Adding class variables as predictors

Examples:

- Add species to an equation to estimate tree height.
- Add gender (male/female) to an equation to estimate weight of adult tailed frogs.
- Add machine type to an equation that predicts output.

Use “dummy” or “indicator” variables to represent the class variable

- e.g. have 3 species. Set up X1 and X2 as dummy variables:

Species	X1	X2
Cedar	1	0
Hemlock	0	1
Douglas fir	0	0

  - Only need two dummy variables to represent the three species (levels-1 dummies).
  - **The two dummy variables as a group represent the species. (→ we can not use a t-test)**

- Add the dummy variables to the equation **– this will alter the intercept**

- To alter the slopes, add an interaction between dummy variables and continuous variable(s)

  e.g. have 3 species, and a continuous variable, dbh

Species	X1	X2	X3=dbh	X4=X1 * dbh	X5=X2*dbh
Cedar	1	0	10	10	0
Hemlock	0	1	22	0	22
Douglas fir	0	0	15	0	0

- NOTE: **The two dummy variables, and the interactions with the continuous variable as a group represent the species.**

How does this work?

$$y_i = b_0 + \underbrace{b_1 x_{1i} + b_2 x_{2i}}_{\textit{Dummy variables}} + \underbrace{b_3 x_{3i}}_{\textit{dbh}} + \underbrace{b_4 x_{4i} + b_5 x_{5i}}_{\textit{interactions}} + e_i$$

- For Cedar

intercept

Slope of dbh

$$y_i = b_0 + b_1 + b_3 x_{3i} + b_4 x_{4i} + e_i$$

- For Hemlock

$$y_i = b_0 + b_2 + b_3 x_{3i} + b_5 x_{5i} + e_i$$

- For Douglas Fir

$$y_i = b_0 + b_3 x_{3i} + e_i$$

How does this work?

→Fit one equation using all data, but get different equations for different species.

One can also test for differences among species using a **partial-F test.**

Script 6_MLR_dummycoding.R

Single factor studies: examples

- Example 1: experimental study

Effectiveness of different dosages of drug

30 patients, 3 dosage levels: 10 patients in each dosage level

= completely randomized design based on a single, three-level quantitative factor

= *balanced* design (each treatment replicated the same number of times)

Single factor with *J* levels: TWO approaches

I. Regression model

For example

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_{11} x_{ij}^2 + \varepsilon_{ij}$$

where:

x_{ij} = centered dosage level amount for the ijth case

ONLY possible for a quantitive factor (example 1)

Single factor with *J* levels: TWO approaches

II. Analysis of Variance model (ANOVA)

J-1 dummy variables as predictors

Example 1:
treatment *j* (1->3)
replicate *i* (1->10)

A regression model with *only dummy predictor variables* is called an *analysis of variance model*

$$Y_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \varepsilon_{ij}$$

where:

$$X_{ij1} = \begin{cases} 1 & \text{if treatment 1} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij2} = \begin{cases} 1 & \text{if treatment 2} \\ 0 & \text{otherwise} \end{cases}$$

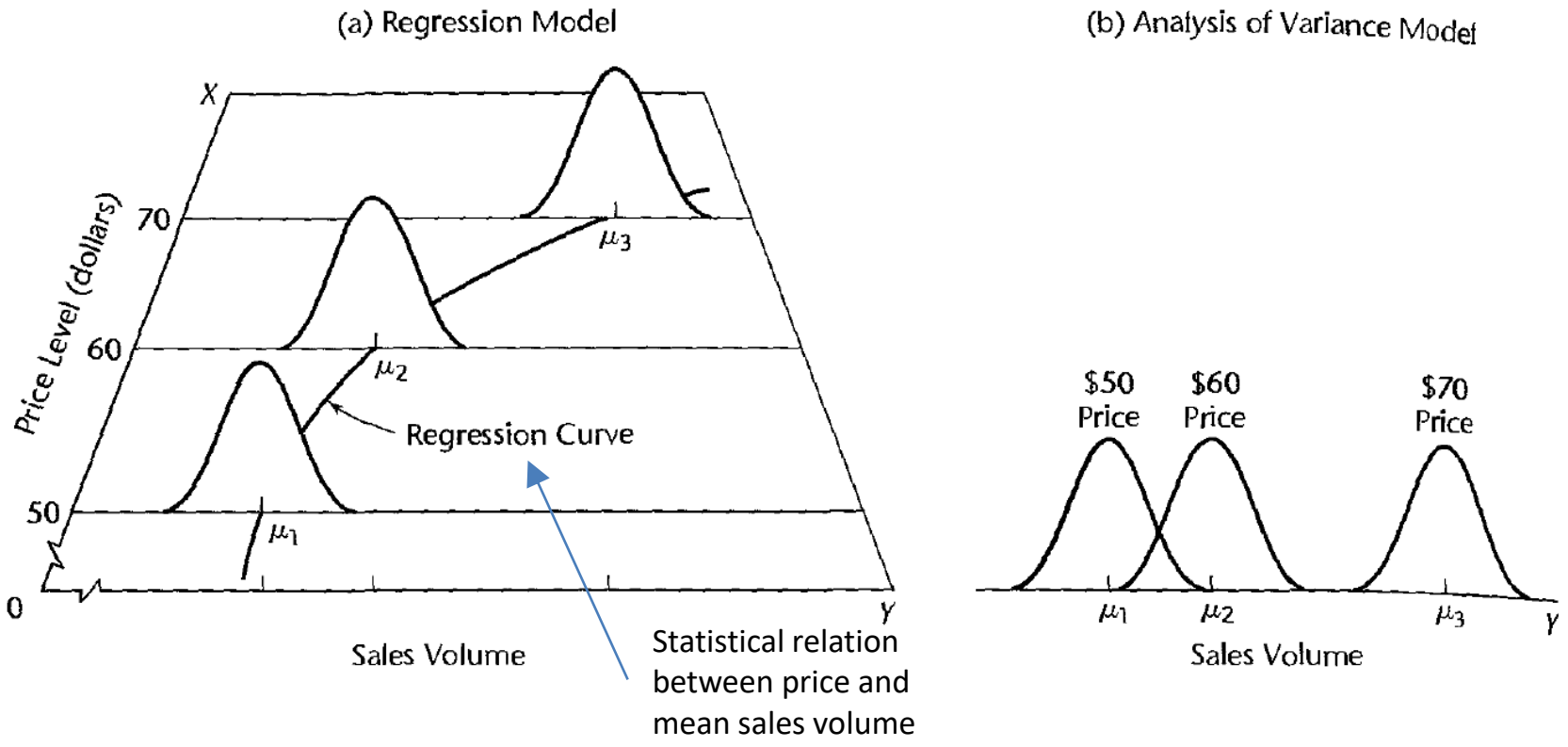
→ The intercept is simply the mean of the reference group. The coefficients for the other groups are the differences in the mean between the reference group and the other groups. *(see TADE I)*

Relation between Regression and ANOVA

Difference ANOVA vs regression:

- Predictor variable may be qualitative
- If predictor variables are quantitative, no assumption is made about the nature of the statistical relation between them and the response variable

Illustration: effect of price levels on sales volume



Assumptions met?

Full:
Common:
Intercept Only:

R Square and SE^E

Full:
Common:
Intercept Only:

Df, SSR, SSE:

Model	df model	SSR	df error	SSE
Full				
Common				
Int. Only				

Full versus Common

H0: Equations are the same for all species

H1: Equations differ

Partial F:

$$partial\ F = \frac{\left(SSreg(\ full) - SSreg(reduced)\right)/r}{SSE/(n-m-1)(\ full)}$$

Compare to:

F distribution for a 1- α percentile with *r* and *n*-*m*-1 (full model) degrees of freedom.

Decision:

If equations differ – could we use the same slope, just different intercepts?

Full versus Intercepts only models

H0: Slopes are the same for all species

H1: Slopes differ

Partial F:

Compare to:

Decision:

Are the differences in intercept significant between the species?

Intercepts only versus common models

H0: intercepts are the same for all species

H1: intercepts differ

Partial F:

Compare to:

Decision:

Categorical variable in R

```
> is.factor(threespec$species)
[1] TRUE
> y <- log(height)
> logdbh <- log10(dbh)
> model <- lm(y~dbh+logdbh+species)#species is a factor with three
   levels
> anova(model)
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq  F value    Pr(>F)
dbh         1 20.6679 20.6679 638.4522 < 2.2e-16 ***
logdbh      1  4.5059  4.5059 139.1905 < 2.2e-16 ***
species     2  0.3907  0.1953   6.0339  0.002688 **
Residuals 309 10.0029  0.0324
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Categorical variable in R

```
> summary(model)

Call:
lm(formula = y ~ dbh + logdbh + species)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80448 -0.09375  0.01841  0.11375  0.50268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9522770  0.3405274   2.796 0.005490 **
dbh         -0.0007072  0.0001968  -3.594 0.000379 ***
logdbh       1.8817861  0.1613111  11.666  < 2e-16 ***
speciesGF    0.0350335  0.0292405   1.19
speciesWC   -0.0432846  0.0284269  -1.52
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1799 on 309 degrees of freedom
Multiple R-squared: 0.7188,     Adjusted R-squared: 0.7151
F-statistic: 197.4 on 4 and 309 DF,  p-value: < 2.2e-16
```

What are these estimates standing for?

Categorical variable in R

```
> #And like this you can model the differences in slopes:
> model3 <- lm(y ~ dbh + logdbh + species + species*dbh + species*logdbh)
> summary(model3)$coefficients
                    Estimate   Std. Error    t value    Pr(>|t|)
(Intercept)       3.2375482444 1.0579012074  3.0603503 0.002407295
dbh               0.0002353346 0.0005380091  0.4374174 0.662118218
logdbh            0.8444593062 0.4921498205  1.7158582 0.087203472
speciesGF        -2.6149630938 1.1901740212 -2.1971267 0.028763267
speciesWC        -2.4059038475 1.1614320579 -2.0714977 0.039152808
dbh:speciesGF    -0.0010876683 0.0006366918 -1.7083122 0.088595852
dbh:speciesWC    -0.0009779391 0.0006048422 -1.6168501 0.106944403
logdbh:speciesGF  1.2038702934 0.5583789885  2.1560093 0.031864672
logdbh:speciesWC  1.0732358636 0.5427323170  1.9774681 0.048888796


> summary(full)$coefficients
                 Estimate   Std. Error    t value    Pr(>|t|)
(Intercept)  3.2375482444 1.0579012074  3.0603503 0.002407295
x1          -2.4059038475 1.1614320579 -2.0714977 0.039152808
x2          -2.6149630938 1.1901740212 -2.1971267 0.028763267
x3           0.8444593062 0.4921498205  1.7158582 0.087203472
x4           0.0002353346 0.0005380091  0.4374174 0.662118218
x5           1.0732358636 0.5427323170  1.9774681 0.048888796
x6           1.2038702934 0.5583789885  2.1560093 0.031864672
x7          -0.0009779391 0.0006048422 -1.6168501 0.106944403
x8          -0.0010876683 0.0006366918 -1.7083122 0.088595852
```

Do you recognize these estimates?

The dummy variables!

Testing for common trend - ANCOVA

Examples:

- Test if the same trend is occurring in a number of locations
- Data from a single site is so poor that trends cannot be detected, but by pooling the sites, a common trend over sites can be detected because of the increased sample size.

Analysis of Covariance (ANCOVA) :

A combinatin of ANOVA and Regression

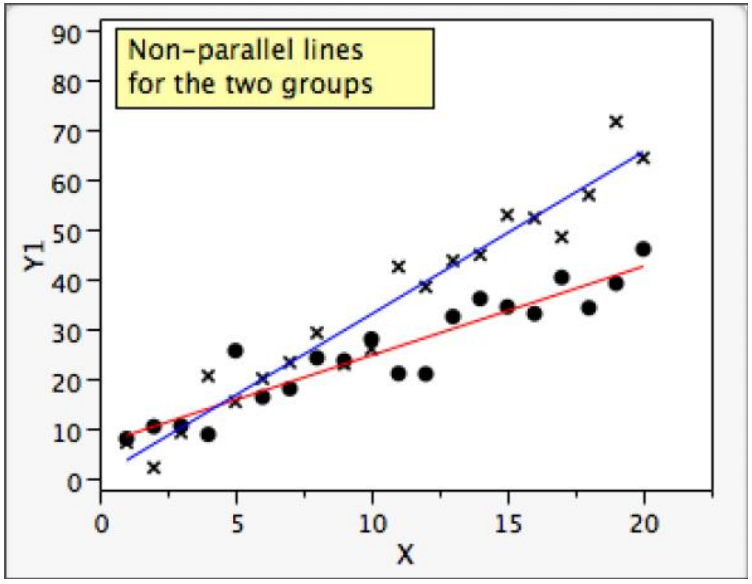
Groups of data (e.g. from the same location) are identified by a nominal or ordinal scale variable

A continuous predictor variable (in trend analysis: time) is also measured for both groups.

Testing for common trend - ANCOVA

1. ANCOVA is used to check if the regression line for the groups are parallel.

If there is evidence that the individual regression lines are not parallel, then a separate regression line (trend line) must be fit for each group for prediction purposes.



Y ~ Group + X + Group*X

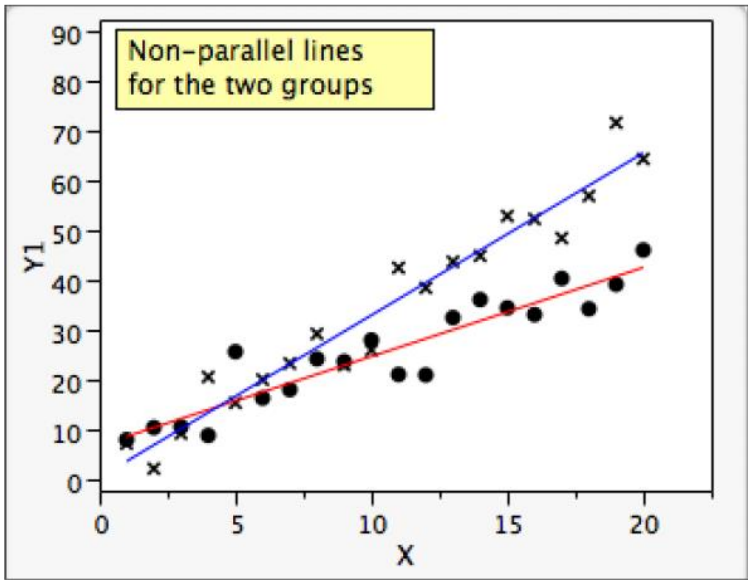
"variation in Y can be explained by a common intercept (never specified) followed by group effects (different intercepts), a common slope (trend) on X, and an "interaction" between Group and X which is interpreted as different slopes (different trends) for each group."

Testing for common trend - ANCOVA

This model is almost equivalent to fitting a separate regression line for each group.

The only advantage to using this joint model for all groups is that all of the groups contribute to a better estimate of residual error.

→ If the number of data points per group is small, this can lead to an improved power to detect trends compared to fitting each group individually .

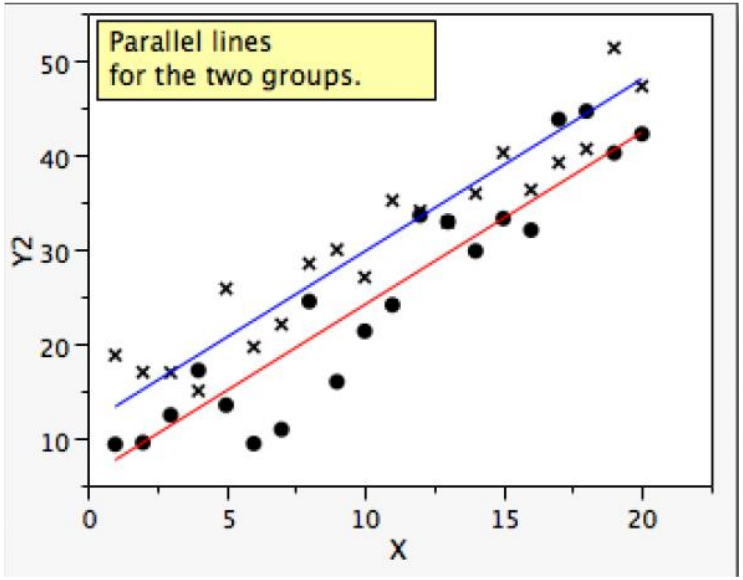


Y ~ Group + X + Group*X

Testing for common trend - ANCOVA

2. If there is no evidence of non-parallelism, then the next task is to see if the lines are co-incident, i.e. have both the same intercept and the same slope.

If there is evidence that the lines are not coincident, then a series of parallel lines are fit to the data. All of the data are used to estimate the common slope.



Y ~ Group + X

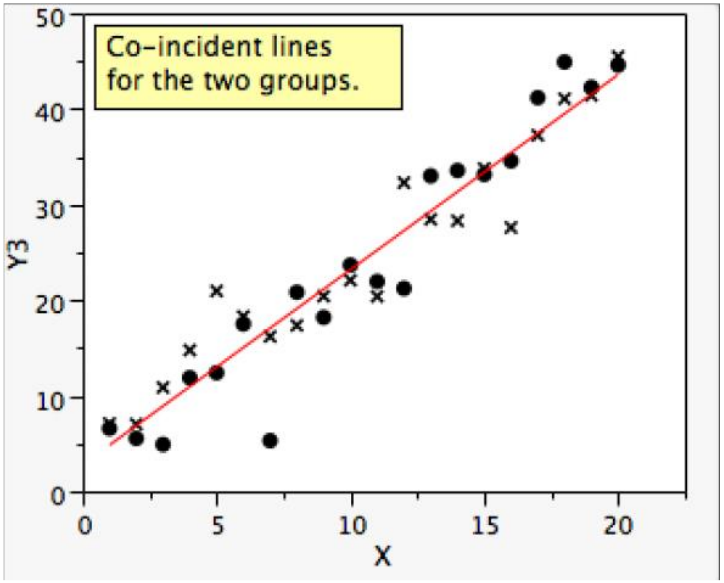
This simpler model lacks the Group*X "interaction term"

→ a statistical test to see if this simpler model is tenable would correspond to examining the p-value of the test on the Group*X term from the complex model.

This is exactly analogous to testing for interaction effects between factors in a two-factor ANOVA.

Testing for common trend - ANCOVA

3. If there is no evidence that the lines are not coincident, then all of the data can be simply pooled together and a single regression line fit for all of the data.



Y ~ X

The Group term that has been dropped.

→ a statistical test to see if this simpler model is tenable would correspond to examining the p-value of the test on the Group term from the previous model.

The test for co-incident lines should only be done if there is insufficient evidence against parallelism. *(after concluding they are parallel)*

Assumptions

As ANCOVA is a combination of ANOVA and Regression, the assumptions are similar.

- The response variable Y is continuous (interval or ratio scaled).
- The Y are a random sample from the various time points measured.
- There must be no outliers. Plot Y vs. X for each group separately to see if there are any points that don't appear to follow the straight line.
- The relationship between Y and X must be linear for each group: Check by looking at the individual plots of Y vs. X for each group.
- The variance must be equal for each group around their respective regression lines: Check that the spread of the points is equal around the range of X and that the spread is comparable between the two groups. This can be formally checked by looking at the MSE from a separate regression line for each group as MSE estimates the variance of the data around the regression line.
- The residuals must be normally distributed around the regression line for each group: Check by examining the residual plots from the fitted model for evidence of non-normality. *For large samples, this is not too crucial; for small sample sizes, you will likely have inadequate power to detect anything but gross departures.*

ANCOVA example 1: Degradation of dioxin - pooling locations

- An unfortunate byproduct of pulp-and-paper production used to be dioxins - a very hazardous material.
- This material was discharged into waterways with the pulp-and-paper effluent where it bioaccumulated in living organisms such a crabs.
- Newer processes have eliminated this by product, but the dioxins in the organisms takes a long time to degrade.
- Government environmental protection agencies take samples of crabs from affected areas each year and measure the amount of dioxins in the tissue.

ANCOVA example: Degradation of dioxin - pooling locations

- Each year, four crabs are captured from two monitoring stations which are situated quite a distance apart on the same inlet where the pulp mill was located.
- The liver is excised and the livers from all four crabs are composited together into a single sample. The dioxins levels in this composite sample is measured.
- As there are many different forms of dioxins with different toxicities, a summary measure, called the Total Equivalent Dose (TEQ) is computed from the sample.

- Is the rate of decline the same for both sites?
- What is the estimated difference or ratio in concentrations between the two sites?

Site	Year	TEQ	$\log(TEQ)$
a	1990	179.05	5.19
a	1991	82.39	4.41
a	1992	130.18	4.87
a	1993	97.06	4.58
a	1994	49.34	3.90
a	1995	57.05	4.04
a	1996	57.41	4.05
a	1997	29.94	3.40
a	1998	48.48	3.88
a	1999	49.67	3.91
a	2000	34.25	3.53
a	2001	59.28	4.08
a	2002	34.92	3.55
a	2003	28.16	3.34
b	1990	93.07	4.53
b	1991	105.23	4.66
b	1992	188.13	5.24
b	1993	133.81	4.90
b	1994	69.17	4.24
b	1995	150.52	5.01
b	1996	95.47	4.56
b	1997	146.80	4.99
b	1998	85.83	4.45
b	1999	67.72	4.22
b	2000	42.44	3.75
b	2001	53.88	3.99
b	2002	81.11	4.40
b	2003	70.88	4.26

The raw data

dioxin2.csv

Read in the data

```
> crabs <- read.csv("../data/dioxin2.csv", header=TRUE,
+                   as.is=TRUE, strip.white=TRUE,
+                   na.string=".")
> crabs$Site   <- factor(crabs$Site)
> crabs$logTEQ <- NULL # drop this and recompute later
> head(crabs)
  Site Year WHO.TEQ
1    a 1990  179.05
2    a 1991   82.39
3    a 1992  130.18
4    a 1993   97.06
5    a 1994   49.34
6    a 1995   57.05
> str(crabs)
'data.frame':   28 obs. of  3 variables:
 $ Site   : Factor w/ 2 levels "a","b": 1 1 1 1 1 1 1 1 1 1 ...
 $ Year   : int  1990 1991 1992 1993 1994 1995 1996 1997 1998
1999 ...
 $ WHO.TEQ: num  179.1 82.4 130.2 97.1 49.3 ...
```

Read in the data

- Year and WHO.TEQ are numeric
- We must declare the Site variable to be a FACTOR, i.e. a categorical variable.

  it is recommended that alphanumeric codes be used for categorical variables, i.e. don't code the sites as 1 and 2 because then there is the possibility that R will treat the sites as a continuous variable if you forget to declare the variable as a factor. With alphanumeric codes, R will either figure it out, or issue and error message if you forget to declare the variable as a factor.
- I also find it convenient to recompute derived variables (e.g. the log() of the TEQ), rather than reading them in. This way I avoid any errors where the derived variables are not in sync with the rest of the data.

initial plot of the data

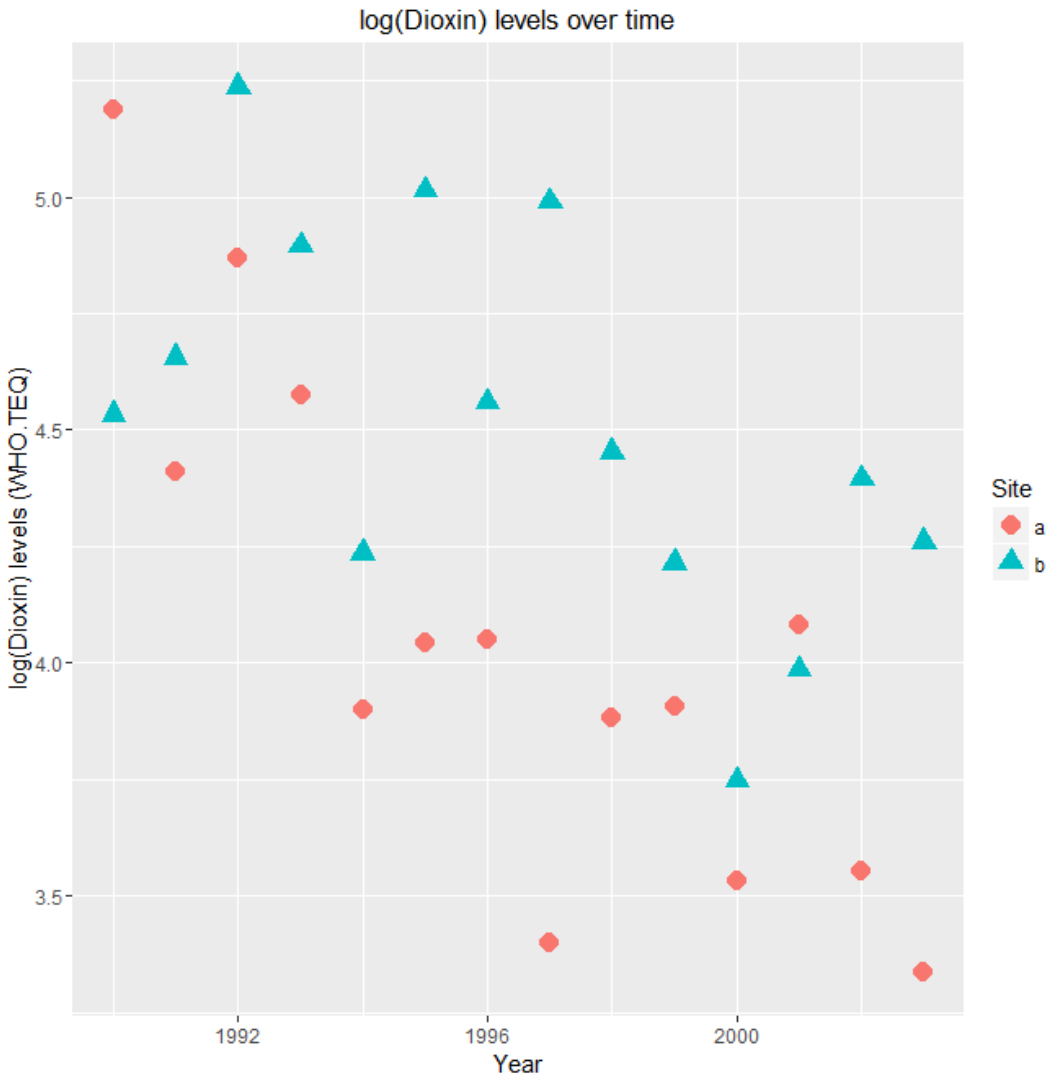
- We already know that we will be plotting on the logscale
- Using a different plotting symbol for each group can be done using the ggplot2 package.

aes() function can specify the different plotting symbols and colors that should be used for the different sites.

ggplot() function creates the legend.

```
> ggplot(data=crabs, aes(x=Year, y=logTEQ, shape=Site,
color=Site))+
+   ggtitle("log(Dioxin) levels over time")+
+   xlab("Year")+ylab("log(Dioxin) levels (WHO.TEQ)")+
+   geom_point(size=4)
```

initial plot of the data



looking for outliers & checking the assumptions

- The initial scatter plot doesn't show any obvious outliers.
- Each year's data is independent of other year's data as a different set of crabs was selected.
- The data from one site are independent from the other site.

Whenever multiple sets of data are collected over time, there is always the worry about common year effects (also known as process error).

For example, if the response variable was body mass of small fish, then poor growing conditions in a single year could depress the growth of fish in all locations. This would then violate the assumption of independence as the residual in one site in a year would be related to the residual in another site in the same year.

In this example, this is unlikely to have occurred. Degradation of dioxin is relatively independent of external environmental factors and the variation that we see about the two regression lines is related solely to sampling error based on the particular set of crabs that were sampled.

Does a single model make sense?

→ Fitting a simple regression to EACH site

- lm() function to fit the regression model to each site.
- d_ply() function is the modern way in R to do by-group processing

```
> d_ply(crabs, "Site", function(x){
+   cat("\n\n***Separate fit for site :",
as.character(x$Site[1]),"\n")
+   model <- lm( logTEQ ~ Year, data=x)
+   print(summary(model))
+   print(confint(model)) # confidence interval on slope
+ })
```

```
***Separate fit for site : a

Call:
lm(formula = logTEQ ~ Year, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59906 -0.16260 -0.01206  0.14054  0.51449

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 218.91364   42.79187   5.116 0.000255 ***
Year         -0.10762    0.02143  -5.021 0.000299 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3233 on 12 degrees of freedom
Multiple R-squared:  0.6775,    Adjusted R-squared:  0.6506
F-statistic: 25.21 on 1 and 12 DF,  p-value: 0.0002986

                 2.5 %       97.5 %
(Intercept) 125.6781579 312.14911470
Year         -0.1543185  -0.06091975
```

```
***Separate fit for site : b

Call:
lm(formula = logTEQ ~ Year, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5567 -0.2399  0.0224  0.2013  0.5059

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 123.23673   46.41606   2.655   0.0210 *
Year         -0.05947    0.02325  -2.558   0.0251 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3507 on 12 degrees of freedom
Multiple R-squared:  0.3528,    Adjusted R-squared:  0.2989
F-statistic: 6.542 on 1 and 12 DF,  p-value: 0.0251

                2.5 %       97.5 %
(Intercept) 22.1048113 224.368641271
Year        -0.1101205  -0.008811465
```

Does a single model make sense?

- The estimated slope for the a site is -0.107 (*se* 0.02) while the estimated slope for the b site is -0.06 (*se* 0.02).
- The 95% confidence intervals overlap considerably so the population slopes could be the same for the two groups.
- The MSE from site a is 0.10 and the MSE from site b is 0.12. This corresponds to standard deviations (RMSE) about the regression line of sqrt(0.10) = 0.32 and sqrt(0.12) = 0.35 which are very similar so that assumption of equal standard deviations about the regression line for the two sites seems reasonable.
- The residual plots (not shown) also look reasonable.

→The assumptions appear to be satisfied, so let us now fit the various models.

1. non-parallel slope model

- Fit the regression line with non-parallel slopes Because lm() produces type I (increment tests), you need to specify the interaction term last in the model sequence.
- Be sure that Site has been declared as a factor.
- The anova() function produces the table that contains the test for the hypothesis (H0) of parallel slopes. (Ha: slopes are not parallel)

```
> crabs.model.np <- lm( logTEQ ~ Site + Year + Year:Site,
data=crabs)
> anova(crabs.model.np)
Analysis of Variance Table

Response: logTEQ
          Df Sum Sq Mean Sq F value    Pr(>F)
Site       1 1.4868  1.4868  13.072  0.001383 **
Year       1 3.1756  3.1756  27.921 2.028e-05 ***
Site:Year  1 0.2638  0.2638   2.319  0.140873
Residuals 24 2.7297  0.1137
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value 0.14 indicates very little evidence against the hypothesis of parallel slopes

2. refit the model, dropping the interaction term

- Fit the regression line with parallel slopes.
- Specify the Site term last to get the proper test for site effects.
- Be sure that Site has been declared as a factor.

```
> crabs.model.p <- lm( logTEQ ~ Year + Site, data=crabs)
> anova(crabs.model.p)
Analysis of Variance Table

Response: logTEQ
          Df Sum Sq Mean Sq F value    Pr(>F)
Year       1 3.1756  3.1756  26.521 2.53e-05 ***
Site       1 1.4868  1.4868  12.417 0.001663 **
Residuals 25 2.9935  0.1197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

small p-value (0.0017) for the Site effect: evidence that two lines are not coincident, i.e. they are parallel with different intercepts.

The rate of decay of the dioxin appears to be equal in both sites, but the initial concentration appears to be different.

3. Estimate Time and Site effects

```
> summary(crabs.model.p)

Call:
lm(formula = logTEQ ~ Year + Site, data = crabs)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61110 -0.18485 -0.04157  0.30391  0.59257

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 170.84475   32.38784   5.275 1.83e-05 ***
Year         -0.08354    0.01622  -5.150 2.53e-05 ***
Siteb         0.46086    0.13079   3.524  0.00166 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.346 on 25 degrees of freedom
Multiple R-squared:  0.609,     Adjusted R-squared:  0.5777
F-statistic: 19.47 on 2 and 25 DF,  p-value: 7.985e-06
```

Two groups → only one dummy variable

```
> coef(crabs.model.p)
 (Intercept)          Year         Siteb
170.84475025  -0.08354254   0.46086209
> sqrt(diag(vcov(crabs.model.p))) # gives the SE
(Intercept)         Year        Siteb
32.38784060  0.01622224  0.13078791
> confint(crabs.model.p)
                  2.5 %       97.5 %
(Intercept) 104.1407439 237.54875661
Year         -0.1169529  -0.05013221
Siteb         0.1914994   0.73022482
> names(summary(crabs.model.p))
 [1] "call"          "terms"         "residuals"
"coefficients"
 [5] "aliased"       "sigma"         "df"              "r.squared"
 [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
> summary(crabs.model.p)$r.squared
[1] 0.6089959
> summary(crabs.model.p)$sigma
[1] 0.3460323
```

!! These results are suitable for any continuous variable (e.g. Year), but be VERY CAUTIOUS about interpreting the estimates for the categorical variable Site as these values depend on the internal parameterization used by R.

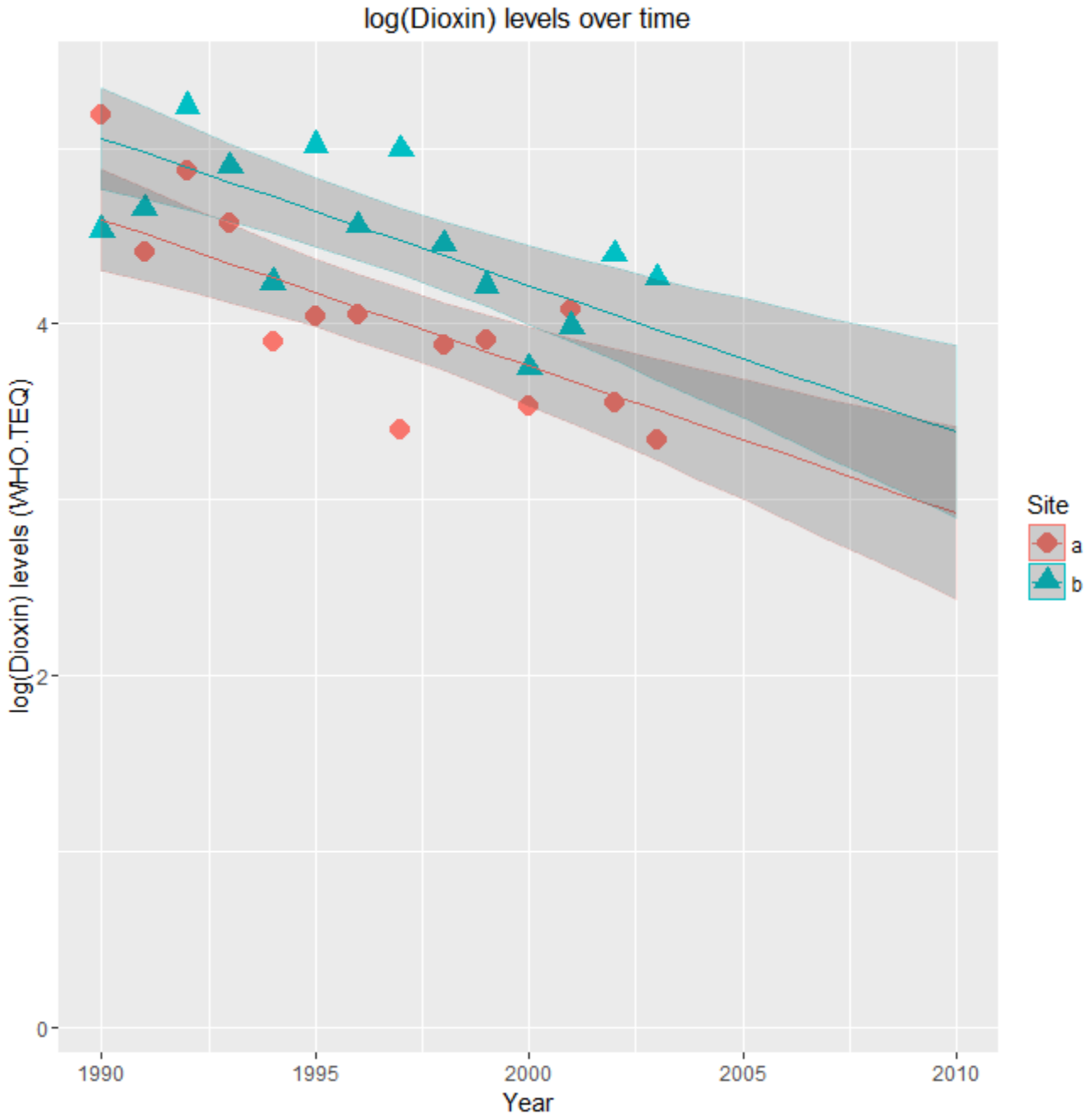
3. Estimate Time and Site effects

- The common slope has a value of -0.083 (*se* 0.016). Because the analysis was done on the log-scale, this implies that the dioxin levels changed by a factor of exp(-0.083) = 0.92 from year to year, *i.e.* about a 8% decline each year.
- The 95% confidence interval for the slope on the log-scale is from (-0.12 -> -0.05) which corresponds to a potential factor between exp(-0.12) = 0.88 to exp(-0.05) = 0.95 per year, i.e. between a 12% and 5% decline per year.
- While it is possible to estimate the difference between the parallel lines from the information produced by the summary() function, this is VERY DANGEROUS as these numbers could change depending on the internal parameterization adopted by R.
- In the case of categorical variables, the preferred method is to use the lsmeans() function in the lsmeans package.

- Caution. There is also a lsmeans() function in the lmerTest package which has different functionality

```
> crabs.model.p.lsmo <- lsmeans::lsmeans(crabs.model.p, ~Site)
> sitediff <- pairs(crabs.model.p.lsmo)
> summary(sitediff, infer=TRUE)
 contrast   estimate        SE df   lower.CL   upper.CL t.ratio
p.value
 a - b    -0.4608621 0.1307879 25 -0.7302248 -0.1914994  -3.524
0.0017

Confidence level used: 0.95
```

- The estimated difference between the lines (on the log-scale) is estimated to be 0.46 (*se* 0.13).
- Because the analysis was done on the log-scale, this corresponds to a ratio of exp(0.46) = 1.58 in median dioxin levels between the two sites, i.e. site b has 1.58 X the dioxin level as site a, on average.
- Because the slopes are parallel and declining, the dioxin levels are falling in both sites, but the 1.58 times ratio remains consistent.


log(Dioxin) levels over time
log(Dioxin) levels (WHO.TEQ)
4
2
0
1990
1995
2000
2005
2010
Year
Site
a
b

ANCOVA example 2: Change in yearly average temperature with regime shifts

- The ANCOVA technique can also be used for trends when there are KNOWN regime shifts in the series.
- The case when the timing of the shift is unknown is more difficult and not covered in this course.
- For example, consider a time series of annual average temperatures measured at Tuscaloosa, Alabama from 1901 to 2001.
- It is well known that shifts in temperature can occur whenever the instrument or location or observer or other characteristics of the station change.

tuscaloosa.csv

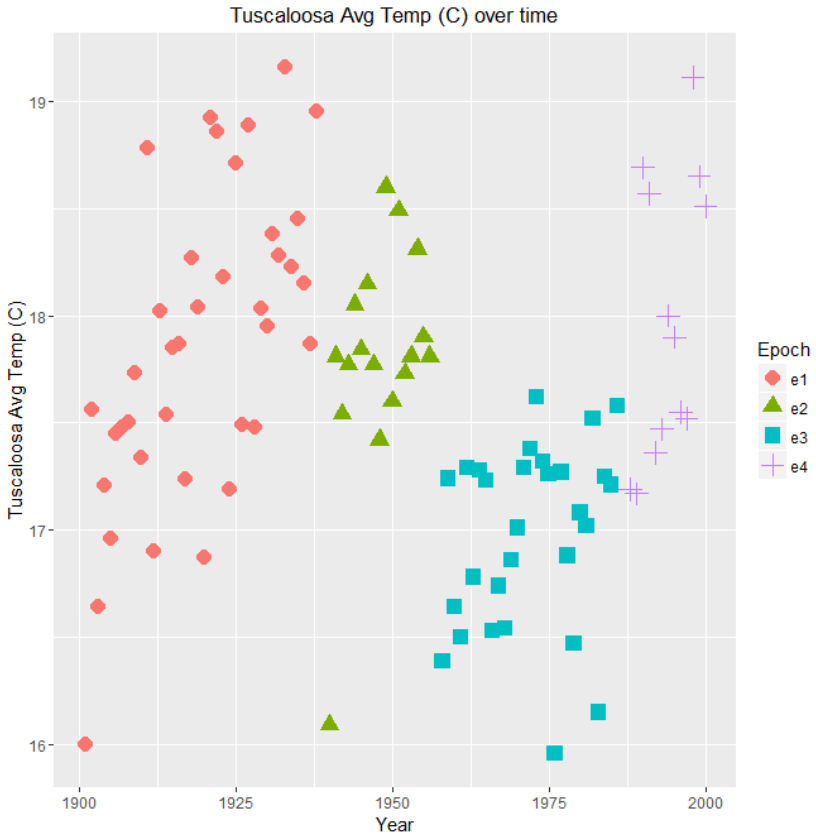
Read in the data

```
> tusctemp <- read.csv("../data/tuscaloosa.csv", header=TRUE,
+                      as.is=TRUE, strip.white=TRUE,
+                      na.string="")  # here missing values are blanks or
null cells
# Years where the average temperature is a mixture of
# reading at two different sites are removed.
> tusctemp <-
tusctemp[complete.cases(tusctemp[,c("Year","Epoch","Avg.Temp..C.")]),]
# Make sure Epoch is a Factor
> tusctemp$Epoch <- factor(tusctemp$Epoch)
```

```
> head(tusctemp)
  Year Avg.Temp..C. Epoch Notes
1 1901        16.00    e1  <NA>
2 1902        17.56    e1  <NA>
3 1903        16.64    e1  <NA>
4 1904        17.21    e1  <NA>
5 1905        16.96    e1  <NA>
6 1906        17.45    e1  <NA>
> str(tusctemp)
'data.frame':   97 obs. of  4 variables:
 $ Year        : int  1901 1902 1903 1904 1905 1906 1907 1908 1909 1910
...
 $ Avg.Temp..C.: num  16 17.6 16.6 17.2 17 ...
 $ Epoch       : Factor w/ 4 levels "e1","e2","e3",..: 1 1 1 1 1 1 1 1
1 1 ...
 $ Notes       : chr  NA NA NA NA ...
```

A time series plot of the data is constructed using the ggplot2 package.

```
> plotprelim <- ggplot(data=tusctemp, aes(x=Year, y=Avg.Temp..C.,
+                         shape=Epoch, color=Epoch))+
+   ggtitle("Tuscaloosa Avg Temp (C) over time")+
+   xlab("Year")+ylab("Tuscaloosa Avg Temp (C)")+
+   geom_point(size=4)
> plotprelim
```



- The plot clearly shows a shift in the readings in 1939 (thermometer changed), 1957 (station moved), and possibly in 1987 (location and thermometer changed).
- There is an obvious outlier around 1940 – this reading needs to be investigated further and the analysis should be repeated with this point removed to see if the results are dramatically different.

Workflow

1. We first run a separate regression line for each epoch
   - to check for outliers
   - to check that the slopes are similar
   - to check that the MSE are comparable among epochs
2. Then we start with the non-parallel slope model to check for evidence against parallelism.

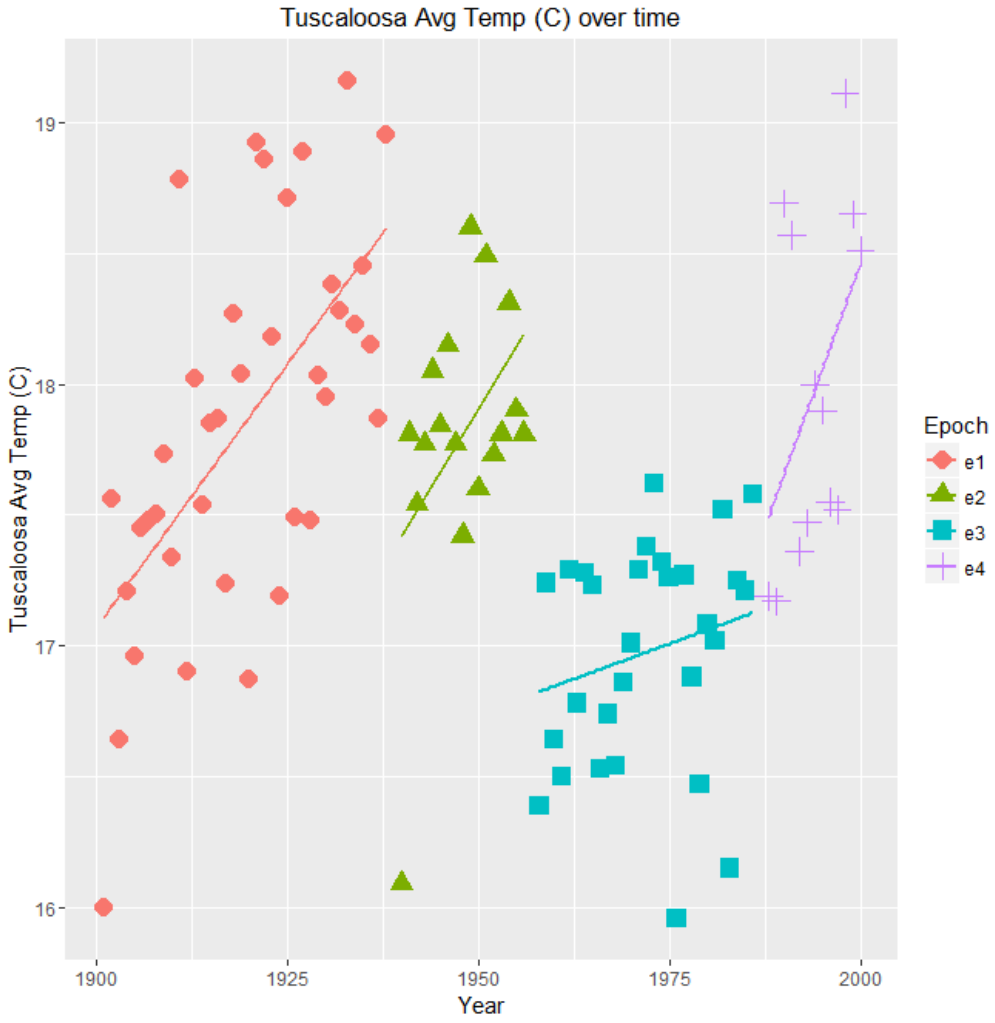
   $$\text{AvgTemp} \sim \text{Year} + \text{Epoch} + \text{Year}*\text{Epoch}$$

   - to check if the change in AvgTemp per year is consistent among Epochs
3. Then we fit the model:

   $$\text{AvgTemp} \sim \text{Year} + \text{Epoch}$$

   - to estimate the common trend

Fit a separate line for each epoch and plot them



- a potentially differential slope in the 3rd epoch

The non-parallel slope model

```
> tusctemp.model.np <- lm( Avg.Temp..C. ~ Epoch + Year + Year:Epoch,
data=tusctemp)
> anova(tusctemp.model.np)
Analysis of Variance Table

Response: Avg.Temp..C.
           Df  Sum Sq Mean Sq F value    Pr(>F)
Epoch       3 16.3005  5.4335 19.6963 7.037e-10 ***
Year        1  8.0230  8.0230 29.0833 5.652e-07 ***
Epoch:Year  3  1.7481  0.5827  2.1123    0.1043
Residuals  89 24.5519  0.2759
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is no strong evidence that the slopes are different among the epochs ($p = 0.10$)

Model with common slopes

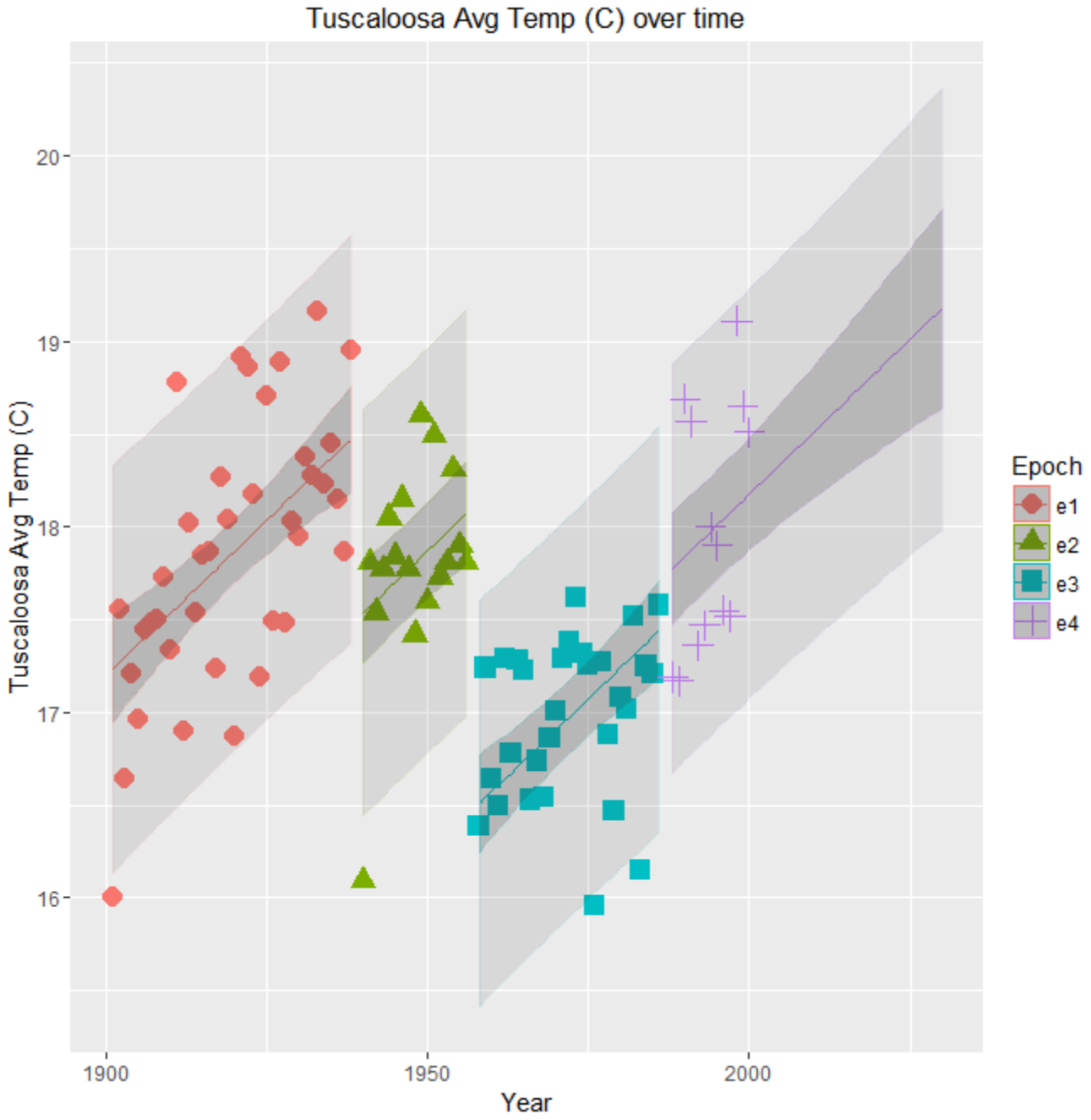
- Specify the Epoch term last to get the proper test for Epoch effects
- Be sure that Epoch has been declared as a factor.
- Notice that the anova() table term for Year is NOT useful

```
> tusctemp.model.p <- lm( Avg.Temp..C. ~ Year + Epoch, data=tusctemp)
```

Model with common slopes

```
> summary(tusctemp.model.p)

Call:
lm(formula = Avg.Temp..C. ~ Year + Epoch, data = tusctemp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44805 -0.26254  0.02385  0.33341  1.21079

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -46.269008  12.104146  -3.823  0.00024 ***
Year          0.033406   0.006306   5.298 7.97e-07 ***
Epoche2      -0.999925   0.237982  -4.202 6.12e-05 ***
Epoche3      -2.631438   0.356335  -7.385 6.72e-11 ***
Epoche4      -2.365726   0.500203  -4.730 8.09e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 0.5347 on 92 degrees of freedom
Multiple R-squared:  0.4805,    Adjusted R-squared:  0.4579
F-statistic: 21.27 on 4 and 92 DF,  p-value: 1.912e-12
```

The estimated change in average temperature is 0.033 (SE 0.006) per year.
The 95% confidence interval does not cover 0.
→Good evidence that the common slope is different from 0


Tuscaloosa Avg Temp (C) over time
Tuscaloosa Avg Temp (C)
20
19
18
17
16
1900
1950
2000
Year
Epoch
e1
e2
e3
e4