# Analysis of data collected under the Squirrel Protocol

2017-03-09

# 1. Summary of protocol:

## 1.1 Basic protocol

As taken from the protocol document:

“Red squirrels regularly emit an audible rattle, especially when their territories are invaded. This protocol involves walking a transect (a section of a trail) and recording the location of rattles heard along the way.

Locate as many transects in a given area as possible (up to 5). Sample them annually but in a different order each year. Sampling involves walking a defined segment of trail and recording squirrel rattles or chattering.

The data collected under this protocol consists of the.

* **Detection Type**. Was a call heard, a visual observation made or other evidence of a squirrel. Only calls will be used in this analysis
* **Distance along the transect**. The location of the observer along the transect when the detection was made.
* **Distance from the transect**. The perpendicular distance from the transect line to the observation along with the side of the transect.

## 1.2 Cautions about the protocol.

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

If no calls were heard on a transect, how is this indicated? For example, in the 2013 data table, there is NO observation for transect 4 on 2013-09-13. Does this mean that the transect was not run, or was it run and no squirrels were detected? The *Transect Information* sheet in the workbook has the GPS co-ordinates of each transect, but does not indicate if the transect was run on each of the days. It will be assume that ALL transects listed on the *Transect Information* worksheet are run on every date listed in the *General Survey* worksheet and if no information is present in the *General Survey* worksheet, then a 0 is imputed.

If would be preferable to create another worksheet indicating which transects were run on each sampling date.

## 1.2.2 No information on transect length available.

There is no information about the length of each transect or if each transect was visited on its entirety each visit. It will be assumed that each transect is approximately equal in length and that the entire transect is visited on a visit.

### 1.2.3 Not suitable for distance sampling to estimate density.

The current data looks very similar to captured by distance sampling methods where RANDOM transects are selected in the study area and the perpendicular distance of observations to the transect are selected. Distance sampling is used to estimate density.

In this protocol, transects are not selected at random. Indeed, according to the protocol,

“Find a location that is not difficult to access and where there are abundant squirrels.”

So the apparent density of squirrels may be biased upwards by the selection of transects.

### 1.2.4 Be careful to document changes in transect over time.

The protocol is silent on how to document changes in transects over time. For example, suppose that a transect is damaged by fire? How is this recorded? Suppose that a transect is abandoned and new transect is chosen. At the very least, the transect label should NOT be recycled over time.

### 1.2.5 Not clear how to group visits.

The date that the transects are visited is also recorded. It is assumed that all transects will be visited on the same date. However, in some cases, the transects are visited over a span of 2 or 3 day – these should presumably be “pooled” into one visit. At the moment, there is no way to decide if all transects were visited on a single day, or if a “visit” corresponds to more than one day. A field should be added to the data base for the “visit”, e.g. if it takes several days to visit all transects, these should be either be recorded on the first date, or all take the same visit indicator.

## 2. Database structure

The database for this protocol is a series of Excel workbooks with multiple sheets in each workbook. The *Transect Information* sheet contains the information on the transects available for this year. It is implicitly assumed that every transect is visited on every date. The *General Survey* sheet contains the information collected. There are multiple lines per transect.

The relevant fields on the *Transect Information* worksheet are:

* *Transect Label.*

The relevant fields on the *General Survey* worksheet are:

* *Transect Label.*
* *Date*. The date the data was collected. The *Year* is extracted from this date.
* *Detect Type*. What type of detection was made. Only calls are of interest. A value of zeros will be imputed for the total number of calls heard on the transect if there is no information on a transect on a particular date.

# 3. Sample Analyses – Single Site.

A sample analysis is presented on the *Purcell Wilderness Conservancy 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 Calls.

The data is first summarized to the transect-year level by finding the mean number of calls on a transect over multiple visitss for each individual transect. This reduces the data to one measurement per transect per site/year. It is implicitly assumed that all transects are run on all days within a so every transect has the same number of days of measurement. If transects are changed over time, that is not a problem, but transects should not be introduced or removed part way through a year.

A summary plot of the mean number of calls on each transect is shown in Figure 1.

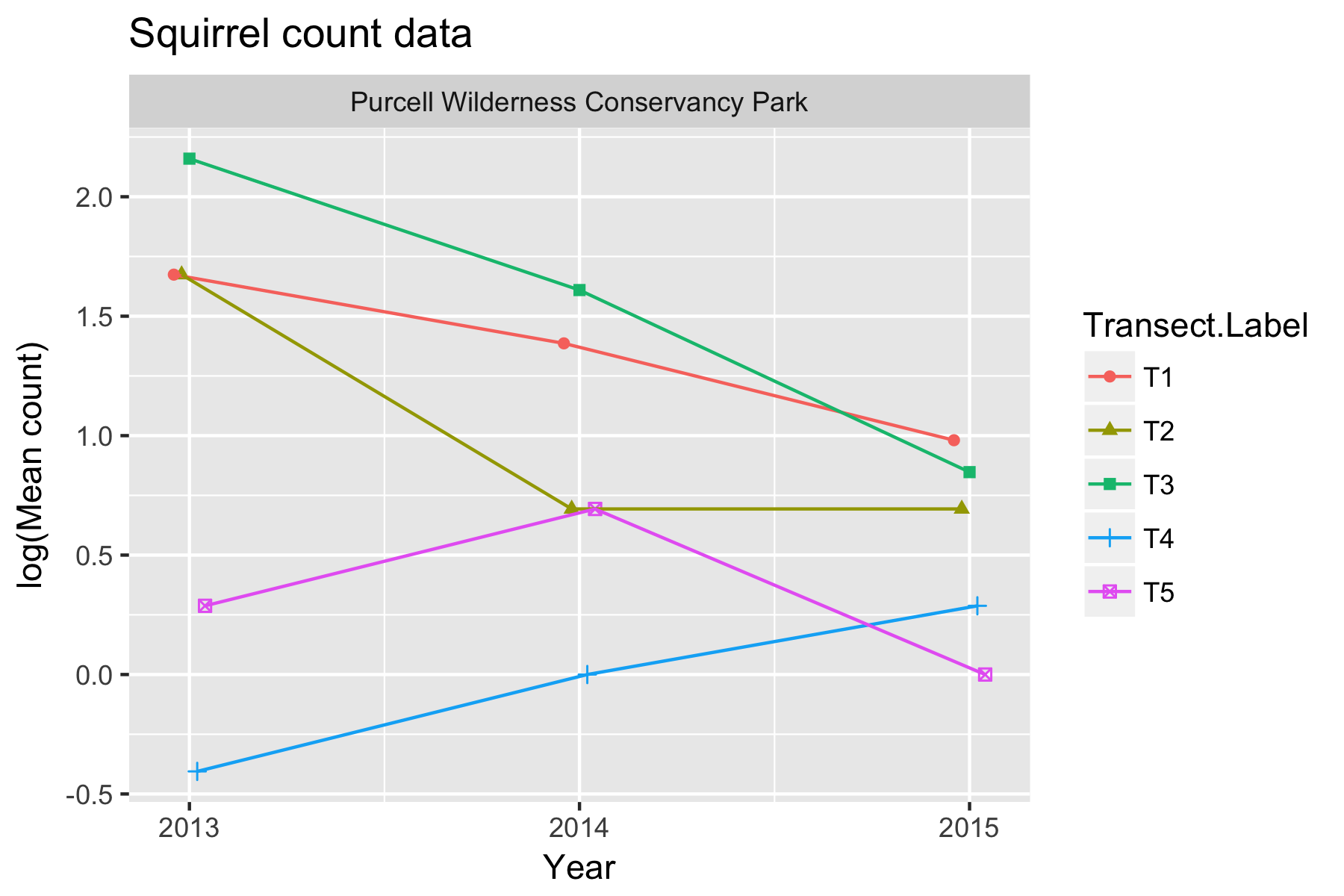


Figure 1. Summary plot of the data. Because the typical counts are small, the data is analyzed on the logarithmic scale.

There is evidence of a transect effect, where, for example, the number of calls at transect T3 generally is higher than at the other transects because of local transect-specific conditions (e.g. better habitat).

Because this is count data, so a linear mixed model is fit to the logarithm of the mean calls per transect. The model is:



where *log(AvgCalls)* is logarithm of the average number of calls for that transect in that year; *TransectF* represents the transect effect; *YearF* represents the year-specific effects (process error), and *Year* represents the calendar year trend over time. The *TransectF* term allows for the fact that transect-specific conditions may tend to affect the counts on this transect consistently over time. The *YearF* term represent the year-specific effects (process error) caused by environmental factors (e.g. a warmer than normal year may elict more calls from squirrels).

Model fit on the logarithmic scale assume that effects are multiplicative over time, so that the when the actual fit is done on the logarithmic scale, the trends are linear. For example, a trend may assume that there is constant 5% change over time rather than a fixed 1-unit change per year. Some caution is needed if any of the values are 0 as log(0) is not defined. In these cases, a small constant (typically ½ of the smallest positive value in the dataset) is added to all values before the analysis proceeds.

The model was fit using the *lmer()* function in *R.* Figure 2 shows asummary plot, along with estimates of the slope, its standard error, and the p-value of the hypothesis of no trend. There is very weak evidence (p=0.08) of a trend with an estimated slope of -0.26 (SE 0.08) /year in the logarithm of the mean response. This corresponds to an approximate exp(-0.26)=0.37x multiplicative change/year, i.e. the mean count in 2014 is about 0.37x the mean count in 2013, and the mean count in 2015 is 0.37x 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.

The reason why the p-value is so large given that the standard errors are small relative to the estimated slope is because there are only 3 years of data and after fitting the line there is only 1 degree of freedom available to estimate the multiplier for the confidence intervals. With 1 degree of freedom, the multiplier is much larger than the usual value of 2 used to convert standard errors to confidence intevals.

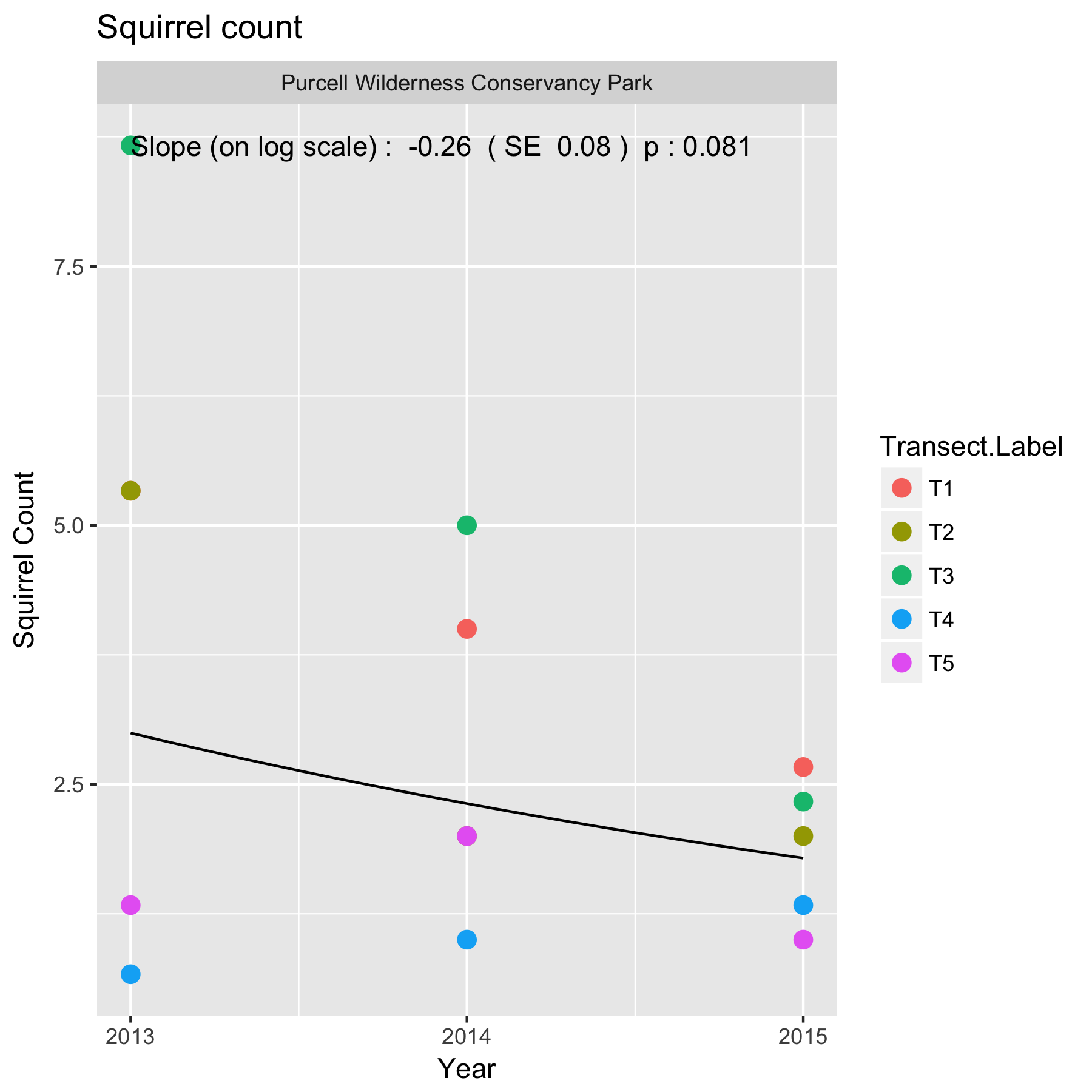


Figure 2. Summary plot of the trend in mean squirrel counts at PWC. Because Poisson regression operates on the logarithmic scale, the fitted trend line is not a straight line but curved.

Following the fit, the diagnostic plots should be examined. An illustration of such a plot is shown in Figure 3.

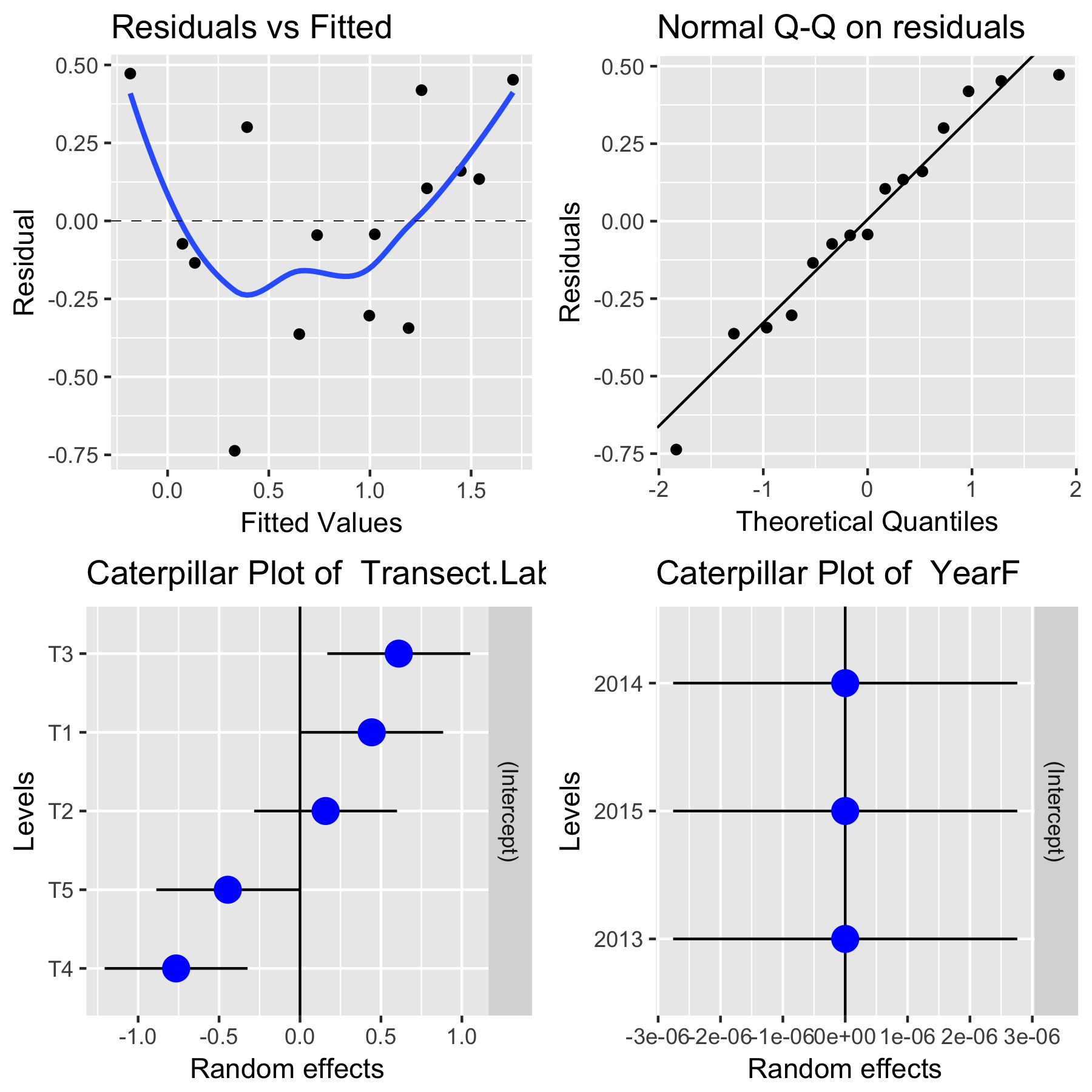


Figure 3. A sample diagnostic plot for the analysis of squirrel calls at PWC

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 of the residual. Points should be close to the dashed reference line. Fortunately, the analysis is fairly robust against non-normality of the residuals (and in fact makes no assumption of normality) so only extreme departures are worrisome. The bottom left plot shows the distribution of the transect effects. The bottom right plot shows the distribution of the year-specific effects (process variation). In this case, the estimated process variation is very small with most of points very close to 0.

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. There was no evidence of autocorrelation – with only three years of data, this test is not very powerful.

This model used the approximate analysis on the logarithm of the average counts per transect. It is possible to analyze the actual raw counts using a generalized linear mixed model – this was not done in this example because of the extreme smallness of the dataset. Once many more years are collected, this may be an alternative analysis that will more naturally deal with 0 counts without having to add a small constant.

# 4. Sample Analysis – Multiple Sites

Now we are interested in comparing the trends in two or more sites. A sample analysis is presented comparing the trends in the *Purcell Wilderness Conservancy Park* and the *Eskers* study area. Data is available from 2013 to 2015 for *PWC* and from 2013 to 2014 for *Eskers*. Usually, at least 3 and preferably 5 years of data are needed for sensible comparison of trend. I created a simulated dataset for Eskers for 2015 so the results below are illustrative only.

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.

## 4.1 Calls.

The data is first summarized to the year level for each transect by finding the mean number of calls over multiple dates in each study area. This reduces the data to one measurement per transect per study area per year. It is implicitly assumed that all transects are run on all days within a so every transect has the same number of days of measurement. If transects are changed over time, that is not a problem, but transects should not be introduced or removed part way through a year.

A summary plot of the mean number of calls is shown in Figure 4.

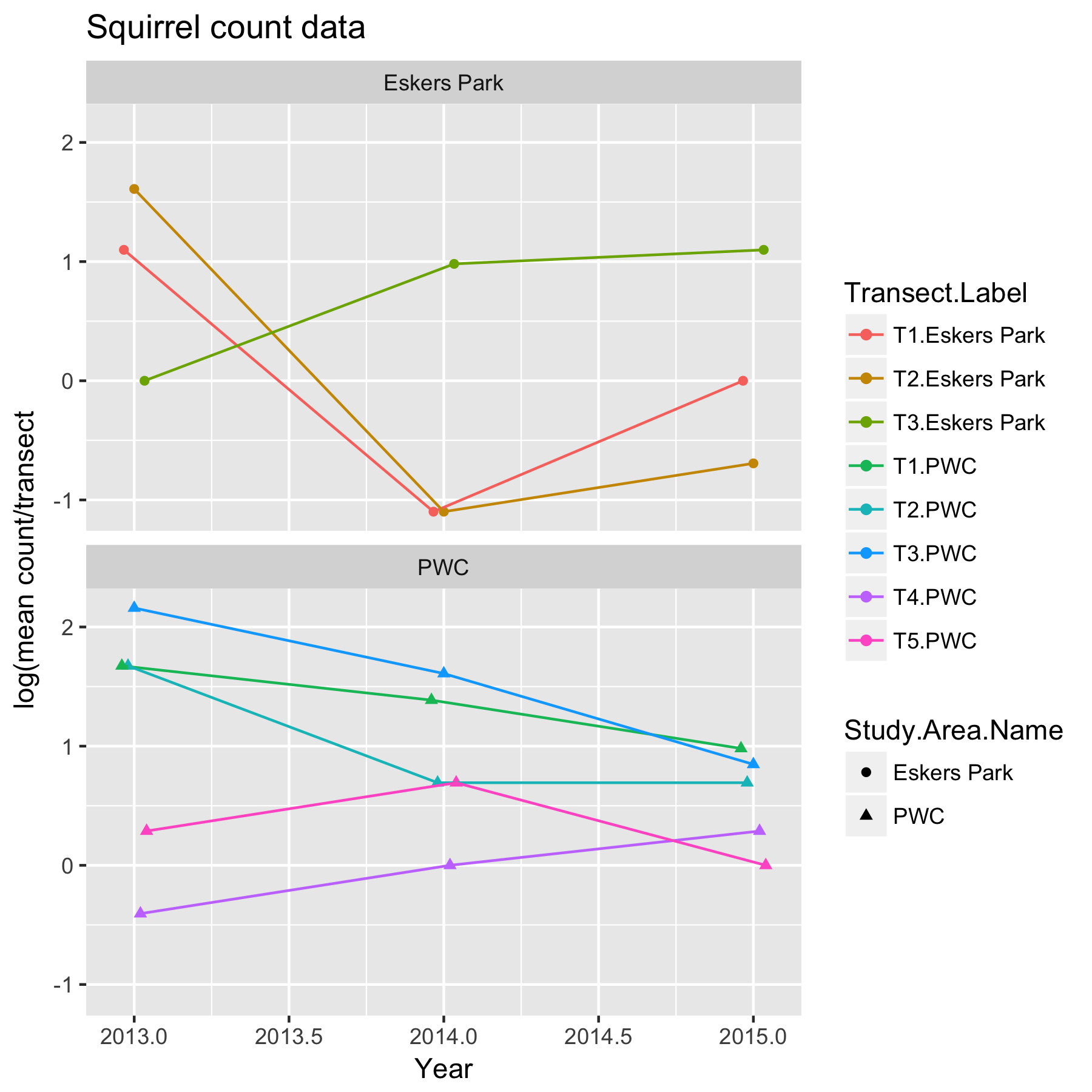


Figure 1. Summary plot of the squirrel calling data.

Notice that there is a definite transect effect in *PWC*, where, for example, the number of calls at transect T3 generally is higher than at the other transects because of local transect-specific conditions (e.g. better habitat). It is not clear if such a transect effect exists for *Eskers*.

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 *log(MeanCalls)* is the logarithm of the mean calls for that transect in that year; *TransectF* represents the transect effect; *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; and *Year:StudyArea* represents the differential trend in the two study areas. The *TransectF* term allows for the fact that transect-specific conditions may tend to affect the counts on this transect consistently over time. The *YearF* term allows for year-specific effects (process error) that affect counts in all study areas simultaneously. The *YearF:StudyArea* term allows the process error to be different for each study area.

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.

These models assume that effects are multiplicative over time, so that the actual fit is done on the logarithmic scale. For example, a trend may assume that there is constant 5% change over time rather than a fixed 1-unit change per year. Some care is needed if some of the calling values are 0 (log(0) is not defined).

Both models can be fit using the *lmer()* function in *R.* Figures 5 and 6 shows asummary plot, along with estimates of the slope, its standard error for each of the models.

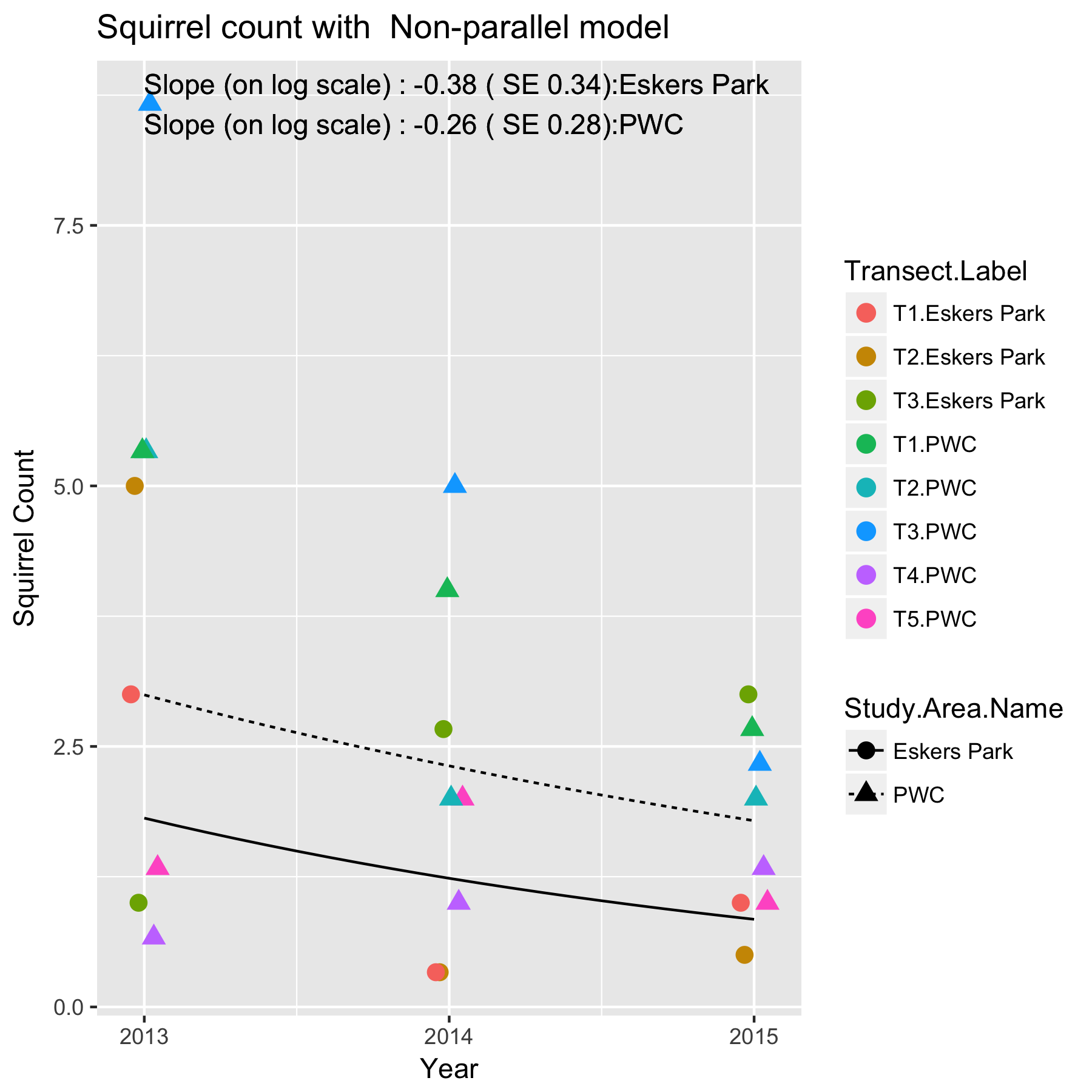


Figure 5. Summary plot of the trend in mean squirrel counts at PWC and Eskers under the non-parallel slope model. Because the model operates on the logarithmic scale, the fitted trend lines are not a straight lines but curved in each case.

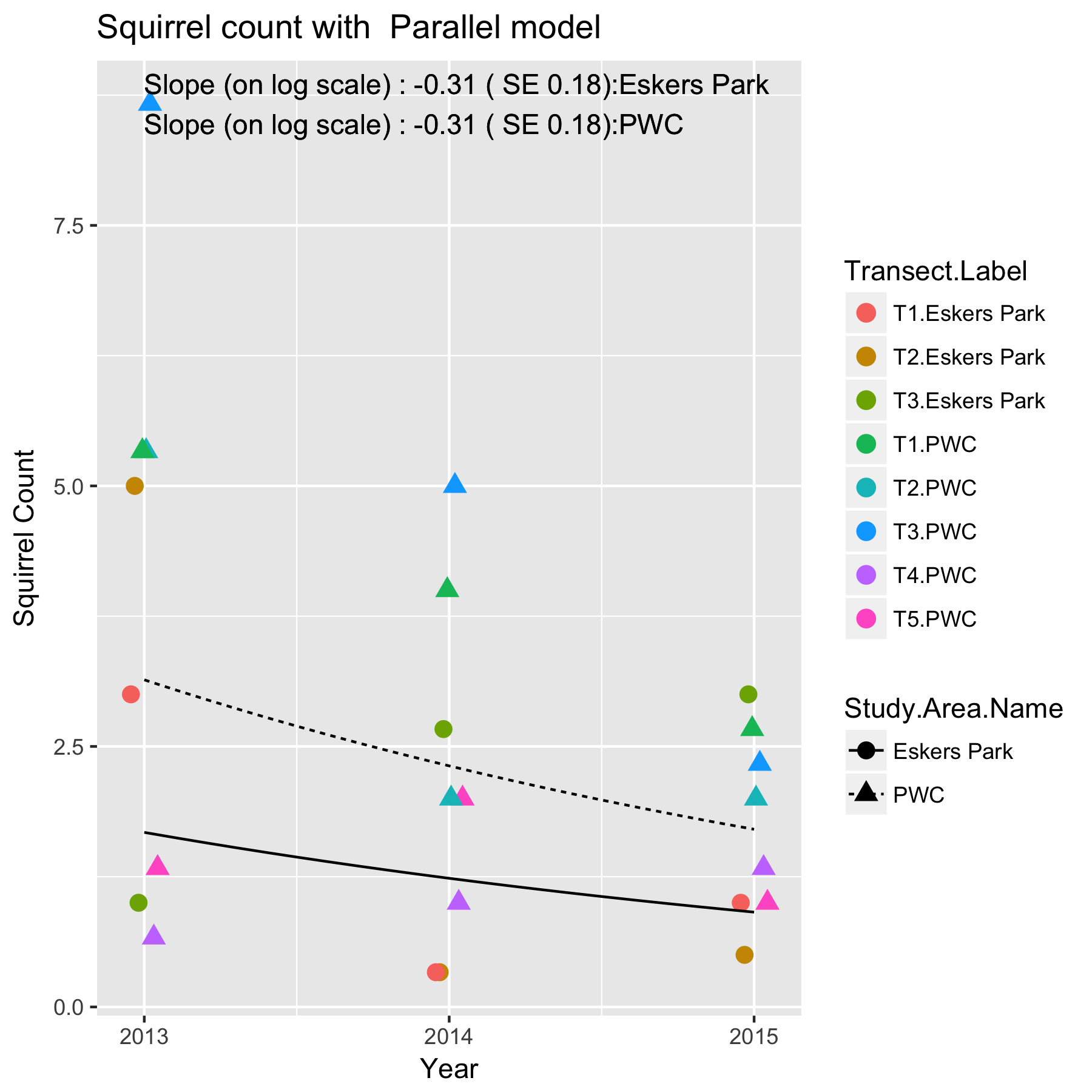


Figure 6. Summary plot of the trend in mean squirrel counts at PWC and Eskers under the parallel slope model. Because the mode operates on the logarithmic scale, the fitted trend line is not a straight line but curved even though the slopes on the logarithmic scale are parallel.

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 -0.31 (SE 0.18) /year for both study areas on the logarithm of the mean response. This corresponds to an approximate exp(-.31)=0.73x multiplicative change/year, i.e. the mean count in 2014 is about 0.73x the mean count in 2013, and the mean count in 2015 is 0.73x 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.

Following the fit, the diagnostic plots should be examined. An illustration of such a plot for the non-parallel slope model is shown in Figure 6.

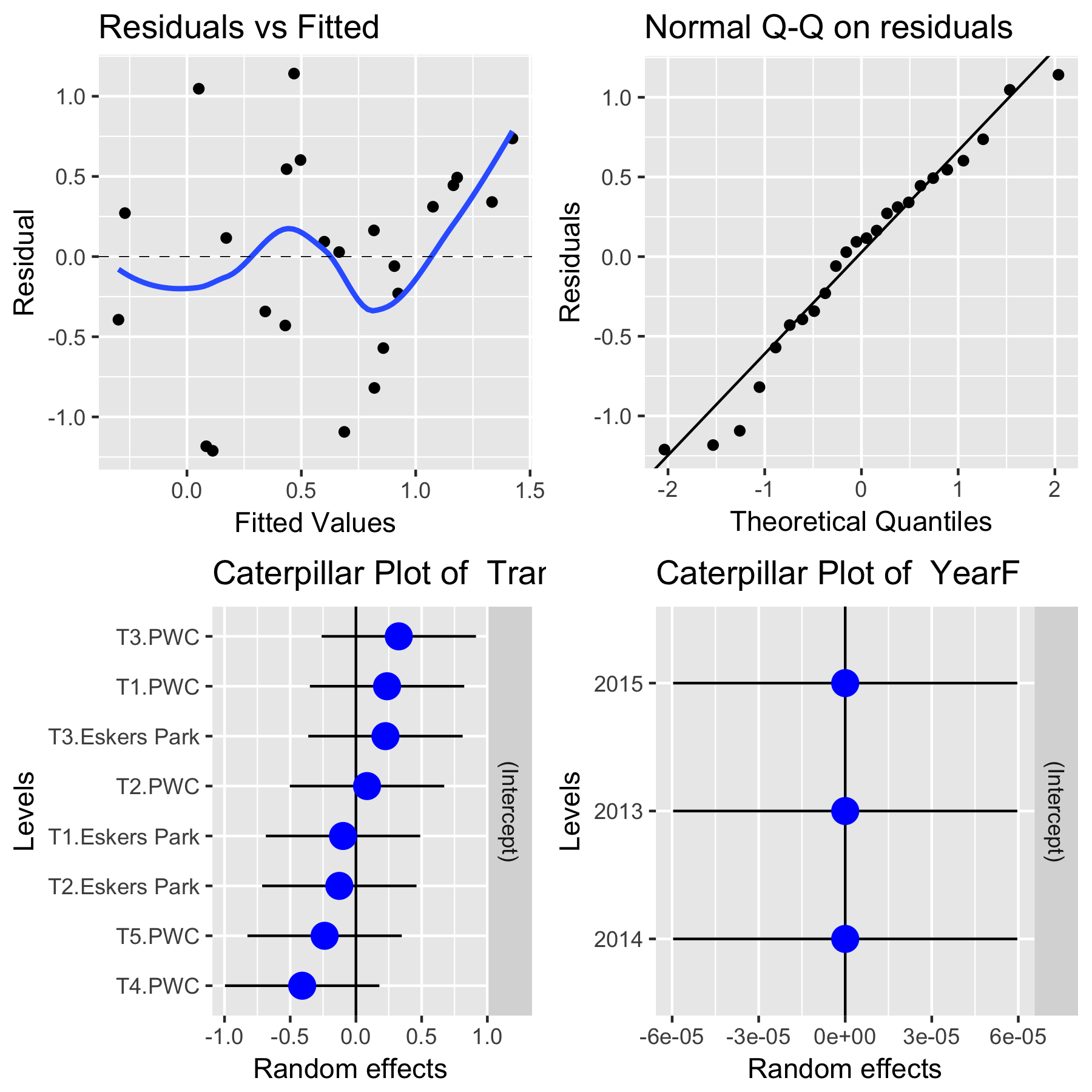


Figure 6. A sample diagnostic plot for the analysis of squirrel calls at PWC and Eskers using the non-parallel trend line model.

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 of the residual. Points should be close to the dashed reference line. Fortunately, the analysis is fairly robust against non-normality of the residuals (and in fact makes no assumption of normality) so only extreme departures are worrisome. The bottom left plot examines the transect effects while the bottom right plot examines the year-specific effects. In this case, the year-specific effects are very small.

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=.73), and so the parallel slope model should be used.

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

Eskers Park -0.38 0.31 -1.04 0.28 1 Non-parallel

PWC -0.25 0.24 -0.77 0.25 1 Non-parallel

Results are averaged over the levels of: Transect.LabelF

Trends are based on the log (transformed) scale

Confidence level used: 0.95

The individual slopes (on the logarithmic scale) 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. [Note that the grouping variables are all single digits; a grouping values of ‘12’ means that this particular study area’s slope belongs to two different groupings of the slopes.] In this case, all of the study areas slopes belong to the same “group”, so there is no evidence of a different in their slopes, i.e. the 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 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. There was no evidence of autocorrelation.

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.

# 5. Summary

Some caution is required to ensure that all transects are run the same number of times in a year. In this balanced design, it is straightforward to simply sum over all measurements of transect in a year and all transects in a year have the same number of visit. It is possible to modify the analysis is only some transects are visited on a particular date with an unequal number of visits to a transect in a year. A simple way to deal with unbalance would be to delete some of the observations, but better methods are available.

References

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 PWC and Eskers study area data.

The following issues were encountered in the databases when a trial analysis on the PWC and Eskers study areas’ data was performed. The spreadsheets for the sample analysis were corrected prior to the analysis.

(a) Inconsistency in transect labels across years and in workbooks.

In 2013, the transects in PWC are labeled as 1, 2, 3, 4, and 5. In other years, the transects in PWC are labeled as T1, T2, T3, T4 and T4.

In 2015, the transects in PWC are labeled as 1, 2, 3, 4, 5 in the *General Survey* sheet, but T1, T2, T3, T4, T5 in the *Transect Information* sheet.

(b) Date formatting. I suggest you always use yyyy-mm-dd as the format for ALL protocols. The documentation in the individual workbooks is inconsistent. This set of workbooks for PWC currently use *dd-mmm-yy* format (with 2 digit years), the set of workbooks for Eskers currently uses dd-mm-yyyy format (with 4 digit years).

(c) The Excel workbook for 2014 for PWC has two rows at the end of the spreadsheet where at least one column has blank (rather than being empty). This causes the programs to insert a row with missing values when reading the data.

(d) Check to see when transects were run. As noted earlier, we assume that ALL transects are run on all days. In 2015, all five transects in PWC were run o 2015-09-03. On 2015-09-04 only T1, T2, and T3 have data – it is assumed that T4 and T4 on these dates are zero. But on 2015-09-05, there is single record for T3 and several for T4. We assume that T1, T2, and T5 were run on these dates and are missing. I suspect that what actually happened is that the 2015-09-04 and 2015-09-05 data should be “combined” into a single “run” on all the transects with only T5 having a 0 count on that “combined” date.