Today

- HW 1: due February 4, 11.59 pm.
- Regression with count data
- Forestry experiment and dose-response modelling
- In the News: “High water mark: the rise in sea levels may be accelerating” Economist, Jan 17
- Cancer bad luck: Data analysis here
- “Prolonged sitting raises the risk of disease”, Globe & Mail, Jan. 21 online

Responses are counts

ELM, Ch. 3

- responses take values $0, 1, 2, \ldots$
- simplest model is $Y \sim \text{Poisson}(\mu)$
- $$f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots ; \mathrm{E}(Y) = \mathrm{var}(Y) = \mu$$
- can be used in preference to Binomial, with large *n* and small *p*
- when events occur at exponentially distributed times, the number of events in a give time period follows a Poisson distributions
- if events occur in a Poisson process, in time or in space, the number of events in a given time interval or spatial area follows a Poisson distribution
- examples: counts of cancer cases in a geographical area; calls arriving at a service centre, occurrence of earthquakes, ...

Poisson regression

- $y_i \sim \text{Poisson}(\mu_i), \quad i = 1, \dots, n$
- $\log(\mu_i) = x_i^{\mathrm{T}}\beta$: log-link
- $$\ell(\beta) = \sum_{i=1}^{n}\{y_i x_i^{\mathrm{T}}\beta - \exp(x_i^{\mathrm{T}}\beta)\}$$
- if count is number falling into some level of a given category then multinomial or binomial is appropriate
- $Y_1 \sim \text{Poisson}(\mu_1), Y_2 \sim \text{Poisson}(\mu_2)$: independent
  - $Y_1 + Y_2 \sim \text{Poisson}(\mu_1 + \mu_2)$
  - $Y_1 | Y_1 + Y_2 \sim \text{Binomial}\{y_1 + y_2, \mu_1/(\mu_1 + \mu_2)\}$
- maximum likelihood estimator:

$$\sum (y_i - e^{x_i^{\mathrm{T}}\hat{\beta}})x_i^{\mathrm{T}} = 0$$

- $$\sum y_i x_i^{\mathrm{T}} = \sum \mu_i(\hat{\beta})x_i^{\mathrm{T}}$$

... Poisson regression

- saturated model $y_i \sim \text{Poisson}(\mu_i)$
- residual deviance

$$2\{\ell(\tilde{\mu}; y) - \ell(\hat{\mu}; y)\} = \sum \{y_i \log y_i - y_i - y_i \log \mu_i(\hat{\beta}) + \mu_i(\hat{\beta})\}$$

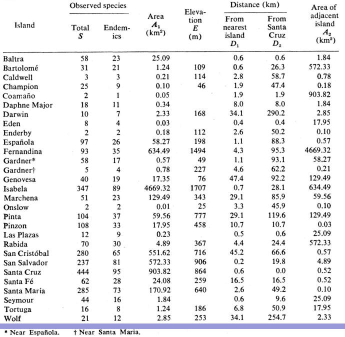
- as with binomial, can be used as a test of model adequacy
- as with binomial, can be approximated by Pearson X^2:

$$X^2 = \sum_{i=1}^{n} \frac{\{y_i - \mu_i(\hat{\beta})\}^2}{\mu_i(\hat{\beta})}$$

Example

ELM p.56ff

```
> library(faraway); data(gala)
> head(gala)
             Species Endemics  Area Elevation Nearest Scruz Adjacent
Baltra            58       23 25.09       346     0.6   0.6     1.84
Bartolome         31       21  1.24       109     0.6  26.3   572.33
Caldwell           3        3  0.21       114     2.8  58.7     0.78
Champion          25        9  0.10        46     1.9  47.4     0.18
Coamano            2        1  0.05        77     1.9   1.9   903.82
Daphne.Major      18       11  0.34       119     8.0   8.0     1.84
> ?gala
> dim(gala)
[1] 30  7
> gala <- gala[,-2] # remove variable "Endemics"
```



Island	Observed species Total S	 Endemics	Area A_1 (km²)	Elevation E (m)	Distance (km) From nearest island D_1	 From Santa Cruz D_2	Area of adjacent island A_2 (km²)
Baltra	58	23	25.09		0.6	0.6	1.84
Bartolomé	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamaño	2	1	0.05		1.9	1.9	903.82
Daphne Major	18	11	0.34		8.0	8.0	1.84
Darwin	10	7	2.33	168	34.1	290.2	2.85
Eden	8	4	0.03		0.4	0.4	17.95
Enderby	2	2	0.18	112	2.6	50.2	0.10
Española	97	26	58.27	198	1.1	88.3	0.57
Fernandina	93	35	634.49	1494	4.3	95.3	4669.32
Gardner*	58	17	0.57	49	1.1	93.1	58.27
Gardner†	5	4	0.78	227	4.6	62.2	0.21
Genovesa	40	19	17.35	76	47.4	92.2	129.49
Isabela	347	89	4669.32	1707	0.7	28.1	634.49
Marchena	51	23	129.49	343	29.1	85.9	59.56
Onslow	2	2	0.01	25	3.3	45.9	0.10
Pinta	104	37	59.56	777	29.1	119.6	129.49
Pinzon	108	33	17.95	458	10.7	10.7	0.03
Las Plazas	12	9	0.23		0.5	0.6	25.09
Rabida	70	30	4.89	367	4.4	24.4	572.33
San Cristóbal	280	65	551.62	716	45.2	66.6	0.57
San Salvador	237	81	572.33	906	0.2	19.8	4.89
Santa Cruz	444	95	903.82	864	0.6	0.0	0.52
Santa Fé	62	28	24.08	259	16.5	16.5	0.52
Santa Maria	285	73	170.92	640	2.6	49.2	0.10
Seymour	44	16	1.84		0.6	9.6	25.09
Tortuga	16	8	1.24	186	6.8	50.9	17.95
Wolf	21	12	2.85	253	34.1	254.7	2.33

* Near Española. † Near Santa Maria.

Species Number and Endemism: The Galápagos Archipelago Revisited Author(s): Michael P. Johnson and Peter H. Raven Source: Science, New Series, Vol. 179, No. 4076 (Mar. 2, 1973), pp. 893-895

... example

see ELM for fit of linear model for `Species` and $\sqrt{\texttt{Species}}$

```
> modp <- glm(Species ~ ., data = gala, family = poisson)
> summary(modp)
...
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68

> modp2 <- glm(Species ~ ., data = gala, family = quasipoisson)
> summary(modp2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.1548079  0.2915901  10.819 1.03e-10 ***
Area        -0.0005799  0.0001480  -3.918 0.000649 ***
Elevation    0.0035406  0.0004925   7.189 1.98e-07 ***
Nearest      0.0088256  0.0102622   0.860 0.398292
Scruz       -0.0057094  0.0035251  -1.620 0.118380
Adjacent    -0.0006630  0.0001653  -4.012 0.000511 ***
---

(Dispersion parameter for quasipoisson family taken to be 31.74921)
```

... example

– see p.61 where dispersion computed directly from pearson residuals

```
> sum(residuals(modp,"pearson")^2/24)
[1] 31.74914
```

– note that using quasi-Poisson gives *p*-values based on *t*-distribution

– this is by analogy with normal theory linear regression

– could also drop terms and compare scaled deviances to F distribution ELM p.61

– another way to handle over dispersion is to use negative binomial model ELM §3.3

```
> library(MASS)
> modn <- glm.nb(Species ~ ., data = gala)
> summary(modn)
```

– gives results broadly consistent with quasi-Poisson

Poisson process

SM Ch. 6.5

- observe process $\{N(t), t \in (0, t_0]\}$ which counts events; i.e. $N(t)$ is the number of events occurring between time 0 and time t
- require:
  1. $\Pr[N(t+h) - N(t) = 1] = \lambda(t)h + o(h)$
  2. $\Pr[N(t+h) - N(t) = 0] = 1 - \lambda(t)h + o(h)$
  3. events in disjoint subsets of $(0, t_0]$ are independent
- then $\{N(t), t \in (0, t_0]\}$ is a (non-homogeneous) Poisson process with rate $\lambda(t)$
- can show that

$$\Pr\{N(t_0) = n\} = \frac{\{\Lambda(t_0)\}^n}{n!} \exp\{-\Lambda(t_0)\},$$

- $\Lambda(t_0) = \int_0^{t_0} \lambda(t) dt$
- if $\lambda(t) = \lambda$, then $\Lambda(t) = \lambda t$, and the number of points in $(0, t_0]$ is Poisson with mean λt_0

Rate models

ELM Ch. 3.2

- number of events in $(0, t_0)$ follows a Poisson with mean λt_0
- i.e. $\mu = \lambda t_0$, $\log(\mu) = \log(\lambda) + \underbrace{\log(t_0)}_{\text{fixed}}$

```
> data(dicentric)
> head(dicentric)
  cells  ca doseamt doserate
1   478  25       1     0.10
2  1907 102       1     0.25
3  2258 149       1     0.50
4  2329 160       1     1.00
5  1238  75       1     1.50
6  1491 100       1     2.00
> ?dicentric

> modr <- glm(ca ~ log(doserate)*factor(doseamt) + offset(log(cells)),
+  family = poisson, data = dicentric)
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -2.74671    0.03426 -80.165  < 2e-16 ***
log(doserate)                     0.07178    0.03518   2.041 0.041299 *
factor(doseamt)2.5                1.62542    0.04946  32.863  < 2e-16 ***
factor(doseamt)5                  2.76109    0.04349  63.491  < 2e-16 ***
log(doserate):factor(doseamt)2.5  0.16122    0.04830   3.338 0.000844 ***
log(doserate):factor(doseamt)5    0.19350    0.04243   4.561  5.1e-06 ***
---
(Dispersion parameter for poisson family taken to be 1)
   Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:   21.75  on 21  degrees of freedom
```

Log-linear models

SM §10.5.1

```
> data(soccer)
> head(soccer)
  month day year     team1       team2 score1 score2
1   Aug  19 2000  Charlton ManchesterC      4      0
2   Aug  19 2000   Chelsea     WestHam      4      2
3   Aug  19 2000  Coventry   Middlesbr      1      3
4   Aug  19 2000     Derby Southampton      2      2
5   Aug  19 2000     Leeds     Everton      2      0
6   Aug  19 2000 Leicester  AstonVilla      0      0
> ?soccer
> dim(soccer)
[1] 380   7
> with(soccer, levels(team1))
 [1] "Arsenal"     "AstonVilla"  "Bradford"    "Charlton"    "Chelsea"     "Coventry"
 [7] "Derby"       "Everton"     "Ipswich"     "Leeds"       "Leicester"   "Liverpool"
[13] "ManchesterC" "ManchesterU" "Middlesbr"   "Newcastle"   "Southampton" "Sunderland"
[19] "Tottenham"   "WestHam"
```

$$y_{ij}^h \sim \text{Poisson}(\mu_{ij}^h), \quad y_{ij}^a \sim \text{Poisson}(\mu_{ij}^a)$$

score home/away

$$\mu_{ij}^h = \exp(\Delta + \alpha_i + \beta_j), \quad \mu_{ij}^a = \exp(\alpha_j - \beta_i)$$

α_i: offensive strength β_j: defensive strength Δ: home advantage

… soccer

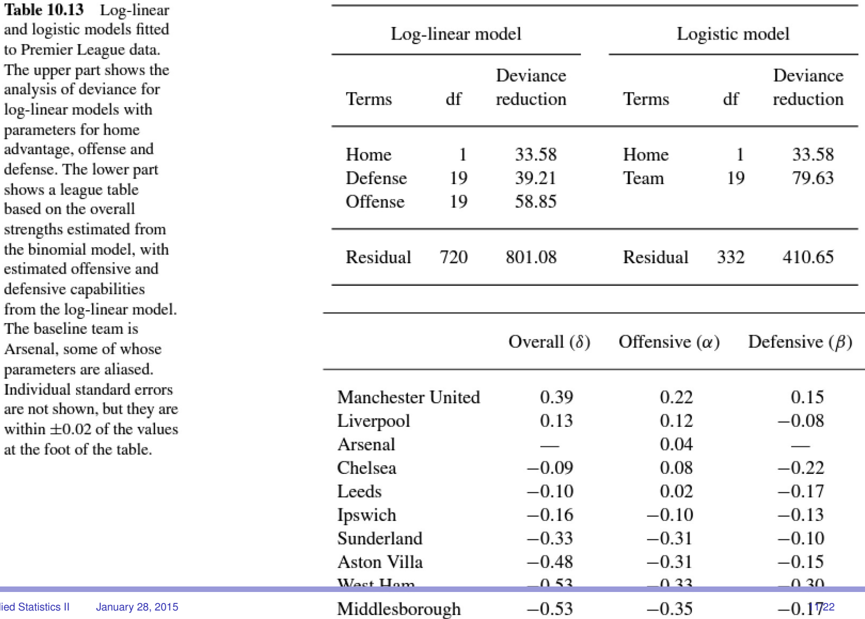


Table 10.13 Log-linear and logistic models fitted to Premier League data. The upper part shows the analysis of deviance for log-linear models with parameters for home advantage, offense and defense. The lower part shows a league table based on the overall strengths estimated from the binomial model, with estimated offensive and defensive capabilities from the log-linear model. The baseline team is Arsenal, some of whose parameters are aliased. Individual standard errors are not shown, but they are within ±0.02 of the values at the foot of the table.

Log-linear model			Logistic model		
Terms	df	Deviance reduction	Terms	df	Deviance reduction
Home	1	33.58	Home	1	33.58
Defense	19	39.21	Team	19	79.63
Offense	19	58.85			
Residual	720	801.08	Residual	332	410.65

	Overall (δ)	Offensive (α)	Defensive (β)
Manchester United	0.39	0.22	0.15
Liverpool	0.13	0.12	−0.08
Arsenal	—	0.04	—
Chelsea	−0.09	0.08	−0.22
Leeds	−0.10	0.02	−0.17
Ipswich	−0.16	−0.10	−0.13
Sunderland	−0.33	−0.31	−0.10
Aston Villa	−0.48	−0.31	−0.15
West Ham	−0.53	−0.33	−0.30
Middlesborough	−0.53	−0.35	−0.17

... soccer

- $y_{ij}^{h} \sim \text{Poisson}(\mu_{ij}^{h}), \quad y_{ij}^{a} \sim \text{Poisson}(\mu_{ij}^{a})$ score home/away
- $\mu_{ij}^{h} = \exp(\Delta + \alpha_i + \beta_j), \quad \mu_{ij}^{a} = \exp(\alpha_j - \beta_i)$

- A different analysis: home score, given total score: $y_{ij}^{h} \mid y_{ij}^{h} + y_{ij}^{a}$

- Binomial with

$$p_{ij} = \frac{\mu_{ij}^{h}}{\mu_{ij}^{h} + \mu_{ij}^{a}} = \frac{\exp\{\Delta + (\overbrace{\alpha_i + \beta_i}^{\delta_i}) - (\alpha_j + \beta_j)\}}{1 + \exp\{\Delta + (\alpha_i + \beta_i) - (\alpha_j + \beta_j)\}}$$

- Games tied at 0 contribute no information
- δ_i is the ‘overall strength’ of team i – can no longer distinguish defensive and offensive
- Analysis based on logistic regression
- R code?

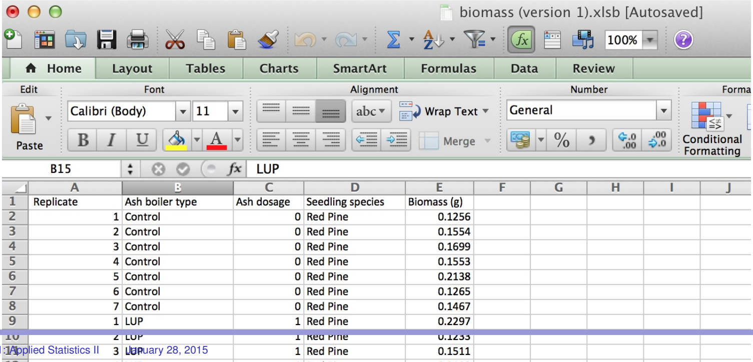
Some data sources

- Iowa State University Stats Dept, Graphics group, has links to the Recovery Act spending
- Kaggle has various competition data sets
- The OECD ran a PISA test visualization contest; the Iowa State Group has some information about it as well.
- ICPSR The Inter-University Consortium for Political and Social Research provides data by topic, geography, etc., including international data
- Canada's Open Government website has an open data portal
- Flowing Data has a handful of haphazard data sets
- World Mapper has links to several data sources
- This ASA site has links to several data sources about 3/4 of the way down the page
- for sports fans

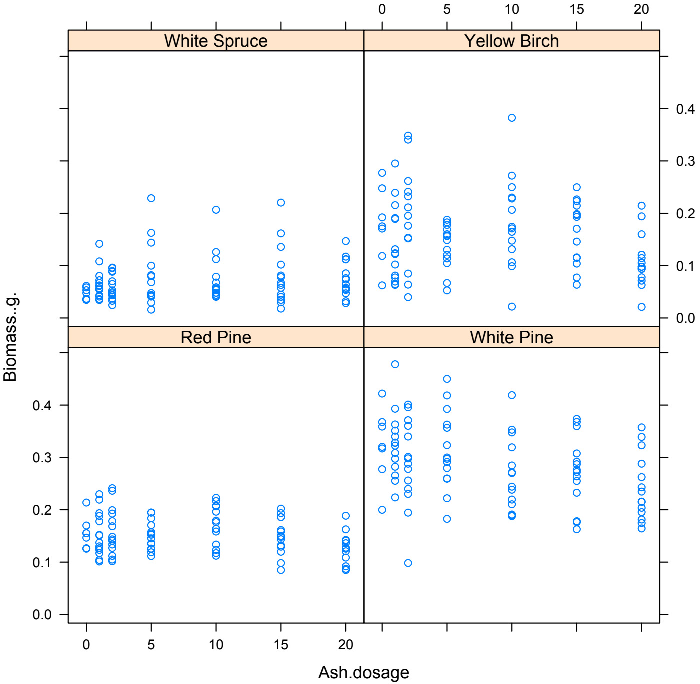
Dose-response curves

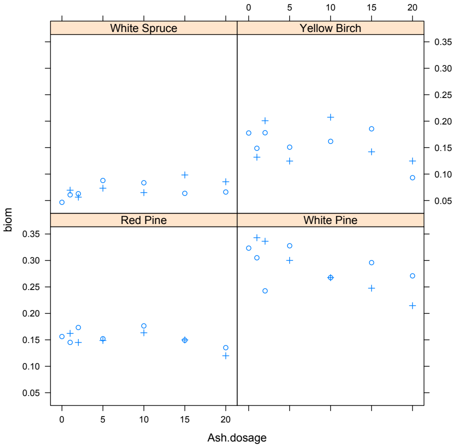
The attached dataset file contains 4 variables: Response: Biomass Explanatory: Dosage(0,1,2,5,10,15,20), Seedling Species(Category), Ash Boiler Type(Category)

The goal of the experiment is to come up with a dose response curve for each seedling species. Since Ash Boiler Type was shown to be non-significant in the regression model, it was neglected.

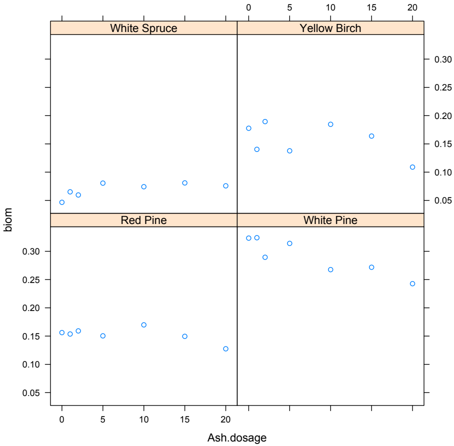


	A	B	C	D	E	F	G	H	I	J
1	Replicate	Ash boiler type	Ash dosage	Seedling species	Biomass (g)					
2	1	Control	0	Red Pine	0.1256					
3	2	Control	0	Red Pine	0.1554					
4	3	Control	0	Red Pine	0.1699					
5	4	Control	0	Red Pine	0.1553					
6	5	Control	0	Red Pine	0.2138					
7	6	Control	0	Red Pine	0.1265					
8	7	Control	0	Red Pine	0.1467					
9	1	LUP	1	Red Pine	0.2297					
10	2	LUP	1	Red Pine	0.1233					
11	3	LUP	1	Red Pine	0.1511					

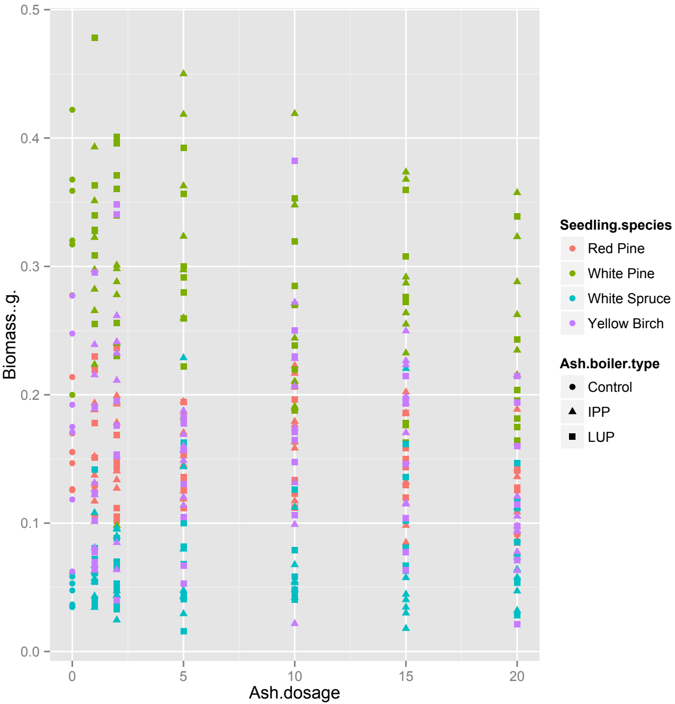

White Spruce
Yellow Birch
Red Pine
White Pine
Biomass..g.
Ash.dosage
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4

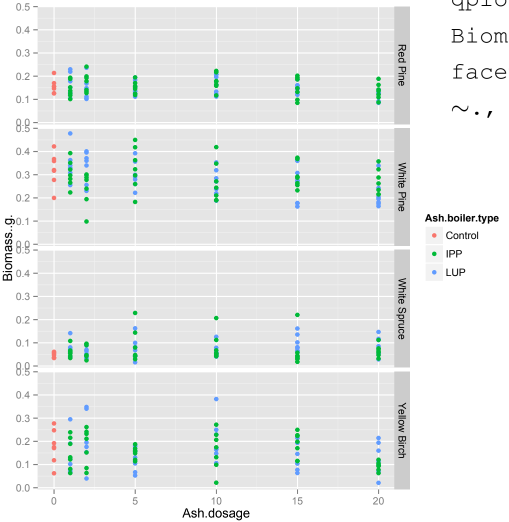


```
meantrees <-
ddply(trees,.(Seedling.species,
Ash.boiler.type, Ash.dosage),summarize,
biom = sum(Biomass..g.))

xyplot(biom ~Ash.dosage |

Seedling.species, data = meantrees)
```

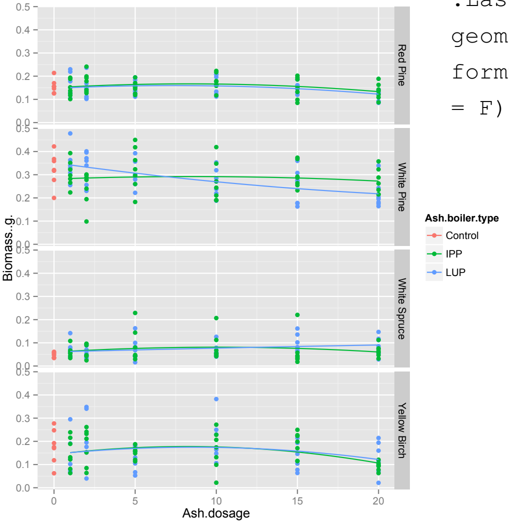


```
meantrees2 <-
ddply(trees,.(Seedling.species,
Ash.dosage),summarize, biom =
mean(Biomass..g.))

xyplot(biom ∼Ash.dosage |

Seedling.species, data = meantrees2)
```


0.5
0.4
0.3
0.2
0.1
0.0
Biomass..g.
0
5
10
15
20
Ash.dosage
Seedling.species
Red Pine
White Pine
White Spruce
Yellow Birch
Ash.boiler.type
Control
IPP
LUP



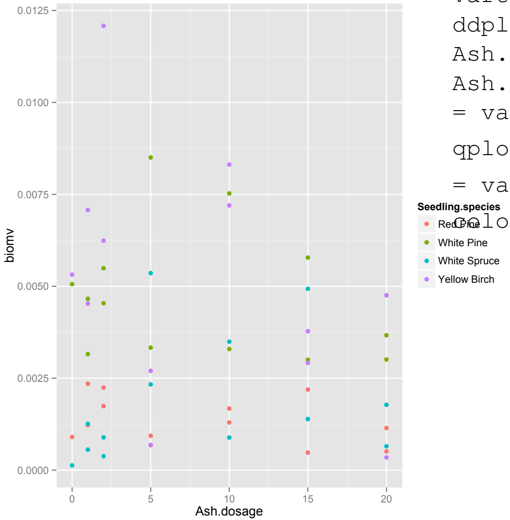
```
qplot(Ash.dosage,
Biomass..g., data = trees,
facets = Seedling.species
~., color = Ash.boiler.type)
```



```
.Last.value +
geom_smooth(method = "lm",
formula = y  ~poly(x,2), se
= F)
```

linear models

```
# straight lines for each species, all with same slope:

trees.lm <- lm(formula = Biomass..g. ~ Ash.dosage + Seedling.species, data = trees) #

# orthogonal polynomials as in class

trees.lm2 <- lm(formula = Biomass..g. ~ as.ordered(Ash.dosage) + Seedling.species,
    data = trees)

# ordinary quadratics (no indication that any higher orders are needed

trees.lm3 <- lm(formula = Biomass..g. ~ Ash.dosage + I(Ash.dosage^2) + Seedling.species,
    data = trees)

# this allows a different slope for each species

trees.lm4 <- lm(formula = Biomass..g. ~ Ash.dosage * Seedling.species, data = trees)

# and a different quadratic for each species

trees.lm5 <- lm(formula = Biomass..g. ~ poly(Ash.dosage, 2) * Seedling.species,
    data = trees)
```



```
vartrees <-
ddply(trees,.(Seedling.species,
Ash.boiler.type,
Ash.dosage),summarize, biomv
= var(Biomass..g.))
qplot(Ash.dosage,biomv,data
= vartrees,
color=Seedling.species)
```