Analyzing binary outcomes, going beyond logistic regression

2018 EHE Forum presentation

James O. Uanhoro

Department of Educational Studies

The Ohio State University

Premise

Obtaining relative risk using Poisson regression

Obtaining risk difference using linear regression (OLS)

Summary

Logistic regression review

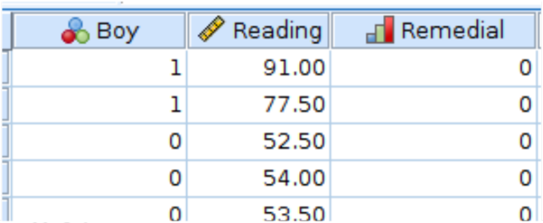
Researchers perform logistic regression to analyze binary outcomes:

- The coefficients obtained from the logistic regression model are logits or log odds
- Logits are difficult to interpret
- Researchers often exponentiate logits to obtain odds ratios to ease interpretation

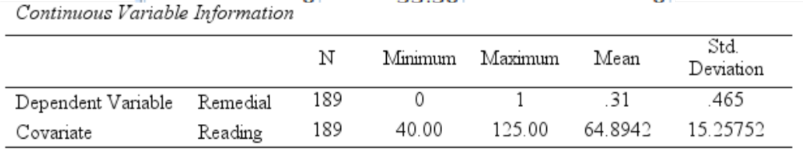
Logistic regression review (example)

We observed the referral rates of 189 children to remedial reading programs. On average, 31% of children were referred.

We would hope that the only determinant of referral would be reading ability, so we test if there are sex differences in assignment, while controlling for reading ability.



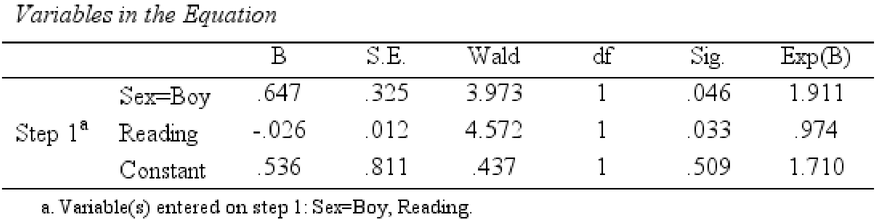
Boy	Reading	Remedial
1	91.00	0
1	77.50	0
0	52.50	0
0	54.00	0
0	53.50	0



Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Remedial	189	0	1	.31	.465
Covariate	Reading	189	40.00	125.00	64.8942	15.25752

Logistic regression review (model results)



Variables in the Equation

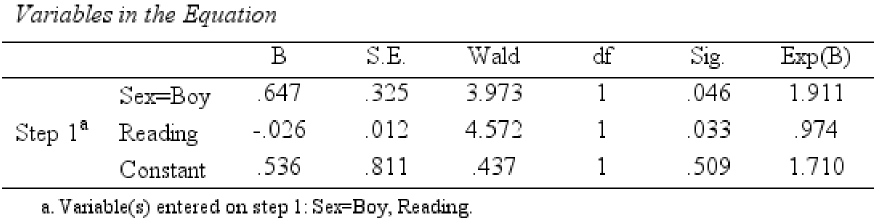
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex=Boy	.647	.325	3.973	1	.046	1.911
	Reading	-.026	.012	4.572	1	.033	.974
	Constant	.536	.811	.437	1	.509	1.710

a. Variable(s) entered on step 1: Sex=Boy, Reading.

Results:

1. As reading has a negative coefficient, children with higher reading abilities were less likely to be recommended to the remedial program.

Logistic regression review (model results)



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex=Boy	.647	.325	3.973	1	.046	1.911
	Reading	-.026	.012	4.572	1	.033	.974
	Constant	.536	.811	.437	1	.509	1.710

a. Variable(s) entered on step 1: Sex=Boy, Reading.

Results:

2. Comparing children with the same reading ability, the odds of boys getting recommended to the program was on average 1.91 times the odds of girls getting recommended to the program, $p = .046$.

Logistic regression review (in closing)

The problem:

1. I'm yet to meet someone who understands odds intuitively - I don't know many gamblers . . .
2. We usually want to compare probabilities. We want to be able to say *boys were x times more likely than girls to . . .* This value is known as the *relative risk* or *risk ratio*.

Logistic regression review (in closing)

The problem:

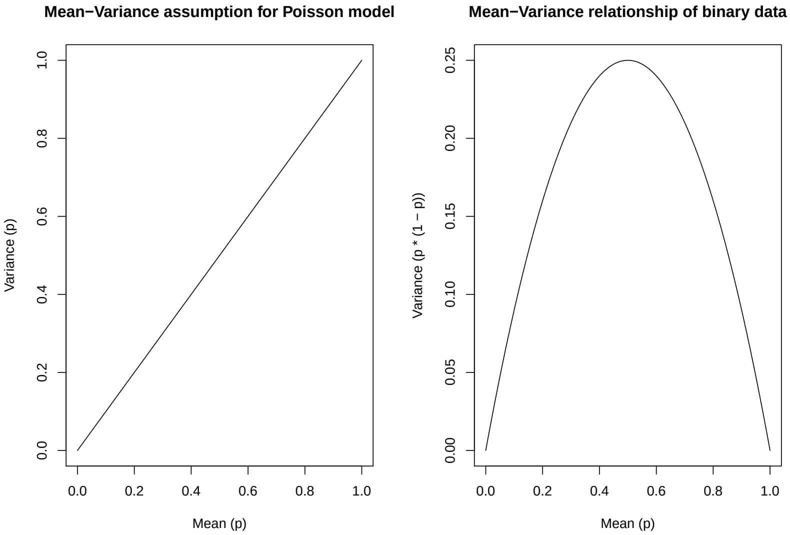
3. Many researchers simply interpret the odds ratio as if it were the risk ratio. This would only be acceptable if the rate of referral to the remedial program was under 5%.
4. This problem exists because in logistic regression, we model log odds. It is possible to calculate the risk ratio in logistic regression, but it requires extra math.

Poisson regression as an alternative

Poisson regression can be used as an alternative to model binary data:

- In Poisson regression, we model the log probabilities. If we exponentiate the model coefficients, we obtain the relative risk.
- However in Poisson regression, we assume the mean of the data equals its variance.

Poisson regression (Mean-variance relationship)



Problems:

- Poisson regression assumes the mean of the data equals the variance. In binary data, the mean exceeds the variance.
- If we proceed, we will be assuming the data are *under-dispersed Poisson*. The regression coefficients are fine, but the standard errors, hence p-values, are incorrect.
- We can attempt to address this problem using robust error variances (or quasi-Poisson models - not available in SPSS).

Poisson regression (sample syntax)

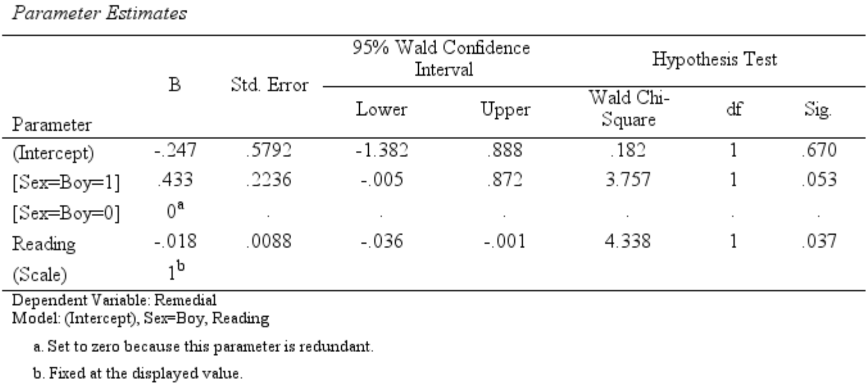
We can use SPSS GENLIN for Poisson regression with robust errors. Sample syntax:

```
GENLIN Remedial BY Boy (ORDER=DESCENDING) WITH Reading
  /MODEL Boy Reading INTERCEPT=YES
 DISTRIBUTION=POISSON LINK=LOG
  /CRITERIA COVB=ROBUST
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

These two lines are particularly important:

```
 DISTRIBUTION=POISSON LINK=LOG
  /CRITERIA COVB=ROBUST
```

Poisson regression (model results)



Parameter Estimates

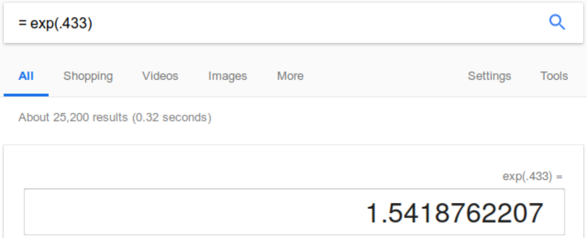
			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-.247	.5792	-1.382	.888	.182	1	.670
[Sex=Boy=1]	.433	.2236	-.005	.872	3.757	1	.053
[Sex=Boy=0]	0[a]	.	.	.	.	.	.
Reading	-.018	.0088	-.036	-.001	4.338	1	.037
(Scale)	1[b]						

Dependent Variable: Remedial
Model: (Intercept), Sex=Boy, Reading

a. Set to zero because this parameter is redundant.

b. Fixed at the displayed value.

Poisson regression (Google to the rescue)



= exp(.433)

About 25,200 results (0.32 seconds)

exp(.433) =

1.5418762207

Results:

1. Comparing children with the same reading ability, boys were on average 1.54 times more likely than girls to be recommended to the program, $p = .053$.
2. Result is not statistically significant. A problem with Poisson alternatives to the logistic regression model is a small loss of statistical power.

Poisson regression (Take-aways)

- If anyone interpreted the odds ratio as a comparison of probabilities, they would overstate the difference in referral between boys and girls.
- Poisson models and modifications may suffer from a loss in statistical power.
- If a study was well-powered to perform logistic regression, Poisson alternatives would likely be sufficiently powered.
- There is one more alternative that does not suffer the loss in statistical power.

Linear regression or the linear probability model

Commonplace linear regression using OLS estimation directly models probabilities when the data are binary. In econometrics, this model is known as the *Linear Probability Model* (LPM).

Linear regression or the linear probability model

Nice features of the LPM:

1. The coefficient of the sex variable would be the difference in probabilities of recommendation to remedial reading between boys and girls.
2. If the rates of recommendation to remedial reading lie between 20% and 80%, the OLS slope approximates the logistic slope.
3. Despite violating homoscedasticity with binary data, standard errors from OLS may still perform adequately.

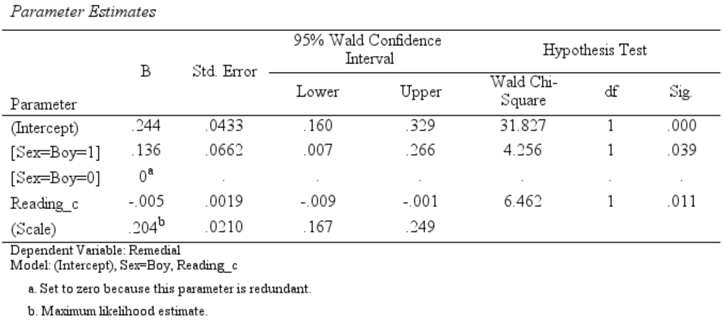
Linear regression (sample syntax)

We can use SPSS GENLIN for linear regression with robust errors (as a safety net). First, we center the reading scores to make the intercept meaningful.

```
COMPUTE Reading_c = Reading - 65.
EXECUTE.

GENLIN Remedial BY Boy (ORDER=DESCENDING) WITH Reading_c
  /MODEL Boy Reading_c INTERCEPT=YES
 DISTRIBUTION=NORMAL LINK=IDENTITY
  /CRITERIA COVB=ROBUST
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

Linear regression (model results)



Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.244	.0433	.160	.329	31.827	1	.000
[Sex=Boy=1]	.136	.0662	.007	.266	4.256	1	.039
[Sex=Boy=0]	0[a]	.	.	.	.	.	.
Reading_c	-.005	.0019	-.009	-.001	6.462	1	.011
(Scale)	.204[b]	.0210	.167	.249			

Dependent Variable: Remedial
Model: (Intercept), Sex=Boy, Reading_c

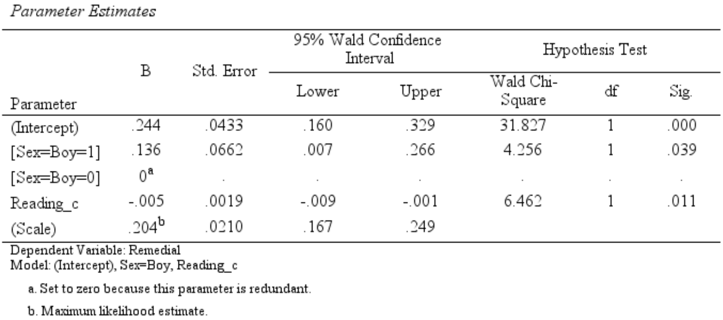
a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Results:

1. The probability of referral to remedial reading for girls with average reading ability was, on average, 24.4%.

Linear regression (model results)



Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.244	.0433	.160	.329	31.827	1	.000
[Sex=Boy=1]	.136	.0662	.007	.266	4.256	1	.039
[Sex=Boy=0]	0[a]	.	.	.	.	.	.
Reading_c	-.005	.0019	-.009	-.001	6.462	1	.011
(Scale)	.204[b]	.0210	.167	.249			

Dependent Variable: Remedial
Model: (Intercept), Sex=Boy, Reading_c

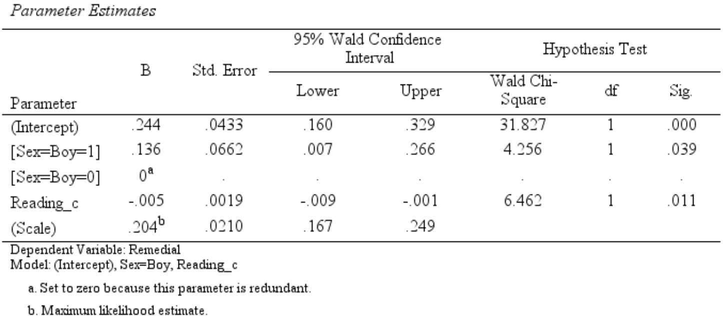
a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Results:

2. The probability of referral to remedial reading for boys with average reading ability was on average 13.6% higher than it was for girls of average reading ability, $p = .039$.

Linear regression (model results)



Parameter Estimates

	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
Parameter			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.244	.0433	.160	.329	31.827	1	.000
[Sex=Boy=1]	.136	.0662	.007	.266	4.256	1	.039
[Sex=Boy=0]	0[a]	.	.	.	.	.	.
Reading_c	-.005	.0019	-.009	-.001	6.462	1	.011
(Scale)	.204[b]	.0210	.167	.249			

Dependent Variable: Remedial
Model: (Intercept), Sex=Boy, Reading_c

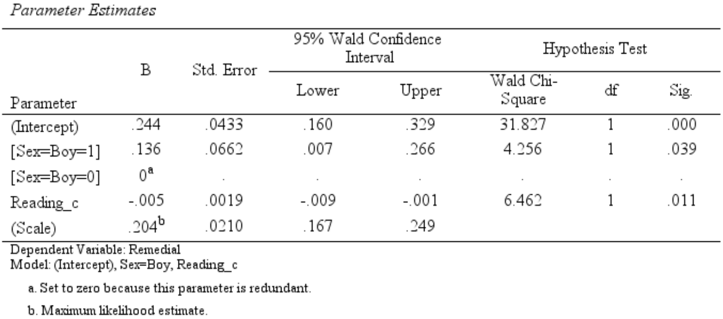
[a] Set to zero because this parameter is redundant.

[b] Maximum likelihood estimate.

Results:

3. The probability of boys with average reading ability to be referred to remedial reading was, on average, 38% (.244 + .136).

Linear regression (model results)



Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test		
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.244	.0433	.160	.329	31.827	1	.000
[Sex=Boy=1]	.136	.0662	.007	.266	4.256	1	.039
[Sex=Boy=0]	0[a]	.	.	.	.	.	.
Reading_c	-.005	.0019	-.009	-.001	6.462	1	.011
(Scale)	.204[b]	.0210	.167	.249			

Dependent Variable: Remedial
Model: (Intercept), Sex=Boy, Reading_c

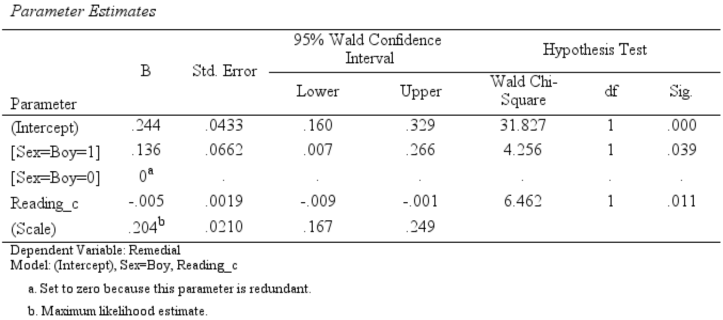
a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

Results:

4. Boys with average reading ability were 56% ($\frac{.136}{.244}$) more likely to be referred to remedial reading than girls with average reading ability. Poisson value was 54% more.

Linear regression (model results)



Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	.244	.0433	.160	.329	31.827	1	.000
[Sex=Boy=1]	.136	.0662	.007	.266	4.256	1	.039
[Sex=Boy=0]	0[a]	.	.	.	.	.	.
Reading_c	-.005	.0019	-.009	-.001	6.462	1	.011
(Scale)	.204[b]	.0210	.167	.249			

Dependent Variable: Remedial
Model: (Intercept), Sex=Boy, Reading_c

a. Set to zero because this parameter is redundant.

b. Maximum likelihood estimate.

You could also get the odds ratio, but why:

$$\frac{\frac{.38}{1-.38}}{\frac{.244}{1-.244}} = 1.90 \tag{1}$$

Logistic regression gave us 1.91.

Why doesn't anyone do this already?

1. They sometimes do, just not in education. Besides, if we have low or high probabilities, it is unreliable.
2. OLS can return predicted probabilities less than 0 and greater than 1. However, this is usually not a problem if we are interested in average effects.

Why doesn't anyone do this already?

3. The error of binary data is not homoskedastic - an assumption for correct standard errors with OLS. However, we can address this using robust standard errors as we have done above. Alternatives include weighted least squares.
4. Predictor variables with high leverage (outliers) can easily bias the model; the logistic model is more robust to outliers on the predictor variables. Most software packages can return leverage values on request, after conducting regression analysis.

Practical advice

1. Calculate the average probability of the outcome. If it is close to 50%, a linear regression model would work just as fine as a logistic regression model. If it is less than 20% or greater than 80%, a linear model would probably be inadequate. However, beware of cases with high leverage. They can throw off your linear model very easily.

Practical advice

2. If the probability of the outcome is less than 5%, you can freely interpret the exponentiated logistic regression coefficient using “likely” language. At low probabilities, odds approximately equal probabilities, so the odds ratio approximates the risk ratio.

Practical advice

3. If you run a Poisson regression so you can interpret your results using "likely" language, but your p-values are not statistically significant and they are stat. sig. in the logistic regression model, take the extra minute or two to do the math to obtain the risk ratio using the logistic regression model. You should probably center continous predictors before doing this.

The intercept changed to -1.165 on doing this:

$$\frac{\frac{1}{1+exp(-(-1.165+.647))}}{\frac{1}{1+exp(-(-1.165))}} = 1.57 \tag{2}$$

Practical advice

4. These results extend to multilevel models. If you find that your multilevel logistic model is taking too much time, a linear multilevel model might suffice.

Questions??

https://ghostbin.com/paste/ou6ds

Useful references

- Cheung, Y. B. (2007). A modified least-squares regression approach to the estimation of risk difference. *American Journal of Epidemiology, 166*(11), 1337–1344. `https://doi.org/10.1093/aje/kwm223`
- Hellevik, O. (2009). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity, 43*(1), 59–74. `https://doi.org/10.1007/s11135-007-9077-3`
- Zou, G. (2004). A modified Poisson regression approach to prospective studies with binary data. *American Journal of Epidemiology, 159*(7), 702–706. `https://doi.org/10.1093/aje/kwh090`
- Zou, G., & Donner, A. (2013). Extension of the modified Poisson regression model to prospective studies with correlated binary data. *Statistical Methods in Medical Research, 22*(6), 661–670. `https://doi.org/10.1177/0962280211427759`

Table of Contents

Premise

Obtaining relative risk using Poisson regression

Obtaining risk difference using linear regression (OLS)

Summary