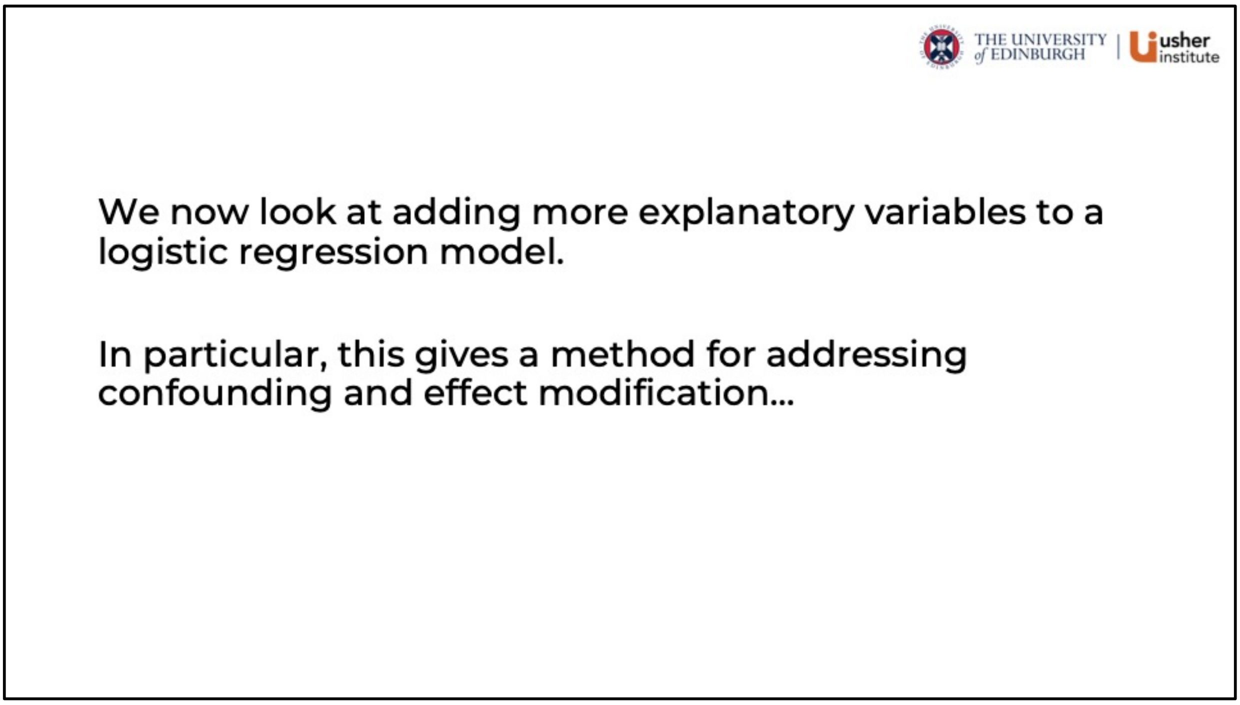
THE UNIVERSITY *of* EDINBURGH | usher institute

Extending Logistic Regression

Niall Anderson

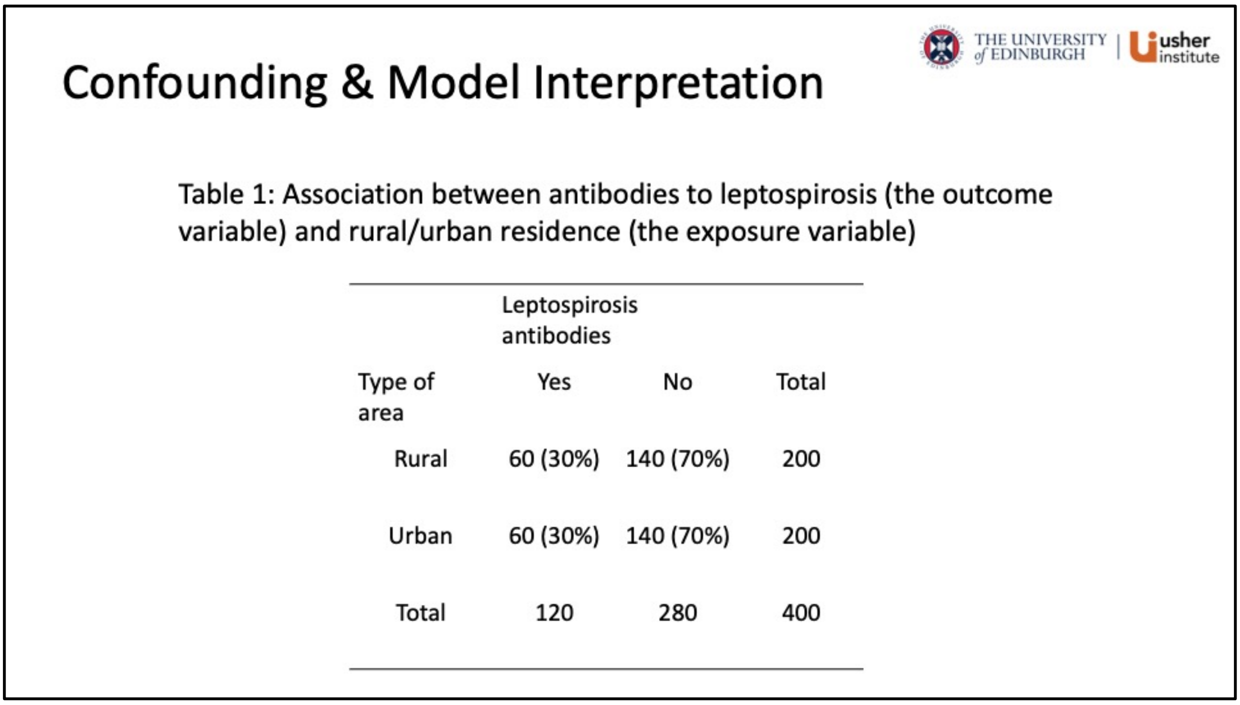
Supported by

THE UNIVERSITY *of* EDINBURGH | Data-Driven Innovation



We now look at adding more explanatory variables to a logistic regression model.

In particular, this gives a method for addressing confounding and effect modification…



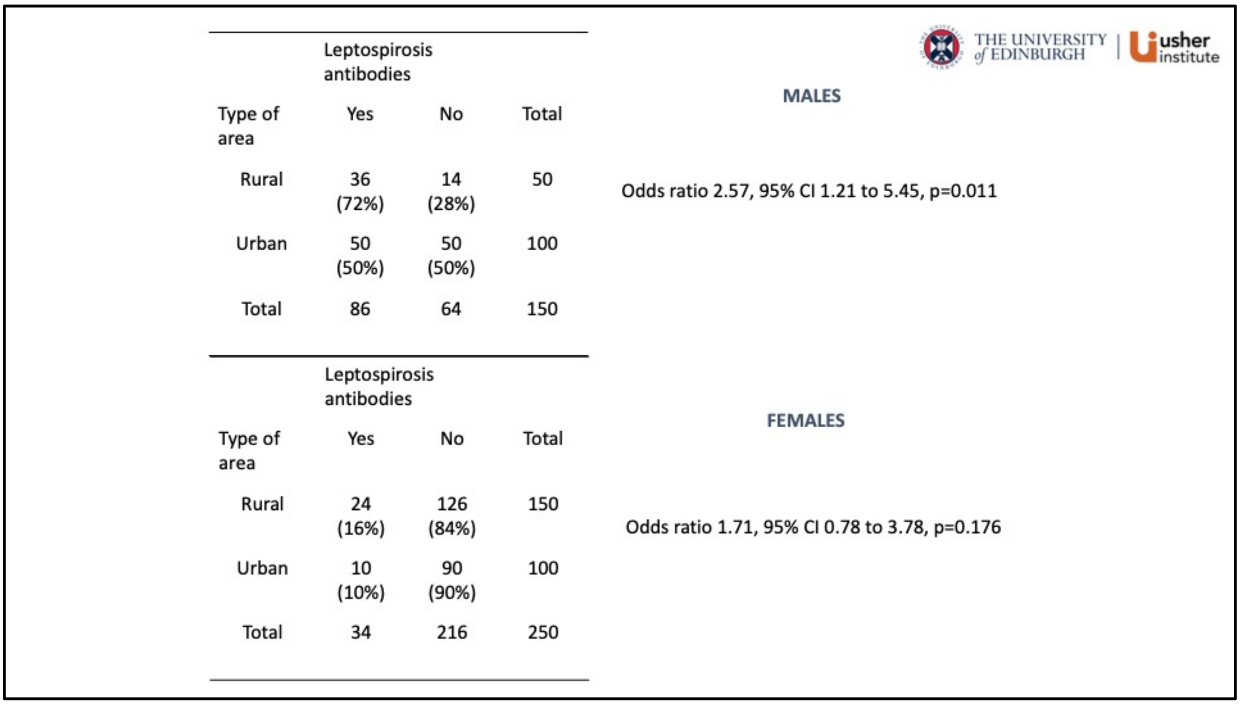
Confounding & Model Interpretation

Table 1: Association between antibodies to leptospirosis (the outcome variable) and rural/urban residence (the exposure variable)

	Leptospirosis antibodies		
Type of area	Yes	No	Total
Rural	60 (30%)	140 (70%)	200
Urban	60 (30%)	140 (70%)	200
Total	120	280	400

This (artificial) example allows us to see logistic regression used with categorical explanatory variables, and to practice interpreting the resulting odds ratios. It also provides a good way of understanding the structure of the model, how the parameters allow estimates of the probability of outcome, and lets us discuss confounding and interactions.

The table on this slide summarises the result of a fictitious study to investigate association of rural or urban residence on the presence of leptospirosis antibodies. As can be seen from this 2x2 table, there appears to be absolutely no association of any sort.



THE UNIVERSITY of EDINBURGH | usher institute

MALES

	Leptospirosis antibodies		
Type of area	Yes	No	Total
Rural	36 (72%)	14 (28%)	50
Urban	50 (50%)	50 (50%)	100
Total	86	64	150

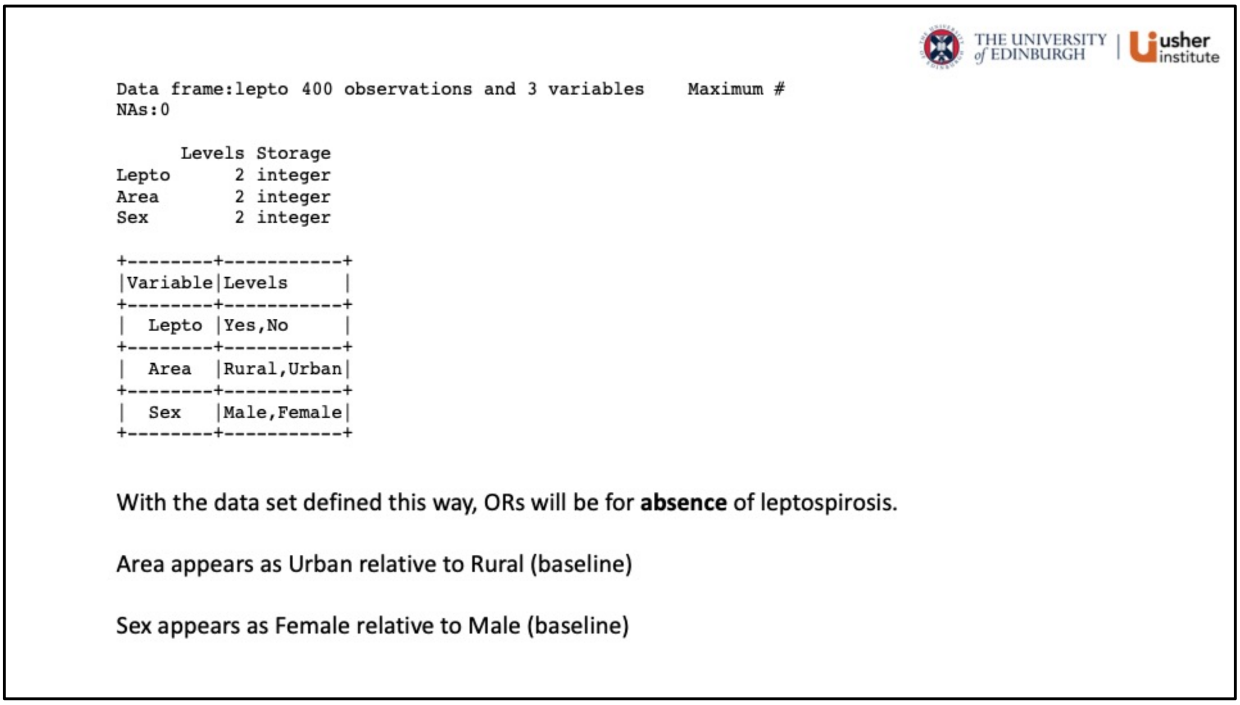
Odds ratio 2.57, 95% CI 1.21 to 5.45, p=0.011

FEMALES

	Leptospirosis antibodies		
Type of area	Yes	No	Total
Rural	24 (16%)	126 (84%)	150
Urban	10 (10%)	90 (90%)	100
Total	34	216	250

Odds ratio 1.71, 95% CI 0.78 to 3.78, p=0.176

However, when we break the data down by sex, we discover that, in fact, in both men and women there is a slightly higher rate of antibody presence in rural residents, but that the antibodies are very prevalent in men but relatively uncommon in women. The combined effect of these two processes is to cancel out the association, and leave no evidence for it at the higher level. This is an example of something called Simpson’s Paradox, and represents a specialised type of confounding!



```
Data frame:lepto 400 observations and 3 variables    Maximum #
NAs:0

      Levels Storage
Lepto      2 integer
Area       2 integer
Sex        2 integer

+--------+-----------+
|Variable|Levels     |
+--------+-----------+
|  Lepto |Yes,No     |
+--------+-----------+
|  Area  |Rural,Urban|
+--------+-----------+
|  Sex   |Male,Female|
+--------+-----------+
```

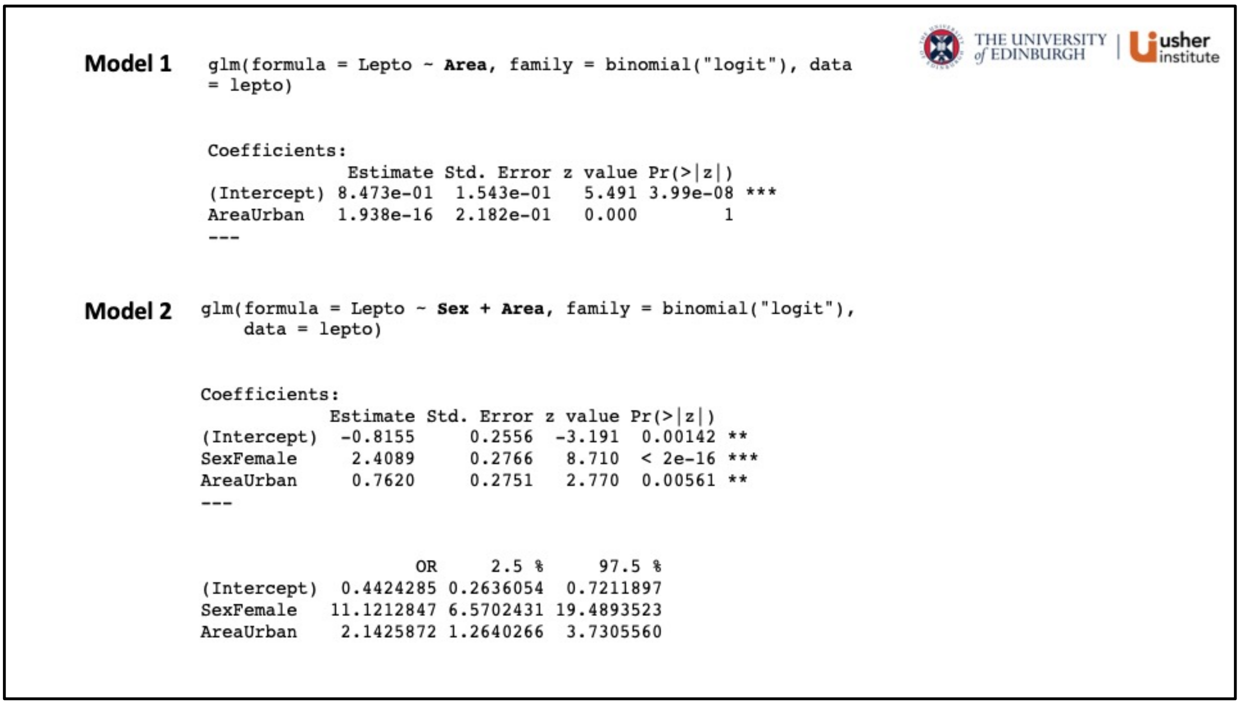
With the data set defined this way, ORs will be for **absence** of leptospirosis.

Area appears as Urban relative to Rural (baseline)

Sex appears as Female relative to Male (baseline)

With this data set of 400 observations, the outcome variable is coded as {Yes, No}, so that R will fit logistic regression models for the **absence** of leptospirosis, unless we change the coding. For simplicity, let’s leave this as it is…

Similarly, the Area and Sex explanatory variables will use their respective first levels as baselines – Area ORs will be calculated for Urban relative to Rural, and Sex ORs for Female relative to Male.



Model 1

```
glm(formula = Lepto ~ Area, family = binomial("logit"), data
= lepto)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 8.473e-01  1.543e-01   5.491 3.99e-08 ***
AreaUrban   1.938e-16  2.182e-01   0.000        1
---
```

Model 2

```
glm(formula = Lepto ~ Sex + Area, family = binomial("logit"),
    data = lepto)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8155     0.2556  -3.191  0.00142 **
SexFemale     2.4089     0.2766   8.710  < 2e-16 ***
AreaUrban     0.7620     0.2751   2.770  0.00561 **
---


                    OR     2.5 %     97.5 %
(Intercept)  0.4424285 0.2636054  0.7211897
SexFemale   11.1212847 6.5702431 19.4893523
AreaUrban    2.1425872 1.2640266  3.7305560
```

The 2 models presented on this slide correspond to the 2 levels of crosstabulation shown on the previous slides.

If we simply take account of Area, we see absolutely no association with the absence of antibodies – the parameter estimate is approximately 0, with a p-value of 1 (therefore corresponding to an OR of 1). However, when we first adjust for Sex, then fit Area, we see that, first, the odds on antibody absence are substantially greater in women relative to men (OR = 11), after which the adjusted odds of antibody absence in Urban areas are about double that of Rural areas.

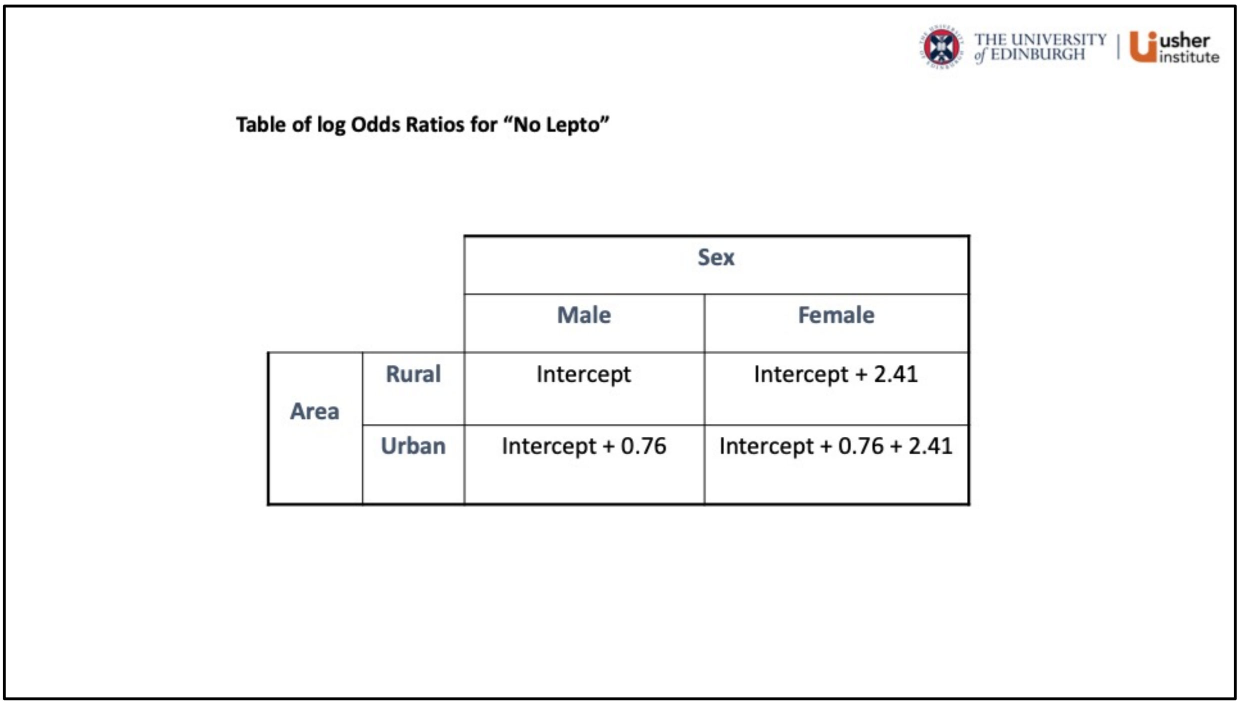


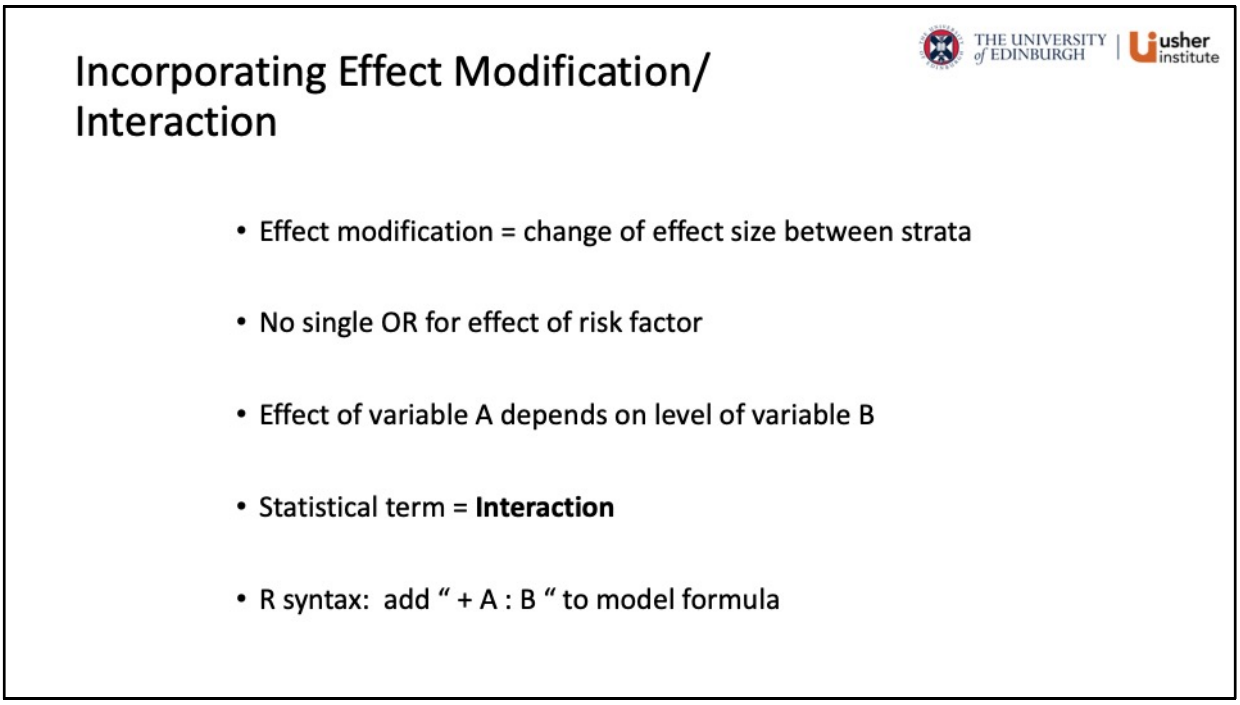
Table of log Odds Ratios for "No Lepto"

		Sex	
		Male	Female
Area	Rural	Intercept	Intercept + 2.41
	Urban	Intercept + 0.76	Intercept + 0.76 + 2.41

We can use Model 2 to see how the parameters combine to produce the estimate of the log odds ratio. A Male study participant from a Rural area is in the baseline categories of both factors. Accordingly, neither of the non-Intercept parameters need be considered (as these represent the difference from baseline of the other levels of each factor). If we instead look at a Female participant, but again from a Rural area, then now we must add in the value of the Area parameter, and the log odds are the total of the Intercept plus 2.41 = 1.5945.

Similarly, if we look at an Urban area study participant rather than a Rural one, then we add a further contribution of 0.76 to our estimate. Thus, the Area and Sex factors are modelled as entirely independent factors in Model 2 – the level of one does not affect the calculation of the odds estimate based on the level of the other.

This need not always be the case, however…

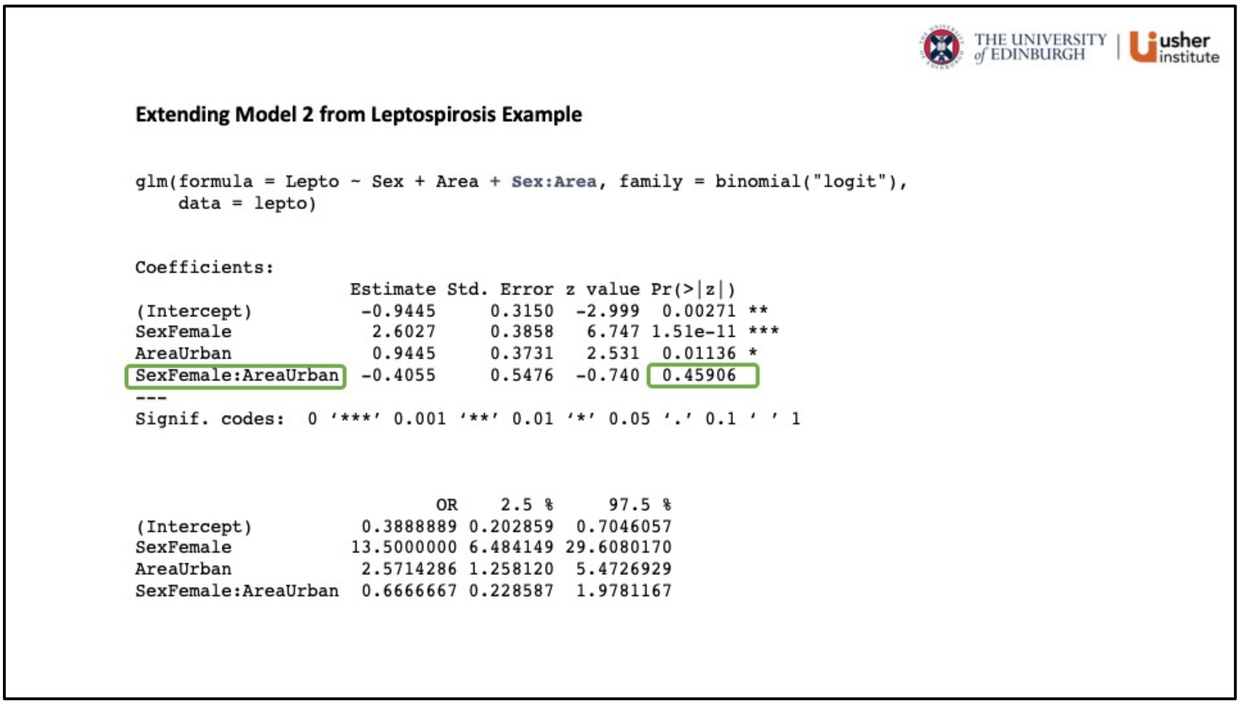


Incorporating Effect Modification/ Interaction

- Effect modification = change of effect size between strata
- No single OR for effect of risk factor
- Effect of variable A depends on level of variable B
- Statistical term = **Interaction**
- R syntax: add “ + A : B “ to model formula

Effect modification is the term used to describe such a dependence between the effect estimates of two factors. This means that there is no single OR for the effect of either risk factor – there will be a separate estimate depending on the level of the modifying factor. The statistical term used to describe this behaviour is **interaction**.

In R, we add a term of syntax “+ A:B” to describe the interaction between factor A and factor B, and can then judge if it is a substantial contributor to the model and directly estimate the magnitude of its contribution in the usual ways.



Extending Model 2 from Leptospirosis Example

```
glm(formula = Lepto ~ Sex + Area + Sex:Area, family = binomial("logit"),
    data = lepto)


Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.9445     0.3150  -2.999  0.00271 **
SexFemale             2.6027     0.3858   6.747 1.51e-11 ***
AreaUrban             0.9445     0.3731   2.531  0.01136 *
SexFemale:AreaUrban  -0.4055     0.5476  -0.740  0.45906
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



                            OR    2.5 %     97.5 %
(Intercept)          0.3888889 0.202859  0.7046057
SexFemale           13.5000000 6.484149 29.6080170
AreaUrban            2.5714286 1.258120  5.4726929
SexFemale:AreaUrban  0.6666667 0.228587  1.9781167
```

We can investigate effect modification in Model 2 for the Leptospirosis data in this way, and the results from fitting a model with the addition of a Sex by Area interaction are shown here.

The interaction term (the last line of the ***Coefficients:*** table) has a p-value of 0.459, and therefore it is not significantly different from zero. We would therefore conclude that there is no interaction between these two variables that we can observe in these data, and we would be justified in removing this term from the model.

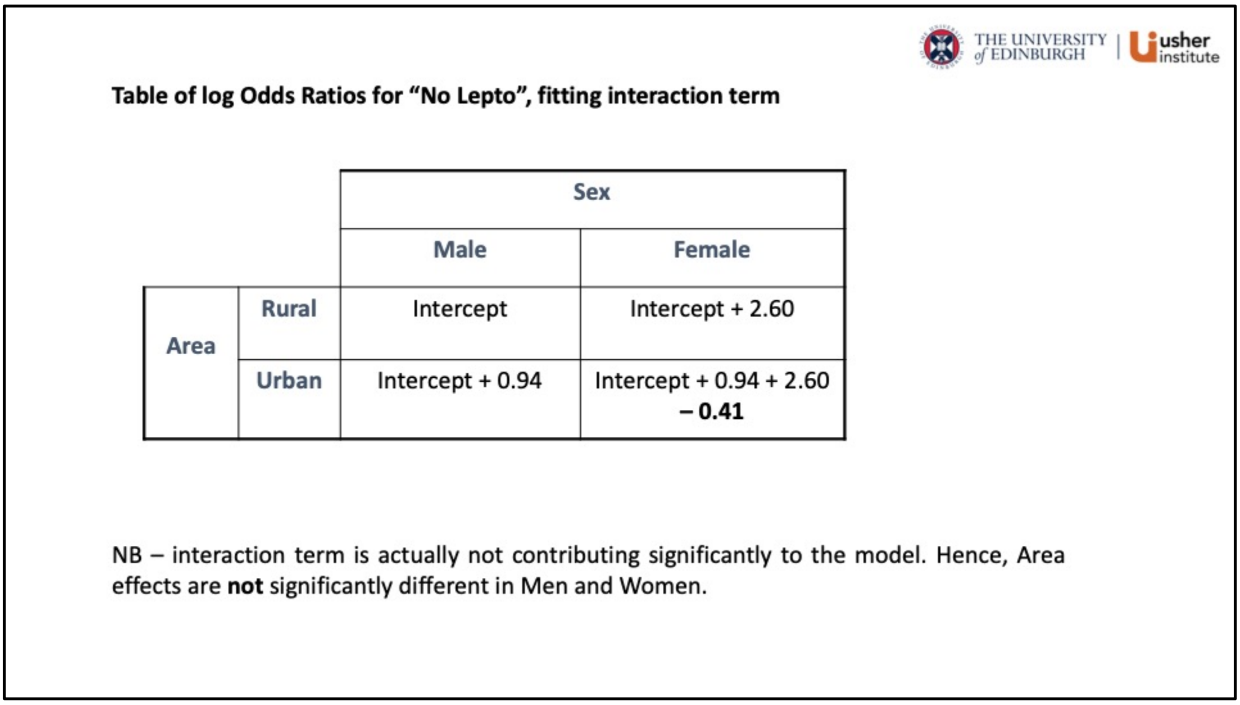


Table of log Odds Ratios for “No Lepto”, fitting interaction term

		Sex	
		Male	Female
Area	Rural	Intercept	Intercept + 2.60
	Urban	Intercept + 0.94	Intercept + 0.94 + 2.60 **– 0.41**

NB – interaction term is actually not contributing significantly to the model. Hence, Area effects are **not** significantly different in Men and Women.

By way of comparison to the earlier slide looking at the parameter estimates, here is the equivalent table including the interaction term. The estimates of log odds are built up in the same way as before, but the difference in this case is that the interaction parameter modifies the estimate in the Female/ Urban cell: the effect of Sex, for example, is no longer adding 2.60 as we move from the Male to Female Columns, as we need to know whether we are discussing a Rural or Urban study participant, because the latter will require the interaction parameter to be included in the calculation.

As noted on the previous slide, this is for illustration only. In this case we do not actually require an interaction term in Model 2.