R16 – Logistic Regression
Prof Colleen F. Moore
Montana State University

Overview:

This handout covers the basics of logistic regression using R's 'glm' function and the 'binomial' family of cumulative density functions. Logistic regression is appropriate for data with a dichotomous DV. The predictors can be scaled or factors, etc. 'glm' stands for 'generalized linear models', not 'general linear model'. The models are fit by maximum likelihood, not least squares.

Quick look at code:

```
> library(aod)  # needed only for Wald test
> library(car)  # for Type II or Type III tests
> library(effects)  # for effect plots
> modeln=glm(dichotomousDV~iv1+iv2+iv3, family="binomial")
> summary(modeln)  # shows the coefficients
>  confint(modeln)  # gives 95% confidence interval of coefficients
> exp(confint(modeln)  # gives conf intervals of odds ratios
> exp(coef(modeln))  # gives odds ratios
> exp(cbind(OddRatiomodeln=coef(modeln),confint(modeln)))  # odds & conf
interval at once
```

I.A. Bring in data and poke around at descriptives.

This is a study of how spirituality is related to environmental attitudes and behavior, part of the dissertation of Andrew M. Garfield, University of Wisconsin. Participants were approached at a popular state park and asked to fill out a survey. At the end they were given an opportunity to donate to the Friends of Wisconsin State Parks nonprofit, volunteer for the park, and give their email to be passed on to the state park friends organization.

The research question is whether 'spiritual oneness' measured by a new questionnaire predicts (or is related to) proenvironmental behavior.

The dichotomous DVs we can use are donate (any at all), email (give email to interviewer).

```
> garfield=read.table(pipe("pbpaste"),header=T) # before you hit 'enter'
paste the data to the clipboard
> attach(Garfield)  # some warn not to attach
> garfield[1:5,]  # show first 5 observations
```

	ID	gndr	age	isn	donate	amtdonate	voluntr	volhrs	email	EAS	stewardsh	dombelief
1	A19	2	NA	5	0	0	0	0	0	8.5	9	3
2	A12	NA	NA	2	0	0	0	0	1	6.5	7	4
3	B14	1	NA	6	0	0	0	0	0	8.5	9	1
4	A16	NA	NA	5	0	0	0	0	0	9.0	9	6
5	B15	1	NA	6	0	0	1	2	1	8.0	9	2

	envintent	spiritOne	envprotec	dombel3	vol2
1	7.0	7.00	7.75	2.33	0
2	7.5	4.75	7.00	3.67	0
3	9.0	3.50	9.00	1.33	0
4	7.0	5.75	8.00	4.33	0
5	3.5	2.25	5.75	1.67	1

```
> summary(garfield)  # shows that indeed 'donate' and 'email' are
dichotomous. Look for out-of-range scores too.

      ID           gndr            age             isn            donate
 B14    :  2   Min.   :1.000   Min.   :18.00   Min.   :0.000   Min.   :0.0000
 A01    :  1   1st Qu.:1.000   1st Qu.:25.00   1st Qu.:4.000   1st Qu.:0.0000
 A03    :  1   Median :2.000   Median :34.00   Median :5.000   Median :0.0000
 A04    :  1   Mean   :1.528   Mean   :35.41   Mean   :5.073   Mean   :0.1124
 A05    :  1   3rd Qu.:2.000   3rd Qu.:42.50   3rd Qu.:6.000   3rd Qu.:0.0000
 A06    :  1   Max.   :2.000   Max.   :63.00   Max.   :7.000   Max.   :1.0000
 (Other):173   NA's   :4       NA's   :9       NA's   :1       NA's   :2

   amtdonate          voluntr           volhrs          email             EAS
 Min.   : 0.0000   Min.   :0.0000   Min.   : 0.00   Min.   :0.0000   Min.   :1.000
 1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.00   1st Qu.:0.0000   1st Qu.:7.000
 Median : 0.0000   Median :0.0000   Median : 0.00   Median :0.0000   Median :8.000
 Mean   : 0.4389   Mean   :0.4667   Mean   : 0.75   Mean   :0.2278   Mean   :7.683
 3rd Qu.: 0.0000   3rd Qu.:1.0000   3rd Qu.: 0.00   3rd Qu.:0.0000   3rd Qu.:8.500
 Max.   :20.0000   Max.   :2.0000   Max.   :20.00   Max.   :1.0000   Max.   :9.000

. . . (output truncated)
```

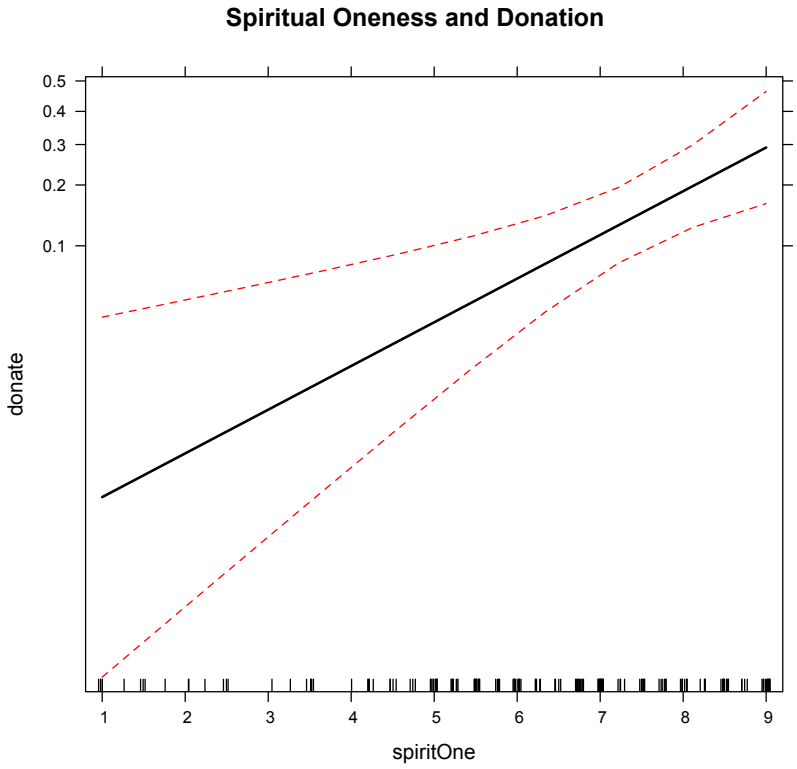
Continue with descriptives until you have 'made friends' with your data. Also, it is a good idea to look at the correlation matrix of predictor variables so that you don't inadvertently have collinearity.

II. Fit models with 'glm', and interpret. Hard Q: What covariates make sense? Should we choose covariates empirically after seeing relationships or on a priori grounds? (probably a bit of both).

II.A. Start simple with a model with a single predictor

```
> mod3=glm(donate~spiritOne,family="binomial",na.action=na.omit)
> summary(mod3)
## the z-value below is the 'Wald' test of significance. It is analogous to
the t-test for the coefficients in linear regression

Call:
glm(formula = donate ~ spiritOne, family = "binomial", na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8307  -0.5265  -0.3597  -0.2261   2.8660

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.1289     1.4059  -4.360  1.3e-05 ***
spiritOne     0.5825     0.1854   3.142  0.00168 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.46  on 175  degrees of freedom
Residual deviance: 106.82  on 174  degrees of freedom
  (4 observations deleted due to missingness)
AIC: 110.82
```

```
Number of Fisher Scoring iterations: 6

-- Get Likelihood Ratio tests, which are preferred to Wald tests given by
default. Could use 'car' package to get Type II or III tests, but with only
one predictor variable it won't matter.
> anova(mod3,test=c("LR"))  # 'LR' means Likelihood Ratio
Analysis of Deviance Table

Model: binomial, link: logit

Response: donate

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        175     120.46
spiritOne  1   13.638       174     106.82 0.0002216 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



-- Make odds ratio and confidence interval by taking the exponential
>  exp(cbind(coef(mod3),confint(mod3)))  # outputs both odds ratio & conf
intervals
Waiting for profiling to be done...
                               2.5 %     97.5 %
(Intercept) 0.00217891 9.731009e-05 0.02500139
spiritOne   1.79045825 1.284167e+00 2.66846570

## Interpretation: A one-unit increase in Spiritual Oneness increases the
odds of donating by 1.8 times.

-- Graph the effect
> library(effects)  # load this package
> plot(allEffects(mod3), main="Spiritual Oneness and Donation")
## the dashes on the x-axis are observations, and notice that the y-axis is
non-linear. The Y-axis is the fitted probability, and the overall probability
of donating is pretty low.
```



Calculate a pseudo-R-squared value using Field's function:

```
> logisticRsq <- function(LogModel)
+ {
+ dev = LogModel$deviance
+      nullDev = LogModel$null.deviance
+      modelN = length(LogModel$fitted.values)
+      R.1 <- 1 - dev / nullDev
+      R.cs <- 1- exp(-(nullDev-dev)/modelN)
+      R.n <- R.cs /(1-(exp(-(nullDev / modelN))))
+      cat("Pseudo R-sq for logistic regression\n")
+      cat("Hosmer & Lemeshow R-sq   ", round(R.1, 4),"\n")
+      cat("Cox and Snell R-sq        ", round(R.cs, 4), "\n")
+      cat("Nagelkerke R-sq           ", round(R.n,  4), "\n")
}

## delete the +'s before you paste this in. I had to retype it all in
TextWrangler

> logisticRsq(mod3)
Pseudo R-sq for logistic regression
Hosmer & Lemeshow R-sq    0.1132
Cox and Snell R-sq         0.0746
Nagelkerke R-sq            0.1504
```

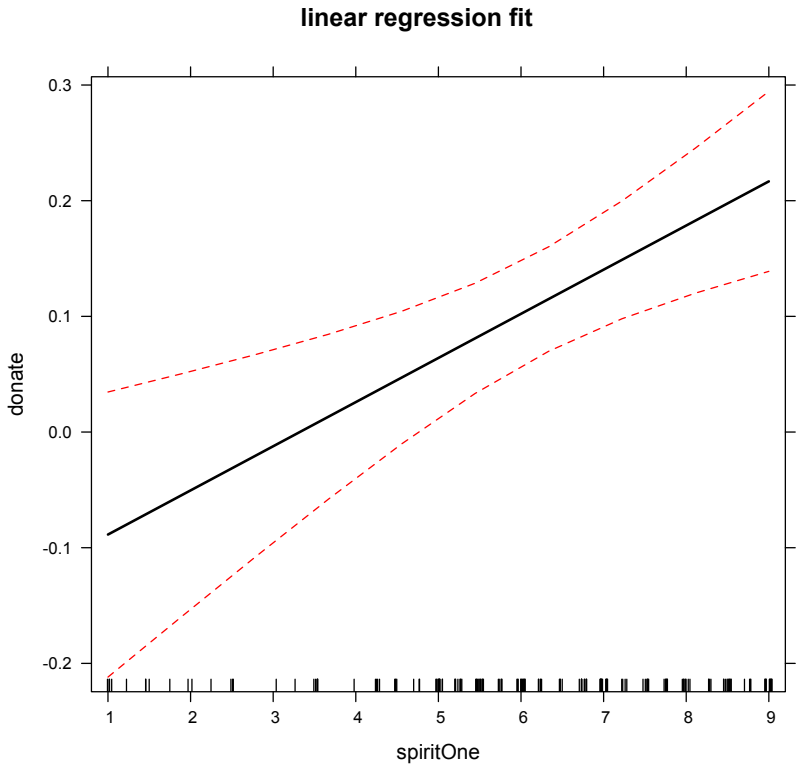
II. B. Compare to regular least squares linear regression. See your lecture notes for reasons not to use 'lm' to fit dichotomous variables.

```
> mod4=lm(donate~spiritOne,na.action=na.omit)
> summary(mod4)
```

```
Call:
lm(formula = donate ~ spiritOne, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21671 -0.14992 -0.09744 -0.03542  0.99321

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12679    0.07306  -1.735 0.084431 .
spiritOne    0.03817    0.01129   3.382 0.000889 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3023 on 174 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.06168,  Adjusted R-squared: 0.05628
F-statistic: 11.44 on 1 and 174 DF,  p-value: 0.0008892
```

interpretation: for each 1 unit increase in Spiritual Oneness, the probability of donating increases by .038. But notice that the intercept is negative.
also notice that the logistic regression gives a higher pseudo-R-sq.

```
> plot(allEffects(mod4), main="linear regression fit")  # plot effect
```
The negative probability is noticeable in the graph. A person would need spiritual oneness score of about 3.5 in order to have a zero probability of donating. Doesn’t make sense.



III. More predictor variables.
III.A. Fit a model with most available covariates, including an environmental attitude measure (EAS), gender, age, and 'dominance beliefs' (beliefs that humans should rule over nature), and 'stewardship' beliefs.

```
> mod2=glm(donate~gndr+age+EAS+stewardsh+envintent+dombelief+spiritOne,
family="binomial")
> summary(mod2)

Call:
glm(formula = donate ~ gndr + age + EAS + stewardsh + envintent +
    dombelief + spiritOne, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3376  -0.3824  -0.1803  -0.0566   2.7419

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.62295    3.94304  -2.694 0.007058 **
gndr         -1.18892    0.65734  -1.809 0.070501 .
age           0.13507    0.03509   3.849 0.000119 ***
EAS          -0.31860    0.29151  -1.093 0.274422
stewardsh    -0.16135    0.34634  -0.466 0.641302
envintent     0.78386    0.36194   2.166 0.030335 *
dombelief    -0.36932    0.21291  -1.735 0.082808 .
spiritOne     0.59368    0.21365   2.779 0.005457 **
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.346  on 166  degrees of freedom
Residual deviance:  78.754  on 159  degrees of freedom
  (13 observations deleted due to missingness)
AIC: 94.754

Number of Fisher Scoring iterations: 7
```

-- Get Likelihood Ratio tests, which are preferred to Wald tests given by default. Will use ‘car’ package to get Type II or III tests.

```
> anova(mod2,test=c("LR"))  ## order-dependent Type I tests
Analysis of Deviance Table

Model: binomial, link: logit

Response: donate

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        166    118.346
gndr       1   0.2436       165    118.103  0.621631
age        1  16.2485       164    101.854 5.555e-05 ***
EAS        1   0.9023       163    100.952  0.342174
stewardsh  1   0.1167       162    100.835  0.732612
envintent  1   8.7449       161     92.090  0.003105 **
dombelief  1   3.6477       160     88.443  0.056146 .
spiritOne  1   9.6885       159     78.754  0.001854 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> library(car)
> Anova(mod2, test=c("LR"))  # capital-A ‘Anova’ in ‘car’ package
## Notice that the p-values differ from the Type I tests above

Analysis of Deviance Table (Type II tests)

Response: donate
          LR Chisq Df Pr(>Chisq)
gndr        3.5047  1   0.061196 .
age        22.1984  1  2.459e-06 ***
EAS         1.1231  1   0.289250
stewardsh   0.1968  1   0.657319
envintent   5.2408  1   0.022063 *
dombelief   3.4885  1   0.061797 .
spiritOne   9.6885  1   0.001854 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

-- **Make odds ratio** and confidence interval by taking the exponential

If the confidence interval includes 1, it is nonsig. This is in scientific notation.

```
> exp(cbind(coef(mod2),confint(mod2)))
Waiting for profiling to be done...
                                 2.5 %     97.5 %
(Intercept) 2.435068e-05 3.447447e-09 0.02648321
gndr        3.045490e-01 7.746665e-02 1.05609886
age         1.144611e+00 1.075862e+00 1.23750589
EAS         7.271630e-01 4.094397e-01 1.34309420
stewardsh   8.509909e-01 4.689522e-01 1.96835922
envintent   2.189909e+00 1.115217e+00 4.71835162
dombelief   6.912026e-01 4.372840e-01 1.01712623
spiritOne   1.810639e+00 1.227426e+00 2.85810245
```

Interpretation: for each 1 unit increase in Spiritual Oneness score, the odds of donating increases by 1.8 times, after adjusting for gender, age, environmental attitude, stewardship attitudes, intent to act pro-environmentally, and dominion beliefs.

The other important variables are: age (for each 1 yr increase in age, the odds of donation go up by 1.14 times), and 'envintent' (for each one unit increase in env intent, the odds of donating go up by 2.19 times).

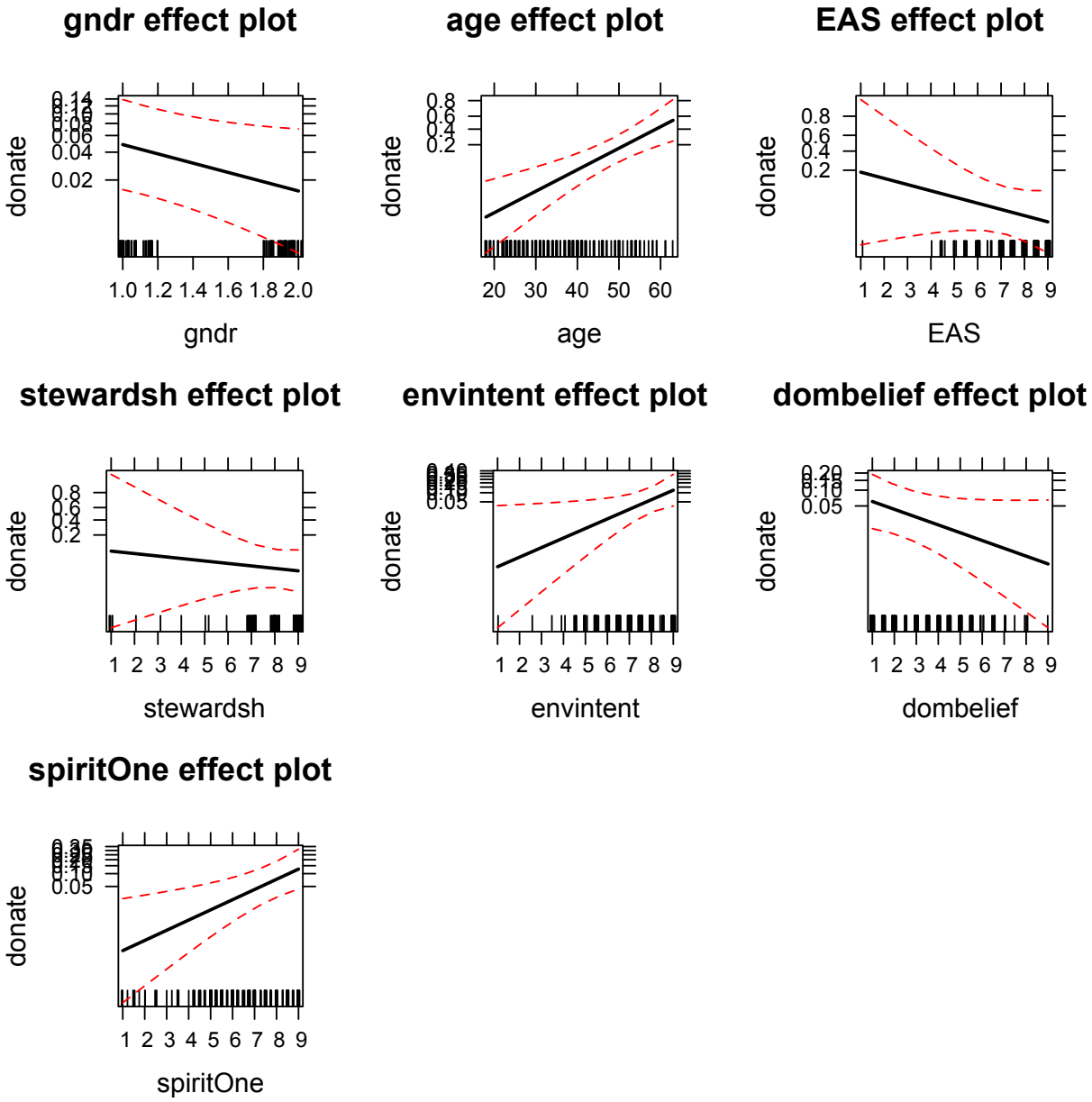
--Find pseudo-R-sq values:

```
> logisticRsq(mod2)
Pseudo R-sq for logistic regression
Hosmer & Lemeshow R-sq    0.3345
Cox and Snell R-sq         0.2111
Nagelkerke R-sq            0.4157
```

Interpretation: the pseudo-Rsq values are higher with more variables in the model, but the odds ratio for Spiritual Oneness remains about 1.8.

-- Plot effects:

```
> plot(allEffects(mod2)) # wish I could set the y-axis the same on all
graphs!
```



III. B. Compare to least squares regression, just for kicks (see lecture notes for advantages of doing things properly when you have a dichotomous outcome variable).

-- Fit linear regression with 'glm' by using the gaussian family. The next model will show he equivalence of glm with gaussian to the 'lm' function.

```
> mod5=glm(donate~gndr+age+EAS+stewardsh+envintent+dombelief+spiritOne,
family="gaussian")
> summary(mod5)
```

```
Call:
glm(formula = donate ~ gndr + age + EAS + stewardsh + envintent +
    dombelief + spiritOne, family = "gaussian")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.37769  -0.15772  -0.07585   0.03671   0.94357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.210306   0.213944  -0.983   0.3271
gndr        -0.050181   0.046827  -1.072   0.2855
age          0.009313   0.002053   4.536 1.12e-05 ***
EAS         -0.018702   0.021504  -0.870   0.3858
stewardsh   -0.017276   0.017342  -0.996   0.3207
envintent    0.029100   0.021990   1.323   0.1876
dombelief   -0.016087   0.013555  -1.187   0.2371
spiritOne    0.034285   0.013348   2.568   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.08636368)

    Null deviance: 16.838  on 166  degrees of freedom
Residual deviance: 13.732  on 159  degrees of freedom
  (13 observations deleted due to missingness)
AIC: 74.713

Number of Fisher Scoring iterations: 2
```

fit with 'lm' function gives same result as 'glm' gaussian, but NOT the same as logistic regression (glm with binomial).

```
> mod6=lm(donate~gndr+age+EAS+stewardsh+envintent+dombelief+spiritOne)
> summary(mod6)

Call:
lm(formula = donate ~ gndr + age + EAS + stewardsh + envintent +
    dombelief + spiritOne)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37769 -0.15772 -0.07585  0.03671  0.94357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.210306   0.213944  -0.983   0.3271
gndr        -0.050181   0.046827  -1.072   0.2855
age          0.009313   0.002053   4.536 1.12e-05 ***
EAS         -0.018702   0.021504  -0.870   0.3858
stewardsh   -0.017276   0.017342  -0.996   0.3207
envintent    0.029100   0.021990   1.323   0.1876
dombelief   -0.016087   0.013555  -1.187   0.2371
spiritOne    0.034285   0.013348   2.568   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2939 on 159 degrees of freedom
```

```
  (13 observations deleted due to missingness)
Multiple R-squared: 0.1845,   Adjusted R-squared: 0.1486
F-statistic: 5.139 on 7 and 159 DF,  p-value: 2.734e-05
```

-- Find Type II SS for both of these models:

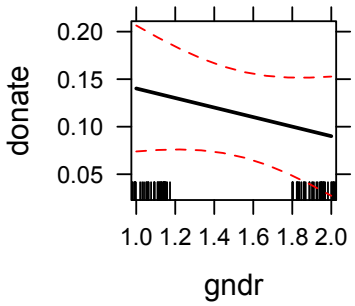
```
> Anova(mod5, test=c("F"))  ## this is the 'glm' gaussian model
Analysis of Deviance Table (Type II tests)

Response: donate
              SS  Df       F    Pr(>F)
gndr       0.0992   1  1.1484   0.28552
age        1.7769   1 20.5751 1.124e-05 ***
EAS        0.0653   1  0.7564   0.38576
stewardsh  0.0857   1  0.9924   0.32067
envintent  0.1512   1  1.7513   0.18762
dombelief  0.1216   1  1.4084   0.23710
spiritOne  0.5698   1  6.5972   0.01113 *
Residuals 13.7318 159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> Anova(mod6)  # this is the 'lm' model
Anova Table (Type II tests)

Response: donate
           Sum Sq  Df F value    Pr(>F)
gndr       0.0992   1  1.1484   0.28552
age        1.7769   1 20.5751 1.124e-05 ***
EAS        0.0653   1  0.7564   0.38576
stewardsh  0.0857   1  0.9924   0.32067
envintent  0.1512   1  1.7513   0.18762
dombelief  0.1216   1  1.4084   0.23710
spiritOne  0.5698   1  6.5972   0.01113 *
Residuals 13.7318 159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-- Notice the increased power in the logistic model compared to the linear model.

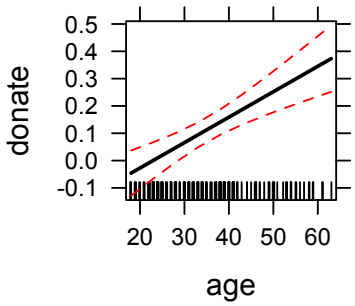
-- plot effects. Notice that the -.2 intercept makes for uninterpretable effects.

```
> plot(allEffects(mod5))
```

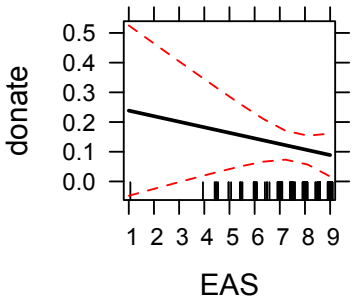
gndr effect plot



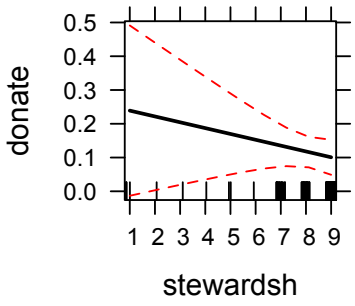
age effect plot



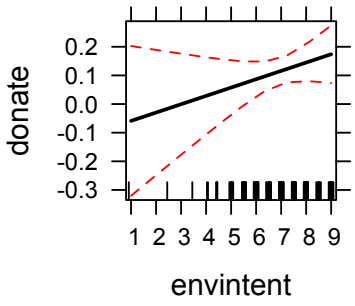
EAS effect plot



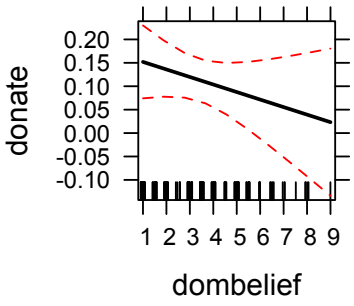
stewardsh effect plot



envintent effect plot



dombelief effect plot



spiritOne effect plot