Lecture 13: Project Guideline

General Procedure

- Problem specification: What problems are you interested in? What questions will you answer? Why the questions are valuable, i.e., motivation?
- Data preparation: cleaning, sorting, missing data, etc.
- Exploratory data analysis: summary, scatter plot, correlation plot, boxplot, basic modeling. Tell the basic story of data via this analysis, and what models may be suitable for this analysis.
- Model specification: must be GLM!
  - Continuous data: Traditional normal model (not preferred!), non-normal data, longitudinal or continuous data
  - Categorical data: (Quasi-) Poisson model, binomial model, negative binomial model, nominal logistic model, ordinary logistic model, proportional odds model, log-linear model etc.
- Estimation: What method(s) you will use to estimate the parameters? Briefly explain the idea.
- Model selection: Which model(s) is the best to answer the question(s)? How do you select the model(s)?
- Adequacy: Checking the adequacy of the model–how well it fits or summarizes the data: deviance, residuals, goodness-of-fit statistics, qq plot, etc.
- Inference: Calculating confidence intervals, testing hypotheses about the parameters in the model and interpreting the results.
- Prediction: Predict based on the model you developed, and assess the prediction accuracies based on different metrics. For the continuous data or binary outcomes, the prediction is recommended to assess the prediction power of your model.
- Summary: You need to summarize your model results, and present the relevant discussion. For example, would your model can answer the question of interest? Any possible drawbacks of your model? Any possible improvement?

Case study 1: Stress and Smoking

For many people who smoke, the most natural thing to do in the midst of a stressful situation is to reach for a cigarette. Many smokers will explain that smoking helps them to relax and relieves their feeling of stress. Their adamant belief that this truly works has introduced the question of whether smoking does indeed relieved the amount of stress perceived by a smoker.

- Goal: Investigate the relationship between smoking and the amount of recent life stress perceived
- Response: Stress at three levels
- Variables: Age (three levels), Gender (two levels), Smoking (three levels)
- **smoking status** (1 = smoker, 2 = quitter, 3 = non-smoker)
- **gender** (1 = male, 2 = female)
- **age** (1 = young, 2 = middle, 3 = old)
- **perceived stress level** (1 = severe / a lot, 2 = moderate / some, 3 = mild / none)

Exploratory data analysis

```
  > data.frame(smoke1,smoke2,smoke3)
          stress=1  stress=2  stress=3
smoke=1     61        40        14
smoke=2      39        27        18
smoke=3      39        24        31

> smoke_freq=data.frame(level=c(rep(1,3),rep(2,3),rep(3,3)),
smoke=c(1,2,3,1,2,3,1,2,3),freq=c(smoke1,smoke2,smoke3))
> smoke_freq
  level smoke freq
1     1     1   61
2     1     2   39
3     1     3   39
4     2     1   40
5     2     2   27
6     2     3   24
7     3     1   14
8     3     2   18
9     3     3   31
> smoke_freq=data.frame(level=c(rep(1,3),rep(2,3),rep(3,3)),
smoke=c(1,2,3,1,2,3,1,2,3),freq=c(smoke1,smoke2,smoke3))
> smoke1_glm<-glm(freq~level+smoke,family=poisson(),data=smoke_freq)
```

```
> summary(smoke1_glm)

Call:
glm(formula = freq ~ level + smoke, family = poisson(), data = smoke_freq)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0226  -0.7066  -0.5944   0.9529   2.6383

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.44073    0.19833  22.391  < 2e-16 ***
level       -0.39934    0.07441  -5.367 8.01e-08 ***
smoke       -0.10772    0.07176  -1.501    0.133
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 48.521  on 8  degrees of freedom
Residual deviance: 16.307  on 6  degrees of freedom
AIC: 69.467

Number of Fisher Scoring iterations: 4

>
>
> smoke2_glm<-glm(freq~level*smoke,family=poisson(),data=smoke_freq)
> summary(smoke2_glm)

Call:
glm(formula = freq ~ level * smoke, family = poisson(), data = smoke_freq)

Deviance Residuals:
       1         2         3         4         5         6         7         8         9
-0.06541  -0.88639   1.08172   1.45385  -0.62133  -1.00440  -0.54186  -0.57956   0.90244

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.38589    0.36023  14.951  < 2e-16 ***
level       -0.94963    0.19803  -4.795 1.62e-06 ***
smoke       -0.59726    0.17606  -3.392 0.000693 ***
level:smoke  0.28024    0.09164   3.058 0.002228 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 48.5210  on 8  degrees of freedom
Residual deviance:  6.9125  on 5  degrees of freedom
AIC: 62.073

Number of Fisher Scoring iterations: 4
```

Model specification

```
##################propotional odds model 1##############
> glm1=polr(factor(stress)~factor(gender)+factor(smoke)
+factor(age),weights=count,data=stress)
> summary(glm1)

Re-fitting to get Hessian

Call:
polr(formula = factor(stress) ~ factor(gender) + factor(smoke) +
    factor(age), data = stress, weights = count)

Coefficients:
                  Value Std. Error t value
factor(gender)2  0.7519     0.2351   3.199
factor(smoke)2   0.5869     0.2832   2.072
factor(smoke)3   0.7216     0.2783   2.593
factor(age)2    -0.3734     0.2693  -1.386
factor(age)3    -0.9347     0.2880  -3.245

Intercepts:
    Value   Std. Error t value
1|2  0.2393  0.2963     0.8075
2|3  1.7512  0.3144     5.5698

Residual Deviance: 584.2929
AIC: 598.2929
>


##################propotional odds model 2##############
> glm2=polr(factor(stress)~factor(gender)+factor(smoke)
+factor(gender):factor(smoke)+factor(age),weights=count,data=stress)
> summary(glm2)

Re-fitting to get Hessian
```

```
Call:
polr(formula = factor(stress) ~ factor(gender) + factor(smoke) +
    factor(gender):factor(smoke) + factor(age), data = stress,
    weights = count)

Coefficients:
                               Value Std. Error t value
factor(gender)2                0.7973     0.3998  1.9943
factor(smoke)2                 0.7279     0.4336  1.6788
factor(smoke)3                 0.6625     0.4416  1.5002
factor(age)2                  -0.3752     0.2707 -1.3862
factor(age)3                  -0.9396     0.2880 -3.2619
factor(gender)2:factor(smoke)2 -0.2996     0.5786 -0.5179
factor(gender)2:factor(smoke)3  0.1413     0.5670  0.2493

Intercepts:
    Value   Std. Error t value
1|2  0.2685  0.3707     0.7244
2|3  1.7835  0.3840     4.6439

Residual Deviance: 583.6929
AIC: 601.6929
>

#################propotional odds model 3#############
> glm3=polr(factor(stress)~factor(gender)+factor(smoke)
+factor(smoke):factor(age)+factor(age),weights=count,data=stress)
> summary(glm3)

Re-fitting to get Hessian

Call:
polr(formula = factor(stress) ~ factor(gender) + factor(smoke) +
    factor(smoke):factor(age) + factor(age), data = stress, weights = count)

Coefficients:
                              Value Std. Error t value
factor(gender)2              0.7738     0.2370  3.2645
factor(smoke)2               0.7634     0.5130  1.4880
factor(smoke)3               0.3936     0.4331  0.9088
factor(age)2                -0.6240     0.4459 -1.3992
factor(age)3                -0.9878     0.4425 -2.2322
factor(smoke)2:factor(age)2 -0.1305     0.6960 -0.1875
factor(smoke)3:factor(age)2  0.7432     0.6281  1.1833
factor(smoke)2:factor(age)3 -0.3191     0.6856 -0.4653
```

```
factor(smoke)3:factor(age)3  0.3541     0.7126  0.4969

Intercepts:
    Value   Std. Error t value
1|2  0.1485  0.3552     0.4181
2|3  1.6709  0.3699     4.5177

Residual Deviance: 582.0157
AIC: 604.0157


########################multinomial model####################
> library("nnet")
> glm4=multinom(factor(stress)~factor(gender)+factor(smoke)
+factor(age), weights=count,data=stress)
# weights:  21 (12 variable)
initial  value 321.893401
iter  10 value 287.328931
final  value 286.226760
converged
> summary(glm4)
Call:
multinom(formula = factor(stress) ~ factor(gender) + factor(smoke) +
    factor(age), data = stress, weights = count)

Coefficients:
  (Intercept) factor(gender)2 factor(smoke)2 factor(smoke)3 factor(age)2 factor(age)3
2  -0.8193217       1.0792960      0.3887932      0.0286234   -0.2187583   -0.6104496
3  -1.3192935       0.7745189      1.0055519      1.1598633   -0.5125820   -1.4478623

Std. Errors:
  (Intercept) factor(gender)2 factor(smoke)2 factor(smoke)3 factor(age)2 factor(age)3
2   0.3595797       0.2915069      0.3451086      0.3494905    0.3465622    0.3544223
3   0.4236983       0.3304450      0.4359565      0.4022682    0.3676794    0.4407481

Residual Deviance: 572.4535
AIC: 596.4535
```

Model selection and evaluation

```
> M2 <- logLik(glm4)
> G <- -2*(M1[1] - M2[1])
>
> pchisq(G,(12-7),lower.tail = FALSE)
[1] 0.03705554
```

```
> M1
'log Lik.' -292.1465 (df=7)
> M2
'log Lik.' -286.2268 (df=12)
> G <- -2*(M1[1] - M2[1])
>
> pchisq(G,(12-7),lower.tail = FALSE)
[1] 0.03705554
```

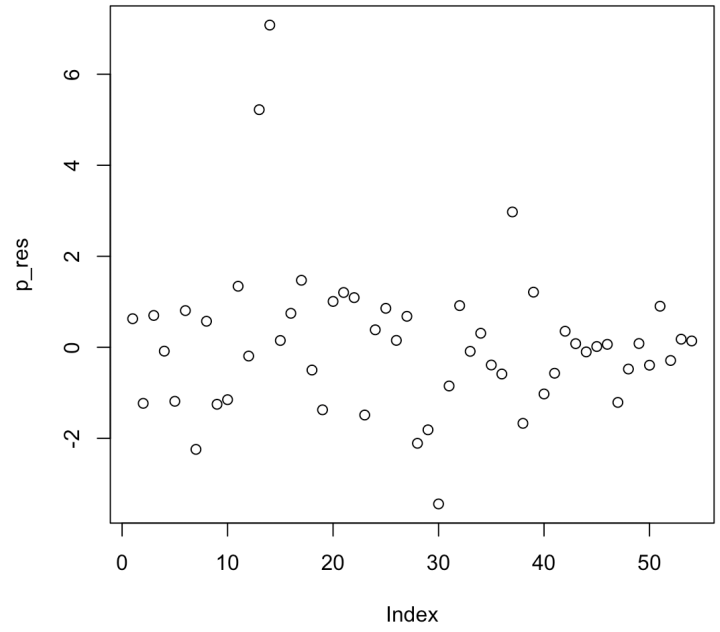
```
> fitted(glm1)
           1         2         3
1  0.5595350 0.2925742 0.1478908
2  0.5595350 0.2925742 0.1478908
3  0.5595350 0.2925742 0.1478908
4  0.6485396 0.2447309 0.1067296
5  0.6485396 0.2447309 0.1067296
6  0.6485396 0.2447309 0.1067296
7  0.7638564 0.1723334 0.0638102
8  0.7638564 0.1723334 0.0638102
9  0.7638564 0.1723334 0.0638102
10 0.3745718 0.3563515 0.2690767
11 0.3745718 0.3563515 0.2690767
12 0.3745718 0.3563515 0.2690767
13 0.4652307 0.3325794 0.2021899
14 0.4652307 0.3325794 0.2021899
15 0.4652307 0.3325794 0.2021899
16 0.6039646 0.2697244 0.1263110
17 0.6039646 0.2697244 0.1263110
18 0.6039646 0.2697244 0.1263110
19 0.4139572 0.3481623 0.2378805
20 0.4139572 0.3481623 0.2378805
```

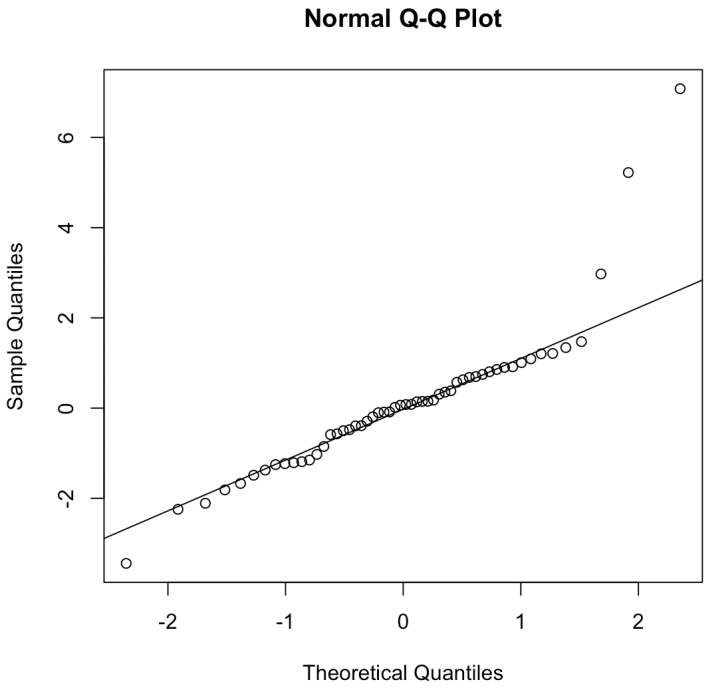
```
> aggregate(x=stress$count, by=stress[,c(1,2,3)], FUN=sum)
   smoke gender age  x
1      1      1   1 14
2      2      1   1 11
3      3      1   1 23
4      1      2   1 20
5      2      2   1  9
6      3      2   1 22
7      1      1   2 10
8      2      1   2 19
9      3      1   2 19
10     1      2   2 27
11     2      2   2 10
```

```
12     3      2   2 14
13     1      1   3 18
14     2      1   3 22
15     3      1   3  8
16     1      2   3 26
17     2      2   3 13
18     3      2   3  8
#################
> countagg$x*prob_unqi
           1        2          3
1   7.833490 4.096038  2.0704716
4   7.133935 2.692039  1.1740253
7  17.568698 3.963667  1.4676346
10  7.491436 7.127030  5.3815335
13  4.187076 2.993215  1.8197094
16 13.287221 5.933937  2.7788426
19  4.139572 3.481623  2.3788050
22  9.622166 6.017267  3.3605670
25 12.211005 4.714308  2.0746873
28  6.745212 9.499739 10.7550489
31  3.260275 3.608922  3.1308030
34  6.424177 4.686930  2.8888932
37  6.870732 6.392559  4.7367093
40 10.401329 7.257209  4.3414618
43  4.889596 2.126404  0.9840002
46  5.861464 8.932262 11.2062740
49  3.863008 4.681292  4.4557000
52  3.405299 2.760479  1.8342214

############pearson residuals#######################
> p_res=(stress$count-stress$expected)/sqrt(stress$expected*as.numeric(t(1-prob_unqi)))
> p_res
 [1]  0.62799351 -1.23133658  0.69981038 -0.08458478 -1.18663977  0.80655976 -2.24302774  0.
 [9] -1.25206540 -1.15100835  1.34138158 -0.19237283  5.22125536  7.07987928  0.14963123  0.
[17]  1.47287631 -0.49984953 -1.37367660  1.00790354  1.20404939  1.09111533 -1.48801685  0.
[25]  0.85646083  0.15174539  0.68064500 -2.10948200 -1.81338970 -3.44171623 -0.85019226  0.
[33] -0.08919426  0.30883698 -0.38903230 -0.58704202  2.97378839 -1.67092906  1.21146975 -1.
[41] -0.57009098  0.35277428  0.08007262 -0.10116480  0.01722331  0.06501790 -1.21093482 -0.
[49]  0.08313843 -0.39363528  0.90241670 -0.28981064  0.17813516  0.13942912
```

H-L statistics?


p_res
6
4
2
0
-2
0
10
20
30
40
50
Index


Normal Q-Q Plot
Sample Quantiles
6
4
2
0
-2
-2
-1
0
1
2
Theoretical Quantiles