NON-LINEAR REGRESSIONS

PROF. XINXIN LI

POTENTIAL PROBLEMS IN USING LINEAR REGRESSIONS

- The outcome variable is assumed to be normally distributed which may not be true
- Using a linear regression could result in the predicted outcome variables being senseless or outside of allowable range
- For example,
  - When outcome variable is binary, which likely arises when we have fine grained individual level data
  - When outcome variable is a count variable which can only be non-negative

I. LOGISTICS REGRESSION

LOGISTIC REGRESSION

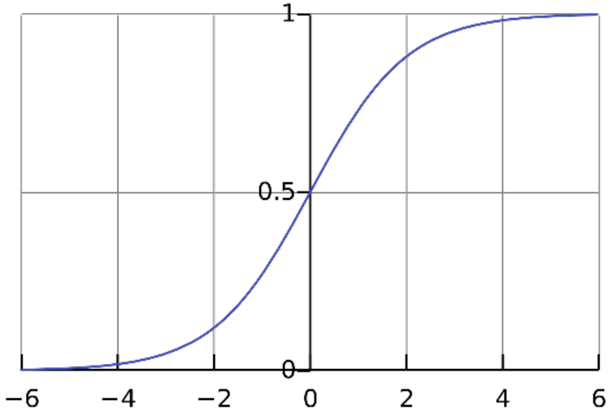
- A widely used statistical model to model a binary outcome variable Y).
  - e.g., whether to click an ad, whether to make a purchase, whether to exit etc.
- How to transform a binary variable (Y) to an unbounded continuous variable?
  - First, consider the probability of Y=1 instead of Y itself as the outcome variable, but it is still bounded between 0 and 1
  - Second, consider the odds ratio: $\frac{Prob\{Y=1\}}{1-Prob\{Y=1\}}$, which measures the probability of “success” over “failure”, but it is still non-negative
  - Third, take a log transformation: $\ln\left(\frac{Prob\{Y=1\}}{1-Prob\{Y=1\}}\right)$, which is now an unbounded continuous variable

LOGISTICS REGRESSION

- We can then model $\ln\left(\frac{Prob\{Y=1\}}{1-Prob\{Y=1\}}\right)$ using a linear function:

$$\ln\left(\frac{Prob\{Y=1|X\}}{1-Prob\{Y=1|X\}}\right) = \alpha + \beta X$$

- This is equivalent to: $Prob\{Y = 1|X\} = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$



- Use Maximum Likelihood estimation to find the coefficients that maximize the chance of observing the data we observed.

LOGISTICS REGRESSION

- How to interpret the coefficients?
  - The coefficient (β) measures the change in log odds ratio of the outcome variable (Y) upon one unit increase in the explanatory variable (X):

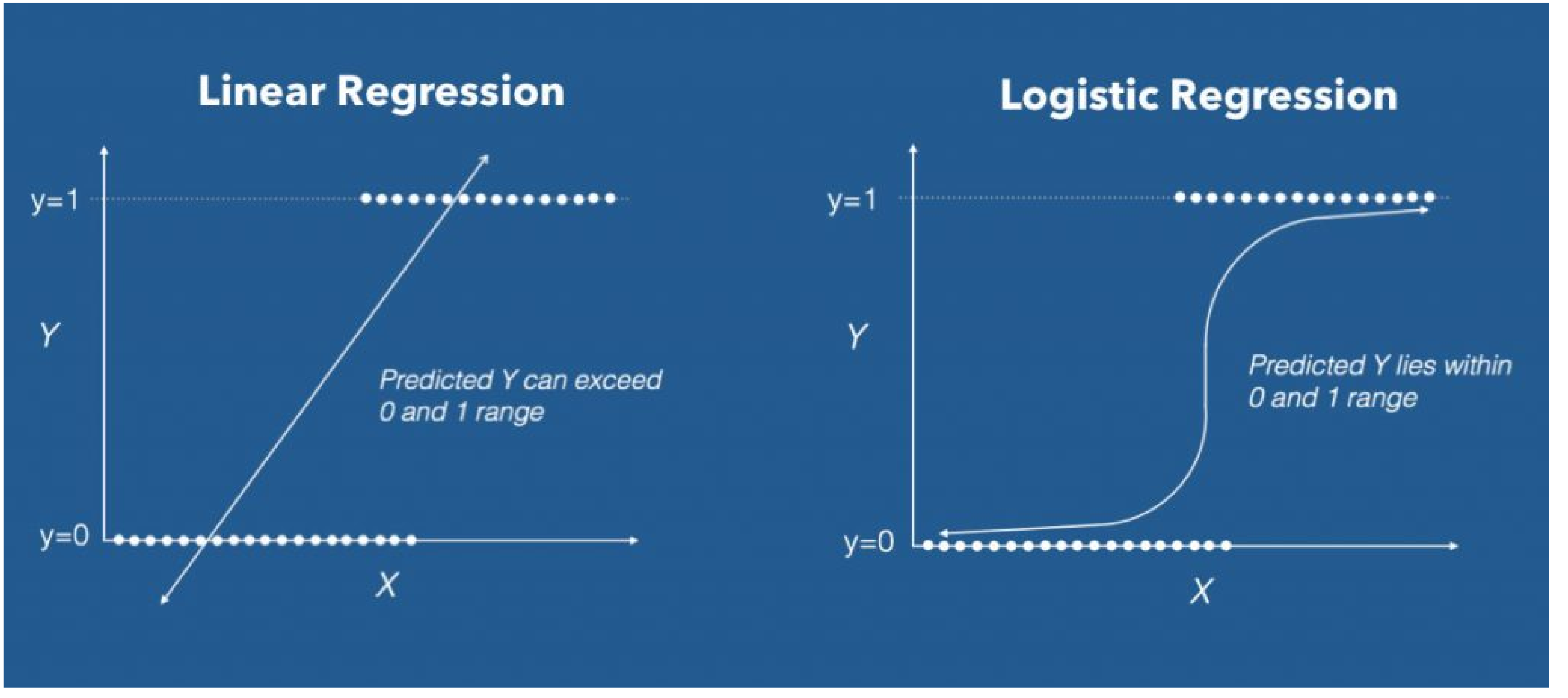
$$\ln\left(\frac{Prob\{Y=1|X\}}{1-Prob\{Y=1|X\}}\right) - \ln\left(\frac{Prob\{Y_0=1|X_0\}}{1-Prob\{Y_0=1|X_0\}}\right) = \beta(X - X_0)$$

  - We can also take exponent of the coefficient to get the percent change in odds ratio of the outcome variable upon one unit increase in X.

$$\frac{\dfrac{Prob\{Y=1|X\}}{1-Prob\{Y=1|X\}}}{\dfrac{Prob\{Y_0=1|X_0\}}{1-Prob\{Y_0=1|X_0\}}} = e^{\beta(X-X_0)}$$

  - If X increases by 1 unit, then the odds of the event increases by a factor of e^{β}

LINEAR VS. LOGISTICS REGRESSION



THE WEATHER IMPACT EXAMPLE – INDIVIDUAL DATA

- Does weather (rainy, sunny or cloudy) have any impact on people's propensity to click an ad?
- Data (MobileAd.xls)
  - ID: individual ID
  - $\text{Responded} = \begin{cases} 1 & \text{if responded to the ad and purchased} \\ 0 & \text{otherwise} \end{cases}$
  - $\text{Sunny} = \begin{cases} 1 & \text{if weather is sunny} \\ 0 & \text{otherwise} \end{cases}$
  - $\text{Rainy} = \begin{cases} 1 & \text{if weather is rainy} \\ 0 & \text{otherwise} \end{cases}$
  - $\text{AdVersion} = \begin{cases} 1 & \text{if ad has prevention framing "Do not miss the opportunity to take advantage of this special deal! "} \\ 0 & \text{otherwise} \end{cases}$
  - Location: id of the location
- How to set up the model?

DOES WEATHER HAVE AN IMPACT?

```
> summary(ml<-glm(Responded~Sunny+Rainy+AdVersion+as.factor(Location), family=binomial("
logit“), data=MobileAd))

Call:
glm(formula = Responded ~ Sunny + Rainy + AdVersion + as.factor(Location),
    family = binomial("logit"), data = MAData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7274  -0.8785  -0.6940   0.9524   2.2732

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.94474    0.25433  -3.715 0.000204 ***
Sunny                   0.31664    0.06464   4.898 9.66e-07 ***
Rainy                  -0.36505    0.07350  -4.967 6.80e-07 ***
AdVersion               0.59158    0.07974   7.419 1.18e-13 ***
as.factor(Location)2    0.93061    0.32419   2.871 0.004097 **
......
(omitted)
......
as.factor(Location)31   0.67400    0.31287   2.154 0.031223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16355  on 11955  degrees of freedom
Residual deviance: 14347  on 11922  degrees of freedom
AIC: 14415

Number of Fisher Scoring iterations: 11
```

DOES WEATHER HAVE AN IMPACT?

```
> summary(ml<-glm(Responded~Sunny+Rainy+AdVersion+as.factor(Location), family=binomial("
logit“), data=MobileAd))

Call:
glm(formula = Responded ~ Sunny + Rainy + AdVersion + as.factor(Location),
    family = binomial("logit"), data = MAData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7274  -0.8785  -0.6940   0.9524   2.2732

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.94474    0.25433  -3.715 0.000204 ***
Sunny                   0.31664    0.06464   4.898 9.66e-07 ***
Rainy                  -0.36505    0.07350  -4.967 6.80e-07 ***
AdVersion               0.59158    0.07974   7.419 1.18e-13 ***
as.factor(Location)2    0.93061    0.32419   2.871 0.004097 **
......
(omitted)
......
as.factor(Location)31   0.67400    0.31287   2.154 0.031223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16355  on 11955  degrees of freedom
Residual deviance: 14347  on 11922  degrees of freedom
AIC: 14415

Number of Fisher Scoring iterations: 11
```

Yes, compared to cloudy days, people are more likely to respond to ad on sunny days and less likely to respond on rainy days

DOES WEATHER IMPACT VARY WITH AD DESIGN?

```
> summary(iml<-glm(Responded~Sunny+Rainy+AdVersion+Sunny:AdVersion+Rainy:AdVersion+as.fa
ctor(Location),family=binomial("logit"), data=MobileAd))

Call:
glm(formula = Responded ~ Sunny + Rainy + AdVersion + Sunny:AdVersion +
    Rainy:AdVersion + as.factor(Location), family = binomial("logit"),
    data = MAData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6085  -0.8837  -0.6475   0.9299   2.3990

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.10028    0.26109  -4.214 2.51e-05 ***
Sunny                   0.61253    0.09939   6.163 7.14e-10 ***
Rainy                  -0.47510    0.13836  -3.434 0.000595 ***
AdVersion               0.76349    0.10268   7.436 1.04e-13 ***
Sunny:AdVersion        -0.57788    0.13417  -4.307 1.65e-05 ***
Rainy:AdVersion         0.17569    0.15814   1.111 0.266584
as.factor(Location)2    1.03439    0.32826   3.151 0.001627 **
......
(omitted)
......
as.factor(Location)31   0.66223    0.31728   2.087 0.036867 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16355  on 11955  degrees of freedom
Residual deviance: 14323  on 11920  degrees of freedom
AIC: 14395

Number of Fisher Scoring iterations: 11
```

DOES WEATHER IMPACT VARY WITH AD DESIGN?

```
> summary(iml<-glm(Responded~Sunny+Rainy+AdVersion+Sunny:AdVersion+Rainy:AdVersion+as.fa
ctor(Location),family=binomial("logit"), data=MobileAd))

Call:
glm(formula = Responded ~ Sunny + Rainy + AdVersion + Sunny:AdVersion + 
    Rainy:AdVersion + as.factor(Location), family = binomial("logit"), 
    data = MAData)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6085  -0.8837  -0.6475   0.9299   2.3990  

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -1.10028    0.26109  -4.214 2.51e-05 ***
Sunny                   0.61253    0.09939   6.163 7.14e-10 ***
Rainy                  -0.47510    0.13836  -3.434 0.000595 ***
AdVersion               0.76349    0.10268   7.436 1.04e-13 ***
Sunny:AdVersion        -0.57788    0.13417  -4.307 1.65e-05 ***
Rainy:AdVersion         0.17569    0.15814   1.111 0.266584    
as.factor(Location)2    1.03439    0.32826   3.151 0.001627 ** 
......
(omitted)
......
as.factor(Location)31   0.66223    0.31728   2.087 0.036867 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16355  on 11955  degrees of freedom
Residual deviance: 14323  on 11920  degrees of freedom
AIC: 14395

Number of Fisher Scoring iterations: 11
```

Sunny weather effect varies with ad version, but rainy weather effect does not vary with ad version

WEATHER IMPACT FOR AD WITHOUT PREVENTION FRAMING

```
> summary(iml<-glm(Responded~Sunny+Rainy+AdVersion+Sunny:AdVersion+Rainy:AdVersion+as.factor(Location),family
=binomial("logit"), data=MobileAd))

Call:
glm(formula = Responded ~ Sunny + Rainy + AdVersion + Sunny:AdVersion +
    Rainy:AdVersion + as.factor(Location), family = binomial("logit"),
    data = MAData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6085  -0.8837  -0.6475   0.9299   2.3990

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.10028    0.26109  -4.214 2.51e-05 ***
Sunny                   0.61253    0.09939   6.163 7.14e-10 ***
Rainy                  -0.47510    0.13836  -3.434 0.000595 ***
AdVersion               0.76349    0.10268   7.436 1.04e-13 ***
Sunny:AdVersion        -0.57788    0.13417  -4.307 1.65e-05 ***
Rainy:AdVersion         0.17569    0.15814   1.111 0.266584
as.factor(Location)2    1.03439    0.32826   3.151 0.001627 **
......
(omitted)
......
as.factor(Location)31   0.66223    0.31728   2.087 0.036867 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16355  on 11955  degrees of freedom
Residual deviance: 14323  on 11920  degrees of freedom
AIC: 14395

Number of Fisher Scoring iterations: 11
```

Sunny and rainy effect compared to cloudy

```
> linearHypothesis(iml,"Sunny=Rainy")
Linear hypothesis test

Hypothesis:
Sunny - Rainy = 0

Model 1: restricted model
Model 2: Responded ~ Sunny + Rainy + AdVersion + Sunny:AdVersion + Rainy:AdVersion +
    as.factor(Location)

  Res.Df Df  Chisq Pr(>Chisq)
1  11921
2  11920  1 59.821  1.039e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sunny compared to rainy

WEATHER IMPACT FOR AD WITH PREVENTION FRAMING

```
> linearHypothesis(iml,"Sunny+Sunny:AdVersion=0")
Linear hypothesis test

Hypothesis:
Sunny  + SunnyxAdVersion = 0

Model 1: restricted model
Model 2: Responded ~ Sunny + Rainy + AdVersion + SunnyxAdVersion + RainyxAdVersion +
    as.factor(Location)

  Res.Df Df Chisq Pr(>Chisq)
1  11921
2   11920  1 0.155     0.6938
```

Sunny effect is the same as cloudy

```
> linearHypothesis(iml,"Rainy+Rainy:AdVersion=0")
Linear hypothesis test

Hypothesis:
Rainy  + RainyxAdVersion = 0

Model 1: restricted model
Model 2: Responded ~ Sunny + Rainy + AdVersion + SunnyxAdVersion + RainyxAdVersion +
    as.factor(Location)

  Res.Df Df  Chisq Pr(>Chisq)
1  11921
2  11920  1 12.424   0.000424 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rainy effect is different from cloudy

```
> linearHypothesis(iml,"Sunny+Sunny:AdVersion=Rainy+Rainy:AdVersion")
Linear hypothesis test

Hypothesis:
Sunny - Rainy  + Sunny:AdVersion - Rainy:AdVersion = 0

Model 1: restricted model
Model 2: Responded ~ Sunny + Rainy + AdVersion + Sunny:AdVersion + Rainy:AdVersion +
    as.factor(Location)

  Res.Df Df  Chisq Pr(>Chisq)
1  11921
2  11920  1 8.6555   0.003261 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sunny effect is different from rainy

II. POISSON AND NEGATIVE BINOMIAL REGRESSION

POISSON REGRESSION

- A statistical model to model count data (non-negative integers).
  - e.g., number of ad clicks, number of purchases, number of pageviews etc.
- It assumes that the outcome variable (Y) follows a Poisson distribution:

$$\Pr(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \text{ where } y = 0,1,2,\ldots$$

  - λ is the mean or the expected value of the poisson distribution
  - λ is also variance of the poisson distribution
- The logarithm of the expected value of the outcome variable (Y) can be modeled as a linear function of explanatory variables (X):

$$Ln\big(E(Y|\mathrm{X})\big) = \alpha + \beta X$$

NEGATIVE BINOMIAL REGRESSION

- A generalization of poisson regression by relaxing the assumption that variance equals mean
  - Poisson distribution is restrictive in that it requires the variance to be similar to the mean.
  - This may not be true when overdispersion is present in the data.
- It assumes that the outcomes variable (Y) follows a negative binomial distribution:

$$\Pr(Y = y|m, \alpha) = \frac{\Gamma\left(\frac{1}{\alpha}+y\right)}{y!\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{\frac{1}{\alpha}}{\frac{1}{\alpha}+m}\right)^{\frac{1}{\alpha}}\left(\frac{m}{\frac{1}{\alpha}+m}\right)^{y}, \text{ where } y = 0,1,2,\ldots$$

  - m is the mean or the expected value of the negative binomial distribution
  - α is the over dispersion parameter. When $\alpha = 0$, the negative binomial distribution is the same as the poisson distribution
  - Γ is the gamma function: $\Gamma(n) = (n-1)!$
- Again, the logarithm of the expected value of the outcome variable (Y) can be modeled as a linear function of explanatory variables (X):

$$Ln\big(E(Y|\mathrm{X})\big) = \alpha + \beta X$$

INTERPRETATION OF THE COEFFICIENTS

- The coefficient (β) measures the change in log of expected count (Y) upon one unit increase in the explanatory variable (X):

$$\ln\big(E(Y|\mathrm{X})\big) - \ln\big(E(Y_0|X_0)\big) = \beta(X - X_0)$$

- We can also take exponent of the coefficient to get the percent change in the expected count upon one unit increase in X.

$$\frac{E(Y|\mathrm{X})}{E(Y_0|X_0)} = e^{\beta(X-X_0)}$$

THE WEATHER IMPACT EXAMPLE – AGGREGATE DATA

- Does weather (rainy, sunny or cloudy) have any impact on people's propensity to click an ad?
- Data (MobileAdAggregate.xlsx)
  - $\text{Sunny} = \begin{cases} 1 & \text{if weather is sunny} \\ 0 & \text{otherwise} \end{cases}$
  - $\text{Rainy} = \begin{cases} 1 & \text{if weather is rainy} \\ 0 & \text{otherwise} \end{cases}$
  - $\text{AdVersion} = \begin{cases} 1 & \text{if ad has prevention framing "Do not miss the opportunity to take advantage of this special deal!"} \\ 0 & \text{otherwise} \end{cases}$
  - Location: id of the location
  - Purchases: the number of purchases
- Poisson or Negative Binomial?
  - Depends on the nature of the data
  - In the presence of overdispersion, negative binomial would be more appropriate.

DOES WEATHER HAVE AN IMPACT?

```
> summary(nml<-glm(Purchases~Sunny+Rainy+AdVersion+as.factor(Location),family=negative.b
inomial(1),data=MobileAdAggregate,control=glm.control(maxit=500)))

Call:
glm(formula = Purchases ~ Sunny + Rainy + AdVersion + as.factor(Location),
    family = negative.binomial(1), data = MAData2, control = glm.control(maxit = 500))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1712  -1.9504  -1.2295  -0.6605   4.0109

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.207e-01  8.825e-01   0.930  0.35312
Sunny                  1.096e-01  3.681e-01   0.298  0.76612
Rainy                 -1.100e+00  3.906e-01  -2.816  0.00518 **
AdVersion              1.867e+00  3.507e-01   5.325 1.96e-07 ***
as.factor(Location)2   1.245e+00  1.303e+00   0.955  0.34032
 ......
(omitted)
......

as.factor(Location)31  1.348e+00  1.333e+00   1.011  0.31278
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1) family taken to be 4.985853)

    Null deviance: 1920.8  on 339  degrees of freedom
Residual deviance: 1060.7  on 306  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 1760.5

Number of Fisher Scoring iterations: 18
```

DOES WEATHER HAVE AN IMPACT?

```
> summary(nml<-glm(Purchases~Sunny+Rainy+AdVersion+as.factor(Location),family=negative.b
inomial(1),data=MobileAdAggregate,control=glm.control(maxit=500)))

Call:
glm(formula = Purchases ~ Sunny + Rainy + AdVersion + as.factor(Location),
    family = negative.binomial(1), data = MAData2, control = glm.control(maxit = 500))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1712  -1.9504  -1.2295  -0.6605   4.0109

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            8.207e-01  8.825e-01   0.930  0.35312
Sunny                  1.096e-01  3.681e-01   0.298  0.76612
Rainy                 -1.100e+00  3.906e-01  -2.816  0.00518 **
AdVersion              1.867e+00  3.507e-01   5.325 1.96e-07 ***
as.factor(Location)2   1.245e+00  1.303e+00   0.955  0.34032
 ......
(omitted)
......

as.factor(Location)31  1.348e+00  1.333e+00   1.011  0.31278
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1) family taken to be 4.985853)

    Null deviance: 1920.8  on 339  degrees of freedom
Residual deviance: 1060.7  on 306  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 1760.5

Number of Fisher Scoring iterations: 18
```

DOES WEATHER IMPACT VARY WITH AD DESIGN?

```
> nml2<-glm(Purchases~Sunny+Rainy+AdVersion+SunnyxAdVersion+RainyxAdVersion+as.factor(Lo
cation),family=negative.binomial(1),data=MAData2,control=glm.control(maxit=500))
> summary(nml2)

Call:
glm(formula = Purchases ~ Sunny + Rainy + AdVersion + SunnyxAdVersion +
    RainyxAdVersion + as.factor(Location), family = negative.binomial(1),
    data = MAData2, control = glm.control(maxit = 500))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2522  -1.8230  -1.2427  -0.5357   4.3238

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.581e-01  9.192e-01   0.825 0.410167
Sunny                  2.968e-01  4.126e-01   0.719 0.472505
Rainy                 -1.712e+00  5.040e-01  -3.397 0.000772 ***
AdVersion              1.575e+00  5.395e-01   2.920 0.003765 **
Sunny:AdVersion       -9.509e-01  8.980e-01  -1.059 0.290509
Rainy:AdVersion        1.495e+00  8.609e-01   1.736 0.083546 .
as.factor(Location)2   1.321e+00  1.341e+00   0.986 0.325136
......
(omitted)
......
as.factor(Location)31  1.315e+00  1.409e+00   0.933 0.351458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1) family taken to be 5.180961)

    Null deviance: 1920.8  on 339  degrees of freedom
Residual deviance: 1039.7  on 304  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 1743.5

Number of Fisher Scoring iterations: 18
```

WEATHER IMPACT FOR AD WITHOUT PREVENTION FRAMING

```
> nml2<-glm(Purchases~Sunny+Rainy+AdVersion+SunnyxAdVersion+RainyxAdVersion+as.factor(Lo
cation),family=negative.binomial(1),data=MAData2,control=glm.control(maxit=500))
> summary(nml2)

Call:
glm(formula = Purchases ~ Sunny + Rainy + AdVersion + SunnyxAdVersion +
    RainyxAdVersion + as.factor(Location), family = negative.binomial(1),
    data = MAData2, control = glm.control(maxit = 500))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2522  -1.8230  -1.2427  -0.5357   4.3238

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.581e-01  9.192e-01   0.825 0.410167
Sunny                  2.968e-01  4.126e-01   0.719 0.472505
Rainy                 -1.712e+00  5.040e-01  -3.397 0.000772 ***
AdVersion              1.575e+00  5.395e-01   2.920 0.003765 **
Sunny:AdVersion       -9.509e-01  8.980e-01  -1.059 0.290509
Rainy:AdVersion        1.495e+00  8.609e-01   1.736 0.083546 .
as.factor(Location)2   1.321e+00  1.341e+00   0.986 0.325136
......
(omitted)
......
as.factor(Location)31  1.315e+00  1.409e+00   0.933 0.351458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1) family taken to be 5.180961)

    Null deviance: 1920.8  on 339  degrees of freedom
Residual deviance: 1039.7  on 304  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 1743.5

Number of Fisher Scoring iterations: 18
```

WEATHER IMPACT FOR AD WITH PREVENTION FRAMING

```
> linearHypothesis(nml2,"Sunny+Sunny:AdVersion=0")
Linear hypothesis test

Hypothesis:
Sunny  + SunnyxAdVersion = 0

Model 1: restricted model
Model 2: Purchases ~ Sunny + Rainy + AdVersion + SunnyxAdVersion + RainyxAdVersion + 
    as.factor(Location)

  Res.Df Df  Chisq Pr(>Chisq)
1    305                     
2    304  1 0.6469     0.4212


> linearHypothesis(nml2,"Rainy+Rainy:AdVersion=0")
Linear hypothesis test

Hypothesis:
Rainy  + RainyxAdVersion = 0

Model 1: restricted model
Model 2: Purchases ~ Sunny + Rainy + AdVersion + SunnyxAdVersion + RainyxAdVersion + 
    as.factor(Location)

  Res.Df Df  Chisq Pr(>Chisq)
1    305                     
2    304  1 0.0918     0.7619
```

III. PROPENSITY SCORE MATCHING

PROPENSITY SCORE MATCHING (PSM)

- Quasi-experiment does not have random assignment between control and treatment groups. As a result, subjects in these two groups are not comparable. What can we do?

- **Idea**: match subjects in control group to those in treatment group with the same (or similar) observable characteristics
  - Find a matched control subject as the counterfactual for the treated subject

- **Challenge**: as the number of characteristics determining selection increases, it becomes more and more difficult to find comparable subjects

- **Solution**: match subjects on a single index (propensity score), reflecting the probability of a subject selecting into the treatment group

PROPENSITY SCORE MATCHING (PSM)

- **Method**
  - Estimate propensity score (PS) for each subject, which is calculated as the probability of participating in the treatment, conditional on the characteristics X

$$PS = Pr\{\text{Treatment}_i = 1|X_i\} = G(\gamma_0 + \gamma_1 X_i) = \frac{e^{\gamma_0+\gamma_1 X_i}}{1+e^{\gamma_0+\gamma_1 X_i}}$$

$$\widehat{PS}_i = \widehat{Pr}\{\text{Treatment}_i = 1|X_i\}$$

  - Match participants (in treatment group) and non-participants (in control group) with equal/similar propensity score
  - Compare outcomes of participants (in treatment group) and matched non-participants (in control group)

- **Assumptions**
  - There are no systematic differences between participants and non-participants in terms of unobserved characteristics that may influence participation
  - All the variables that affect participation decision and outcome simultaneously are observed
  - Matching is feasible, and similar propensity scores are based on similar observed X

PSM EXAMPLE

- Question of interest
  - On a website, some products were featured on the front page, whereas others were not
  - Does "being featured on the front page" increase product sales?
- Data (Feature.xlsx)
  - Product ID: ID of the product
  - Age: how long the product has been offered on the website (in months)
  - Price: price of the product
  - Feature = $\begin{cases} 1 & \text{if product was featured on the front page} \\ 0 & \text{otherwise} \end{cases}$
  - Sales: average daily sales of a product
- How to set up the model?

PSM EXAMPLE

- Treatment: being featured on the front page
- Treatment group: products that were featured on the front page
- Control group: products that were not featured on the front page
- To examine whether "being featured" increases sales:

$$\text{AvgReview}_i = \beta_0 + \beta_1 \text{Feature}_i + \beta_2 \text{Age}_i + \beta_3 \text{Price}_i$$

- But products in the treatment group and products in the control group may not be comparable
- Estimate the propensity score (PS) of each product getting featured (i.e., participating in the treatment):

$$PS = Pr\{\text{Feature}_i = 1 | X_i\} = \frac{e^{(\gamma_0 + \gamma_1 \text{Age}_i + \gamma_2 \text{Price}_i)}}{1 + e^{(\gamma_0 + \gamma_1 \text{Age}_i + \gamma_2 \text{Price}_i)}}$$

- Then, based on propensity score, match products from the treatment group and products with control group

MATCH TREATED AND CONTROL PRODUCTS

```
> summary(psm <- matchit(Feature~Age+Price, data=feature,method = "nearest", ratio=1, distance='logit', cal
iper=0.0001))

Call:
matchit(formula = Feature ~ Age + Price, data = feature, method = "nearest", 
    distance = "logit", ratio = 1, caliper = 1e-04)

Summary of balance for all data:
         Means Treated Means Control SD Control  Mean Diff  eQQ Med  eQQ Mean   eQQ Max
distance         0.490        0.4749     0.0638     0.0151   0.0087    0.0159 1.038e-01
Age              7.394        6.8348     4.8600     0.5592   1.0000    0.5950 3.000e+00
Price         4395.228     6815.3523 13608.3096 -2420.1247 381.0000 3051.9294 8.200e+04


Summary of balance for matched data:
         Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance        0.4917        0.4917     0.0315    0.0000       0   0.0000       0
Age             6.1727        6.1727     4.2968    0.0000       0   0.1439       2
Price        2320.7122     2320.7266  3630.5647   -0.0144       0  64.4317    1000

Percent Balance Improvement:
         Mean Diff. eQQ Med eQQ Mean eQQ Max
distance    99.9995     100  99.9933 99.9943
Age        100.0000     100  75.8167 33.3333
Price       99.9994     100  97.8888 98.7805

Sample sizes:
          Control Treated
All           684     637
Matched       139     139
Unmatched     545     498
Discarded       0       0
```

FEATURE EFFECT REMAINS SIGNIFICANT ON MATCHED SAMPLE

```
lm(formula = Sales ~ Feature + Age + Price, data = feature)

Residuals:
    Min      1Q  Median      3Q     Max 
 -4.361  -2.501  -1.529  -0.016 100.287 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.793e+00  3.629e-01   7.695 2.75e-14 ***
Feature      1.709e+00  3.596e-01   4.752 2.24e-06 ***
Age         -9.507e-02  3.661e-02  -2.596  0.00953 ** 
Price       -5.833e-05  1.477e-05  -3.949 8.26e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.486 on 1317 degrees of freedom
Multiple R-squared:  0.03569,	Adjusted R-squared:  0.03349 
F-statistic: 16.25 on 3 and 1317 DF,  p-value: 2.26e-10
```

Use original data

```
lm(formula = Sales ~ Feature + Age + Price, data = data_psm)

Residuals:
   Min     1Q Median     3Q    Max 
-5.379 -3.015 -2.019  0.149 47.825 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.7342920  0.8724672   4.280 2.58e-05 ***
Feature      2.2010048  0.8335280   2.641  0.00875 ** 
Age         -0.1831139  0.0991073  -1.848  0.06573 .  
Price       -0.0002432  0.0001160  -2.095  0.03706 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.949 on 274 degrees of freedom
Multiple R-squared:  0.05399,	Adjusted R-squared:  0.04363 
F-statistic: 5.213 on 3 and 274 DF,  p-value: 0.001626
```

Use matched data

IV. CONDITIONAL LOGIT MODEL

MODEL CONSUMER CHOICE

- Consumers typically consider and compare multiple options in their choice set before making a choice, e.g.,
  - after browsing multiple humidifiers on Amazon.com, select one humidifier to buy
  - after considering multiple hotels on Expedia.com, make reservation with neither hotel, etc.
- Although the outcome variable for each option is still binary (yes/no), outcomes of different options in the same choice set are not independent.
- Therefore, the standard logit model does not capture this situation well.
- What should we do?

CONDITIONAL LOGIT MODEL

- Estimates how characteristics of different alternatives in a choice set affect a user's choice among these alternatives
  - The user still makes a binary decision (yes/no) for each alternative
  - At most one alternative can have the binary choice of "yes"
  - All the other alternatives have the binary choice of "no"

- The probability of individual i choosing alternative $j = \frac{e^{V_{ij}}}{\sum_{k \in J} e^{V_{ik}}}$,
  - where J represents the set of all alternatives (that can include the "choosing non of the alternatives" option as well), called the individual's choice set.
  - User's utility of selecting alternative j:

$$V_{ij} = a + b * charateristics\ of\ alternative\ j$$

- Use Maximum Likelihood estimation to find the coefficients (a and b) that maximize the chance of observing the data we observed.

CONDITIONAL LOGIT MODEL EXAMPLE

- Suppose we want to examine how coupon offering affects consumer's choice among different restaurants they viewed on a website
- We observe the restaurants each individual browsed on the website before choosing a restaurant to dine in
- Data (Coupon.xlsx)
  - purchased = $\begin{cases} 1 & \text{if the individual dined at the restaurant} \\ 0 & \text{otherwise} \end{cases}$
  - browse_id: id of the browsing session
  - coupon = $\begin{cases} 1 & \text{if the restaurant offers coupon on the website} \\ 0 & \text{otherwise} \end{cases}$
  - review_val: review valence of the restaurant on the website
  - review_vol: review volume of the restaurant on the website (log transformed)
  - price_level: price level of the restaurant (log transformed)
  - coupon_prone: how frequent the individual used coupon in the past

```
> summary(cml<-clogit(purchased ~ review_val + review_vol + coupon + price_le
vel + coupon_prone:coupon + strata(browse_id), data=Coupon))
Call:
coxph(formula = Surv(rep(1, 7577L), purchased) ~ review_val +
    review_vol + coupon + price_level + coupon_prone:coupon +
    strata(browse_id), data = Coupon, method = "exact")

  n= 7577, number of events= 1942

                        coef exp(coef) se(coef)      z Pr(>|z|)
review_val           0.39774   1.48846  0.26704  1.489 0.136369
review_vol           0.13851   1.14856  0.05785  2.394 0.016650 *
coupon              -0.51634   0.59670  0.18492 -2.792 0.005234 **
price_level         -0.94406   0.38905  0.12458 -7.578 3.51e-14 ***
coupon:coupon_prone  0.85532   2.35212  0.24743  3.457 0.000547 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> linearHypothesis(cml,"coupon+coupon:coupon_prone=0")
Linear hypothesis test

Hypothesis:
coupon  + coupon:coupon_prone = 0

Model 1: restricted model
Model 2: Surv(rep(1, 7577L), purchased) ~ review_val + review_vol + coupon +
    price_level + coupon_prone:coupon + strata(browse_id)

  Res.Df Df  Chisq Pr(>Chisq)
1   7573
2   7572  1 4.2141    0.04009 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```