Running Regression in R (I)

R tutorial #2

Eun Jong Kong
(ekong@kau.ac.kr)

Korea Aerospace University

Logistic Regression: glm ()

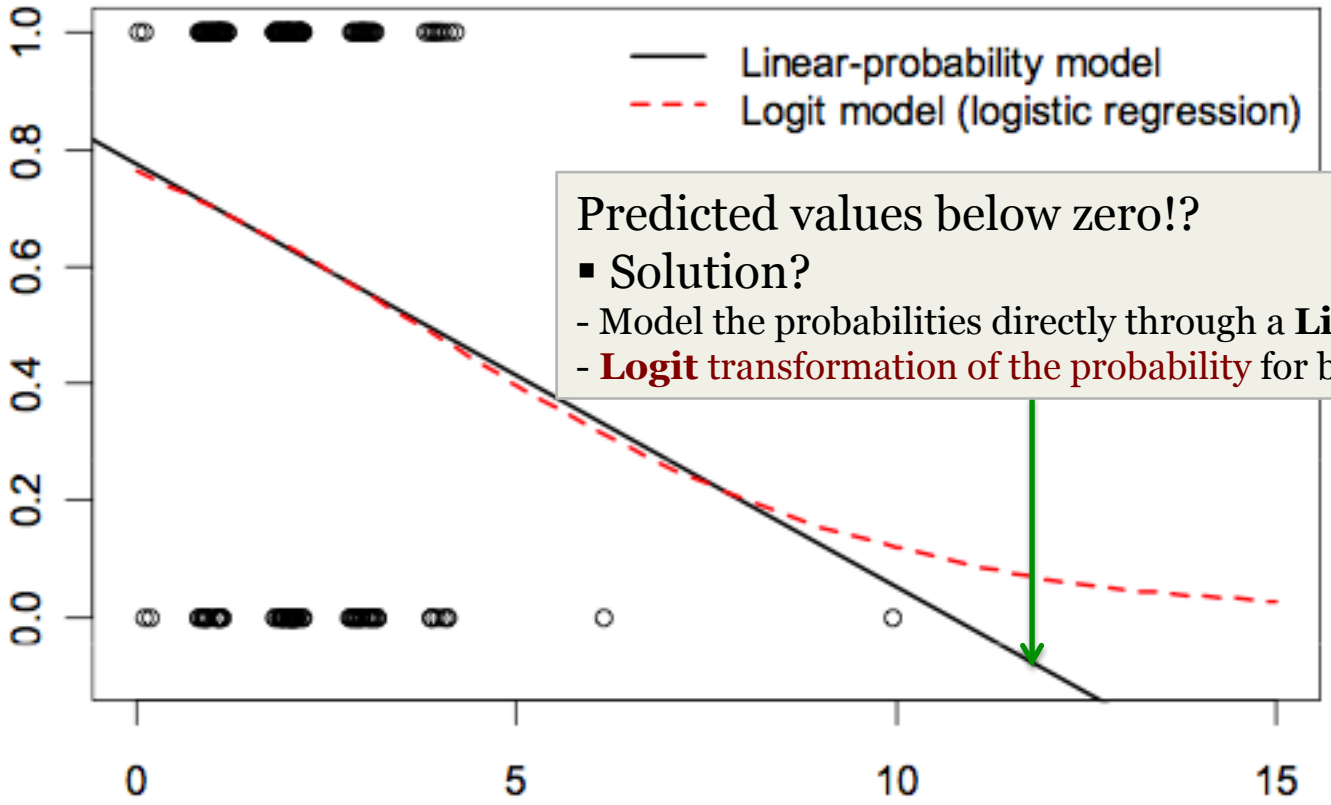
- Analyzing Linguistic Data (Baayen)

Dichotomous response

- Data sets with ***binary (dichotomous)*** dependent variables :
  - correct, incorrect
  - “s”, “ſ”

1. **(Linear) Probability model**

Dichotomous response



Predicted values below zero!?

- Solution?

- Model the probabilities directly through a **Link Function**.
- **Logit** transformation of the probability for binary data.

Dichotomous response

- Data sets with ***binary (dichotomous)*** dependent variables :
  - correct, incorrect
  - “s”, “ʃ”

1. **(Linear) Probability model**

2. **Generalized linear model (GLM)**
   - **Logit** or probit regression models.

Odds, Odds-ratio, Logit

- ***Odds*** is the relative chance of an event (the ratio of two probabilities).

$$\text{odds} = \frac{\text{probability of an event}}{1 - \text{probability of an event}} = \frac{\pi}{1-\pi}$$

- ***Odds-ratio*** is the ratio of two odds.
- ***Log-odds (Logit)*** is the natural logarithm of the odds.
- The log-odds is modeled as a linear function in ***logistic regression***.

Logistic Regression

▪ ***Logistic regression*** (logit model) models the logit (log-odds) as a linear function of the independent variables:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 X_{1i}$$

✓ **Logistic regression coefficients** refer to the log-odds.
✓ Those logistic regression coefficients are exponentiated to be interpretable.

Data in R

Reading objects from files (.cvs)

http://homepages.wmich.edu/~hillenbr/voweldata.html

```
>  dat = read.csv(file.choose(), header= T)  # "reg_dat_logi.csv"
```

```
> head(dat)      # first 6 rows
> tail(dat)      # last 6 rows
> dat[15:20, ]   # 15th~ 20th rows of the data
```

```
> dat[15:20,]
   sub gender dur vow
15 m18      m 179  ih
16 m19      m 194  ih
17 m20      m 221  ih
18 m21      m 145  ih
19 m22      m 220  ih
20 m23      m 207  ih
```

Data in R: Reading objects from files (.cvs)

```
>   str(dat)      # examine the data structure
```

```
> str(dat)
'data.frame': 186 obs. of  4 variables:
 $ sub   : Factor w/ 93 levels "m01","m02","m03",..: 1 2 3 4 5 6 7 8 9
10 ...
 $ gender: Factor w/ 2 levels "m","w": 1 1 1 1 1 1 1 1 1 1 ...
 $ dur   : int  268 183 262 227 159 190 235 221 186 301 ...
 $ vow   : Factor w/ 2 levels "ih","iy": 1 1 1 1 1 1 1 1 1 1 ...
```

```
>   dim(dat)      # dimensions of the dataframe
```

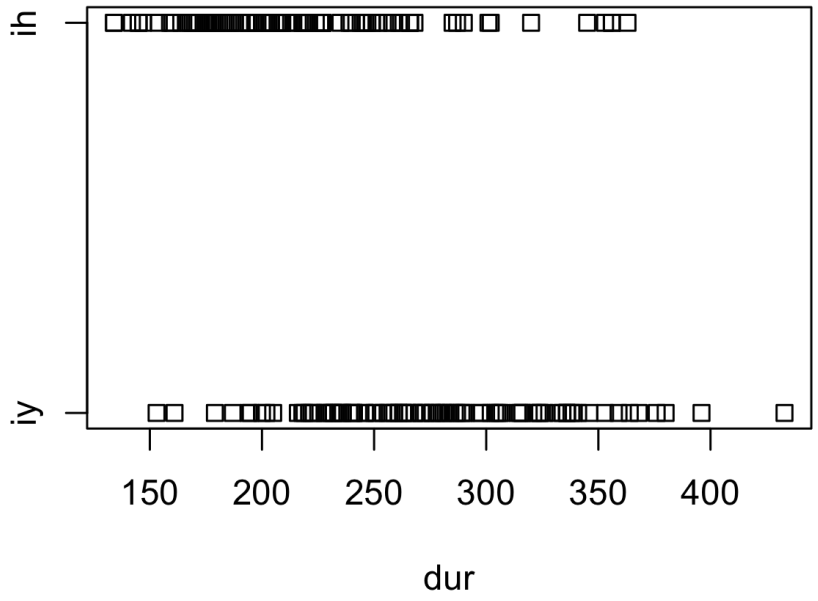
```
>   summary(dat)      # data summary
```

```
> summary(dat)
      sub        gender       dur             vow
 m01    :  2   w:96    Min.   :134.0   iy:93
 m02    :  2   m:90    1st Qu.:202.0   ih:93
 m03    :  2           Median :240.0
 m04    :  2           Mean   :247.1
 m06    :  2           3rd Qu.:286.8
 m07    :  2           Max.   :433.0
 (Other):174
```

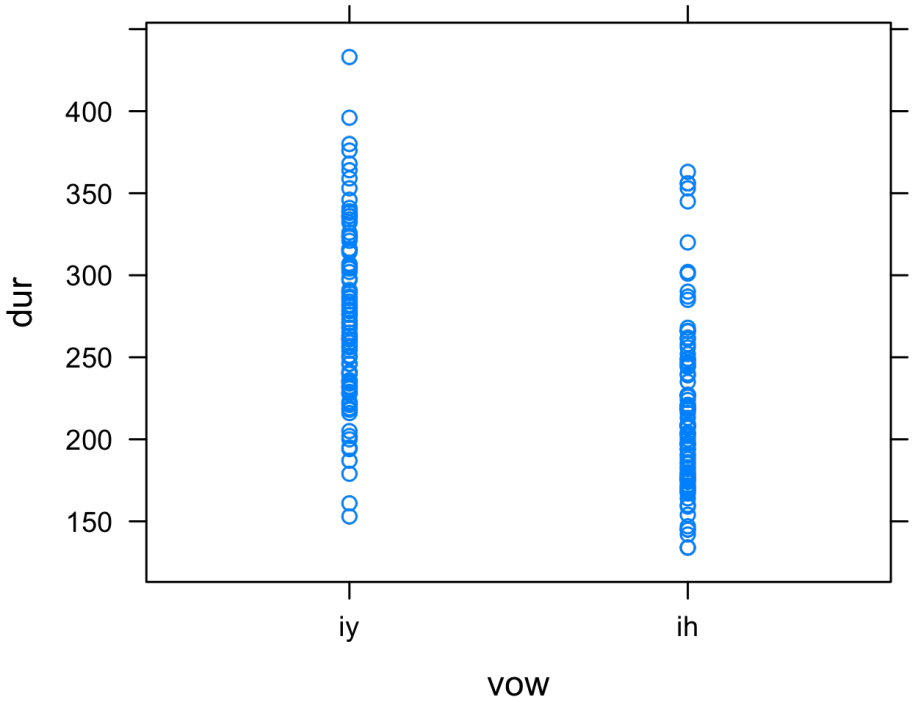
Research Question?

Visual Inspection: plot()

```
> stripchart(dur ~  vow, dat)
```




```
> library(lattice)
> xyplot(dur ~  vow, dat)
```



Logistic Regression: glm()

Dichotomous dependent variable: “1” or “0”

```
> dat$dv[dat$vow == "ih"] = 0
> dat$dv[dat$vow == "iy"] = 1
```

```
> dat$dv[dat$vow == "ih"] = 0
> dat$dv[dat$vow == "iy"] = 1
> dat$dv = as.factor(dat$dv)
> is.numeric(dat$dv)
[1] FALSE
> is.factor(dat$dv)
[1] TRUE
```

Reference level of the dependent variable: “iy” (vs. “ih”)

```
> dat$vow = relevel(as.factor(dat$vow), ref = "iy")
```

Linear regression: glm()

Formula: Generalized Linear Model

```
> md.log = glm(dv ~ dur, data = dat, binomial(link = "logit"))
```

```
> md.log

Call:  glm(formula = dv ~ dur, family = binomial(link = "logit"), data = dat)

Coefficients:
(Intercept)          dur
   -5.01261      0.02046

Degrees of Freedom: 185 Total (i.e. Null);  184 Residual
Null Deviance:     257.9
Residual Deviance: 209   AIC: 213
```

The relative chance in log scale!

Linear regression: glm()

```
> summary(md.log)
```

```
> summary(md.log)

Call:
glm(formula = dv ~ dur, family = binomial(link = "logit"), data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2362  -0.8828  -0.1189   0.9573   2.0119

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.012615   0.858001  -5.842 5.15e-09 ***
dur          0.020461   0.003473   5.891 3.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 209.03  on 184  degrees of freedom
AIC: 213.03

Number of Fisher Scoring iterations: 4
```

The relative chance of being “iy” in log scale:
> **exp()**

Linear regression: glm()

```
> summary(md.log)
> exp(coef(md.log)) # coefficients on odds-scale
```

```
> exp(coef(md.log))
(Intercept)         dur
0.006653483 1.020671766
```

- **Intercept**: With 0 ms. the relative chance (odds) of being "iy" is 0.006.

- **Slope**: With one ms. increase of duration, the relative chance (odds) for "iy" increases by 2%.

- ex: the token with 200ms duration has a relative chance of 0.0067×1.02^{200} of getting "iy".

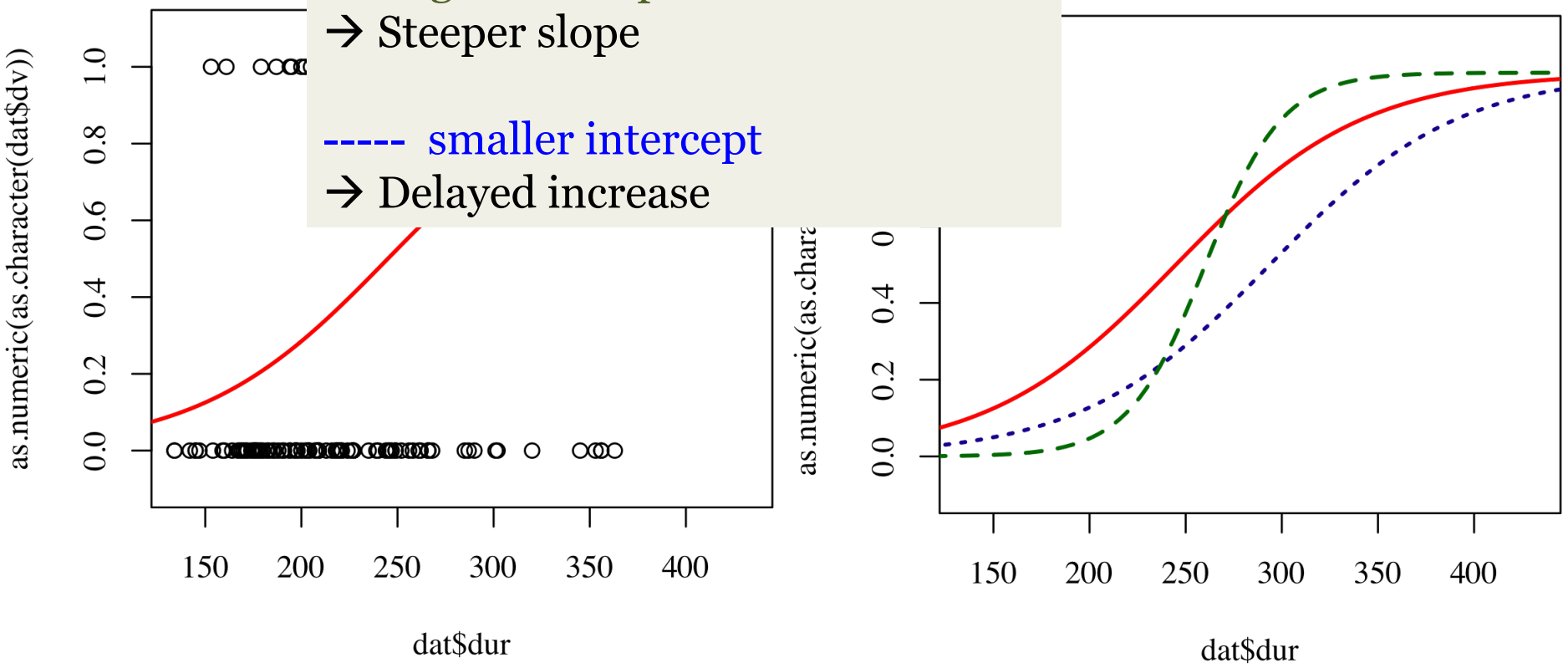
Linear regression: logit plot

```
> plot(dat$dur, as.numeric(as.character(dat$dv)), ylim = c(-0.1,1.1))

> lines(100:450, predict(md.log, data.frame(dur = 100:450), type =
'response'), lwd = 2, col = 'red', lty = 1)
```

- - - greater slope coefficient
→ Steeper slope

----- smaller intercept
→ Delayed increase



Linear regression: glm()

```
> summary(md.log)
```

```
> summary(md.log)

Call:
glm(formula = dv ~

Deviance Residuals
    Min       1Q
-2.2362  -0.8828

Coefficients:
             Estim
(Intercept) -5.012
dur          0.020
---
Signif. codes:  0

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 209.03  on 184  degrees of freedom
AIC: 213.03

Number of Fisher Scoring iterations: 4
```

- **Null deviance**: the deviance estimated from a model with only an intercept.
- **Residual deviance**: the deviance estimated from a model with only an intercept and independent variables. i.e., generalization of the residual sum of squares for a linear model.
- The smaller the deviance, the better the fit (the better the model's predictive power)
- The difference between Null and Residual deviances follows a **chi-square distribution.**
- **anova()**

Linear regression: glm()

```
    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 209.03  on 184  degrees of freedom
```

```
> 1 - pchisq(257.85 - 209.03, 185 - 184)
> anova(md.log, test = "Chisq") # deviance test, likelihood-ratio test
```

```
> (anova(md.log, test = "Chisq"))
Analysis of Deviance Table

Model: binomial, link: logit

Response: dv

Terms added sequentially (first to last)
```

- If the difference in the null and fitted residual deviance is rather large, the model has significant predictive power.
- A very small p-value shows that we have a model with explanatory value.

```
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                   185     257.85
dur   1   48.821       184     209.03 2.805e-12 ***
```

Linear regression: glm()

```
> summary(md.log)
```

```
> summary(md.log)

Call:
glm(formula = dv ~ dur, family = binomial(li

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2362  -0.8828  -0.1189   0.9573   2.0119

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.012615   0.858001  -5.842 5.15e-09 ***
dur          0.020461   0.003473   5.891 3.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 209.03  on 184  degrees of freedom
AIC: 213.03

Number of Fisher Scoring iterations: 4
```

Akaike Information Criterion:

- An index of fit that adds a penalty term to the deviance which takes the number of parameters into account.
- The smaller, the better model.

Logistic Regression: glm ()

interaction term

Data in R

Reading objects from files (.cvs)

```
> dat = read.csv(file.choose(), header= T)  # “reg_dat_logi.csv”
> summary(dat)
```

```
> summary(dat)
      sub        gender       dur             vow
 m01    :  2   w:96    Min.   :134.0   iy:93
 m02    :  2   m:90    1st Qu.:202.0   ih:93
 m03    :  2           Median :240.0
 m04    :  2           Mean   :247.1
 m06    :  2           3rd Qu.:286.8
 m07    :  2           Max.   :433.0
 (Other):174
```

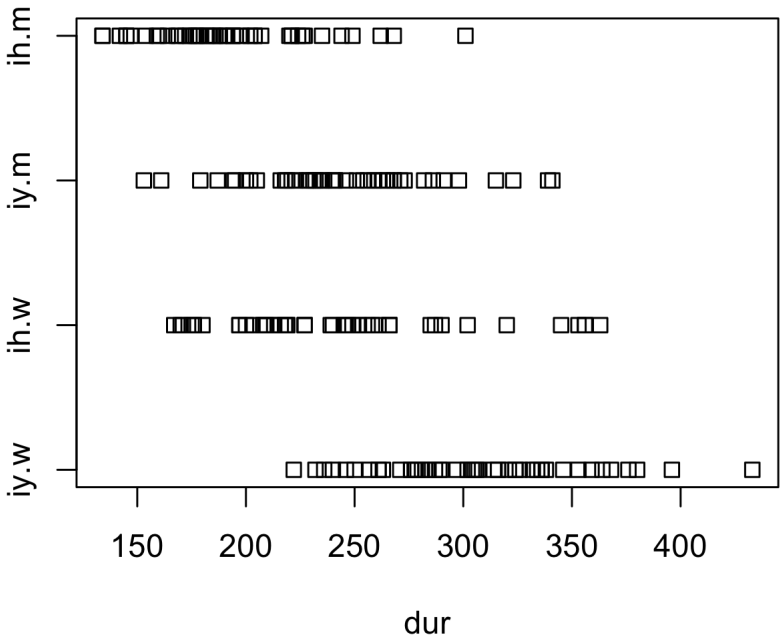
Data in R

```
> tapply(dat$dur, dat$vow, mean)
> tapply(dat$dur, list(vow = dat$vow, gender = dat$gender), mean)
```

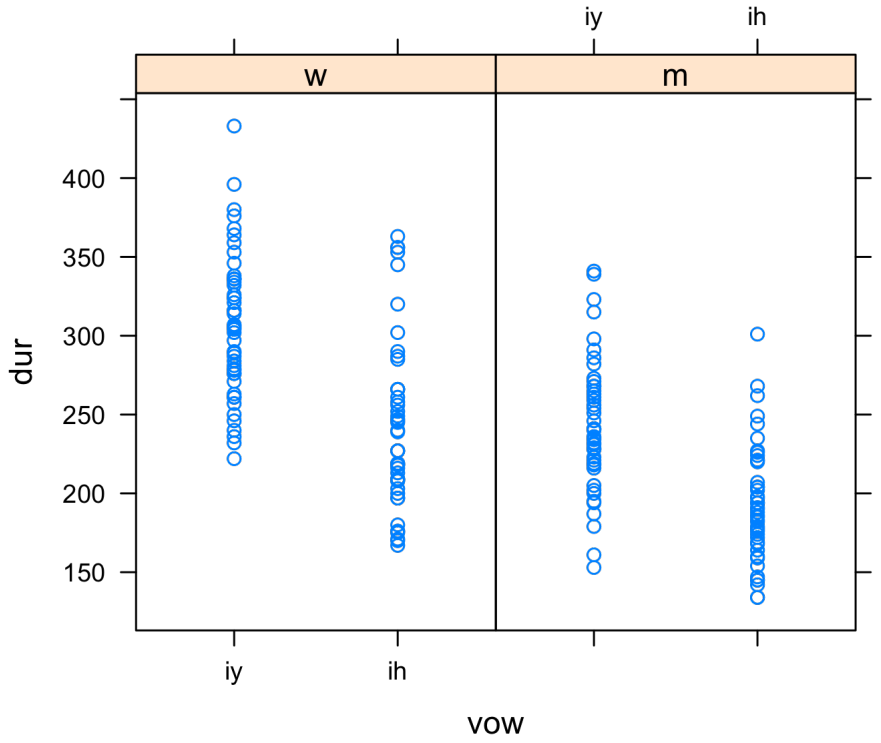
```
> tapply(dat$dur, dat$vow, mean)
      iy       ih
276.1935 217.9247
> tapply(dat$dur, list(vow = dat$vow, gender = dat$gender), mean)
    gender
vow          w        m
  iy 306.8333 243.5111
  ih 241.3958 192.8889
```

Visual Inspection: plot()

```
> stripchart(dur ~  vow, dat)
```




```
> xyplot(dur ~  vow, dat)
```



Logistic Regression: glm()

Dichotomous variable: “1” or “0”

```
> dat$dv[dat$vow == "ih"] = 0
> dat$dv[dat$vow == "iy"] = 1
```

```
> dat$dv[dat$vow == "ih"] = 0
> dat$dv[dat$vow == "iy"] = 1
> dat$dv = as.factor(dat$dv)
> is.numeric(dat$dv)
[1] FALSE
> is.factor(dat$dv)
[1] TRUE
```

Reference level of the dependent variable: “iy” (vs. “ih”)

```
> dat$vow = relevel(as.factor(dat$vow), ref = "iy")
```

Linear regression: glm()

Formula: Generalized Linear Model (interaction)

```
> dat$gender = relevel(as.factor(dat$gender), ref = "w")
> md.log.int = glm(dv ~ dur * gender,
                   data = dat,
                   family = 'binomial')
```

- The relative chance in log scale:
  - ✓ **exp()**
- reference level : "W"

```
> md.log.int

Call:  glm(formula = dv ~ dur * gender, family = "binomial", data = dat)

Coefficients:
(Intercept)          dur      genderm  dur:genderm
  -6.962466     0.025503    -0.099709     0.007122

Degrees of Freedom: 185 Total (i.e. Null);  182 Residual
Null Deviance:     257.9
Residual Deviance: 192.7   AIC: 200.7
```

Linear regression: glm()

```
> summary(md.log.int)
```

```
> summary(md.log.int)

Call:
glm(formula = dv ~ dur * gender, family = "binomial", data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.37459  -0.79081  -0.08848   0.86138   2.09255

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.962466   1.488865  -4.676 2.92e-06 ***
dur          0.025503   0.005403   4.720 2.36e-06 ***
genderm     -0.099709   2.188005  -0.046    0.964
dur:genderm  0.007122   0.009136   0.780    0.436
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 192.75  on 182  degrees of freedom
AIC: 200.75

Number of Fisher Scoring iterations: 4
```

Linear regression: glm()

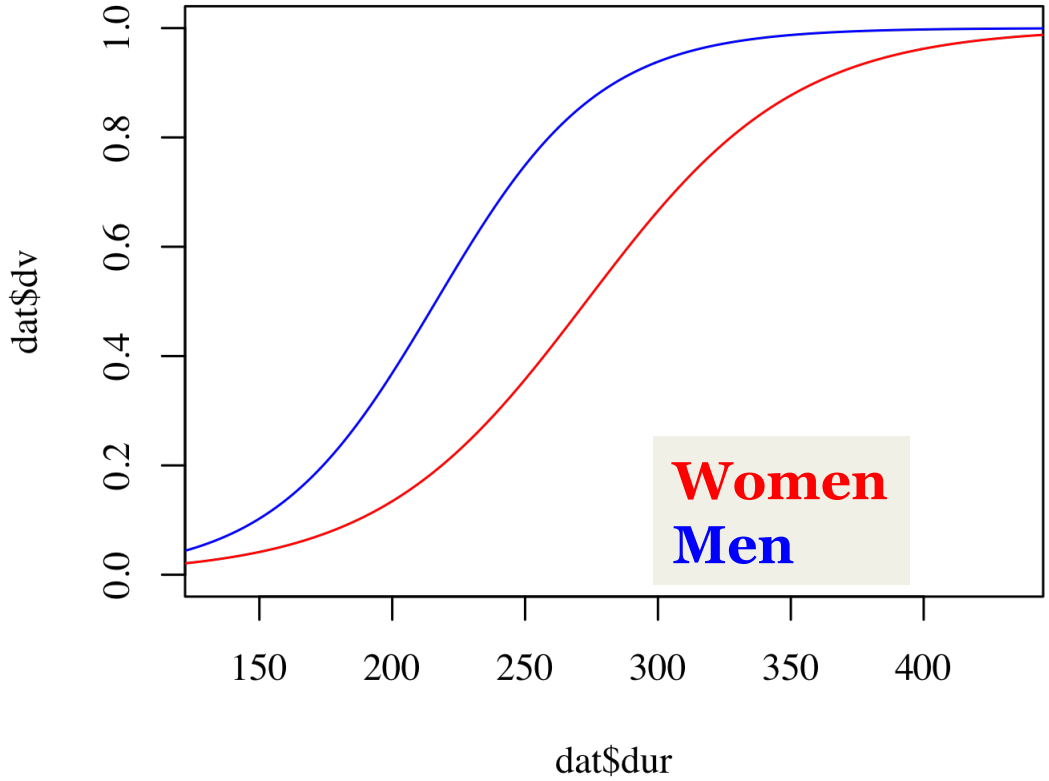
```
> coef(md.log.int)
 (Intercept)          dur      genderm  dur:genderm
-6.962466270  0.025502649 -0.099708607  0.007121983
> exp(coef(md.log.int))
 (Intercept)          dur      genderm  dur:genderm
0.0009467587 1.0258306236 0.9051011196 1.0071474048
```

- **Intercept (women)**: With 0 ms. the relative chance (odds) of being "iy" produced by women is 0.0009.
- **Slope (women)**: With one ms. increase of duration, the relative chance (odds) for "iy" produced by women increases by 2.5%.

- **Intercept (men)**: With 0 ms. the relative chance (odds) of being "iy" produced by men is 0.0008569124 (0.0009467587 * 0.9051011196).
- **Slope (men)**: With one ms. increase of duration, the relative chance (odds) for "iy" produced by men increases by 3.3% (1.0258306236 * 1.0071474048)

Linear regression: glm()

```
> plot(dat$dur, dat$dv, ylim = c(0,1), type = "n")
> lines(100:450, predict(md.log.int, data.frame(dur = 100:450, gender
= "w"), type = 'response'), lwd = 1, col = 'red', lty = 1)
> lines(100:450, predict(md.log.int, data.frame(dur = 100:450, gender
= "m"), type = 'response'), lwd = 1, col = 'blue', lty = 1)
```



	women	men
slope	0.025503	0.032624632
intercept	-6.962466	-7.062174877

Linear regression: glm()

```
> summary(md.log.int)

Call:
glm(formula = dv ~ dur * gender, family = "binomial", data = dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.37459  -0.79081  -0.08848   0.86138   2.09255

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.962466   1.488865  -4.676 2.92e-06 ***
dur          0.025503   0.005403   4.720 2.36e-06 ***
genderm     -0.099709   2.188005  -0.046    0.964
dur:genderm  0.007122   0.009136   0.780    0.436
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 192.75  on 182  degrees of freedom
AIC: 200.75

Number of Fisher Scoring iterations: 4
```

Deviance difference:
257.85 – 192.75 (df = 3)
→ **anova()**

Linear regression: glm()

```
    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 192.75  on 182  degrees of freedom
```

```
> 1 - pchisq(257.85 - 192.75, 185 - 182)
```

```
>    1 - pchisq(257.85-192.75, 185-182)
[1] 4.773959e-14
```

- Large difference between null and fitted residual deviance & a very small p-value

→ the model has significant predictive power.

- Note that the model has a meaningful difference when duration factor, gender factor and interaction factor were added.

Linear regression: glm()

```
> md.log.int.0 = glm(dv ~ 1, data = dat, family = 'binomial')
> md.log.int.1 = glm(dv ~ dur , data = dat, family = 'binomial')
> md.log.int.2 = glm(dv ~ dur + gender, data = dat, family = 'binomial')
> md.log.int = glm(dv ~ dur + gender + dur:gender, data = dat, family =
'binomial')
```

```
Analysis of Deviance Table

Model 1: dv ~ 1
Model 2: dv ~ dur
Model 3: dv ~ dur + gender
Model 4: dv ~ dur + gender + dur:gender
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       185     257.85
2       184     209.03  1   48.821 2.805e-12 ***
3       183     193.37  1   15.662 7.572e-05 ***
4       182     192.75  1    0.621    0.4307
```

- Addition of duration factor
- Addition of gender factor
- Addition of gender and duration interaction

Linear regression: glm()

```
> summary(md.log)
```

```
> summary(md.log)

Call:
glm(formula = dv ~ dur, family = binomial(li

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2362  -0.8828  -0.1189   0.9573   2.0119

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.012615   0.858001  -5.842 5.15e-09 ***
dur          0.020461   0.003473   5.891 3.85e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 257.85  on 185  degrees of freedom
Residual deviance: 209.03  on 184  degrees of freedom
AIC: 213.03

Number of Fisher Scoring iterations: 4
```

Akaike Information Criterion:

- An index of fit that adds a penalty term to the deviance which takes the number of parameters into account.
- The smaller, the better model.