Statistical Analysis Using Structural Equation Models

EPsy 8266

Christopher David Desjardins

Research Methodology Consulting Center

2/7/19

Topics

- Logistic regression
- Probit regression

Motivation for logistic regression/probit regression

Alternative models

- Multiple regression is inappropriate for data that are not continuous (i.e, either interval or ratio)
- For dichotomous models, logistic regression or probit regression can be used.
- For data with more than 2 categories, multiple regression still is not appropriate. Consider multinomial or proportional odds model depending on scale.
- How many categories is enough for regression?

Logistic Regression Model

- In our example of schizophrenia, an individual could either be schizophrenic or not.
- This is akin to flipping a coin once.
- In both cases, we could say the outcome has a Bernoulli distribution, $Y \sim Bern(\pi)$.
- Equivalently, it has a Binomial distribution with a single trial, $Y \sim Bin(n = 1, \pi)$.
- π is the probability of a success (e.g., being schizophrenic or the coin being a heads) - the expected value of Y.

Logistic Regression Model - 2

- Presently, we are stuck at a 0 (no schizophrenia) or a 1 (schizophrenia).

Logistic Regression Model - 2

- Presently, we are stuck at a 0 (no schizophrenia) or a 1 (schizophrenia).
- It would be ideal if we could take the 0s and 1s and **link** them to the real line.

Logistic Regression Model - 2

- Presently, we are stuck at a 0 (no schizophrenia) or a 1 (schizophrenia).
- It would be ideal if we could take the 0s and 1s and **link** them to the real line.
- We could convert to an **odds ratio**.
- Odds ratio, $\Omega = \frac{\pi}{(1-\pi)}$ - the ratio of successes to failures.

Logistic Regression Model - 2

- Presently, we are stuck at a 0 (no schizophrenia) or a 1 (schizophrenia).
- It would be ideal if we could take the 0s and 1s and **link** them to the real line.
- We could convert to an **odds ratio**.
- Odds ratio, $\Omega = \frac{\pi}{(1-\pi)}$ - the ratio of successes to failures.
- This doesn't quite get us there.

Logistic Regression Model - 2

- Presently, we are stuck at a 0 (no schizophrenia) or a 1 (schizophrenia).
- It would be ideal if we could take the 0s and 1s and **link** them to the real line.
- We could convert to an **odds ratio**.
- Odds ratio, $\Omega = \frac{\pi}{(1-\pi)}$ - the ratio of successes to failures.
- This doesn't quite get us there.
- What if we take the log?

Logistic Regression Model - 3

- Log odds or logit of a success, $\log \Omega = \log\left[\frac{\pi}{(1-\pi)}\right]$

Logistic Regression Model - 3

- Log odds or logit of a success, $\log \Omega = \log \left[\frac{\pi}{(1-\pi)} \right]$
- By applying some algebra, we can also recover our probability of success (**inverse link**):

Logistic Regression Model - 3

- Log odds or logit of a success, $\log\Omega = \log\left[\frac{\pi}{(1-\pi)}\right]$
- By applying some algebra, we can also recover our probability of success (**inverse link**):

$$\pi = \frac{\exp(\log\Omega)}{1+\exp(\log\Omega)}$$

Logistic Regression Model - 3

- Log odds or logit of a success, $\log \Omega = \log \left[\frac{\pi}{(1-\pi)} \right]$
- By applying some algebra, we can also recover our probability of success (**inverse link**):

$$\pi = \frac{\exp(\log \Omega)}{1+\exp(\log \Omega)}$$

- The log odds can be any real number.

Logistic Regression Model - 4

Suppose we want to add some explanatory variables of schizophrenia (e.g., paranoia, which we'll call x_1).

Then, we can let the log odds of success (being schizophrenic) be represented by the linear function: $\beta_0 + \beta_1 x_1$.

We can plug this back into our equation:

$$\log \Omega = \beta_0 + \beta_1 x_1$$

Logistic Regression Model - 4

Suppose we want to add some explanatory variables of schizophrenia (e.g., paranoia, which we'll call x_1).

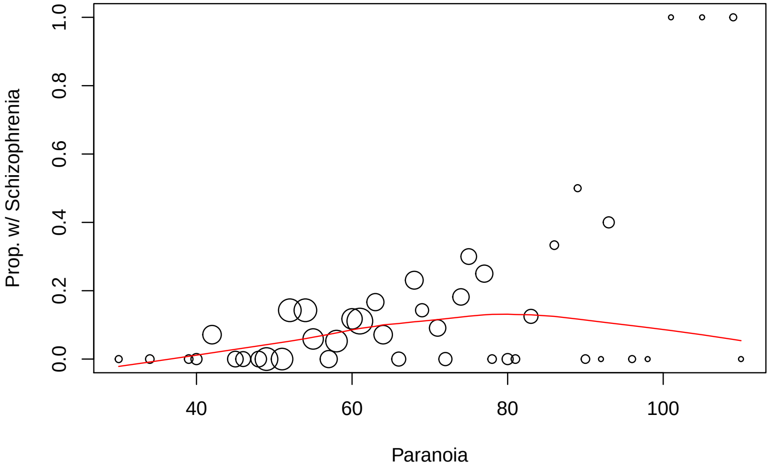
Then, we can let the log odds of success (being schizophrenic) be represented by the linear function: $\beta_0 + \beta_1 x_1$.

We can plug this back into our equation:

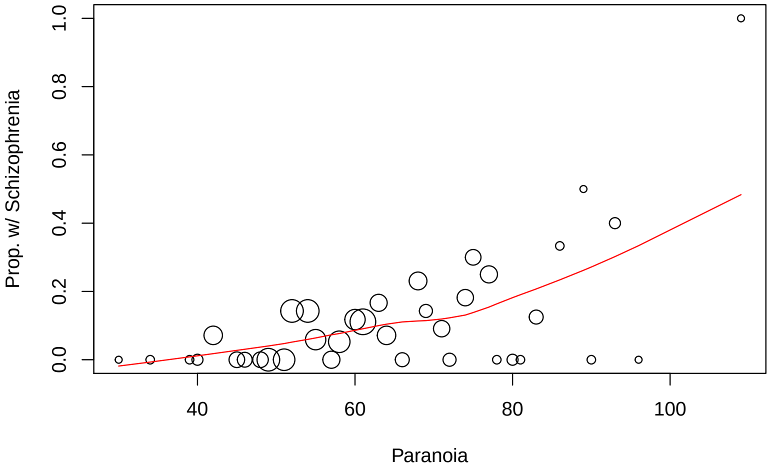
$$\log \Omega = \beta_0 + \beta_1 x_1$$

- The log-odds that a person with a paranoia score of x will be schizophrenic is $\beta_0 + \beta_1 x_1$.
- The odds that a person with a paranoia score of x will be schizophrenic is $\exp(\beta_0 + \beta_1 x_1)$.
- The probability that a person with a paranoia score of x will be schizophrenic is $\frac{\exp(\beta_0+\beta_1 x_1)}{1+\exp(\beta_0+\beta_1 x_1)}$.

```r
file <- "https://raw.githubusercontent.com/cddesja/epsy8266/master/course_materials/data/wuschiz.csv"
wuschiz <- read.csv(file)
means <- aggregate(Schizo ~ Pa, data = wuschiz, FUN = mean)
N <- aggregate(Schizo ~ Pa, data = wuschiz, FUN = length)
means$N <- N$Schizo
plot(Schizo ~ Pa, data = means, xlab = "Paranoia", cex = sqrt(N / pi),
     ylab = "Prop. w/ Schizophrenia")
lines(lowess(means$Pa, means$Schizo), col = "red")
```



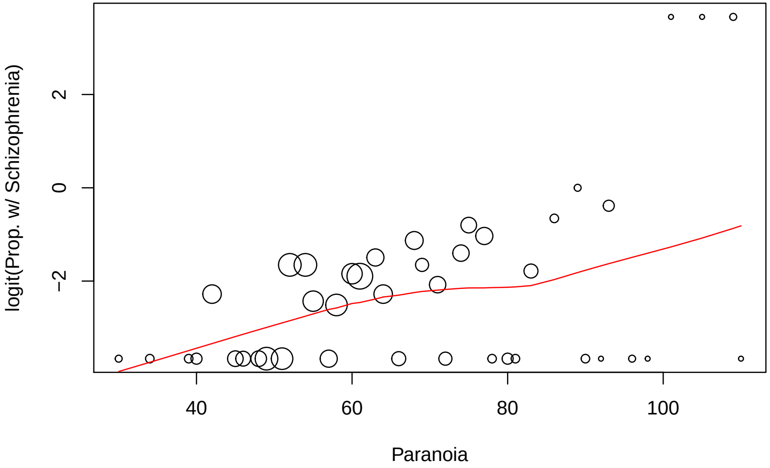
```r
means.n1 <- subset(means, N > 1)
plot(Schizo ~ Pa, data = means.n1, xlab = "Paranoia", cex = sqrt(N / pi),
     ylab = "Prop. w/ Schizophrenia")
lines(lowess(means.n1$Pa, means.n1$Schizo), col = "red")
```



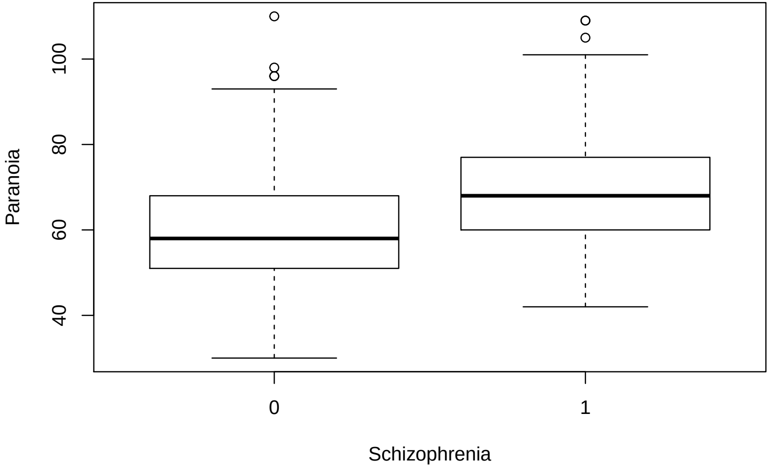
```
means$logit <- car::logit(means$Schizo)

## Warning in car::logit(means$Schizo): proportions remapped to (0.025, 0.975)

plot(logit ~ Pa, data = means, xlab = "Paranoia", cex = sqrt(N / pi),
     ylab = "logit(Prop. w/ Schizophrenia)")
lines(lowess(means$Pa, means$logit, f = 3/4), col = "red")
```



```r
boxplot(Pa ~ Schizo, data = wuschiz,
        xlab = "Schizophrenia",
        ylab = "Paranoia")
```



Schizophrenia logistic regression

```
mod.lr <- glm(Schizo ~ Pa, data = wuschiz, family = "binomial")
summary(mod.lr)

##
## Call:
## glm(formula = Schizo ~ Pa, family = "binomial", data = wuschiz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2560  -0.4778  -0.3818  -0.3098   2.6072
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.55592    0.80613  -6.892 5.50e-12 ***
## Pa           0.05217    0.01141   4.572 4.83e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 250.56  on 375  degrees of freedom
## Residual deviance: 229.23  on 374  degrees of freedom
## AIC: 233.23
##
## Number of Fisher Scoring iterations: 5
```

Schizophrenia model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\log \hat{\Omega} = -5.56 + .052 x_1$$

Schizophrenia model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\log \hat{\Omega} = -5.56 + .052 x_1$$

- How do interpret $\hat{\beta}_0$, $\hat{\beta}_1$?

Schizophrenia model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\log \hat{\Omega} = -5.56 + .052 x_1$$

- How do interpret $\hat{\beta}_0$, $\hat{\beta}_1$?

  A one-unit increase in paranoia **increases** the log-odds of developing schizophrenia by .052 ($\hat{\beta}_1$).

Schizophrenia model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\log \hat{\Omega} = -5.56 + .052 x_1$$

- How do interpret $\hat{\beta}_0$, $\hat{\beta}_1$?

  A one-unit increase in paranoia **increases** the log-odds of developing schizophrenia by .052 ($\hat{\beta}_1$).
- What if we exponentiate $\hat{\beta}_0$,$\hat{\beta}_1$?

Schizophrenia model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\log \hat{\Omega} = -5.56 + .052 x_1$$

- How do interpret $\hat{\beta}_0$, $\hat{\beta}_1$?

  A one-unit increase in paranoia **increases** the log-odds of developing schizophrenia by .052 ($\hat{\beta}_1$).
- What if we exponentiate $\hat{\beta}_0$,$\hat{\beta}_1$?

  A one-unit increase in paranoia **multiplies** the odds of success by 1.05 ($\hat{\beta}_1$).

Schizophrenia model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

$$\log \hat{\Omega} = -5.56 + .052 x_1$$

- How do interpret $\hat{\beta}_0$, $\hat{\beta}_1$?

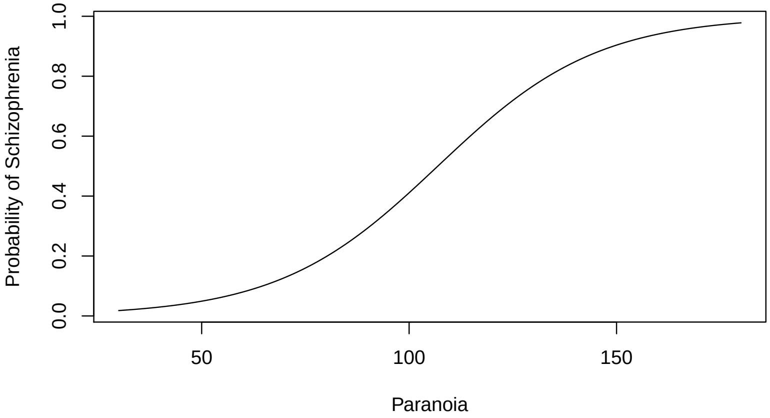
  A one-unit increase in paranoia **increases** the log-odds of developing schizophrenia by .052 ($\hat{\beta}_1$).

- What if we exponentiate $\hat{\beta}_0$,$\hat{\beta}_1$?

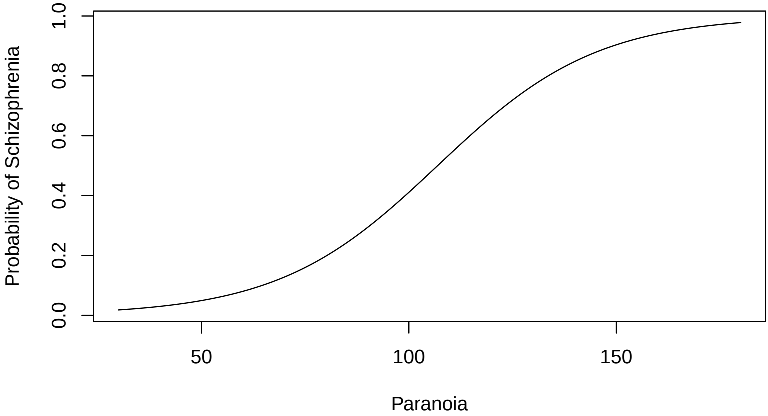
  A one-unit increase in paranoia **multiplies** the odds of success by 1.05 ($\hat{\beta}_1$).

  What are the odds of developing schizophrenia for participants with paranoia of 20, 30, and 40?

Logistic Curve



Logistic Curve



Where is the greatest rate of change in the probability of schizophrenia?

Important notes

- Increase is linear only in the log odds
- Increase is not linear for probability
  - Difference in the probability of schizophrenia is not the same between participants with paranoia of 50 and 60 and 100 and 110.
- Increase is multiplicative for the odds

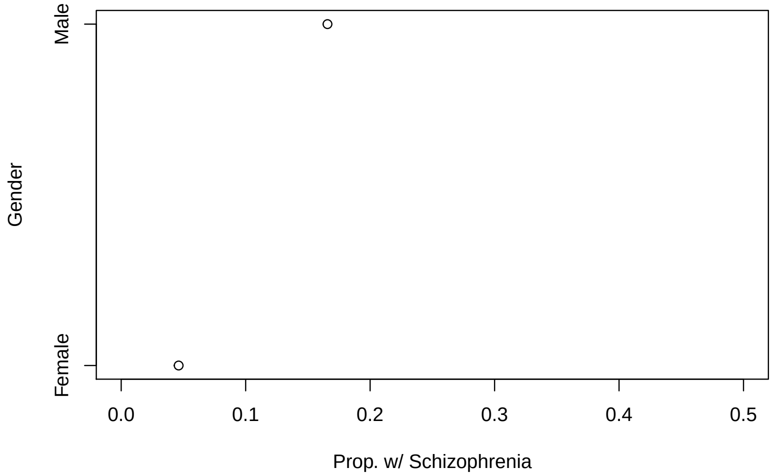
Multiple logistic regression

Let's now try to predict the probability of being schizophrenia given paranoia and gender (coded 1 as male and 0 as female) (x_2).

We can write this model as:

$$\log \Omega = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

```
means <- aggregate(Schizo ~ male, data = wuschiz, FUN = mean)
plot(male ~ Schizo, means, xlim = c(0, .5), yaxt = "n",
     ylab = "Gender", xlab = "Prop. w/ Schizophrenia")
axis(2, at=c(0, 1),labels=c("Female", "Male"))
```



```r
mod.lr2 <- glm(Schizo ~ Pa + male, data = wuschiz, family = "binomial")
summary(mod.lr2)

##
## Call:
## glm(formula = Schizo ~ Pa + male, family = "binomial", data = wuschiz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1566  -0.5041  -0.3357  -0.2658   2.6947
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.27497    0.81450  -6.476  9.4e-11 ***
## Pa           0.03979    0.01273   3.125  0.00178 **
## male         0.84938    0.44376   1.914  0.05561 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 250.56  on 375  degrees of freedom
## Residual deviance: 225.39  on 373  degrees of freedom
## AIC: 231.39
##
## Number of Fisher Scoring iterations: 5
```

The fitted model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log \hat{\Omega} = -5.274 + .039 x_1 + 0.849 x_2$$

The fitted model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log \hat{\Omega} = -5.274 + .039 x_1 + 0.849 x_2$$

- How do we interpret $\hat{\beta}_2$?

The fitted model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log \hat{\Omega} = -5.274 + .039 x_1 + 0.849 x_2$$

- How do we interpret $\hat{\beta}_2$?
- The log odds for a male being schizophrenic are .849 higher than for a female holding paranoia constant.

The fitted model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log \hat{\Omega} = -5.274 + .039 x_1 + 0.849 x_2$$

- How do we interpret $\hat{\beta}_2$?
- The log odds for a male being schizophrenic are .849 higher than for a female holding paranoia constant.
- How do we interpret $\exp(\hat{\beta}_2)$?

The fitted model

$$\log \hat{\Omega} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log \hat{\Omega} = -5.274 + .039 x_1 + 0.849 x_2$$

- How do we interpret $\hat{\beta}_2$?
- The log odds for a male being schizophrenic are .849 higher than for a female holding paranoia constant.
- How do we interpret $\exp(\hat{\beta}_2)$?
- The odds of of a male being schizophrenic are 2.33 times the odds of being schizophrenic for a female

Probit regression

For probit regression, the outcome is analyzed using a **probit function**.

$$Pr(Y = 1|X) = \phi(\beta_0 + \beta_1 x_2 + ...)$$

ϕ is the cumulative distribution function of the standard normal distribution.

Your book also motivates the use of a probit model as a normal latent variable, Y^*, such that

$$Y = \begin{cases} 1 & \text{if } Y^* \geq 0 \\ 0 & \text{if } Y^* < 0 \end{cases}$$

where $\hat{Y^*}$ is the metric of z-scores and

$$\hat{\pi} = \phi(\hat{Y^*})$$

This last equation is the **normal ogive model**.

Probit Regression

```
mod.pb <- glm(Schizo ~ Pa + male, data = wuschiz,
              family = "binomial"(link = "probit"))
summary(mod.pb)

##
## Call:
## glm(formula = Schizo ~ Pa + male, family = binomial(link = "probit"),
##     data = wuschiz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1041  -0.5081  -0.3446  -0.2602   2.7385
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.894533   0.428276  -6.759 1.39e-11 ***
## Pa           0.021635   0.006944   3.116  0.00184 **
## male         0.405230   0.215896   1.877  0.06052 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 250.56  on 375  degrees of freedom
## Residual deviance: 225.71  on 373  degrees of freedom
## AIC: 231.71
##
## Number of Fisher Scoring iterations: 5
```

The fitted probit model

$$\hat{\pi} = \phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right)$$

$$\hat{\pi} = \phi\left(-2.89 + .021 x_1 + 0.405 x_2\right)$$

The fitted probit model

$$\hat{\pi} = \phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right)$$

$$\hat{\pi} = \phi\left(-2.89 + .021 x_1 + 0.405 x_2\right)$$

- How do we interpret $\hat{\beta}_1$?

The fitted probit model

$$\hat{\pi} = \phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right)$$

$$\hat{\pi} = \phi\left(-2.89 + .021x_1 + 0.405x_2\right)$$

- How do we interpret $\hat{\beta}_1$?
- For one-unit increase in paranoia, the z-score for being schizophrenic increases .021.

The fitted probit model

$$\hat{\pi} = \phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right)$$

$$\hat{\pi} = \phi\left(-2.89 + .021 x_1 + 0.405 x_2\right)$$

- How do we interpret $\hat{\beta}_1$?
- For one-unit increase in paranoia, the z-score for being schizophrenic increases .021.
- How do we interpret $\hat{\beta}_2$?

The fitted probit model

$$\hat{\pi} = \phi\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2\right)$$

$$\hat{\pi} = \phi\left(-2.89 + .021 x_1 + 0.405 x_2\right)$$

- How do we interpret $\hat{\beta}_1$?
- For one-unit increase in paranoia, the z-score for being schizophrenic increases .021.
- How do we interpret $\hat{\beta}_2$?
- Being male increases the z-score of being schizophrenic by .405 relative to females.

Activity

Rerun the regression of predicting schizophrenia given hypochondriasis, hypomania, and gender as a logistic regression and probit regression.

How are the results similar?

How are the results different?

Practice interpreting the parameters