CSSS/SOC/STAT 536:
Logistic Regression and Log-linear Models

Log-linear Models of Contingency Tables: 2D Tables

Christopher Adolph*

University of Washington, Seattle

March 8, 2005

*Assistant Professor, Department of Political Science and Center for Statistics and the Social Sciences.

Outline

Now that we are comfortable with contingency tables, we want to model tabular data

- Consider interaction effects of rows and columns (non-independence)
- Compare the fit of various models
- Calculate expected cell counts and residuals

We'll start today with $I \times J$ tables.

We'll use this case to get a handle on notation and concepts

Next time: $I \times J \times K \times \ldots$ tables, which are potentially much more interesting

Notation for Log-linear models

Recall that under independence,

$$\mathrm{E}(\mu_{ij}) = n\hat{\pi}_{i.}\hat{\pi}_{.j}$$

Let's take logs

$$\ln \mathrm{E}(\mu_{ij}) = \ln n + \ln \hat{\pi}_{i.} + \ln \hat{\pi}_{.j}$$

Independence makes for an additive model of the logged expected count.

Now, let's introduce new notation for the last equation

$$\ln \mathrm{E}(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Notation for Log-linear models

$$\ln \mathrm{E}(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Recall that the marginals of the contingency table summed to 1, by the basic rules of probability.

This meant that a set of I row marginals only had $k - 1$ degrees of freedom

In the same way, the I λ_i^X's only have $k - 1$ degrees of freedom

To identify them, we impose the following constraints

$$\sum_i^I \lambda_i^X = 0 \qquad \sum_j^J \lambda_j^X = 0$$

Notation for Log-linear models

$$\sum_{i}^{I} \lambda_i^X = 0 \qquad \sum_{j}^{J} \lambda_j^X = 0$$

Note that this achieves identification in the same way dropping one of a set of dummy regressors does.

Both techniques are equivalent to fixing one λ_i^X at some value:

$$\sum_{i}^{I-1} \lambda_i^X + \lambda_I^X = 0$$

$$\sum_{i}^{I-1} \lambda_i^X = -\lambda_I^X$$

Notation for Log-linear models

Let's get an intuitive grasp of the log-linear specification of independence

$$\ln \mathrm{E}(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

There are $1 + I + J$ parameters on the RHS, but implicitly two are fixed.

For any given cell, only three parameters matter.

1. The baseline count
2. The row probability
3. The column probability

We just add them up

Notation for Log-linear models

Independence is a boring model. What if the effect of X depends on the level of Y?

Then the conditional probability of an event is no longer the product of the marginal probabilities

We need an extra (set of) terms: interaction(s) between X and Y

$$\ln \mathrm{E}(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

We will talk much more about this specification next time, when we talk about 3+ dimensional tables

For a 2D table, interactions *saturate* the model. That is:

- They use up all the degrees of freedom (consider the 2×2)

Notation for Log-linear models

Independence is a boring model. What if the effect of X depends on the level of Y?

Then the conditional probability of an event is no longer the product of the marginal probabilities

We need an extra (set of) terms: interaction(s) between X and Y

$$\ln \mathrm{E}(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

We will talk much more about this specification next time, when we talk about 3+ dimensional tables

For a 2D table, interactions *saturate* the model. That is:

- They use up all the degrees of freedom (consider the 2×2)
- They perfectly predict the counts (equivalent to a dummy for each cell)

Notation for Log-linear models

Independence is a boring model. What if the effect of X depends on the level of Y?

Then the conditional probability of an event is no longer the product of the marginal probabilities

We need an extra (set of) terms: interaction(s) between X and Y

$$\ln \mathrm{E}(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

We will talk much more about this specification next time, when we talk about 3+ dimensional tables

For a 2D table, interactions *saturate* the model. That is:

- They use up all the degrees of freedom (consider the 2×2)
- They perfectly predict the counts (equivalent to a dummy for each cell)
- They perfectly fit the data ($G^2 = 0$)

Notation for Log-linear models

One final model to consider. Suppose that events are equally likely to fall in any cell

$$\mathrm{E}(\mu_{ij}) = n\hat{\pi}$$

Taking logs

$$\ln \mathrm{E}(\mu_{ij}) = \ln n + \ln \hat{\pi}$$

We will rewrite this to hava a single parameter

$$\ln \mathrm{E}(\mu_{ij}) = \lambda$$

This is call the *null* model.

It is the least interesting possible specification, with the worst possible fit.

Note that all models of contingency tables have $G^2_{\text{saturated}} \geq G^2 \geq G^2_{\text{null}}$

Notation for Log-linear models

Note that we have two uninteresting models (the null and independence) and one infeasible model (saturation)

So for the $I \times J$ case, fits, parameters, and the like aren't *too* interesting

They will be for $I \times J \times K \ldots$ tables, where we can have interaction without saturation

But for now, we're mainly stuck with rejecting or accepting independence

Unless we get creative. . .

Estimating Log-linear models

Loglinear models are estimated just like other Poisson models.

The log of the likelihood is

$$\ln \mathcal{L}(\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) = \sum_{i=1}^{N} y_i X_i \beta - \exp(X_i \beta)$$

which we maximize by numerical means. You could use you old `optim()` function.

If you want to analyze data in tabular form, try `loglm` in the `MASS` library of `R`

Interpreting Log-linear models

Poisson parameters represent factor changes in Y given level changes in X.

With LLM, the level change in X is always 1.

So at first blush, we might think that given $X = i$, Y increases by $\exp \lambda_i^X)$ times. . .

But that would be wrong.

When we turn “on” $X = i$, we turn off some other $X = \sim i$.

The constraints on λ_i^X and λ_j^Z make them hard to interpret directly.

FWIW, the difference $\lambda_1 - \lambda_2$ is the log of the odds of being in Row 1 versus Row 2

I recommend showing fitted values,
or first differences, or factor changes under particular counterfactuals

Fitting Log-linear models

Much of the effort in LLM seems to go into choosing the best model

Unlike most modeling exercises, it is possible to consider every LLM against every other

Selection then rests heavily on the choice of criteria

LR tests will tend to favor large models

BIC and other penalized tests will favor parsimony. If n is large, BIC is probably a much safer bet

Refresher on the BIC (for a single model):

$$BIC_k = G^2 - \mathrm{df}\,\ln(n)$$

where n is the sum of the table's cells.

The BIC of the saturated model is 0. BIC< 0 is preferred.

Fitting Log-linear models

We can calculate residuals of a LLM easily.

The Pearson residuals are

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\hat{\mu}_{ij}^{1/2}}$$

Investigating the table of residuals can help identify sources of mis-fit.

(Sidenote: the Pearson residuals sum to the Pearson X^2)

In small tables, residual analysis is complicated by masking:
outliers are skewing the fit, and appear to be less outlying

Can deal with this using “deleted residuals”

Example: Occupational Status Mobility

We will examine a table of social mobility from postwar Britain (Glass 1954; see King 1989)

The table is square; rows give the father's occupational "status", columns give the son's

The 8 classes, in (presumed) order within the status hierarchy, are:

Professional
Manager/executive
High supervisor
Low supervisor
Routine non-manual
Skilled manual
Semi-skilled
Unskilled manual

The dependent variable is the "count" in each cell, corresponding to the number of families with a particular career status trajectory

Example: Occupational Status Mobility

The data (note that it fits easily on one page):

	prof	mana	hsup	lsup	rout	skil	sskl	uskl
prof	50	19	26	8	7	11	6	2
mana	16	40	34	18	11	20	8	3
hsup	12	35	65	66	35	88	23	21
lsup	11	20	58	110	40	183	64	32
rout	2	8	12	23	25	46	28	12
skil	12	28	102	162	90	554	230	177
sskl	0	6	19	40	21	158	143	71
uskl	0	3	14	32	15	126	91	106

What's the dependent variable? What are the independent variables?

How many observations are there?

What distribution should we assume?

Example: Occupational Status Mobility

Occupational mobility tables are a typical example for LLMs

(Another typical example is assortative mating)

The blurring of independent and dependent variables may be an asset in such data

Our hypotheses are really about joint distributions; e.g.,

- Are occupational statuses of father and sons correlated?
- Are sons upwardly or downwardly mobile?
- Are these patterns uniform across the hierarchy?

We begin with a specification assuming independence of father and son status

Example: Occupational Status Mobility

We obtain estimated parameters from `loglm`, which takes in the table above, and spits out. . .

	Father	Son
Professional	-0.929	-1.196
Manager/executive	-0.778	-0.761
High supervisor	0.055	-0.031
Low supervisor	0.461	0.299
Routine non-manual	-0.739	-0.333
Skilled manual	1.423	1.248
Semi-skilled	0.338	0.555
Unskilled manual	0.170	0.219
baseline	3.459	

Enlightening, eh?

Example: Occupational Status Mobility

We observe the following fit, relative to the null & saturated models

	df	G^2	BIC
Null model	63	4679	4165
Independence	49	954	555
Saturation	0	0	0

Recall, the BIC here is, e.g.,

$$BIC = 954 - 49 \times \ln(3498)$$

where the sum over the table $n = 3498$

How do we interpret these results?

Example: Occupational Status Mobility

We could estimate the model using our old Poisson function

But first we'll have to reorganize the data into 64 observations

(Show Excel sheet)

We impose the identifying restriction on λ^X and λ^Y by omitting λ_I^X and λ_J^Y

Recall this equivalent to assuming the λs sum to 1, though the parameterization differs

Example: Occupational Status Mobility

Because of the different identifying assumptions, the estimates from `loglm` and `optim()` look different. But they are exactly equivalent

Parameter	Optim	Loglm
Father prof	-1.0986	-0.9290
Father man	-0.9478	-0.7782
Difference	-0.1508	-0.1508

(Recall that differences of λs are log odds ratios, which are invariant to the identifying restrictions)

It doesn't matter which set of estimates we use; if we do our math right, we'll get the same

- likelihoods
- fitted values
- first difference
- anything of substantive interest

Example: Occupational Status Mobility

Here is a table of the fitted values from the Poisson model

	prof	mana	hsup	lsup	rout	skil	sskl	uskl
prof	3.8	5.9	12.2	16.9	9.0	43.7	21.9	15.6
mana	4.4	6.8	14.2	19.7	10.5	50.9	25.4	18.2
hsup	10.2	15.7	32.5	45.3	24.1	117.0	58.5	41.8
lsup	15.3	23.5	48.9	68.0	36.1	175.6	87.8	62.8
rout	4.6	7.1	14.7	20.5	10.9	52.9	26.4	18.9
skil	39.9	61.6	127.8	177.8	94.5	459.4	229.7	164.2
sskl	13.5	20.8	43.2	60.1	31.9	155.3	77.6	55.5
uskl	11.4	17.6	36.5	50.8	27.0	131.2	65.6	46.9

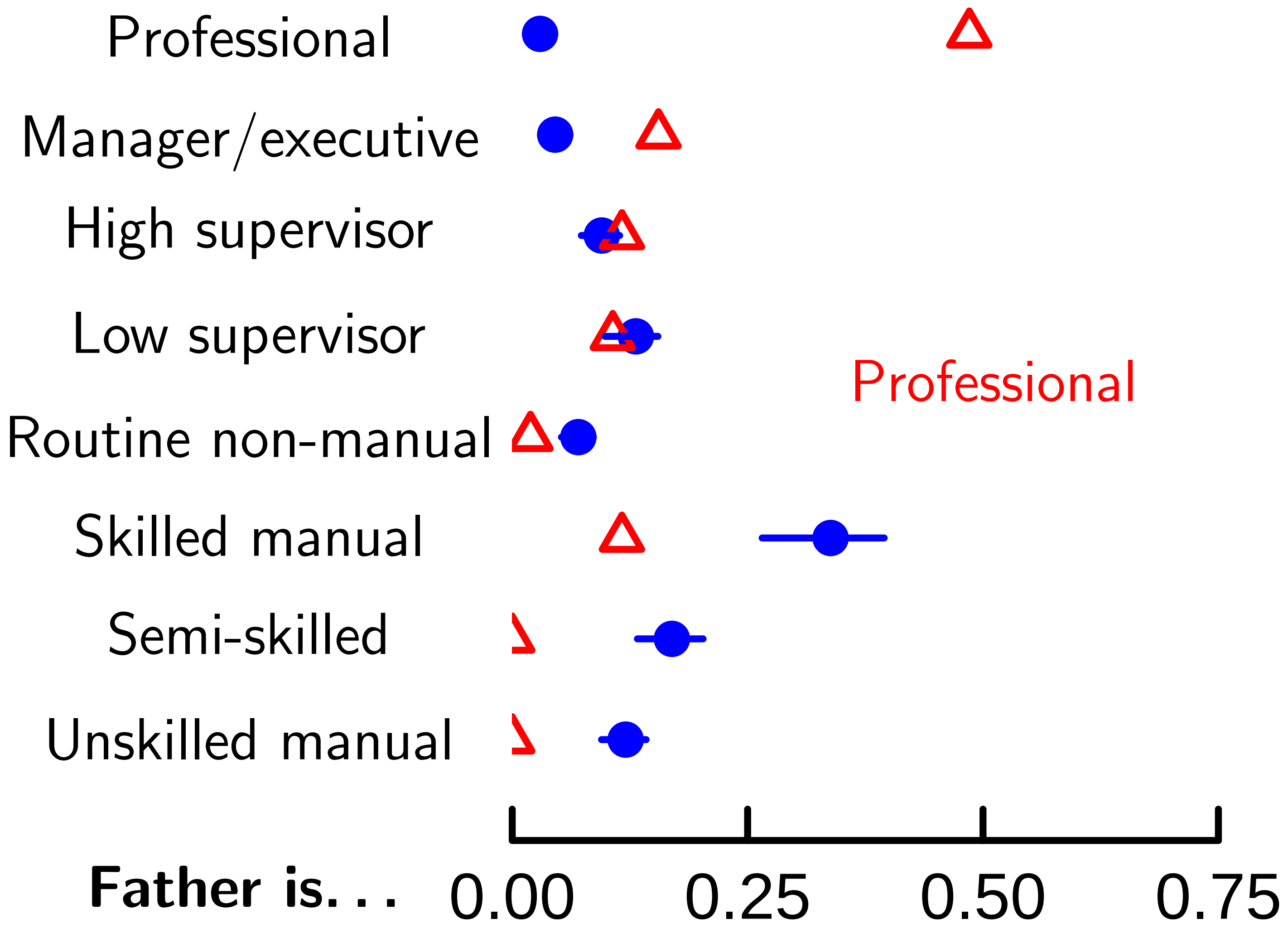
Ugh. Bet you'd like a graphical alternative?

Mosaic plots can be very useful here.

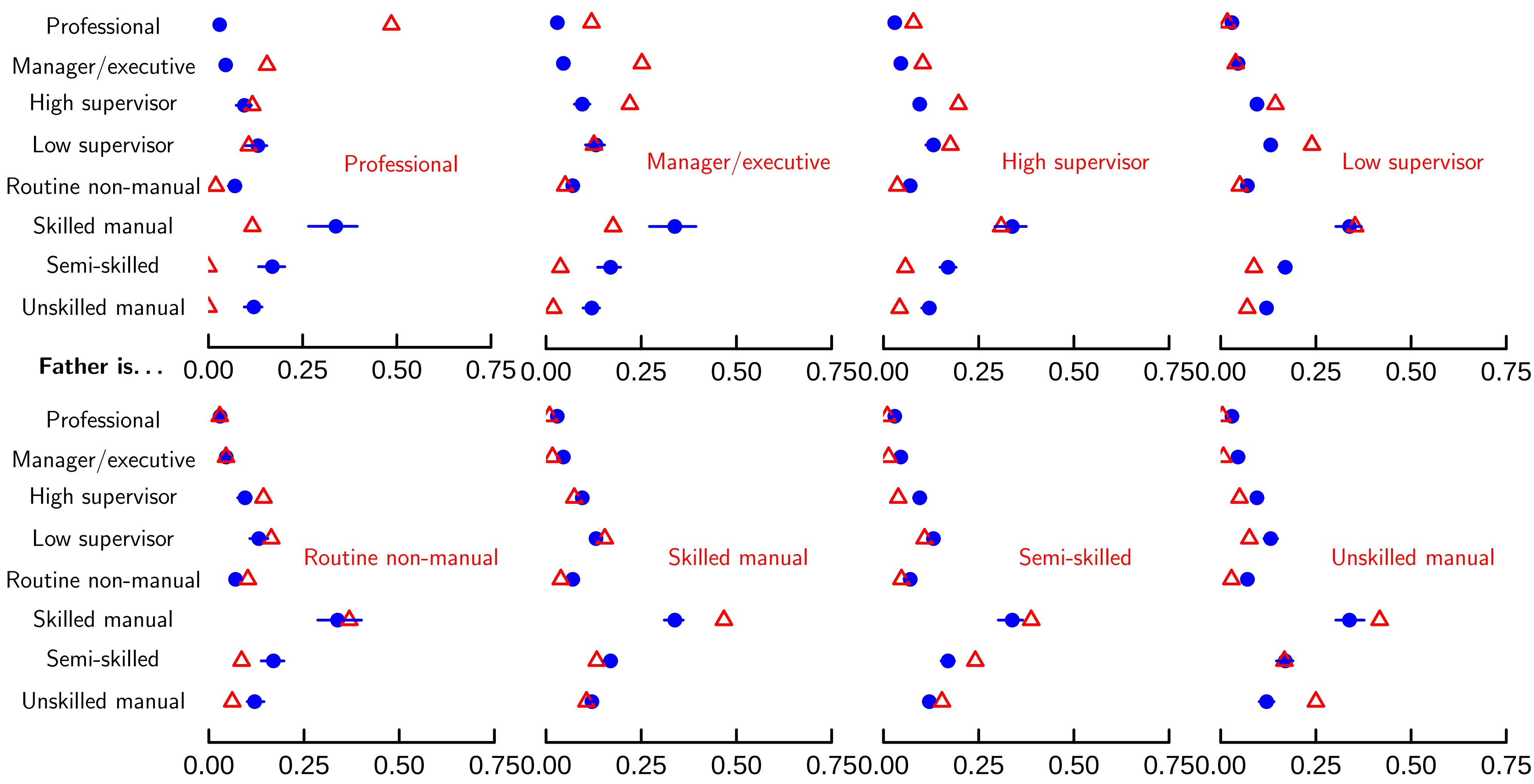
But we'll look at another alternative, the "propeller" plot

We will plot expected probability a son falls in a category given the father's category

Occupational Status: Poisson Fits, with 95% CI & Actual Data



Occupational Status: Poisson Fits, with 95% CI & Actual Data



A little too good. . .

The Poisson estimates seem suspiciously precise. What could be causing this?

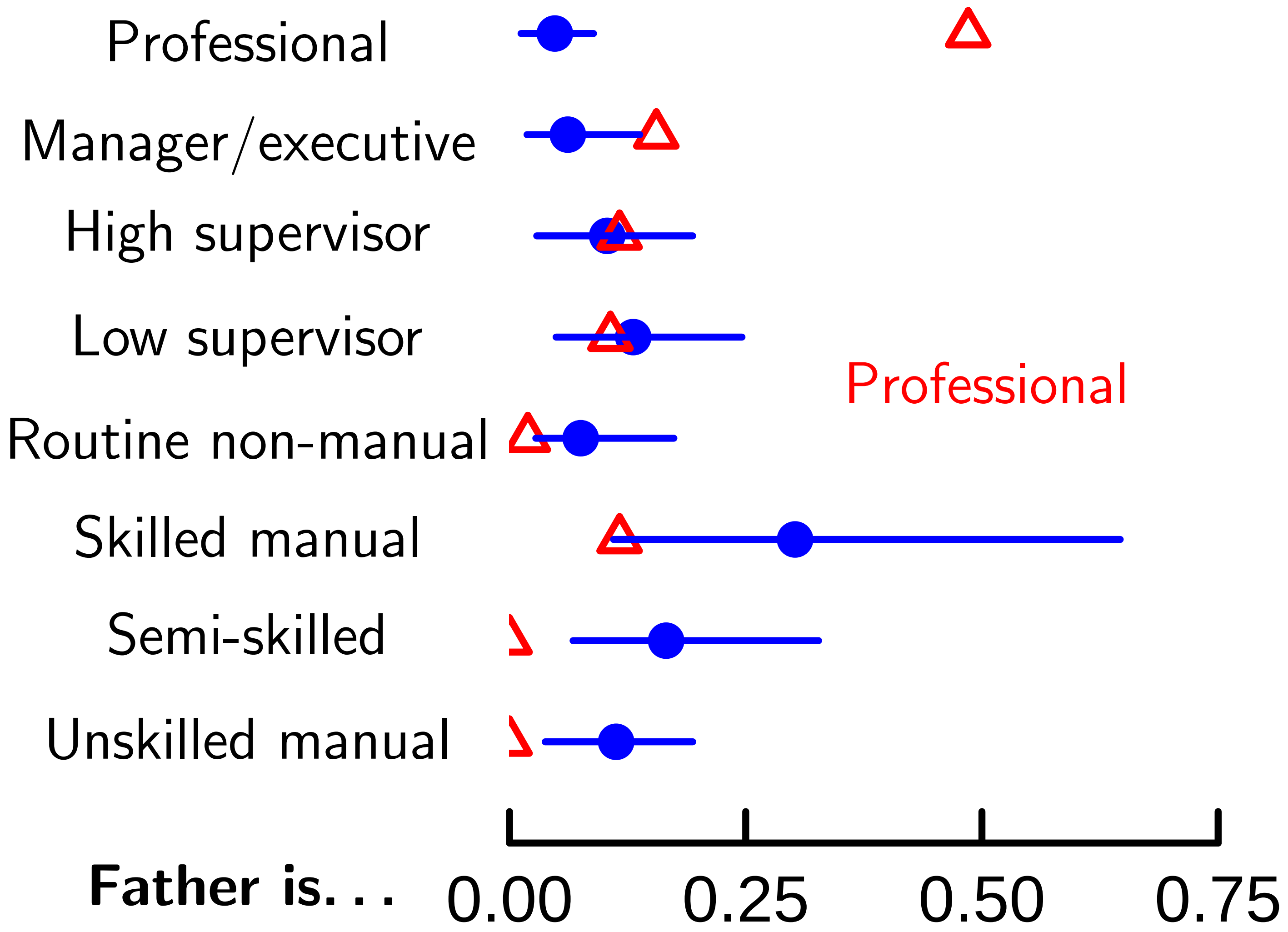
Perhaps there is overdispersion in these data, leading to biased standard errors?

(What would overdispersion mean in this case?)

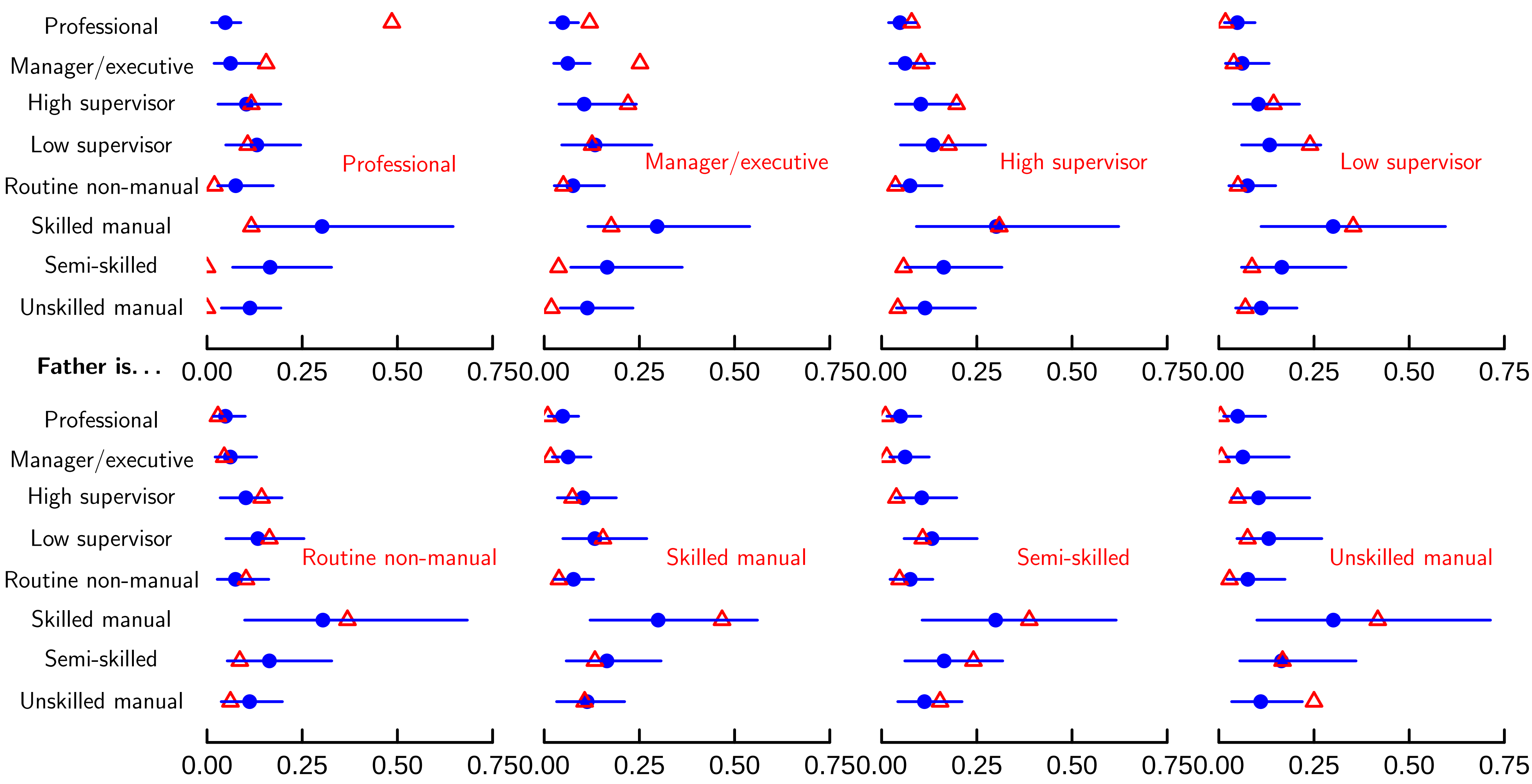
(How do we cope with overdispersion?)

Let's re-estimate with the negative binomial

Negative Binomial Fits, with 95% CI & Actual Data



Negative Binomial Fits, with 95% CI & Actual Data



Give up?

Yikes!

It looks like we can't be sure of anything. . .

But maybe we're just asking too much of the data.

Trying to estimate 15 parameters from 64 datapoints is rather greedy

Can we put build a simpler, theoretically sharpened specification?

What might it be?

Transforming the variables

Sometimes, we'll want a compromise specification that doesn't just dummy out each row or column

We might construct a theoretically interesting new variable from the rows and columns

For example, how about `inherit`, `upward`, and `downward`:

$$\begin{aligned}
\texttt{Inherit} &= 1 \quad \text{if} \quad \text{Son} = \text{Father}, \quad 0 \quad \text{otherwise} \\
\texttt{Upward} &= 1 \quad \text{if} \quad \text{Son} > \text{Father}, \quad 0 \quad \text{otherwise} \\
\texttt{Downward} &= 1 \quad \text{if} \quad \text{Son} < \text{Father}, \quad 0 \quad \text{otherwise}
\end{aligned}$$

We just recode our data so that for each of the 64 cells,
instead of regression on Son and Father, we have regression in Inherit and Upward

This is a simpler model than independence (3 parameters, rather than 15), but more complicated than the null model (1 parameter)

Don't let the tabular frame trap you into a certain style of specification

Example: Inheriting occupational class

Poisson	
Inherit	1.232
	(0.043)
Upward	0.144
	(0.041)
Constant	3.685
	(0.030)
N	64

We'll run the regression using the Poisson model

Note that although the data are all categorical, we're doing the *exact same thing* we did with continous RHS variables.

This is still, in all respects, a Poisson model

Still, the interpretation may be a little confusing, because the distinction between the dependent and independent variables is blurred. . . .

Example: Inheriting occupational class

Poisson	
Inherit	1.232
	(0.043)
Upward	0.144
	(0.041)
Constant	3.685
	(0.030)
N	64

The dependent variable is the "count" in each cell.

Positive coefficients suggest an higher count in a cell, when the explanatory condition is met.

These are still Poisson coefficients, so $\mathrm{E}(Y) = \exp(X\hat{\beta})Y$, but Y is just a "count"

Counts are $\exp(1.232) \approx 3.42$ times bigger when sons inherit occupational status, holding `upward` constant

(what's the problem with the above statement?)

Example: Inheriting occupational class

The statement on the previous page is inaccurate, in these sense that it depends, through a logical constraint, on the other variables

Let's calculate first diffs, taking care to specify the value of the other covariate

		Poisson	
	1st diff	Lower 95%	Upper 95%
Down → Inh	96.8	88.6	105.2
Inh → Up	-90.6	-99.31	-82.3
Down → Up	6.2	2.7	9.6

The average cell count is about 54.7.

Inheritance cells are expected to have 96.8 more members than Downward cells and 90.6 more members than Upward mobility cells

Upward cells are expected to have 6.2 more members than Downward cells and 90.6 fewer members than Inheritance cells

All relationships appear significant, and the Inheritance cells seem precisely estimated

Example: Inheriting occupational class

The simpler model with just two variables seems more informative than the others we've tried

Does it fit as well?

No.

	df	G^2	BIC
Null model	63	4679	4164.9
Inherit, Upward	61	3824	3326.2
Independence	49	955	554.7
R & C Marginals, Inh, Upw	47	657	273.7
Saturation	0	0	0

The best fitting model (on whatever criteria) is not always the most useful

An ideal model simplifies the substance of the data and fits the data well

We can't always have both—sometimes there is a tradeoff

Example: Inheriting occupational class

Under the Poisson, effects appear very significant/precisely estimated.

Maybe suspiciously so. We only have 64 observations.

Example: Inheriting occupational class

Let's reestimate using the negative binomial.

	Poisson	Negative Binomial
Inherit	1.232	1.232
	(0.043)	(0.491)
Upward	0.144	0.144
	(0.041)	(0.329)
Constant	3.685	3.685
	(0.030)	(0.233)
"theta"		0.868
		(0.142)
N	64	64

(What do we make of this table?)

There is evidence of overdispersion

The coefficients are essentially unchanged, but the standard errors are *much* bigger

Substantive conclusions *has* changed: we no longer can conclude that there is upward mobility

Example: Inheriting occupational class

The first differences show the change in precision rather dramatically:

	Poisson			Negative Binomial		
	1st diff	Lower 95%	Upper 95%	1st diff	Lower 95%	Upper 95%
Down → Inh	96.8	88.6	105.2	109.2	14.8	280.3
Inh → Up	-90.6	-99.3	-82.3	-102.4	-271.5	-7.0
Down → Up	6.2	2.7	9.6	6.4	-22.0	35.9

Remember that the average cell count is about 54.7.

Also note that the NB model fits much better than the Poisson.

If we include all marginals, Inherit, and Upward, we get the best model yet by fit: $G^2 = 74.1$, $BIC = -301.2$, with 46 degrees of freedom

Conclusion: Estimating Log-linear models using Poisson is *dangerous*

Always check for overdispersion