Poisson Regression with the Language Development Data

```
> lang =
read.table("http://www.utstat.toronto.edu/~brunner/312f10/code_n_data/language.data
")
> lang[1:5,]
  age    sex vocab subind  mlu errors
1  58   Male    19   1.00 2.33      5
2  58   Male    17   1.04 5.29      0
3  47 Female    14   1.10 7.10      0
4  60 Female    62   1.32 7.45      0
5  58   Male    15   1.00 2.00      0
> table(lang$errors)

 0  1  2  3  4  5 
57 19 10 13  4  3 
> table(lang$sex)

Female   Male 
    56     50 
> lang$sex <- factor(lang$sex,levels=c("Male","Female"))
> table(lang$sex)

  Male Female 
    50     56 
> 
> redmodel <- glm(errors ~ age+sex, data=lang, family=poisson)
> summary(redmodel)

Call:
glm(formula = errors ~ age + sex, family = poisson, data = lang)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7285  -1.3630  -1.1685   0.4816   3.2812  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept)  0.82604    0.50535   1.635   0.1021  
age         -0.01980    0.01022  -1.937   0.0528 .
sexFemale    0.22875    0.19655   1.164   0.2445  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 199.98  on 105  degrees of freedom
Residual deviance: 194.16  on 103  degrees of freedom
AIC: 326.70

Number of Fisher Scoring iterations: 6

> fullmodel <- update(redmodel, . ~ . + vocab+subind+mlu)
> summary(fullmodel)

Call:
glm(formula = errors ~ age + sex + vocab + subind + mlu, family = poisson, 
    data = lang)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8604  -1.3730  -0.9666   0.5994   2.6089

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.5856125  1.1366050   2.275   0.0229 *
age         -0.0003854  0.0124018  -0.031   0.9752
sexFemale    0.2709556  0.1989416   1.362   0.1732
vocab       -0.0107604  0.0090945  -1.183   0.2367
subind      -2.1901794  1.3049854  -1.678   0.0933 .
mlu         -0.0017284  0.0963611  -0.018   0.9857
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 199.98  on 105  degrees of freedom
Residual deviance: 186.17  on 100  degrees of freedom
AIC: 324.70

Number of Fisher Scoring iterations: 6

> anodev = anova(redmodel,fullmodel); anodev
Analysis of Deviance Table

Model 1: errors ~ age + sex
Model 2: errors ~ age + sex + vocab + subind + mlu
  Resid. Df Resid. Dev  Df Deviance
1       103    194.160
2       100    186.168   3    7.992
> G2 = anodev[2,4]; df = anodev[2,3]; pval = 1-pchisq(G2,df)
> cat("\n G-squared = ",G2,"  df = ", df, "  p = ",pval,"\n\n")

 G-squared =  7.992272   df =  3   p =  0.0461717

>
> qchisq(0.95,df=1) # Critical value for alpha=0.05, df=1
[1] 3.841459
> anova(fullmodel)
Analysis of Deviance Table

Model: poisson, link: log

Response: errors

Terms added sequentially (first to last)


        Df  Deviance Resid. Df Resid. Dev
NULL                       105    199.978
age      1     4.446       104    195.532
sex      1     1.371       103    194.160
vocab    1     3.069       102    191.092
subind   1     4.923       101    186.168
mlu      1 0.0003217       100    186.168

> model2 = update(fullmodel, . ~ . - sex - mlu)
> summary(model2)

Call:
glm(formula = errors ~ age + vocab + subind, family = poisson,
    data = lang)
```

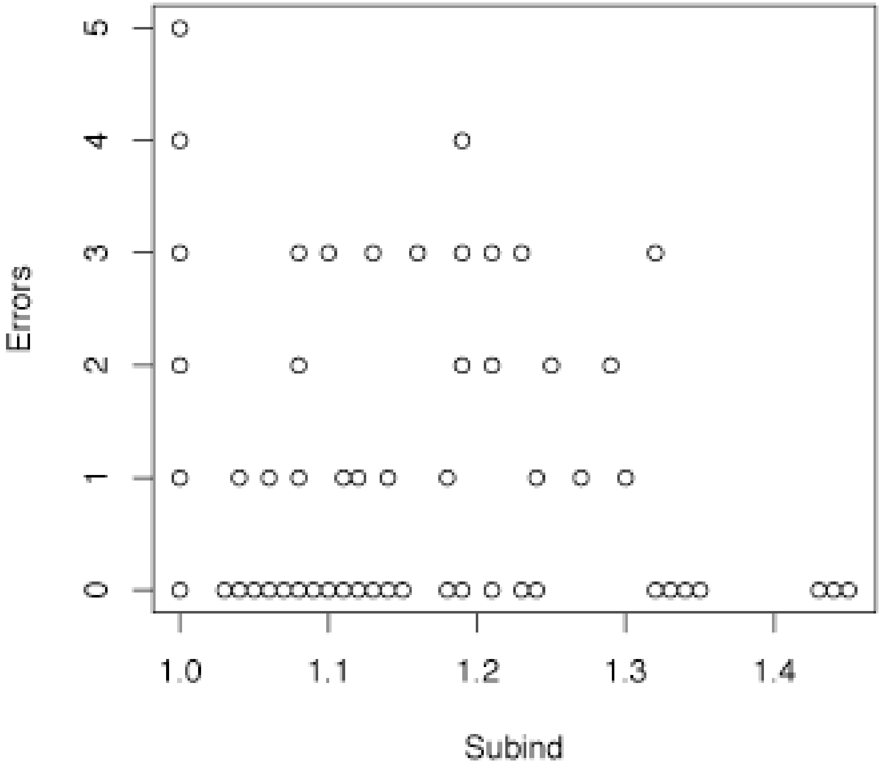
```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7693  -1.3737  -0.9755   0.6870   2.4658

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.687582   1.042445   2.578  0.00993 **
age         -0.002671   0.012124  -0.220  0.82565
vocab       -0.011167   0.008989  -1.242  0.21413
subind      -2.045566   1.028374  -1.989  0.04669 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 199.98  on 105  degrees of freedom
Residual deviance: 188.06  on 102  degrees of freedom
AIC: 322.60

Number of Fisher Scoring iterations: 6
>
> redmodel2 = update(model2, . ~ . - vocab - age)
> formula(redmodel2)
errors ~ subind
> anova(redmodel2,model2)
Analysis of Deviance Table

Model 1: errors ~ subind
Model 2: errors ~ age + vocab + subind
  Resid. Df Resid. Dev  Df Deviance
1       104    190.417
2       102    188.060   2    2.357
> qchisq(0.95,df=2) # Critical value for alpha=0.05, df=2
[1] 5.991465

> # Try stepwise selection
> null = glm(errors ~ 1, data=lang, family=poisson)
> stepmod <- step(null, scope=list(lower=formula(null),upper=formula(fullmodel)),
direction="both")

Start:  AIC= 328.51
 errors ~ 1

         Df Deviance    AIC
+ subind  1   190.42 320.95
+ vocab   1   193.02 323.56
+ mlu     1   194.28 324.81
+ age     1   195.53 326.07
+ sex     1   197.94 328.47
<none>        199.98 328.51

Step:  AIC= 320.95
 errors ~ subind

         Df Deviance    AIC
+ sex     1   188.01 320.54
+ vocab   1   188.11 320.64
<none>        190.42 320.95
+ age     1   189.64 322.17
+ mlu     1   190.35 322.89
- subind  1   199.98 328.51

Step:  AIC= 320.54
 errors ~ subind + sex
```

```
         Df Deviance    AIC
<none>        188.01 320.54
+ vocab   1   186.17 320.71
- sex     1   190.42 320.95
+ age     1   187.63 322.16
+ mlu     1   187.92 322.45
- subind  1   197.94 328.47

> summary(stepmod)

Call:
glm(formula = errors ~ subind + sex, family = poisson, data = lang)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7514  -1.4167  -0.9295   0.5218   2.6642  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)   2.9182     1.0314   2.829  0.00467 **
subind       -2.7915     0.9497  -2.939  0.00329 **
sexFemale     0.3010     0.1954   1.540  0.12347   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 199.98  on 105  degrees of freedom
Residual deviance: 188.01  on 103  degrees of freedom
AIC: 320.54

Number of Fisher Scoring iterations: 6

> poissonmodel = update(stepmod, ~ . - sex); summary(poissonmodel)

Call:
glm(formula = errors ~ subind, family = poisson, data = lang)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6346  -1.4609  -0.9102   0.5347   2.5505  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)   3.0128     1.0238   2.943  0.00325 **
subind       -2.7232     0.9438  -2.885  0.00391 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 199.98  on 105  degrees of freedom
Residual deviance: 190.42  on 104  degrees of freedom
AIC: 320.95

Number of Fisher Scoring iterations: 6
```

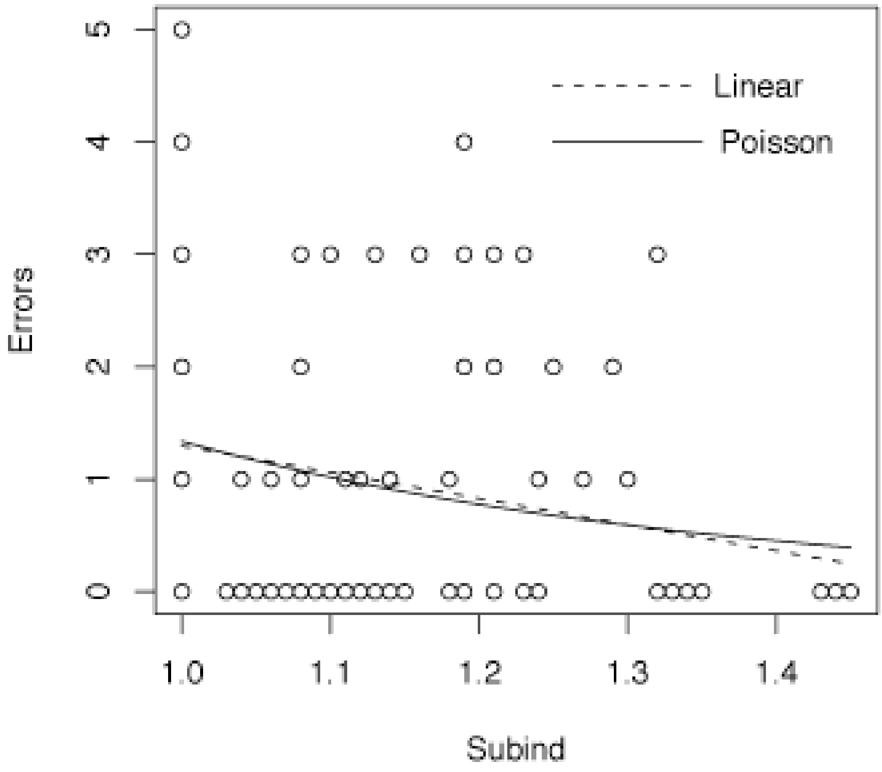
```
> # Compare regression with normal error terms: Had p = 0.0461717
> linfull = lm(formula(fullmodel), data=lang)
> linred = lm(formula(redmodel), data=lang)
> anova(linred,linfull)
Analysis of Variance Table

Model 1: errors ~ age + sex
Model 2: errors ~ age + sex + vocab + subind + mlu
  Res.Df     RSS  Df Sum of Sq      F Pr(>F)
1    103 196.992
2    100 189.808   3     7.184 1.2617 0.2917


> # Try stepwise selection
> nolin = lm(errors~1, data=lang)
> steplin1 = step(nolin,scope=list(lower=formula(nolin),upper=formula(linfull)),
direction="both")
Start:  AIC= 70.83
 errors ~ 1

         Df Sum of Sq     RSS     AIC
+ subind  1     8.755 194.160  68.156
+ vocab   1     6.697 196.218  69.273
+ mlu     1     5.934 196.982  69.685
+ age     1     4.572 198.343  70.415
<none>                202.915  70.831
+ sex     1     2.082 200.834  71.738

Step:  AIC= 68.16
 errors ~ subind

         Df Sum of Sq     RSS     AIC
<none>                194.160  68.156
+ vocab   1     2.447 191.713  68.811
+ sex     1     2.328 191.832  68.877
+ age     1     1.050 193.110  69.581
+ mlu     1     0.169 193.991  70.064
- subind  1     8.755 202.915  70.831
> summary(steplin1)

Call:
lm(formula = errors ~ subind, data = lang)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2952 -1.1028 -0.2952  0.7048  3.7048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.628      1.208   3.004  0.00334 **
subind        -2.333      1.077  -2.166  0.03263 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.366 on 104 degrees of freedom
Multiple R-Squared: 0.04315,  Adjusted R-squared: 0.03395
F-statistic:  4.69 on 1 and 104 DF,  p-value: 0.03263
```

```
> # Starting backwards
> steplin2 = step(linfull,scope=list(lower=formula(nolin),upper=formula(linfull)),
direction="both")
Start:  AIC= 73.75
 errors ~ age + sex + vocab + subind + mlu

         Df Sum of Sq     RSS    AIC
- mlu     1     0.009 189.817 71.758
- age     1     0.020 189.827 71.764
- vocab   1     1.340 191.148 72.498
- sex     1     1.795 191.603 72.751
- subind  1     2.351 192.159 73.057
<none>                189.808 73.753

Step:  AIC= 71.76
 errors ~ age + sex + vocab + subind

         Df Sum of Sq     RSS    AIC
- age     1     0.025 189.842 69.772
- vocab   1     1.393 191.210 70.533
- sex     1     1.786 191.603 70.751
<none>                189.817 71.758
- subind  1     4.378 194.195 72.175
+ mlu     1     0.009 189.808 73.753

Step:  AIC= 69.77
 errors ~ sex + vocab + subind

         Df Sum of Sq     RSS    AIC
- sex     1     1.870 191.713 68.811
- vocab   1     1.990 191.832 68.877
<none>                189.842 69.772
- subind  1     4.881 194.724 70.463
+ age     1     0.025 189.817 71.758
+ mlu     1     0.015 189.827 71.764

Step:  AIC= 68.81
 errors ~ vocab + subind

         Df Sum of Sq     RSS    AIC
- vocab   1     2.447 194.160 68.156
<none>                191.713 68.811
- subind  1     4.505 196.218 69.273
+ sex     1     1.870 189.842 69.772
+ age     1     0.110 191.603 70.751
+ mlu     1     0.004 191.709 70.809

Step:  AIC= 68.16
 errors ~ subind

         Df Sum of Sq     RSS    AIC
<none>                194.160 68.156
+ vocab   1     2.447 191.713 68.811
+ sex     1     2.328 191.832 68.877
+ age     1     1.050 193.110 69.581
+ mlu     1     0.169 193.991 70.064
- subind  1     8.755 202.915 70.831
>
```

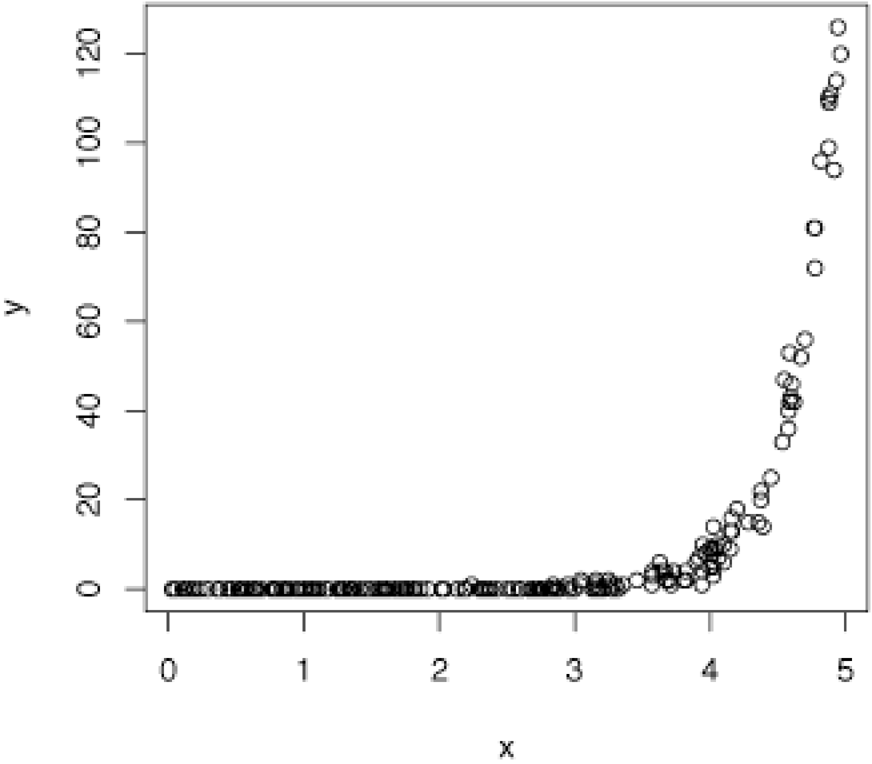
```
> linearmodel = steplin2; summary(linearmodel)

Call:
lm(formula = errors ~ subind, data = lang)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.2952 -1.1028 -0.2952  0.7048  3.7048 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)    3.628      1.208   3.004  0.00334 **
subind        -2.333      1.077  -2.166  0.03263 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 1.366 on 104 degrees of freedom
Multiple R-Squared: 0.04315,  Adjusted R-squared: 0.03395 
F-statistic:  4.69 on 1 and 104 DF,  p-value: 0.03263 


> # Take a look
> Subind = lang$subind; Errors = lang$errors
> plot(Subind,Errors)
```



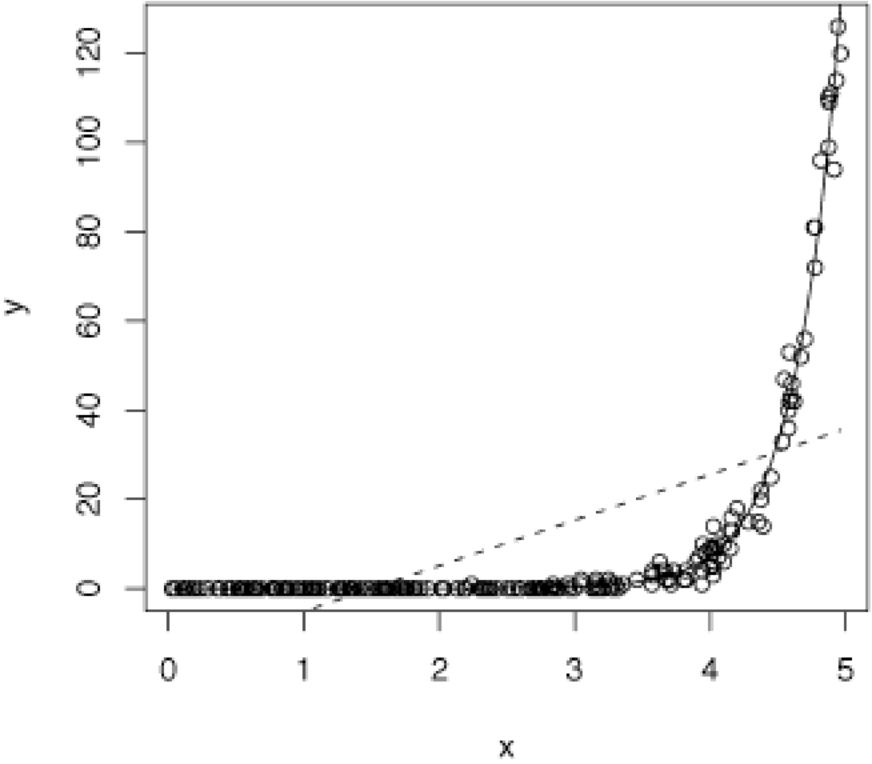
```
> poissonmodel$coefficients
(Intercept)      subind
   3.012800   -2.723178
> Subind[1]
[1] 1
> exp(sum(poissonmodel$coefficients)) # Estimated mean for case 1
[1] 1.335922
> poissonmodel$fitted.values[1] # There are N of them.
       1
1.335922
> #
> sum(linearmodel$coefficients) # b0 + b1*1
[1] 1.295238
> linearmodel$fitted.values[1] # Estimated mean for case 1
       1
1.295238
> # Good! Want to plot these curves.
> kurvdatta = cbind(Subind, poissonmodel$fitted.values,linearmodel$fitted.values)
> kurvdatta = kurvdatta[order(Subind),]; kurvdatta[1:5,]
   Subind
1       1 1.335922 1.295238
5       1 1.335922 1.295238
7       1 1.335922 1.295238
10      1 1.335922 1.295238
15      1 1.335922 1.295238

> lines(kurvdatta[,1],kurvdatta[,2],lty=1) # Solid Line
> lines(kurvdatta[,1],kurvdatta[,3],lty=2) # Dashed line
> x1 <- c(1.25,1.35) ; y1 <- c(4,4) ; lines(x1,y1,lty=1)
> text(1.4,4,"Poisson")
> x2 <- c(1.25,1.35) ; y2 <- c(4.5,4.5) ; lines(x2,y2,lty=2)
> text(1.4,4.5,"Linear    ")
```



```
> # When does it matter?
> beta0 = -10; beta1 = 3   # True parameter values
> N <- 200
> x <- sort(5*runif(N)) # In order for easier plotting
> y <- rpois(N,exp(beta0 + beta1*x)) # rpois(n, lambda)
> plot(x,y)
```



```
> poissonmodel = glm(y ~ x, family=poisson)
> linearmodel = lm(y ~ x)
> lines(x,poissonmodel$fitted.values,lty=1)
> lines(x,linearmodel$fitted.values,lty=2)
```


y
0
20
40
60
80
100
120
x
0
1
2
3
4
5