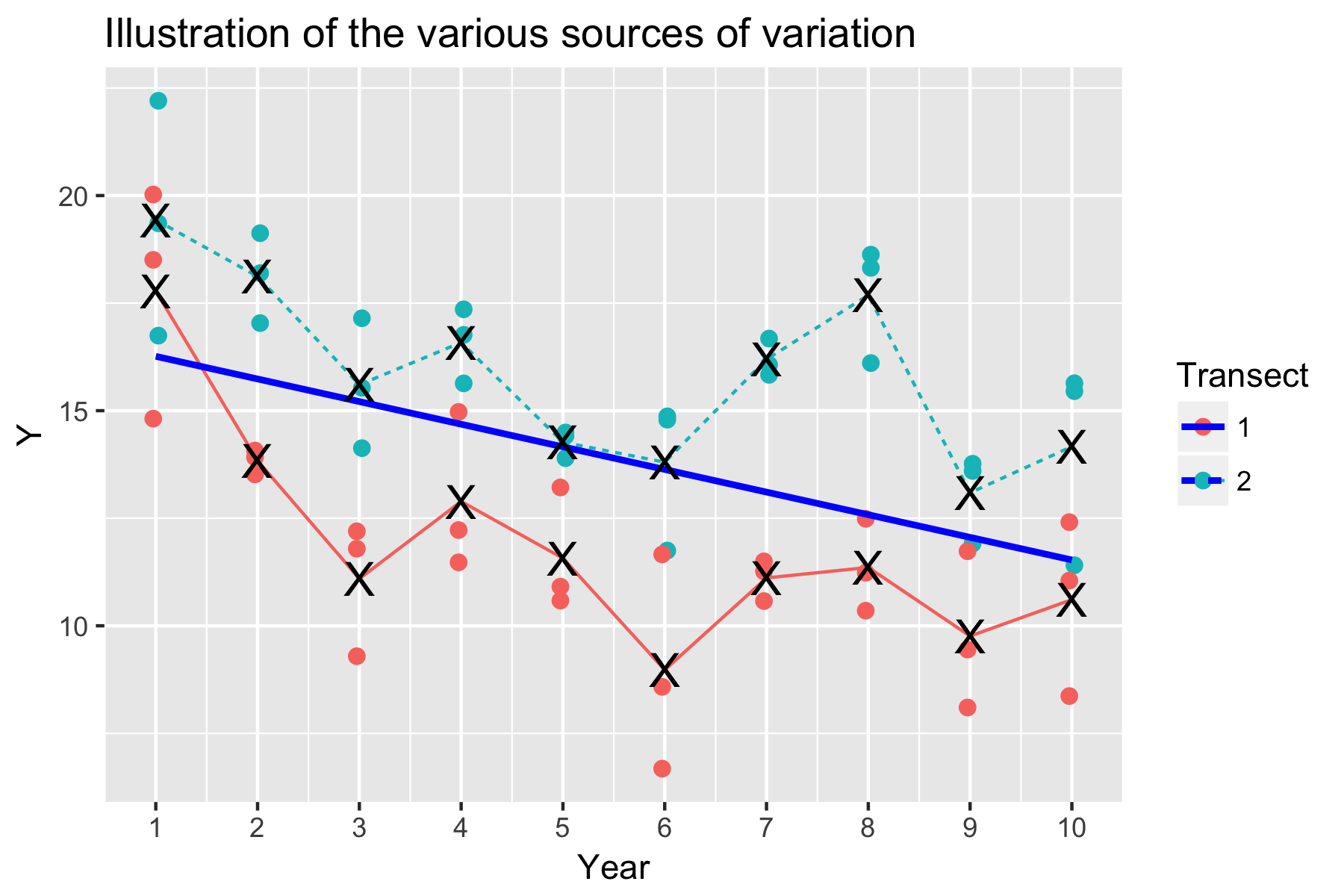
# Fitting Trends with Complex Study Designs.

# 1. Introducton

Many of the designs in the LTEM has a common structure of a site measured over time with fixed transects (or stations) repeated measured in multiple years. Multiple measurments can also be taken on each transect. For example, in the vegetation protocol, two transects are established on the site and 10 fixed plots are measured on each transect repeated over time. In the berry protocol, bushes are tagged, and multiple stems on each bush are repeated measured over time.

This can be represented diagrammatically as:



There are several sources of variation that need to be accounted for when doing the trend analysis:

* Measurement error. This is represented by the variation of the same color dots around the transect mean (denoted by X) for any particular year.
* Transect variation. This is represented by the fact that transect 1 is generally higher than transect 2 in all years. The cause of this is typically unknown and due to transect-specific effects such as local conditions at the transect.
* Process variation (year-specific variation). This is represented by the fact that the transect means (the X’s) tend to move up or down in any year in tandem. For example, in year 6 both transect means are depressed, while in year 4 both transect means are elevated. The cause of process error (year-specific variation) is unknown, but typically caused by external factors that affect the entire site (and all transects) simultaneously.

Year-specific effects cause a very specific violation of the usual regression assumptions. Usually, in a regression setting repeated values taken in the same year should be approximately scattered about the trend line. However, in some case, year specific factors (e.g. a much wetter year or drier year) will force data points to be “clumped” above or below the trend line as illustrated below.

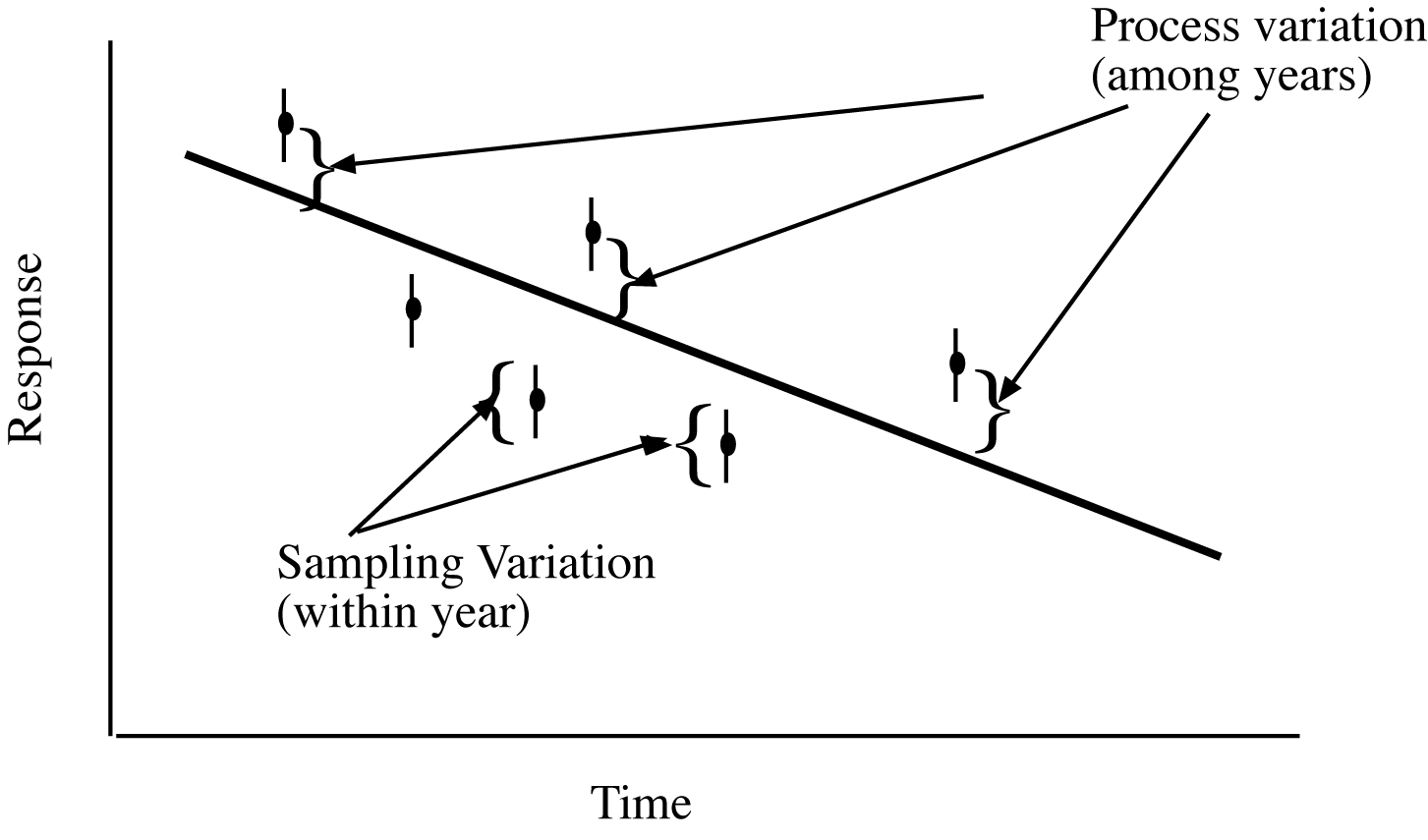


Figure 3. Conceptualization of process error. Year-specific factors force the individual data points above or below the trend line rather than being centered about the trend line as assumed by regression.

# 2. Accounting for Transect and Year-specific effects

A naïve analysis would fit a simple regression through all 10 x 2 x 3 = 60 data points, but this type of data violates a key assumption of regression analysis. A key assumption in regression analysis is that all data points in a year are centered about the trend line as shown below:



Now there is no consistent pattern of one transect tending to always be higher/lower than the other transect, nor is there any evidence of a year-specific effect (process error) where by points are uniformly shifted upwards or downwards.

Ignoring year-specific effects (or process error) is a common error in fitting trends over time. If a naïve fit is done to the data in Figure 1, the following results are obtained from using the *lm()* function in R

naive.fit <- lm(Y ~ Year, data=sim.data)

it gives the following (incorrect) results

Analysis of Variance Table

Response: Y

Df Sum Sq Mean Sq F value Pr(>F)

Year 1 136.90 136.900 16.006 0.0001813 \*\*\*

Residuals 58 496.07 8.553

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.7890 0.8156 20.584 < 2e-16 \*\*\*

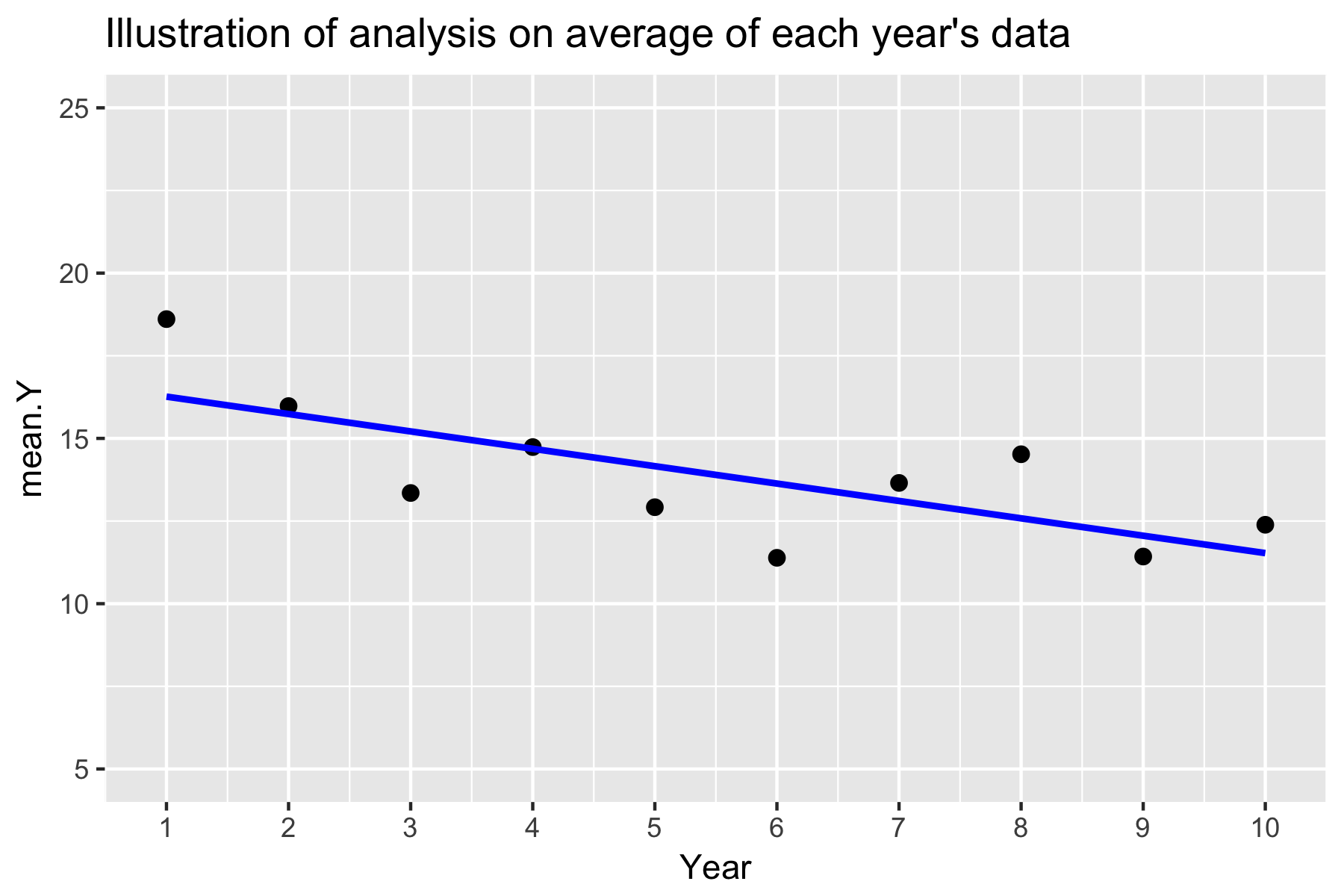
Year -0.5259 0.1314 -4.001 0.000181 \*\*\*

Residual standard error: 2.925 on 58 degrees of freedom

The estimated slope is -0.52 (SE .13, p=.0002). The naïve analyst believes that there is very strong evidence that the slope is negative.

There are two ways to properly analyze such data.

First, if the data are perfectly balanced (i.e. every transect measured every year, and every transect measured the same number of times in each year), then a simple analysis simply analyzes the average reading per year, i.e. average all of the data in a year and do the regression simply on the averaged points.



This gives:

avg.fit <- lm(mean.Y ~ Year, data=sim.data.avg)

Analysis of Variance Table

Response: mean.Y

Df Sum Sq Mean Sq F value Pr(>F)

Year 1 22.817 22.8167 8.7751 0.01808 \*

Residuals 8 20.801 2.6001

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.7890 1.1015 15.241 3.4e-07 \*\*\*

Year -0.5259 0.1775 -2.962 0.0181 \*

Residual standard error: 1.612 on 8 degrees of freedom

The estimated slope is the same as in the naïve fit, but the (correct) standard error is much larger (0.17 rather than 0.13 and the p-value for the slope is not as extreme (.018 rather than .0002).

At first glance, it appears that we have “thrown away” much information. In fact we have not. A key determinant for determining a trend is the number of years of sampling – not how much sampling is done in each year. Here the number of years of sampling is 10 and the slope is determined by the contrast in the MEAN Y value across these 10 values. The variation within a year provides no information on the trend – rather it provides information on transect-to-transect variation (and if it is consistent over time) and measurement error variation which also provides no information on the trend.

The variation in the mean response above is also much less than the variation in the individual data points around the regression line. Hence, averaging does toss out information, but summarizes it in a form most appropriate for a trend analysis. The variation of the mean response around the regression line is a combination of process error, average transect variation, and average measurement variation and captures all three sources of variation.

So if the design is balanced, a simple analysis simply computes the mean response in each year and performs a regression analysis on the mean response. There is no loss of information.

But, it is unlikely that the LTEM will have perfectly balanced data after many years of monitoring. For example, in the berry protocol, plants will be replaced; in the vegetation protocol, some plots may be removed because of damage to the plot.

In these cases, a more complex analysis method is needed, the linear mixed model. A linear mixed model consists of two parts. First, is the structural portion, what is the proposed trend – in this case, it is a trend over time. Second, are the random effects for the sources of variation (identified earlier) being an effect for the transect and an effect for the year. Notice that the effect of year occurs in both the structural and random parts – consequently, most statistical packagesrequire a second copy of the *year* variable to be made and defined as a factor to serve as the random effect of year. The random effect of measurement error is implicit.

The *lmerTest* package in *R* can be used to fit linear mixed models as follows:

re.fit <- lmerTest::lmer(Y~ Year +

(1|TransectF) + (1|YearF), data=sim.data)

Analysis of Variance Table of type III with Satterthwaite

approximation for degrees of freedom

Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)

Year 23.591 23.591 1 8 8.7752 0.01808 \*

Estimate Std. Error df t value Pr(>|t|)

(Intercept) 16.7890 2.2717 1.6520 7.390 0.0293 \*

Year -0.5259 0.1775 8.0000 -2.962 0.0181 \*

Variance Components

Groups Name Std.Dev.

YearF (Intercept) 1.4670

TransectF (Intercept) 2.8098

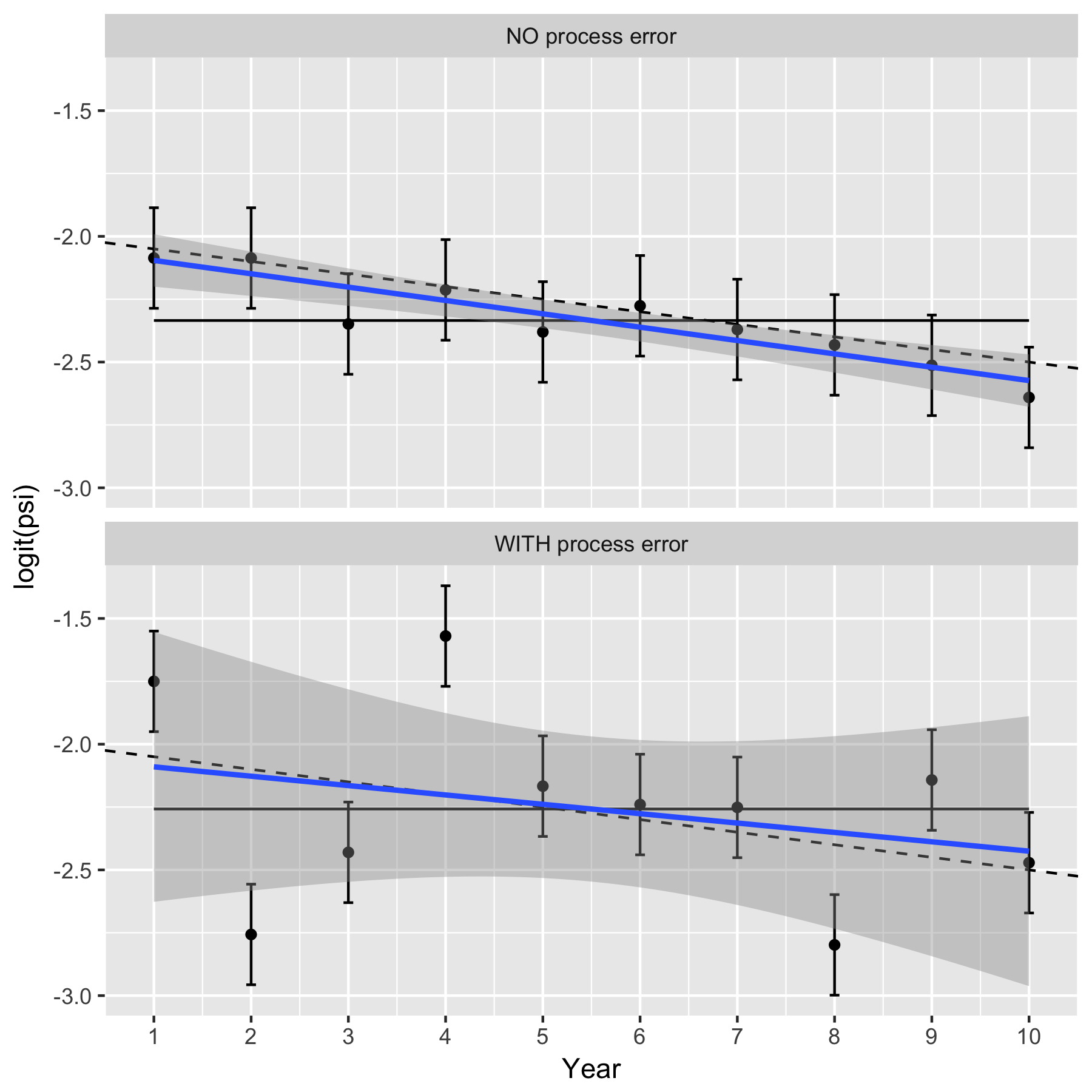
Residual 1.6396

The structural portion of the model is specified as *Y ~ Year* as before. The random effects are specified as (1|TransectF) + (1|YearF)where the *TransectF* and *YearF* variables are defined as factors. The model will account automatically for unbalanced data as long as the transect labels are unique for each transect and not reused, i.e. if transect 1 is replaced, do NOT call the replacement transect as *transect 1.* Instead, refer to it as *transect* 3(assuming that two transects were originally planned).

The slope under this more complex model matches that from the simple analysis on the mean response each year. It will differ if the design is unbalanced, but unless the imbalance is severe, the differences will be slight.

It turns out that fitting linear mixed models is quite difficult for small sample sizes and the *lmerTest*() approach may fail to converge and provide estimates of the trend with standard errors/p-values. In these cases, better starting values for the process may be required. It may be simpler to just do the approximate analysis on the averages.

Process error (year specific effect) also impacts the power to detect effects. Consider the following graph where there is or is not process error present.



In the top panel, there are no year-specific effects (no process error) and the average value of Y changes over time based on the underlying trend line shown as a dashed line. Only sampling variation (measurement error) is present so 95% of the confidence intervals computed each year based on the yearly measurements each year will overlap with the underlying trend line. The fitted line will be close to the true underlying trend line (solid line). The uncertainty is small about the overall trend, and there is clear evidence that the trend is different from a 0 trend (which is shown by the horizontal line).

In the bottom panel, the year-specific effects (process error) add extra variation to the underlying response each year due to effects such as weather, food etc. Now, the 95% confidence intervals for mean response still provide valid estimates for the yearly mean response values, but now may not overlap the true underling trend (in dashed). The fitted line will still be unbiased for the true trend (solid line), but the extra variation makes the uncertainty in the fitted line much larger and now there is no evidence that the trend line differs from 0.

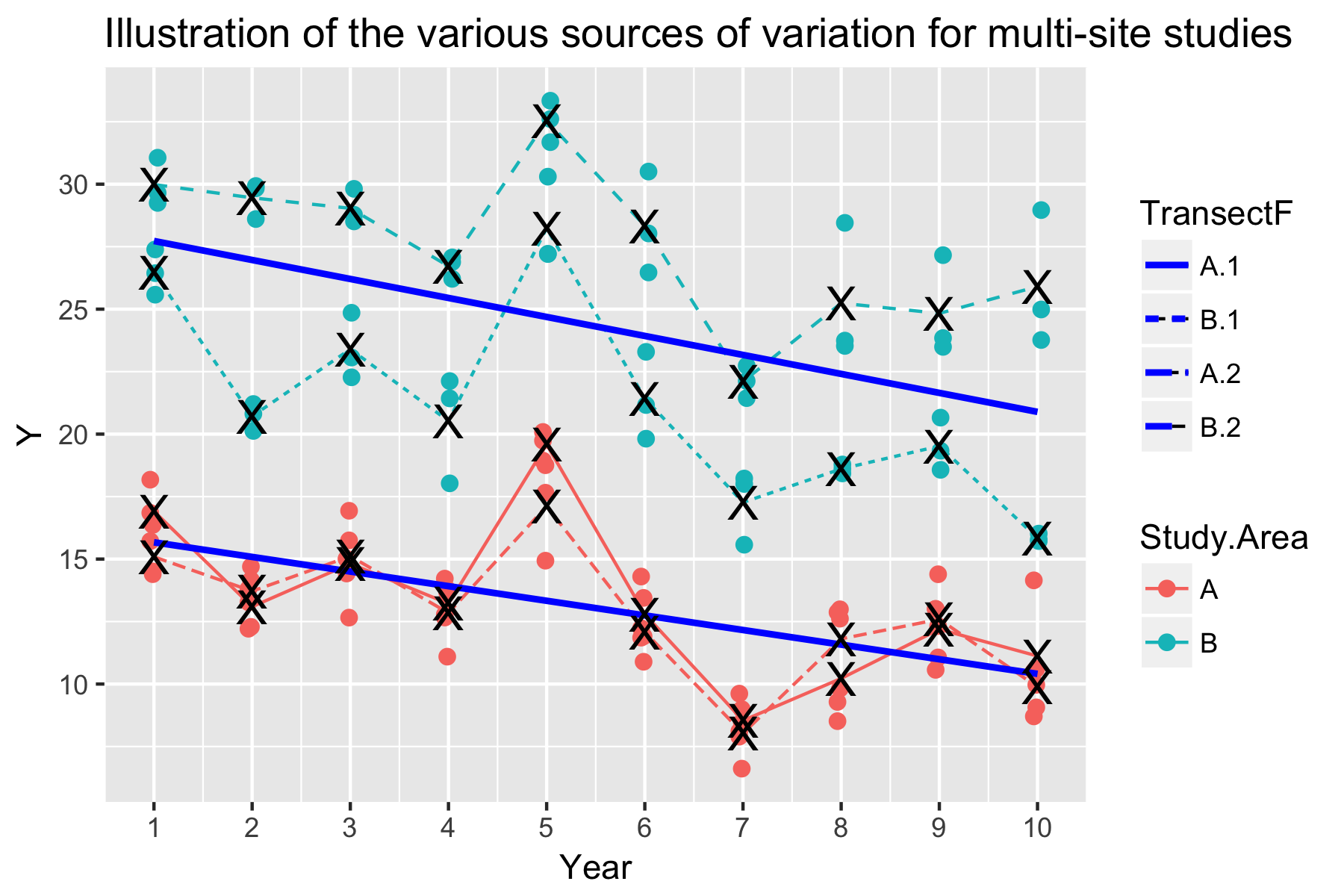
If there is substantial year-specific effects (process error), then performing more sampling in a year will NOT be helpful. Taking more measurements in a year will shrink the size of the confidence intervals in the bottom panel above, but has NO impact on the process error and so the variation around the fitted line will only be reduced slightly. In cases of substantial process error, the limiting factor for detecting trends is likely to be the total years of sampling, and not the number of measurements taken in a year.

# 3. Multi-site analyses – comparing trends

In a multi-site analysis, the focus is now on comparing the trends across multiple study area. For example, you may wish to compare the trend in two different study area that differ in the amount of human impact.

The principle method for comparing trends is a statistical technique called Analysis of Covariance (ANCOV). This is patterned against ANOVA method for comparing mean – the ANCOVA compares the impacts (the trend) of the covariates (in this case, the impact of time).

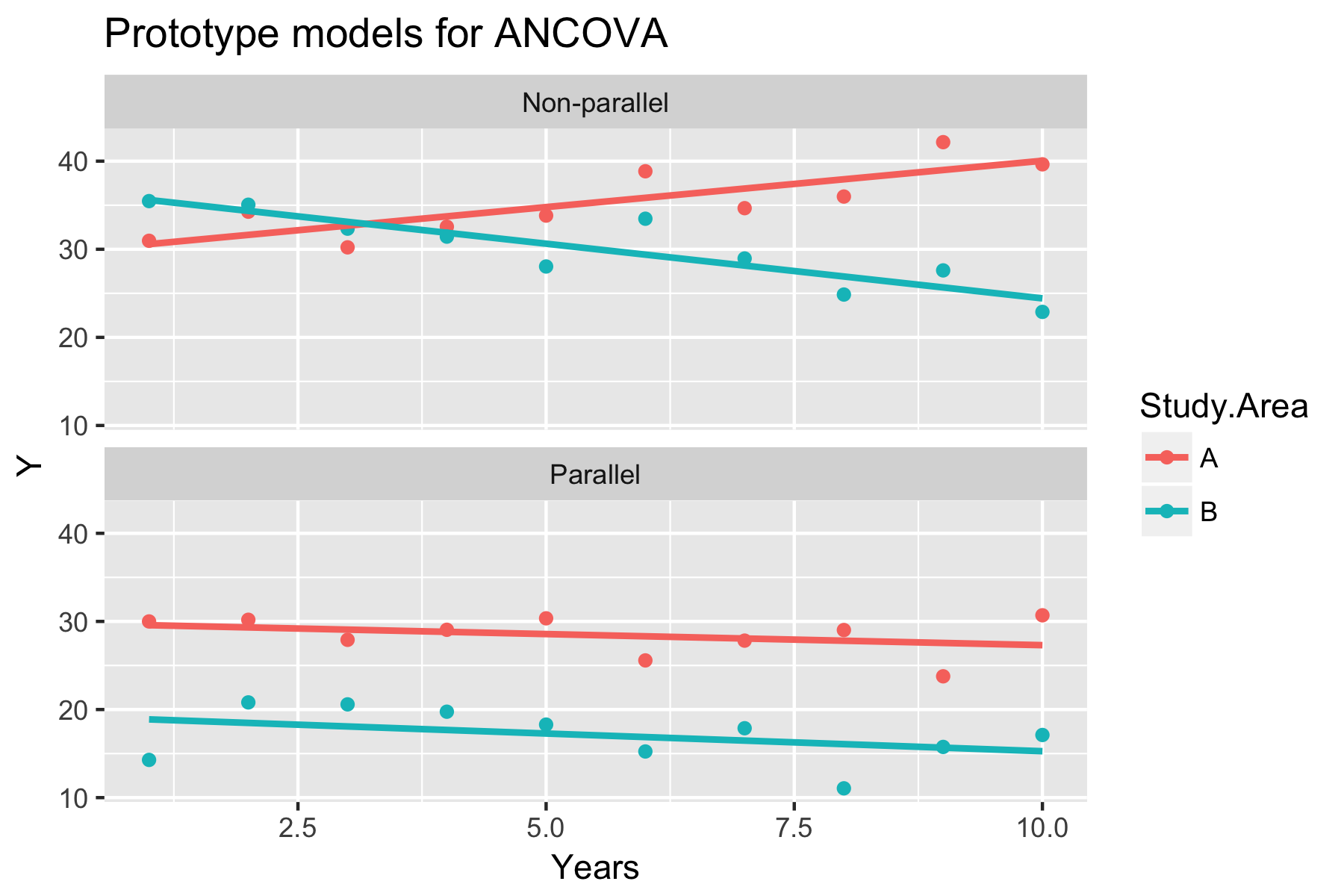
The data now consists of measurement on multiple transects within each study area over time. An illustration of the data is shown below:



There are several sources of variation that need to be accounted for when doing the trend comparison:

* Measurement error. This is represented by the variation of the same color dots around the transect mean (denoted by X) for any particular year.
* Transect variation. This is represented by the fact that on transect within each study area is generally higher than the other transect in all years. The cause of this is typically unknown and due to transect-specific effects such as local conditions at the transect.
* Process variation (year-specific variation). This is represented by the fact that the transect means (the X’s) tend to move up or down in any year across all study areas in tandem. For example, in year 5, all transect means are elevated, while in year 7 all the transect means are depressed. The cause of process error (year-specific variation) is unknown, but typically caused by external factors that affect the entire site (and all transects) simultaneously.
* Different trend. There appears to be a slightly different sloe for the two study areas.

A prototype model of a typical ANCOVA analysis (ignoring transect and year-specific factors) is:



The statistical model for the (very simple) non-parallel slope case is:



where the *StudyArea* term represents different intercepts for the two study areas; the *Year* term represents the average slope for the study areas; and *StudyArea:Year* is an interaction term representing differential slopes for the two study areas.

The statistical model for the (very simple) parallel slope case is similar and simply drop the interaction term



A naïve application of these models to the study designs from these protocols would be incorrect for the same reasons as in the single-site case. A linear mixed model must again be fit accounting for transect and year-specific effects. The two models are now:



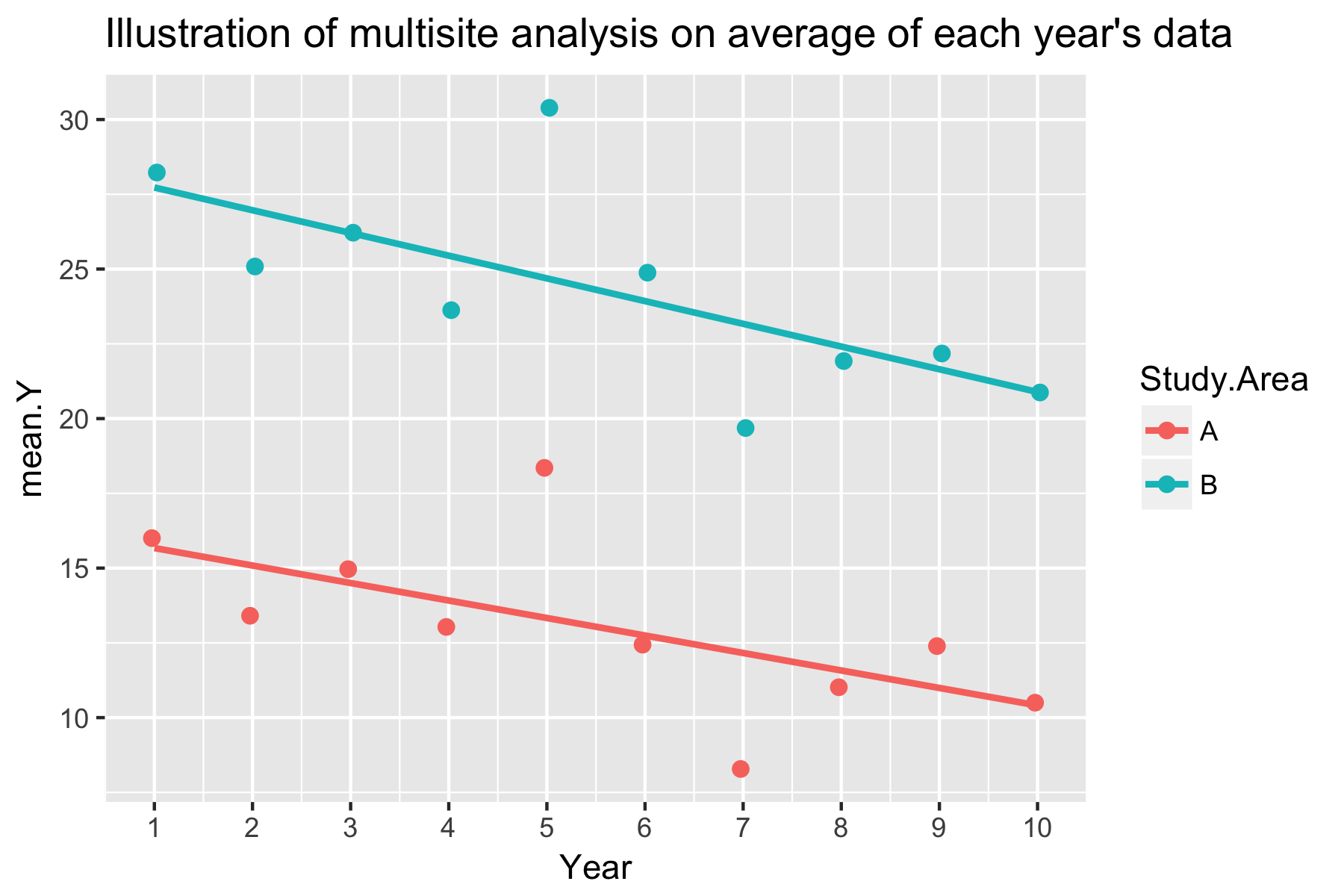
and



respectively with the latter three terms representing the transect and year specific effects that are common and a year-specific effect that could be different for each study area. Note that the transect should be uniquely labeled across study area so that transect T1 in study area A is a different transect than transect T1 in study area B. When implementing these models, the transect label is usually recoded as a combination of study area name and transect code (i.e. as transects A.T1 and B.T1) to make them unique.

There are two ways to properly analyze ANCOVA data with a complex experimental design.

First, if the data are perfectly balanced (i.e. every transect in each study area is measured every year, and every transect in each study area measured the same number of times in each year), then a simple analysis simply analyzes the average reading per year, i.e. average all of the data in a year for each study area, and a simple ANOVA analysis can be done. For example, from the first figure of this section, the averaged data is:



Becase you have averaged the data over all transects in the study area, the non-parallel slope model is simpler:



is fit (using *lmer()*) as in the single-site example

This gives:

> anova(avg.fit, ddf='kenward-roger')

Analysis of Variance Table of type III with Kenward-Roger

approximation for degrees of freedom

Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)

Year 1.581 1.581 1 8 6.11 0.03863 \*

Study.Area 160.284 160.284 1 8 619.17 7.276e-09 \*\*\*

Year:Study.Area 1.261 1.261 1 8 4.87 0.05837 .

The p-value for testing if the slopes are parallel is given by the p-value for the interaction term. In this case, there is some evidence that the slopes are not-parallel.

The estimates of the individual slopes can be obtained from the *lsmeans(*) package:

avg.fit.lsmo <- lsmeans::lstrends(avg.fit,

"Study.AreaF", var="Year")

lsmeans::cld(avg.fit.lsmo)

Study.Area Year.trend SE df lower.CL upper.CL .group

B -0.7597 0.2749 8.34 -1.389214 -0.13024453 1

A -0.5849 0.2749 8.34 -1.214395 0.04457388 1

The estimated slope for study area A is -0.75 (SE .27) and that for study area B is -0.58 (SE .27). The confidence intervals for the slopes overlap considerable. The grouping variable (.group column) indicates which study areas appear to have different slopes than other study areas. If a study area has a different grouping code, then there is evidence that its slope is different from slopes in other study areas with a different grouping code.

At first glance, it appears that we have “thrown away” much information. In fact we have not. A key determinant for comparing trends is the number of years of sampling – not how much sampling is done in each year. Here the number of years of sampling is 10 and the individual slopes is determined by the contrast in the MEAN Y value across these 10 values. The variation within a year provides no information on the trend – rather it provides information on transect-to-transect variation (and if it is consistent over time) and measurement error variation which also provides no information on the trend.

The variation in the mean response above is also much less than the variation in the individual data points around the regression line. Hence, averaging does toss out information, but summarizes it in a form most appropriate for a trend analysis. The variation of the mean response around the regression line is a combination of process error, average transect variation, and average measurement variation and captures all three sources of variation.

So if the design is balanced, a simple analysis simply computes the mean response in each year and performs a regression analysis on the mean response. There is no loss of information.

But, it is unlikely that the LTEM will have perfectly balanced data after many years of monitoring. For example, in the berry protocol, plants will be replaced; in the vegetation protocol, some plots may be removed because of damage to the plot.

In these cases, the more complex linear mixed model that includes additional interaction terms is needed:



The model will account automatically for unbalanced data as long as the transect labels are unique for each transect and not reused, i.e. if transect 1 is replaced, do NOT call the replacement transect as *transect 1.* Instead, refer to it as *transect* 3(assuming that two transects were originally planned).

The *lmerTest* package in *R* can be used to fit this linear mixed models as follows:

re.fit <- lmerTest::lmer(Y~ Study.AreaF + Year +

Year:Study.AreaF +

(1|TransectF) + (1|YearF) +

(1|YearF:Study.AreaF), data=sim.data.B)

> anova(re.fit, dfm='kenward-roger')

Analysis of Variance Table of type III with Satterthwaite

approximation for degrees of freedom

Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)

Study.AreaF 37.110 37.110 1 2.127 14.9561 0.05507 .

Year 15.155 15.155 1 8.000 6.1077 0.03863 \*

Study.AreaF:Year 7.564 7.564 1 106.000 3.0484 0.08371 .

There is slight difference in the interaction p-value because on the standard deviations was estimated as zero (see below).

The estimated slopes are found in the same way:

re.fit.lsmo <- lsmeans::lstrends(avg.fit, "Study.AreaF",

var="Year")

lsmeans::cld(re.fit.lsmo)

Study.Area Year.trend SE df lower.CL upper.CL .group

B -0.759 0.274 8.34 -1.389214 -0.13024453 1

A -0.584 0.274 8.34 -1.214395 0.04457388 1

The slope under this more complex model matches that from the simple analysis on the mean response each year. It will differ if the design is unbalanced, but unless the imbalance is severe, the differences will be slight.

The estimates of the standard deviations of the random effects can also be found:

> VarCorr(re.fit)

Groups Name Std.Dev.

YearF:Study.AreaF (Intercept) 0.0000

YearF (Intercept) 2.4288

TransectF (Intercept) 3.1011

Residual 1.5752

Here the data does not support year-specific factors that are unique to each study area (the standard deviation of the *YearF:Study.AreaF* terms are zero). This will give rise to slightly different p-values for certain terms between the simpler and more complex models because of slight differences in the way the degrees of freedom for the test statistics are computed using the Kenward-Roger methods.

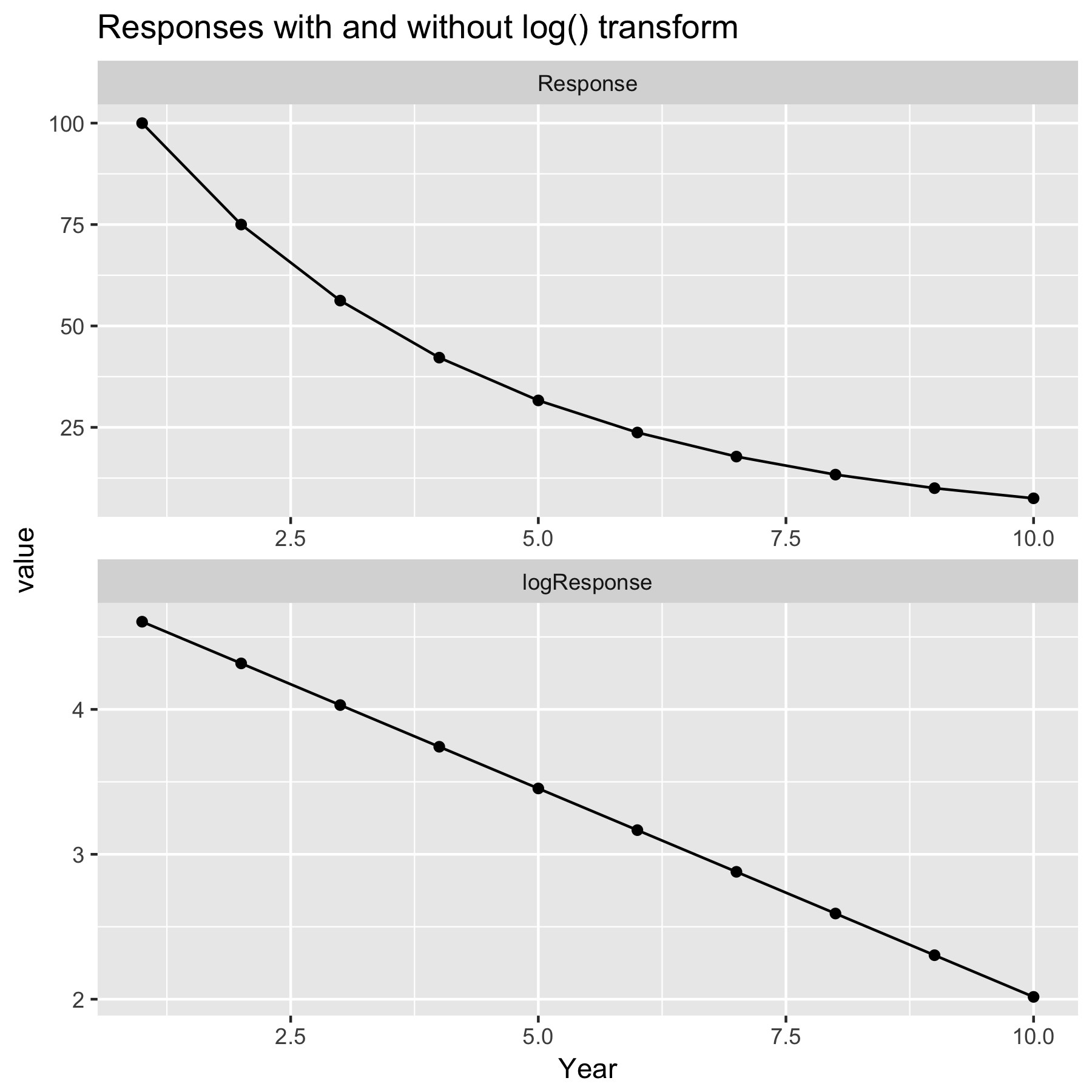
It turns out that fitting linear mixed models is quite difficult for small sample sizes and the *lmerTest*() approach may fail to converge and provide estimates of the trend with standard errors/p-values. In these cases, better starting values for the process may be required. It may be simpler to just do the approximate analysis on the averages.

# 4. Transformations

Many of the protocols have counts as a response variable, e.g. number of squirrel calls, number of berries. There are two approaches to dealing with count data.

(a) Use a regular linear mixed model on the log(counts).

Count data often has two features that make an analysis on the logarithmic scale attractive. First, the variation in the actual data tends to increase with the mean and a logarithmic transform stabilizes the variation. Second, effects often occur on a multiplicative scale, i.e. the trend line is a constant proportional change (e.g. a 5% decline in the mean number of berries per year) rather than a linear change (e.g. 2 fewer berries per year). A proportion change is translated into a linear change after the logarithmic transformation. For example, consider the following plots of a trend on the ordinary and logarithmic scales:



The upper plot shows a proportional change – year’s value is 0.75 of the previous year’s value. This gives rise to the curved line seen. On the logarithmic scale, the trend is now linear with a constant change per year (on the logarithmic scale) of log(0.75)=-.28/year.

Once a linear fit is found on the logarithmic scale, the back transformation can be made and the back transformation shows the trend in the MEDIAN response. When a logarithmic transformation is made, an implicit assumption is that the underlying data (closely) follows a log-normal distribution. The log-normal distribution has long-right tail and the estimated mean on the logarithm scale corresponds to the MEDIAN on the anti-logarithmic scale.

(b) Use a generalized linear mixed model.

A common distribution used for count data is the Poisson distribution. Under a Poisson distribution, the variation increases with the mean, and changes over time are often modeled as occurring on the logarithmic scale. The Poisson distribution naturally deals with 0 counts (these are possible under a Poisson model) and it is not necessary to add a small constant to zero counts.

However, there are many problems in using a generalized linear mixed model that really preclude it from consideration except in special circumstancs.

(a) If the transect data is a mean over multiple readings, then the response variable is no longer an integer and does NOT follow a Poisson distribution , even if the individual counts followed a Poisson distribution.

(b) Many dataset exhibit overdispersion where the variation is actually larger than predicted by a Poisson distribution. Adjustments need to be made to account for overdispersion such a random effects for transect and year effects.

(c) Model fitting is extremely difficult unless Bayesian methods are used. The *glmer()* function in *R* deals with (quasi)-poisson models but often fails to converge because of sparse data.

For this reason, I recommend that a standard linear mixed model be used for count data based on the logarithm of the (mean) response.