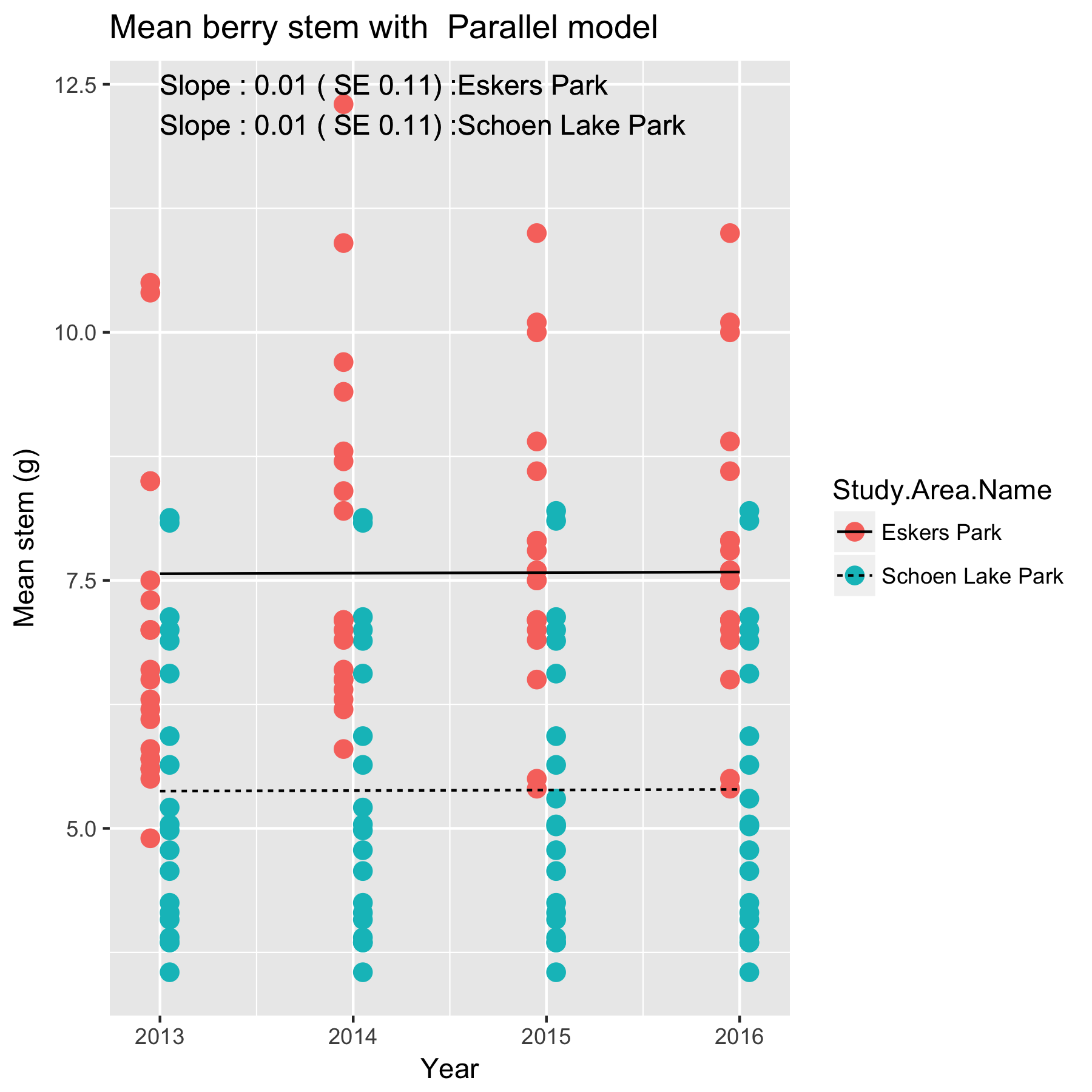


Figure 8. Summary plot of parallel slope model for the mean stem diameter at *Eskers Park* and Schoen Lake.

Which model is preferable? A model comparison of the two models is constructed using the *anova()* function in *R.* There was no evidence that the non-parallel slope model is needed (p=.92), and so the non-parallel slope model should be used. This indicates that there is evidence that the trend lines in the two plots have different slopes.

Whenever an analysis of a trend over time is conducted, the analysis will have to test and adjust for autocorrelation in the year-specific effect. This usually isn’t a problem unless there are 10+ years of data. There was no evidence of autocorrelation.

## 2.3 Berry Count.

This measurement is taken at the stem level and so there is one number per stem/plant/year. The same stem/plant is repeated measured over time. All of the models below automatically will account for branches that are removed or added as long as each stem has a unique label within a site.

The models for the berry count are similar to those from the stem diameter except that the log(count) was used as the response measure. The finally summary plots are shown in Figure 9 and 10. There was very strong evidence that the non-parallel slope was preferred (p<.001).

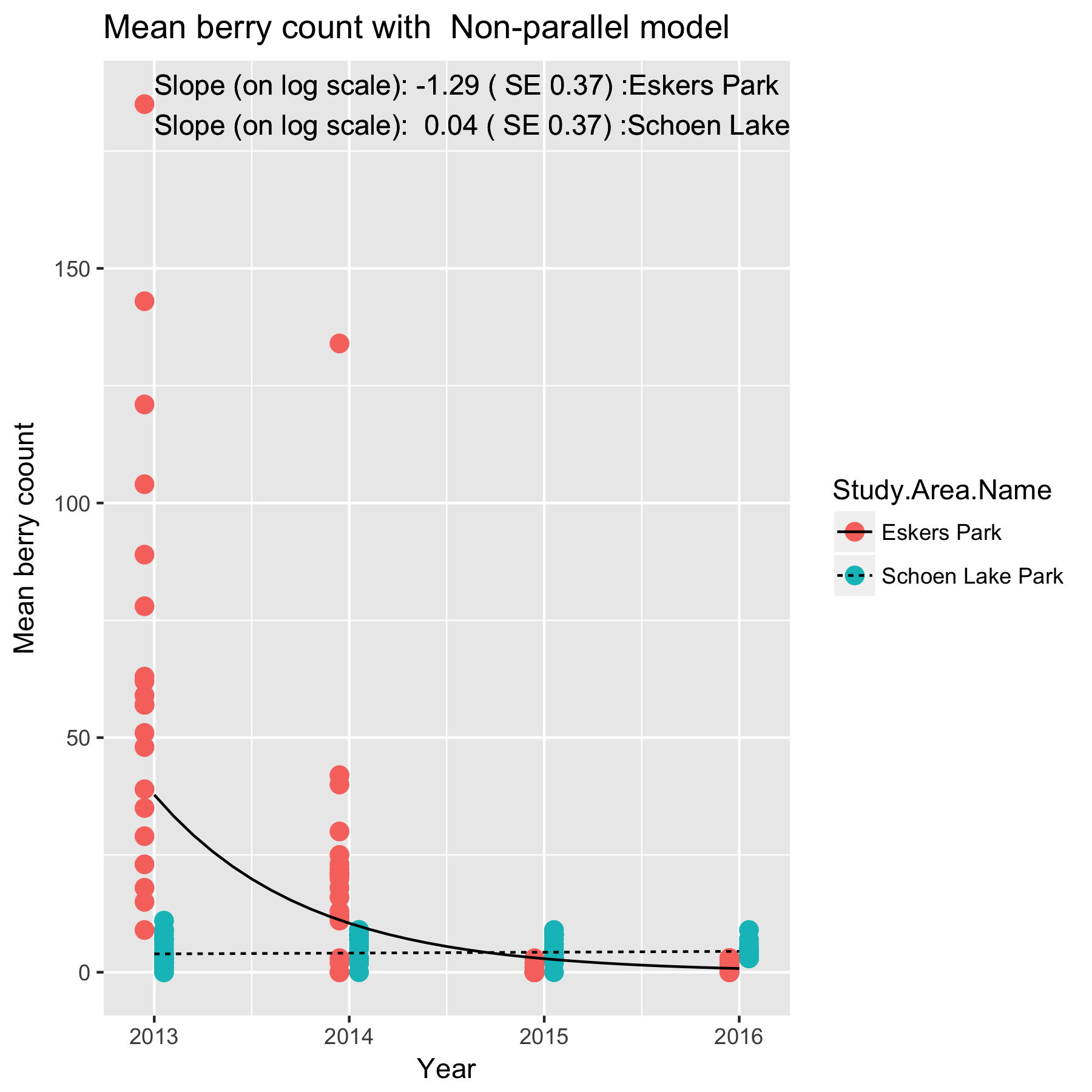


Figure 9. Summary plot of the trend in mean berry counts at *Eskers Park* and *Schoen Lake* under the non-parallel slope model. Because the model operates on the logarithmic scale, the fitted trend lines are not a straight line but curved in each case.

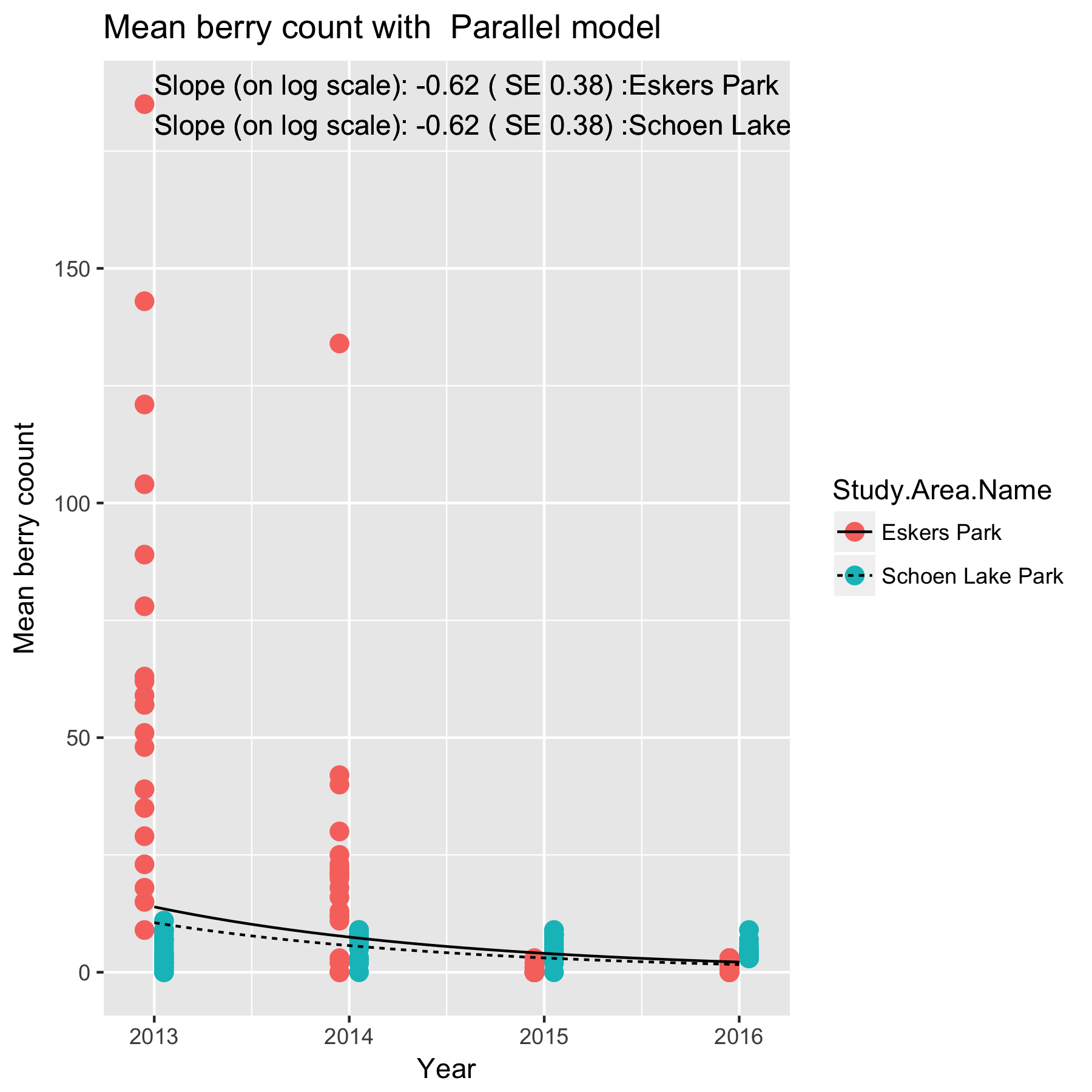


Figure 10. Summary plot of the trend in mean berry counts at *Eskers Park* and *Schoen Lake* under the parallel slope model. Because the model operates on the logarithmic scale, the fitted trend lines are not a straight line but curved in each case.

The slopes are interpreted in the same way as in the single site case. For example, the estimated slope under the parallel slope model is of -0.62 (SE 0.38) /year for both study areas on the logarithm of the mean response. This corresponds to an approximate exp(-.62)=0.54x multiplicative change/year, i.e. the mean count in 2014 is about 0.54x the mean count in 2013, and the mean count in 2015 is 0.54x the mean count in 2014. Because the analysis is done on the logarithmic scale, the fitted trend line looks non-linear on the original (non-transformed) scale.

# 5. Summary

Some caution is needed about the interpretation of the slope in these short time series. Here most of the stems/bushes were measured for at least two years, so the above trend lines may represent nothing more than normal growth in stems as the bushes mature or senescence in berry production and berry weight as stem age. As an analogy, consider measuring the heights of children as they grow up. If the same cohort of children is repeated measured, you would expect to see a positive trend for many years simply as an artifact of the maturation. It may turn out that stem diameter is a surrogate for the stem age which, in turn could be a predictor for the berry count but at the moment there is insufficient information to know if this happens.

One way to deal with this problem is to NOT repeatedly sample the same bushes/stems over time. By taking a new sample each year (hopefully randomly selected), the variation in age in the sample should be similar over time (assuming that the soap berry plants are in “steady state”).

For this reason, this analysis protocol should be revisited after about 10 years to see if a more appropriate analysis may be more suitable when more data are collected and there are many plants that have been measured for many years. At the moment, there appears to be sufficient turnover in the three years that this issue may be moot.

References

Kuznetsova, A., Brockhoff, P.B. and Christensen , R.H.B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-33. https://CRAN.R-project.org/package=lmerTest

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Appendix A.

Issues encountered when doing a trial analysis on the *Eskers* and *Schoen Lake* study areas data.

The following issues were encountered in the databases when a trial analysis on the Eskers study area data was performed.

(a) Inconsistent spelling of Study Area Name across worksbooks in *Eskers Lake*

(b) Inconsistent formatting of dates within and across workbooks. I suggest you always use yyyy-mm-dd as the format. The documentation in the individual workbooks is inconsistent.

(c) Species names is not consistent. The values of

SHEPCAN and Sheperdia Canadensis

are both used.

(d) Need to standardized the bush-stem labeling

In 2013/2014 the format is *SV01-stem1* but in 2015 it changed to *bush1-stem1*

I also want to make sure that in 2015 many new bushes were sampled/and or new

stems were sampled. Is this correct?

(e) Be consistent in labeling stems, especially when switching to a new stem. In 2014, *SV02-stem2* is present, but the other (fate) codes says the a new stem was used. This would then be reusing the same stem number on two different stems.

(f) Missing *fate* code field in 2013. Even though this is the first year of the study, the fate codes shuld be listed.

(g) Some berry counts are given as 0 even if the branch is heaving browsed. I manually replaced the berry count by NA (missing)

(h) The number of berries used in mean count is currently given in the comment field.

It would be better to make a separate field for this rather than having to parse this out because the comment fields are used for many other purposes.

This value is missing for the 2015 data.

(i) I noticed that UTM zone, Easting and northing is not always given in the 2015 dataset.

(j) The berry count for weight for Schoen Lake 2013 is wrong as it is increasing from 20 to 39 by 1. This looks like an Excel expansion area.