SNU - PSIR
Regression in R, with Hands-On Exercises

Instructor: **Hong Min Park**

Associate Professor
Department of Political Science
University of Wisconsin-Milwaukee
hmpark1@uwm.edu

Summarize Data

- Among others, we tend to use:
  - Average or mean
  - Median
  - Variance and standard deviation
  - Covariance and correlation

```
> ## Uploading data
>
> require(UsingR)
> attach(MLBattend)  ## attach data so tha we can use variable names
> head(MLBattend)  ## look at first several entries of data
  franchise league division year attendance runs.scored
1       BAL     AL     EAST   69    1062069         779
2       BOS     AL     EAST   69    1833246         743
3       CLE     AL     EAST   69     619970         573
4       DET     AL     EAST   69    1577481         701
5       NYA     AL     EAST   69    1067996         562
6       WAS     AL     EAST   69     918106         694
  runs.allowed wins losses games.behind
1          517  109     53          0.0
2          736   87     75         22.0
3          717   62     99         46.5
4          601   90     72         19.0
5          587   80     81         28.5
6          644   86     76         23.0
> stl.win <- wins[franchise == "STL"]
> stl.off <- runs.scored[franchise == "STL"]
> stl.def <- runs.allowed[franchise == "STL"]
```

```
> ## Mean and median
>
> MLB.win <- cbind(mean(wins), mean(stl.win),
+                  median(wins), median(stl.win))
> MLB.off <- cbind(mean(runs.scored), mean(stl.off),
+                  median(runs.scored), median(stl.off))
> MLB.def <- cbind(mean(runs.allowed), mean(stl.def),
+                  median(runs.allowed), median(stl.def))
> MLB.record <- rbind(MLB.win, MLB.off, MLB.def)
> rownames(MLB.record) <- c("win", "offense", "defense")
> colnames(MLB.record) <- c("MLB-mean", "STL-mean",
+                           "MLB-median", "STL-median")
>
> print(MLB.record)
         MLB-mean  STL-mean MLB-median STL-median
win      78.84964  80.15625       79.0       82.5
offense 694.94033 674.68750      691.5      669.5
defense 694.89141 663.00000      693.0      664.5
```

```
> ## Variance and standard deviation
>
> MLB.win.spread <- cbind(var(wins), var(stl.win),
+                         sd(wins), sd(stl.win))
> MLB.off.spread <- cbind(var(runs.scored), var(stl.off),
+                         sd(runs.scored), sd(stl.off))
> MLB.def.spread <- cbind(var(runs.allowed), var(stl.def),
+                         sd(runs.allowed), sd(stl.def))
> MLB.spread <- rbind(MLB.win.spread, MLB.off.spread, MLB.def.spread)
> rownames(MLB.spread) <- rownames(MLB.record)
> colnames(MLB.spread) <- c("MLB-var", "STL-var", "MLB-sd", "STL-sd")
>
> print(MLB.spread)
           MLB-var   STL-var    MLB-sd   STL-sd
win       160.6130  110.2006  12.67332 10.49765
offense 11061.4875 8532.2863 105.17361 92.37038
defense 11135.3920 6172.2581 105.52437 78.56372
```

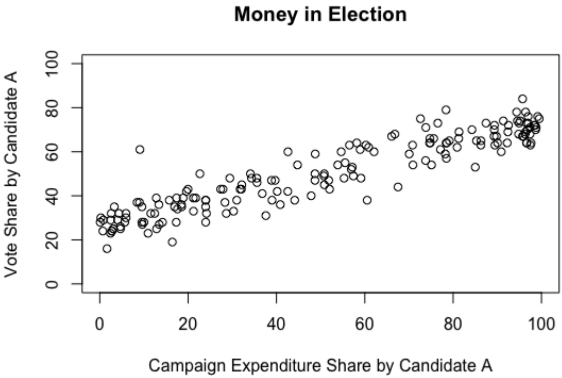
```
> ## Covarianze and Correlation
>
> head(kid.weights)   # another data: kid's weight and height
  age weight height gender
1  58     38     38      M
2 103     87     43      M
3  87     50     48      M
4 138     98     61      M
5  82     47     47      F
6  52     30     24      F
>
> cov(kid.weights$weight, kid.weights$height)
[1] 218.7377
>
> cor(kid.weights$weight, kid.weights$height)
[1] 0.8237564
```

Non-Technical Introduction to Regression (1)

- Example: Consider the role of money in election.
  - Predicted variable = vote share by candidate A
  - Predictor variable = campaign expenditure share by candidate A

Non-Technical Introduction to Regression (1)

- Example: Consider the role of money in election.
  - Predicted variable = vote share by candidate A
  - Predictor variable = campaign expenditure share by candidate A

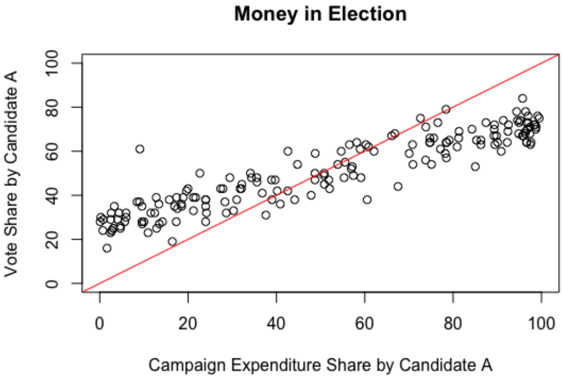


Non-Technical Introduction to Regression (2)

- Can we think of a trend line?

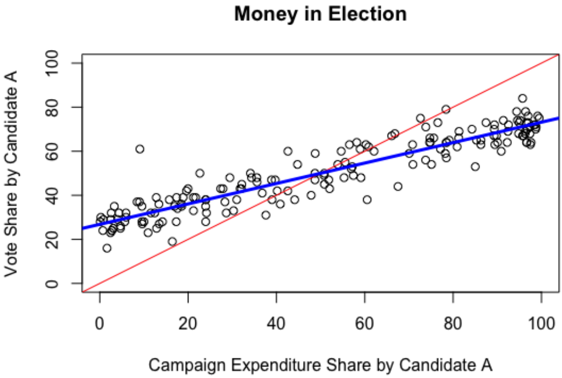
Non-Technical Introduction to Regression (2)

- Can we think of a trend line?



Non-Technical Introduction to Regression (3)

- What about this?



Non-Technical Introduction to Regression (4)

- What are we doing here?
  1. Come up with one **linear** line.
  2. **Move** the line so that it **best fits** the trend.

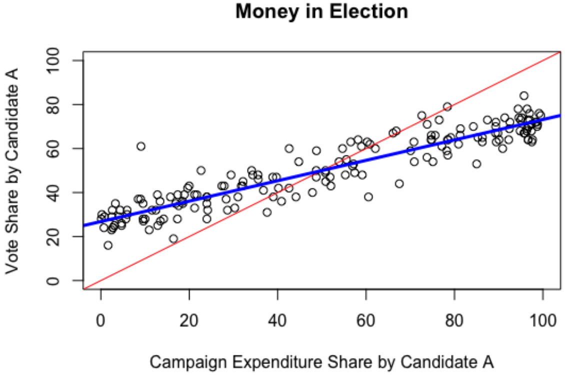
Non-Technical Introduction to Regression (4)

- What are we doing here?
  1. Come up with one **linear** line.
  2. **Move** the line so that it **best fits** the trend.

- If we may use (harmless) math...
  1. Consider the equation $Y = \alpha + \beta X$.
  2. Start with $\alpha = 0, \beta = 1$.
  3. Adjust the values for α and β.
  4. Come up with $\alpha = 26.81, \beta = 0.46$.

Non-Technical Introduction to Regression (5)

$$Y = 26.81 + 0.46X$$



```
> require(foreign)
> vote1 <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/vote1.dta")
> head(vote1)
  state district democA voteA    expendA     expendB prtystrA  lexpendA
1    NA        7      1    68 328.299988   8.7399998       41 5.7939162
2    NA        1      0    62 626.380005 402.4800110       60 6.4399519
3    NA        2      1    73  99.610001   3.0699999       55 4.6012330
4    NA        3      0    69 319.690002  26.2800007       64 5.7673521
5    NA        3      0    75 159.220001  60.0499992       66 5.0702929
6    NA        4      1    69 570.159973  21.3899994       46 6.3459082
   lexpendB    shareA
1 2.1675670 97.410004
2 5.9976382 60.880001
3 1.1200480 97.010002
4 3.2688460 92.400002
5 4.0952439 72.610001
6 3.0630641 96.379997
```

```
> m1 <- lm(voteA ~ shareA, data=vote1)
> print(m1)

Call:
lm(formula = voteA ~ shareA, data = vote1)

Coefficients:
(Intercept)       shareA
   26.81254      0.46382
```

```
> m1 <- lm(voteA ~ shareA, data=vote1)
> print(m1)

Call:
lm(formula = voteA ~ shareA, data = vote1)

Coefficients:
(Intercept)       shareA
   26.81254      0.46382

> plot(vote1$voteA ~ vote1$shareA,
+      xlim=c(0,100), ylim=c(0,100),
+      main="Money in Election",
+      xlab="Campaign Expenditure Share by Candidate A",
+      ylab="Vote Share by Candidate A")
> abline(m1, col=4, lwd=3)
```

Interpretation of Regression

- What can we say from the regression result?

Interpretation of Regression

- What can we say from the regression result?
- $Y = 26.81 + 0.46X$
  - When we change X value from 0 to 1,
  - Y changes from 26.81 to 26.81 + 0.46

Interpretation of Regression

- What can we say from the regression result?
- $Y = 26.81 + 0.46X$
  - When we change X value from 0 to 1,
  - Y changes from 26.81 to 26.81 + 0.46
- In other words,
  - One unit increase in X results in 0.46 unit increase in Y.
  - Or, when a candidate spend money for 10 percentage point more, he/she receives votes for 4.6 percentage point more.

```
> summary(m1)

Call:
lm(formula = voteA ~ shareA, data = vote1)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8924  -4.0649  -0.1697   3.4972  29.9759

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.81254    0.88719   30.22   <2e-16 ***
shareA       0.46382    0.01454   31.90   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 6.385 on 171 degrees of freedom
Multiple R-squared: 0.8561,Adjusted R-squared: 0.8553
F-statistic:  1018 on 1 and 171 DF,  p-value: < 2.2e-16
```

Multiple Regression

- Multiple regression model allows us to *explicitly* control for many other factors that simultaneously affect y.
- Consider $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Then, we have:

$$\Delta y = \beta_1 \Delta x_1 + \beta_2 \Delta x_2$$

  - β_1 measures the change in y with respect to x_1, holding other factors fixed (i.e. $\Delta x_2 = 0$). Similarly, β_2 measures the change in y with respect to x_2, holding other factors fixed (i.e. $\Delta x_1 = 0$).

```
> mod2 <- lm(voteA ~ shareA + prtystrA, data=vote1)
> coef(mod2)
(Intercept)      shareA    prtystrA
 19.8504237   0.4508902   0.1531982
>
> ## try "summary(mod2)" as well
```

```
> mod2 <- lm(voteA ~ shareA + prtystrA, data=vote1)
> coef(mod2)
(Intercept)      shareA    prtystrA
 19.8504237   0.4508902   0.1531982
>
> ## try "summary(mod2)" as well
```

```
> cbind(coef(mod2), c(coef(m1), NA))
                  [,1]       [,2]
(Intercept) 19.8504237 26.8125373
shareA       0.4508902  0.4638239
prtystrA     0.1531982         NA
```

Functional Form: Quadratics

- Quadratic functions are used to capture decreasing or increasing effects of independent variables.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

  - β_1 does not measure the change in y with respect to x: it doesn't make sense to hold x^2 fixed while changing x.
  - Rather, we have $\frac{\Delta \hat{y}}{\Delta x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$: the slope of the relationship between x and y depends on the value of x.

Functional Form: Quadratics

- Quadratic functions are used to capture decreasing or increasing effects of independent variables.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

  - β_1 does not measure the change in y with respect to x: it doesn't make sense to hold x^2 fixed while changing x.
  - Rather, we have $\frac{\Delta \hat{y}}{\Delta x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$: the slope of the relationship between x and y depends on the value of x.

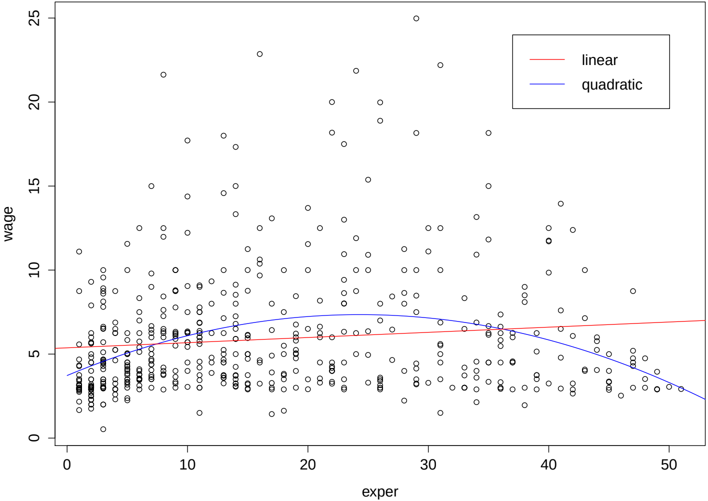
- Interpretation
  - What if $\beta_1 > 0$ and $\beta_2 < 0$?
  - What if $\beta_1 < 0$ and $\beta_2 > 0$?

- The turning point is achieved at $x = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$.

```
> ## Quadratics
> wage1 <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
> wage.mod3 <- lm(wage ~ exper + I(exper^2), data=wage1)
> summary(wage.mod3)

Call:
lm(formula = wage ~ exper + I(exper^2), data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5916 -2.1440 -0.8603  1.1801 17.7649

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.7254058  0.3459392  10.769  < 2e-16 ***
exper        0.2981001  0.0409655   7.277 1.26e-12 ***
I(exper^2)  -0.0061299  0.0009025  -6.792 3.02e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 3.524 on 523 degrees of freedom
Multiple R-squared: 0.09277,Adjusted R-squared: 0.0893
F-statistic: 26.74 on 2 and 523 DF,  p-value: 8.774e-12
```

```
> peak.mod3 <- - coef(wage.mod3)[2]/(2*coef(wage.mod3)[3])
> print(peak.mod3)
  exper
24.3153
```

```
> plot(wage ~ exper, data=wage1)
> abline(lm(wage ~ exper, data=wage1), col=2)
> x <- seq(0, 55, 0.001)
> y <- predict(wage.mod3, newdata=data.frame(exper=x))
> lines(y ~ x, col=4)
> legend(37, 24, c("linear", "quadratic"), col=c(2, 4), lty=1)
```



linear
quadratic
wage
exper
0
5
10
15
20
25
0
10
20
30
40
50

Dummy Variables

- We sometimes have binary information.
  - This zero-one variable is called **dummy variable**.
  - And, we can just add it as an IV:

$$y = \beta_0 + \gamma_0 z + \beta_1 x + \epsilon$$

where z is dummy.

Dummy Variables

- We sometimes have binary information.
  - This zero-one variable is called **dummy variable**.
  - And, we can just add it as an IV:

$$y = \beta_0 + \gamma_0 z + \beta_1 x + \epsilon$$

    where z is dummy.

- Then, we have $\gamma_0 = E(y|z = 1, x) - E(y|z = 0, x)$.
  - Can be understood as an **intercept shift** between $z = 0$ and $z = 1$.

Dummy Variables

- We sometimes have binary information.
  - This zero-one variable is called **dummy variable**.
  - And, we can just add it as an IV:

    $$y = \beta_0 + \gamma_0 z + \beta_1 x + \epsilon$$

    where z is dummy.

- Then, we have $\gamma_0 = E(y|z=1,x) - E(y|z=0,x)$.
  - Can be understood as an **intercept shift** between $z=0$ and $z=1$.

- The use of dummies
  - Make sure to use *one* dummy for *two* mutually exclusive groups $\rightarrow$ perfect collinearity if one dummy for each group.
  - $z=0$ becomes "control" group and $z=1$ is "experiment" group $\rightarrow$ very useful for policy analysis and program evaluation.

- We can use a set of dummy variables for multiple categories.
  - For example, we have four (4) groups: married men, married women, single men, and single women
  - Select a base group: single men
  - Define three (3) dummies for the other groups

- The estimate on the three dummy variables measure the **proportionate** difference in y **relative to** single men.

- We can use a set of dummy variables for multiple categories.
  - For example, we have four (4) groups: married men, married women, single men, and single women
  - Select a base group: single men
  - Define three (3) dummies for the other groups

- The estimate on the three dummy variables measure the **proportionate** difference in y **relative to** single men.

- What about ordinal variable?
  - One alternative is to use it as "continuous."
  - However, the dummy would be a better choice *if* the difference between 1 and 2 is not the same as the difference between 2 and 3.

```
> ## Dummy
> wage.mod5 <- lm(log(wage) ~ educ + exper + female, data=wage1)
> summary(wage.mod5)

Call:
lm(formula = log(wage) ~ educ + exper + female, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89584 -0.26362 -0.03871  0.26765  1.28241

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.480836   0.105016   4.579 5.86e-06 ***
educ         0.091290   0.007123  12.816  < 2e-16 ***
exper        0.009414   0.001449   6.496 1.93e-10 ***
female      -0.343597   0.037667  -9.122  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 0.4289 on 522 degrees of freedom
Multiple R-squared: 0.3526,Adjusted R-squared: 0.3488
F-statistic: 94.75 on 3 and 522 DF,  p-value: < 2.2e-16
```

What if the x-y relationship is NOT linear?

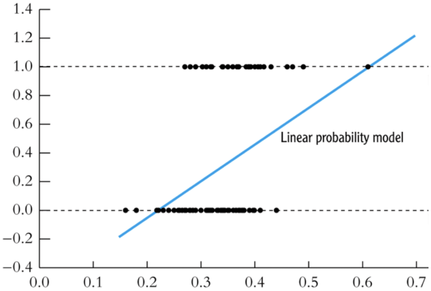
Will Come Back After SHORT Break

Binary Response Models

- There are naturally binary social outcomes:
  - A citizen votes or does not
  - A cabinet forms or does not
  - A child is born or not
  - A refrigerator is bought or not

Binary Response Models

- There are naturally binary social outcomes:
  - A citizen votes or does not
  - A cabinet forms or does not
  - A child is born or not
  - A refrigerator is bought or not
- How to characterize binary outcomes via OLS?



Logit and Probit Models

- Instead, we take an alternative approach:

$$y_i \sim \begin{cases} 1 & \Pr = \pi_i \\ 0 & \Pr = 1 - \pi_i \end{cases}$$

where $\pi_i = f(\beta, x_i)$

Logit and Probit Models

- Instead, we take an alternative approach:

$$y_i \sim \begin{cases} 1 & \Pr = \pi_i \\ 0 & \Pr = 1 - \pi_i \end{cases}$$

  where $\pi_i = f(\beta, x_i)$

- But, π_i represents a *probability* so it must be bounded by $[0, 1]$.
  - So, $\pi_i = x_i\beta$ is a bad idea since this linear function is unbounded and so might well fall outside the $[0, 1]$ interval.

Logit and Probit Models

- Instead, we take an alternative approach:

$$y_i \sim \begin{cases} 1 & \Pr = \pi_i \\ 0 & \Pr = 1 - \pi_i \end{cases}$$

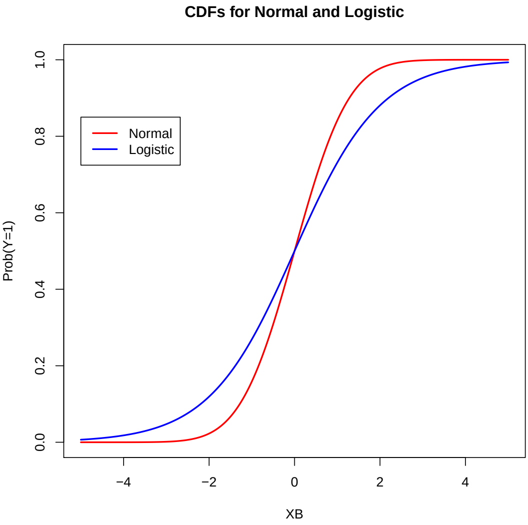
  where $\pi_i = f(\beta, x_i)$

- But, π_i represents a *probability* so it must be bounded by $[0, 1]$.
  - So, $\pi_i = x_i\beta$ is a bad idea since this linear function is unbounded and so might well fall outside the $[0, 1]$ interval.

- Actually, we can take *any* probability distribution function, but...
  - **CDF** of normal and logistic

$$\pi_i = f(\beta, x_i) = \mathit{CDF} \text{ of normal or logistic}$$

- Very little difference between the two
- When considering heteroskedasticity probit becomes more tractable
- When extended to multiple outcomes logit becomes more tractable



```
> library(car)
> head(Mroz)    ## We are using Mroz (1987) data
  lfp k5 k618 age  wc hc       lwg    inc
1 yes  1    0  32  no no 1.2101647 10.910
2 yes  0    2  30  no no 0.3285041 19.500
3 yes  1    3  35  no no 1.5141279 12.040
4 yes  0    3  34  no no 0.0921151  6.800
5 yes  1    2  31 yes no 1.5242802 20.100
6 yes  0    0  54  no no 1.5564855  9.859
>
> (n <- nrow(Mroz))
[1] 753
>
> ?Mroz
starting httpd help server ... done
```

```
U.S. Women's Labor-Force Participation

Description

The Mroz data frame has 753 rows and 8 columns. The observations, from the
Panel Study of Income Dynamics (PSID), are married women.

:

lfp
labor-force participation; a factor with levels: no; yes.

k5
number of children 5 years old or younger.

k618
number of children 6 to 18 years old.

age
in years.
```

```
wc
wife's college attendance; a factor with levels: no; yes.

hc
husband's college attendance; a factor with levels: no; yes.

lwg
log expected wage rate; for women in the labor force, the actual wage rate;
for women not in the labor force, an imputed value based on the regression
of lwg on the other variables.

inc
family income exclusive of wife's income.
```

wc
wife's college attendance; a factor with levels: no; yes.

hc
husband's college attendance; a factor with levels: no; yes.

lwg
log expected wage rate; for women in the labor force, the actual wage rate; for women not in the labor force, an imputed value based on the regression of lwg on the other variables.

inc
family income exclusive of wife's income.

```
> mroz.probit <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc,
+                      family = binomial(link=probit), data=Mroz)
> mroz.logit <- glm(lfp ~ k5 + k618 + age + wc + hc + lwg + inc,
+                     family = binomial(link=logit), data=Mroz)
```

```
> summary(mroz.probit)

Call:
glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial(link = probit),
    data = Mroz)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1358  -1.1024   0.5967   0.9746   2.2236

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.918418   0.382356   5.017 5.24e-07 ***
k5          -0.874712   0.114423  -7.645 2.10e-14 ***
k618        -0.038595   0.040950  -0.942 0.345942
age         -0.037824   0.007605  -4.973 6.58e-07 ***
wcyes        0.488310   0.136731   3.571 0.000355 ***
hcyes        0.057172   0.124207   0.460 0.645306
lwg          0.365635   0.089992   4.063 4.85e-05 ***
inc         -0.020525   0.004852  -4.230 2.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.39  on 745  degrees of freedom
AIC: 921.39

Number of Fisher Scoring iterations: 4
```

```
> summary(mroz.logit)

Call:
glm(formula = lfp ~ k5 + k618 + age + wc + hc + lwg + inc, family = binomial(link = logit),
    data = Mroz)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1062  -1.0900   0.5978   0.9709   2.1893

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.182140   0.644375   4.938 7.88e-07 ***
k5          -1.462913   0.197001  -7.426 1.12e-13 ***
k618        -0.064571   0.068001  -0.950 0.342337
age         -0.062871   0.012783  -4.918 8.73e-07 ***
wcyes        0.807274   0.229980   3.510 0.000448 ***
hcyes        0.111734   0.206040   0.542 0.587618
lwg          0.604693   0.150818   4.009 6.09e-05 ***
inc         -0.034446   0.008208  -4.196 2.71e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  905.27  on 745  degrees of freedom
AIC: 921.27

Number of Fisher Scoring iterations: 4
```

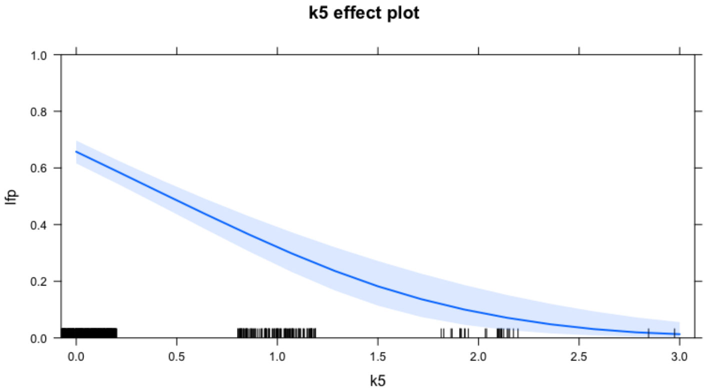
Plot of Predicted Probabilities

- How to interpret the size of coefficients?

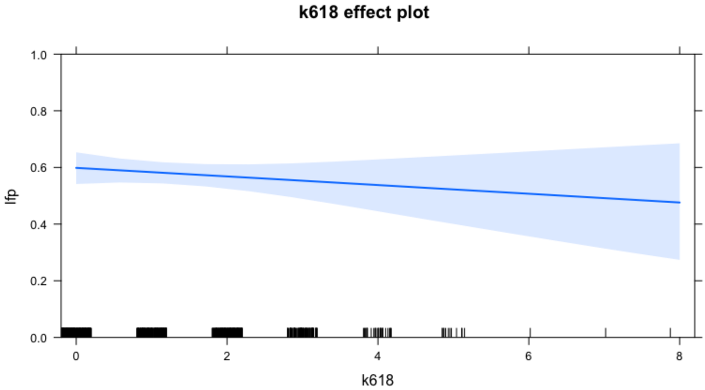
Plot of Predicted Probabilities

- How to interpret the size of coefficients?
- Consider "potential" predicted probabilities based on a set of **hypothetical** (but theoretically interesting) X values.
  - Pick ONE independent variable of interest. Make it vary from its observed minimum to its observed maximum.
  - Choose "average" values for all the other independent variables.
  - Then, see how the predicted probability changes as our IV of interest moves from min to max.
- Visualizing this process is the most popular way

```
> require(effects)
> mroz.eff <- allEffects(mroz.probit)
> plot(mroz.eff, 'k5',
+       rescale.axis=FALSE, ylim=c(0,1))
```



```
> plot(mroz.eff, 'k618',
+       rescale.axis=FALSE, ylim=c(0,1))
```



Event Count Responses

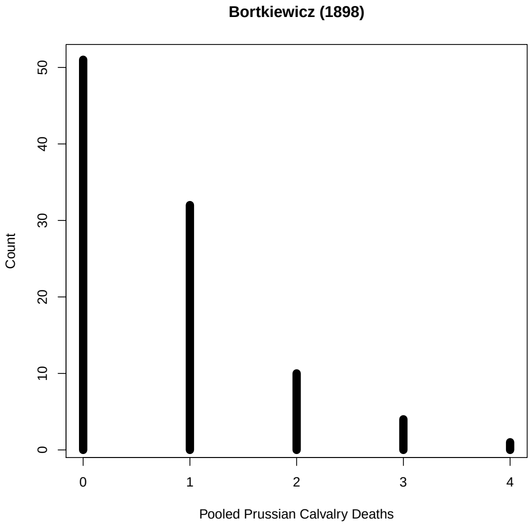
- Suppose the social system produces events over time:
  - Coup d'etat
  - Collapse of cabinet government
  - Outbreak of war
  - Veto exercised by the president

Event Count Responses

- Suppose the social system produces events over time:
  - Coup d'etat
  - Collapse of cabinet government
  - Outbreak of war
  - Veto exercised by the president
- Simeon-Denis Poisson, 1837: Poisson distribution
  - Describes the probability that a **random** event occurs in a time or space **interval** when the probability of the event occurring is **very small**, but the number of trials is **very large**.

Event Count Responses

- Suppose the social system produces events over time:
  - Coup d'etat
  - Collapse of cabinet government
  - Outbreak of war
  - Veto exercised by the president
- Simeon-Denis Poisson, 1837: Poisson distribution
  - Describes the probability that a **random** event occurs in a time or space **interval** when the probability of the event occurring is **very small**, but the number of trials is **very large**.
  - Example: Bortkiewicz (1898) *The Law of Small Numbers*.
    : the number of members of 14 prussian cavalry units killed by being kicked by a horse from 1875-1894.


Bortkiewicz (1898)
Count
0
10
20
30
40
50
0
1
2
3
4
Pooled Prussian Calvalry Deaths

The Poisson Model

- If the process generates events **independently** and at a **fixed rate** within time periods, then the result is a Poisson process:

$$f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

where $E(y) = \lambda_i$ and $\mathrm{Var}(y) = \lambda_i$.

The Poisson Model

- If the process generates events **independently** and at a **fixed rate** within time periods, then the result is a Poisson process:

$$f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

  where $E(y) = \lambda_i$ and $\text{Var}(y) = \lambda_i$.

- Note that:
  - λ_i should be non-negative and continuous.
  - It should be via X_i and β.

The Poisson Model

- If the process generates events **independently** and at a **fixed rate** within time periods, then the result is a Poisson process:

$$f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

  where $E(y) = \lambda_i$ and $\mathrm{Var}(y) = \lambda_i$.

- Note that:
  - λ_i should be non-negative and continuous.
  - It should be via X_i and β.

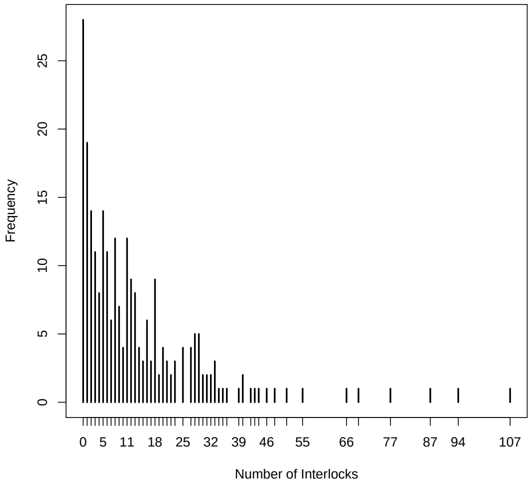
- $\lambda_i = e^{X_i\beta}$, which gives us $f(y_i|\lambda) = \frac{e^{-e^{X_i\beta}}(e^{X_i\beta})^{y_i}}{y_i!}$

```
> ## Data
> head(Ornstein)  ## from car package
> nrow(Ornstein)
> ?Ornstein
```

```
Interlocking Directorates Among Major Canadian Firms

Description

The Ornstein data frame has 248 rows and 4 columns. The observations are the 248 largest Canadian firms
with publicly available information in the mid-1970s.

:

assets
Assets in millions of dollars.

sector
Industrial sector. A factor with levels: AGR, agriculture, food, light industry; BNK, banking;
CON, construction; FIN, other financial; HLD, holding companies; MAN, heavy manufacturing;
MER, merchandizing; MIN, mining, metals, etc.; TRN, transport; WOD, wood and paper.

nation
Nation of control. A factor with levels: CAN, Canada; OTH, other foreign; UK, Britain; US, United States.

interlocks
Number of interlocking director and executive positions shared with other major firms.
```

```
> ## See the event count response
>
> tab <- xtabs(~ interlocks, data=Ornstein)
> tab2 <- table(Ornstein$interlocks)    ## alternative way
> print(tab)
interlocks
 0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
28  19  14  11   8  14  11   6  12   7   4  12   9   8   4   3   6   3   9
19  20  21  22  23  25  27  28  29  30  31  32  33  34  35  36  39  40  42
 2   4   3   2   3   4   4   5   5   2   2   2   3   1   1   1   1   2   1
43  44  46  48  51  55  66  69  77  87  94 107
 1   1   1   1   1   1   1   1   1   1   1   1

> plot(tab, type="h", main="Ornstein's Interlocks",
+      xlab="Number of Interlocks", ylab="Frequency")
```

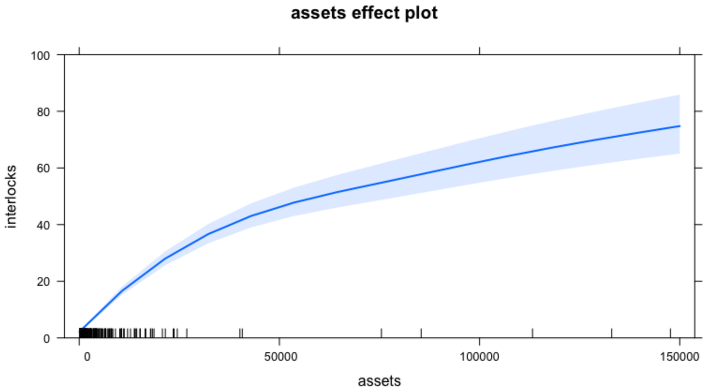
Ornstein's Interlocks



```
> ## Poisson
> mod.ornstein <- glm(interlocks ~ log2(assets) + nation + sector,
+                     family=poisson, data=Ornstein)
> summary(mod.ornstein)

Call:
glm(formula = interlocks ~ log2(assets) + nation + sector, family = poisson,
data = Ornstein)

Deviance Residuals:
Min       1Q   Median      3Q     Max
-6.7111  -2.3159  -0.4595   1.2824   6.2849

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.83938    0.13664  -6.143 8.09e-10 ***
log2(assets)  0.31292    0.01177  26.585  < 2e-16 ***
nationOTH    -0.10699    0.07438  -1.438 0.150301
nationUK     -0.38722    0.08951  -4.326 1.52e-05 ***
nationUS     -0.77239    0.04963 -15.562  < 2e-16 ***
sectorBNK    -0.16651    0.09575  -1.739 0.082036 .
sectorCON    -0.48928    0.21320  -2.295 0.021736 *
```

```
sectorFIN    -0.11161    0.07571  -1.474 0.140457
sectorBNK    -0.16651    0.09575  -1.739 0.082036 .
sectorCON    -0.48928    0.21320  -2.295 0.021736 *
sectorFIN    -0.11161    0.07571  -1.474 0.140457
sectorHLD    -0.01491    0.11924  -0.125 0.900508
sectorMAN     0.12187    0.07614   1.600 0.109489
sectorMER     0.06157    0.08670   0.710 0.477601
sectorMIN     0.24985    0.06888   3.627 0.000286 ***
sectorTRN     0.15181    0.07893   1.923 0.054453 .
sectorWOD     0.49825    0.07560   6.590 4.39e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3737.0  on 247  degrees of freedom
Residual deviance: 1547.1  on 234  degrees of freedom
AIC: 2473.1

Number of Fisher Scoring iterations: 5
```

```
> ornstein.eff <- allEffects(mod.ornstein)
> plot(ornstein.eff, 'log2(assets)',
      rescale.axis=FALSE, ylim=c(0,100))
```



Other Options for Count Responses

- What if the fixed rate ($= \lambda_i$) cannot be fully explained by $X_i\beta$?
  $\rightarrow$ Negative binomial model

```
> nb.ornstein <- glm.nb(interlocks ~ log2(assets) + nation + sector,
+                        data=Ornstein)
> summary(nb.ornstein)
```

Other Options for Count Responses

- What if the fixed rate ($= \lambda_i$) cannot be fully explained by $X_i\beta$?
  $\rightarrow$ Negative binomial model

```
> nb.ornstein <- glm.nb(interlocks ~ log2(assets) + nation + sector,
+                        data=Ornstein)
> summary(nb.ornstein)
```

- What if there are too many zeros?
  $\rightarrow$ Zero-inflated model

```
> library(pscl)
> z.mod.ornstein <- zeroinfl(interlocks ~ log2(assets) + nation + sector,
+                             data=Ornstein)
> z.nb.ornstein <- zeroinfl(interlocks ~ log2(assets) + nation + sector,
+                            dist = "negbin", data=Ornstein)
> summary(z.mod.ornstein)
> summary(z.nb.ornstein)
```

More Models

- This is not the end of the story!
- Depending on the characteristics of dependent variable, we need to choose an appropriate model.

More Models

- This is not the end of the story!
- Depending on the characteristics of dependent variable, we need to choose an appropriate model.
  - Ordinal response: presidential approval rate, political interest, ...
    $\rightarrow$ Ordered logit model

More Models

- This is not the end of the story!
- Depending on the characteristics of dependent variable, we need to choose an appropriate model.
  - Ordinal response: presidential approval rate, political interest, ...
    $\rightarrow$ Ordered logit model
  - Nominal response: vote for Moon, Hong, or Ahn, ...
    $\rightarrow$ Multinomial logit model

More Models

- This is not the end of the story!
- Depending on the characteristics of dependent variable, we need to choose an appropriate model.
  - Ordinal response: presidential approval rate, political interest, ...
    $\rightarrow$ Ordered logit model
  - Nominal response: vote for Moon, Hong, or Ahn, ...
    $\rightarrow$ Multinomial logit model
  - Duration response: length of peace (prior to conflict), ...
    $\rightarrow$ Duration model (Cox Proportional-Hazard model)

More Models

- This is not the end of the story!
- Depending on the characteristics of dependent variable, we need to choose an appropriate model.
  - Ordinal response: presidential approval rate, political interest, ...
    $\rightarrow$ Ordered logit model
  - Nominal response: vote for Moon, Hong, or Ahn, ...
    $\rightarrow$ Multinomial logit model
  - Duration response: length of peace (prior to conflict), ...
    $\rightarrow$ Duration model (Cox Proportional-Hazard model)
- Time-series data?

More Models

- This is not the end of the story!
- Depending on the characteristics of dependent variable, we need to choose an appropriate model.
    - Ordinal response: presidential approval rate, political interest, ...
      $\rightarrow$ Ordered logit model
    - Nominal response: vote for Moon, Hong, or Ahn, ...
      $\rightarrow$ Multinomial logit model
    - Duration response: length of peace (prior to conflict), ...
      $\rightarrow$ Duration model (Cox Proportional-Hazard model)
- Time-series data?
- Panel data?

You are now INDEED an R expert!!!

Make sure you practice it again at home

Questions? - hmpark1@uwm.edu