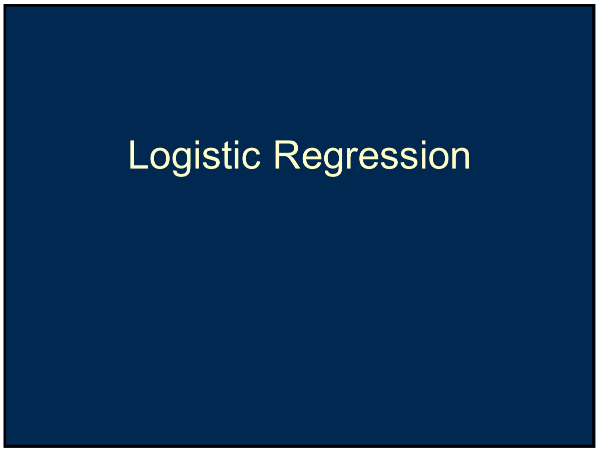
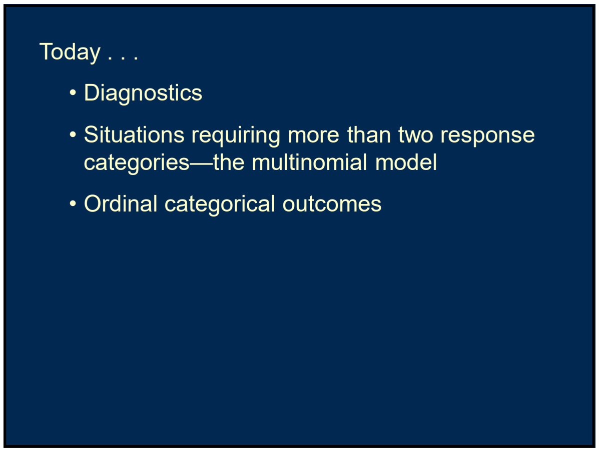
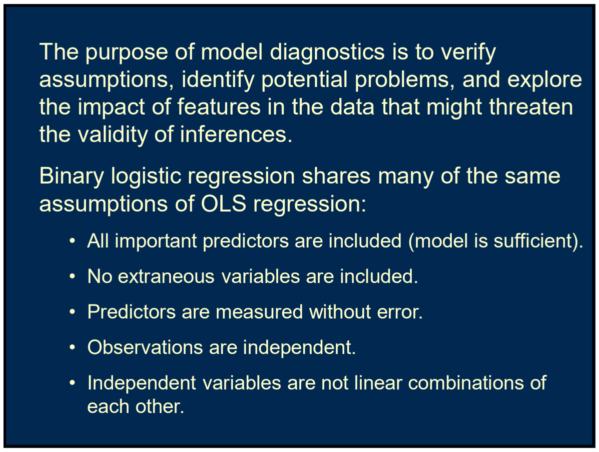


Logistic Regression



Today . . .

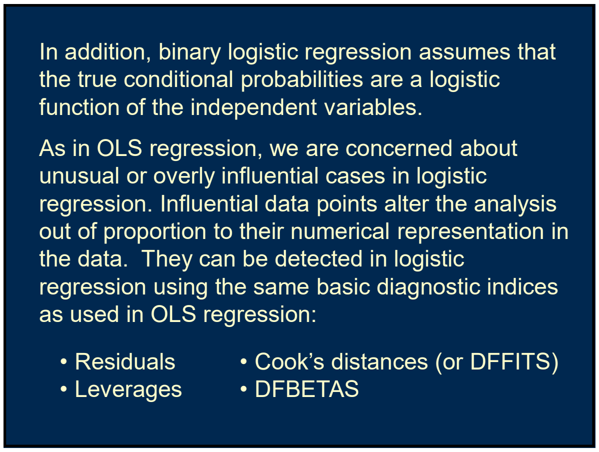
- Diagnostics
- Situations requiring more than two response categories—the multinomial model
- Ordinal categorical outcomes



The purpose of model diagnostics is to verify assumptions, identify potential problems, and explore the impact of features in the data that might threaten the validity of inferences.

Binary logistic regression shares many of the same assumptions of OLS regression:

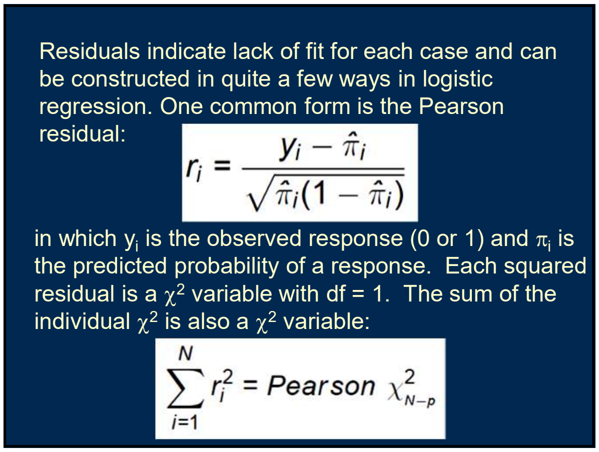
- All important predictors are included (model is sufficient).
- No extraneous variables are included.
- Predictors are measured without error.
- Observations are independent.
- Independent variables are not linear combinations of each other.



In addition, binary logistic regression assumes that the true conditional probabilities are a logistic function of the independent variables.

As in OLS regression, we are concerned about unusual or overly influential cases in logistic regression. Influential data points alter the analysis out of proportion to their numerical representation in the data. They can be detected in logistic regression using the same basic diagnostic indices as used in OLS regression:

- Residuals
- Leverages
- Cook's distances (or DFFITS)
- DFBETAS

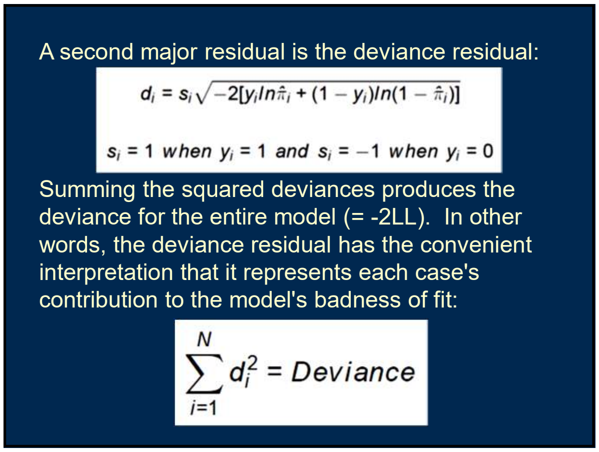


Residuals indicate lack of fit for each case and can be constructed in quite a few ways in logistic regression. One common form is the Pearson residual:

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

in which y_i is the observed response (0 or 1) and π_i is the predicted probability of a response. Each squared residual is a χ^2 variable with df = 1. The sum of the individual χ^2 is also a χ^2 variable:

$$\sum_{i=1}^{N} r_i^2 = Pearson\ \chi^2_{N-p}$$



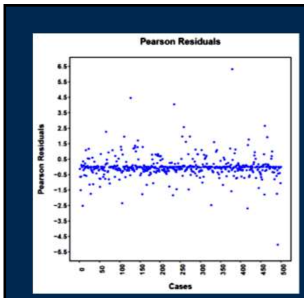
A second major residual is the deviance residual:

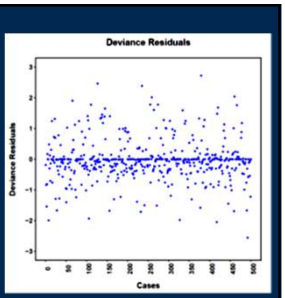
$$d_i = s_i\sqrt{-2[y_i ln\hat{\pi}_i + (1-y_i)ln(1-\hat{\pi}_i)]}$$

$$s_i = 1\ when\ y_i = 1\ and\ s_i = -1\ when\ y_i = 0$$

Summing the squared deviances produces the deviance for the entire model (= -2LL). In other words, the deviance residual has the convenient interpretation that it represents each case's contribution to the model's badness of fit:

$$\sum_{i=1}^{N} d_i^2 = Deviance$$





Nothing too awful is apparent in the case-wise display of these residuals. These residuals can also be standardized and studentized to help isolate oddballs more easily.

Standardization accounts for each case's profile of predictor values as represented by the leverage. The leverage is a measure of a case's potential influence on a regression model. It represents how different a case is from the remaining cases in the multivariate space of the predictors.

$$r_{si} = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

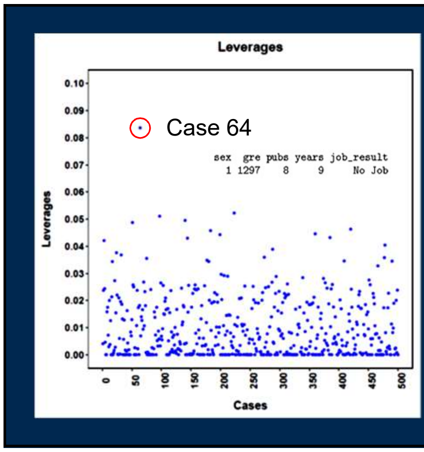
$$d_{si} = \frac{d_i}{\sqrt{1 - h_{ii}}}$$

These residuals are sometimes further adjusted by calculating them from a model that excludes each case so that it cannot influence the model on which its own residual is based. These are known as studentized residuals.

The leverage is a measure of a case's potential influence on a regression model. It comes from the diagonal of the "hat" matrix and represents how different a case is from the remaining cases in the multivariate space of the predictors. They are sometimes called the “hat” values.

$$H = W^{.5}X(X^TWX)^{-1}X^TW^{.5}$$

As leverage gets larger it indicates that a case has the potential to have more influence than it should. The leverage has a close relationship with the Mahalanobis distances for the predictors.

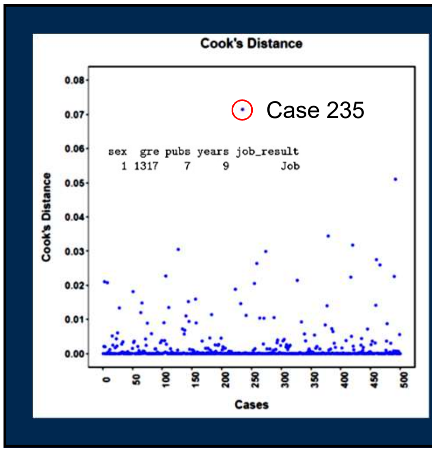


sex	gre	pubs	years	job_result
1	1297	8	9	No Job

One case stands out as a bit more extreme than the others.

Cook's distance is a general measure of influence that combines information about residuals and leverages. It is a scaled change in the fitted values that arises from eliminating each case in turn from the model. The scaling takes into account the leverage for each case.

Cases with unusual Cook's distances are cause for concern because they are influencing the model fit out of proportion to their numerical representation in the data.



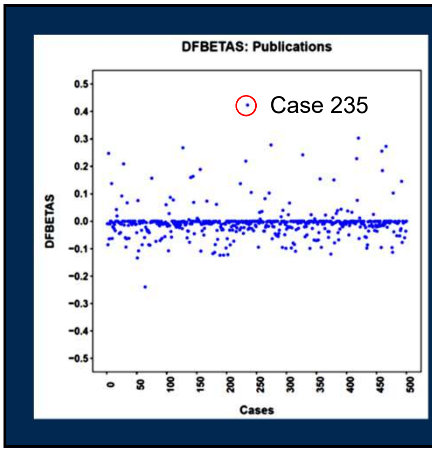
sex	gre	pubs	years	job_result
1	1317	7	9	Job

Here, too, there is one case that stands out a bit.

Cook's distance is a general measure of influence. DFBETAS is coefficient-specific influence:

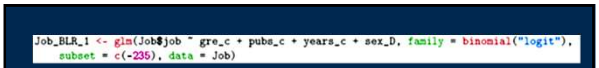
$$DFBETAS_{j(i)} = \frac{\beta_j - \beta_{j(i)}}{SE_{\beta_{j(i)}}}$$

In other words, it is the standardized amount by which a coefficient changes when a case is excluded from the model.

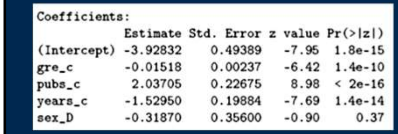


Same case as for Cook's distance.

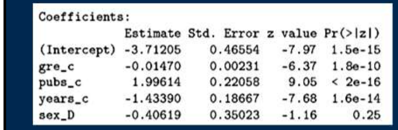
To see the effect on inferences, we could remove the case and re-run the analysis.



```
Job_BLR_1 <- glm(Job$job ~ gre_c + pubs_c + years_c + sex_D, family = binomial("logit"),
    subset = c(-235), data = Job)
```



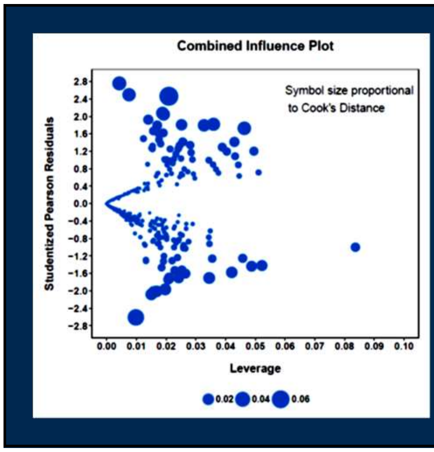
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.92832    0.49389   -7.95  1.8e-15
gre_c       -0.01518    0.00237   -6.42  1.4e-10
pubs_c       2.03705    0.22675    8.98  < 2e-16
years_c     -1.52950    0.19884   -7.69  1.4e-14
sex_D       -0.31870    0.35600   -0.90     0.37
```



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.71205    0.46554   -7.97  1.5e-15
gre_c       -0.01470    0.00231   -6.37  1.8e-10
pubs_c       1.99614    0.22058    9.05  < 2e-16
years_c     -1.43390    0.18667   -7.68  1.6e-14
sex_D       -0.40619    0.35023   -1.16     0.25
```

Original

No change in the inferences when the most unusual case is removed.

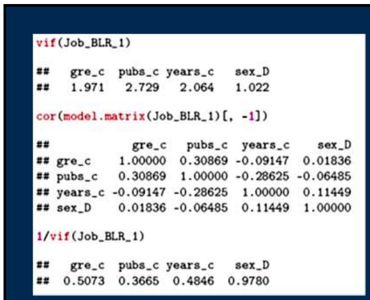


All of the information can be displayed at once.

We also need to consider multicollinearity and overdispersion. The variance inflation factor is an index of multicollinearity. It represents by how much the variance for a coefficient is inflated compared to a model in which the predictors are uncorrelated. It is defined as:

$$VIF_j = \frac{1}{1 - R_j^2}$$

in which R^2_j is the coefficient of determination for a particular predictor (the proportion of variance that a predictor shares with the other predictors). As VIF increases, there is increasing concern about multicollinearity. The tolerance is the reciprocal of VIF.



```
vif(Job_BLR_1)

##   gre_c  pubs_c years_c   sex_D
##   1.971   2.729   2.064   1.022

cor(model.matrix(Job_BLR_1)[, -1])

##              gre_c    pubs_c  years_c    sex_D
## gre_c     1.00000  0.30869 -0.09147  0.01836
## pubs_c    0.30869  1.00000 -0.28625 -0.06485
## years_c -0.09147 -0.28625  1.00000  0.11449
## sex_D     0.01836 -0.06485  0.11449  1.00000

1/vif(Job_BLR_1)

##   gre_c  pubs_c years_c   sex_D
##  0.5073  0.3665  0.4846  0.9780
```

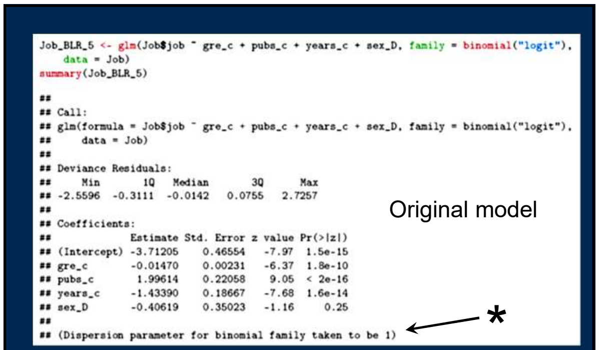
The predictors show low to moderate multicollinearity.

By selecting the binomial model, we assume that the dispersion is consistent with that model. In other words, in a binomial distribution we assume the mean and variance are related:

$$\mu = p$$

$$\sigma^2 = p(1-p)$$

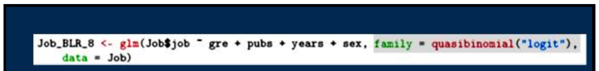
Most commonly the variance will be larger than expected, known as overdispersion. Heterogeneous samples and unmodeled predictors can inflate variance. Underdispersion is also possible; both create inaccuracies.



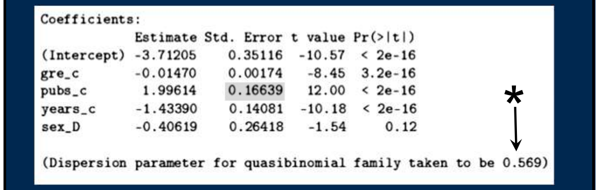
```
Job_BLR_5 <- glm(Job$job ~ gre_c + pubs_c + years_c + sex_D, family = binomial("logit"),
    data = Job)
summary(Job_BLR_5)

##
## Call:
## glm(formula = Job$job ~ gre_c + pubs_c + years_c + sex_D, family = binomial("logit"),
##     data = Job)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5596  -0.3111  -0.0142   0.0755   2.7257
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.71205    0.46554   -7.97  1.5e-15
## gre_c       -0.01470    0.00231   -6.37  1.8e-10
## pubs_c       1.99614    0.22058    9.05  < 2e-16
## years_c     -1.43390    0.18667   -7.68  1.6e-14
## sex_D       -0.40619    0.35023   -1.16     0.25
##
## (Dispersion parameter for binomial family taken to be 1)
```

Original model

*

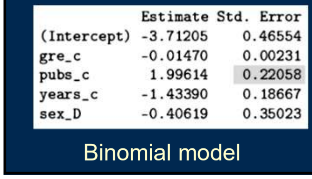


```
Job_BLR_8 <- glm(Job$job ~ gre + pubs + years + sex, family = quasibinomial("logit"),
    data = Job)
```



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.71205    0.35116  -10.57  < 2e-16
gre_c       -0.01470    0.00174   -8.45  3.2e-16
pubs_c       1.99614    0.16639   12.00  < 2e-16
years_c     -1.43390    0.14081  -10.18  < 2e-16
sex_D       -0.40619    0.26418   -1.54     0.12

(Dispersion parameter for quasibinomial family taken to be 0.569)
```

*



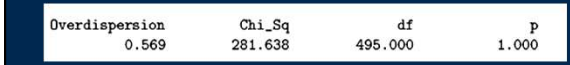
```
            Estimate Std. Error
(Intercept) -3.71205    0.46554
gre_c       -0.01470    0.00231
pubs_c       1.99614    0.22058
years_c     -1.43390    0.18667
sex_D       -0.40619    0.35023
```

$$\sigma^2_{Quasi-Binomial} = .569\sigma^2_{Binomial}$$

Binomial model

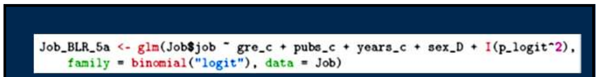
$$\sum_{i=1}^{N} r_i^2 = Pearson\ \chi^2_{N-p}$$

If the model follows the expected binomial distribution, this χ^2 has an expected value of N-p, its degrees of freedom. This can be used as a test of the acceptability of the underlying model. The most common violation is overdispersion.

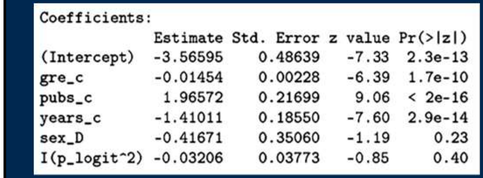


```
Overdispersion        Chi_Sq            df            p
         0.569       281.638       495.000        1.000
```

Is the logistic model adequate? To test this, we can save the fitted logistic values, square them, and then enter them as an additional predictor in the model. This tests if there is any relationship beyond a linear logistic in the data. The coefficient for this new predictor should be clearly nonsignificant.



```
Job_BLR_5a <- glm(Job$job ~ gre_c + pubs_c + years_c + sex_D + I(p_logit^2),
    family = binomial("logit"), data = Job)
```



```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.56595    0.48639   -7.33  2.3e-13
gre_c         -0.01454    0.00228   -6.39  1.7e-10
pubs_c         1.96572    0.21699    9.06  < 2e-16
years_c       -1.41011    0.18550   -7.60  2.9e-14
sex_D         -0.41671    0.35060   -1.19     0.23
I(p_logit^2)  -0.03206    0.03773   -0.85     0.40
```

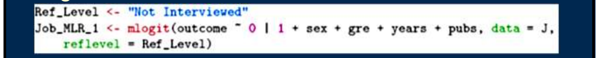
No need to consider a different link function. The logit link is working just fine.

Overall, the data are fit well by a binary logistic model. There are no truly unusual cases. The predictors are reasonably independent. If anything, the model is underdispersed, and so is conservative compared to a truly binomial distribution.

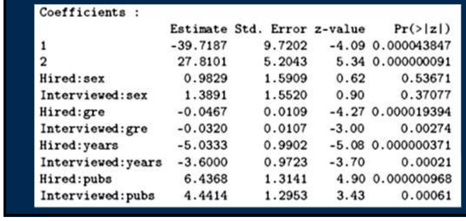
When there are more than two categories for the outcome variable, multinomial logistic regression is used. It might be useful for other problems as well.

- Rating scales with more than 2 (crude) levels
- Doubts about the interval or ordinal nature of a rating scale
- Interest in particular levels of a rating scale (e.g., neutral).
- Messy outcomes
- Patterns of outcomes (including some repeated measures)

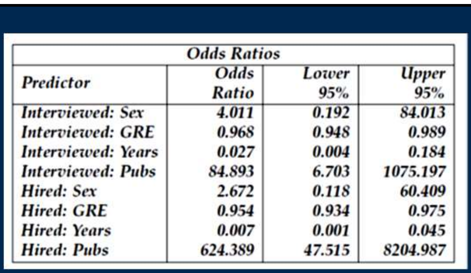
In multinomial logistic regression, one of the outcome categories is chosen as the reference. Each of the remaining outcome categories are compared to the reference, akin to a collection of binary logistic regressions.



```
Ref_Level <- "Not Interviewed"
Job_MLR_1 <- mlogit(outcome ~ 0 | 1 + sex + gre + years + pubs, data = J,
    reflevel = Ref_Level)
```

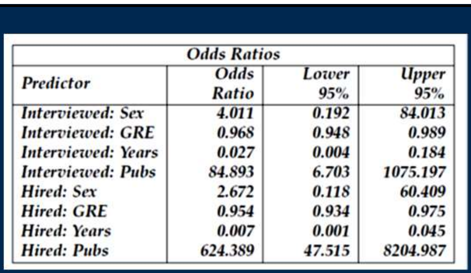


```
Coefficients :
                 Estimate Std. Error z-value    Pr(>|z|)
1                -39.7187     9.7202   -4.09 0.000043847
2                 27.8101     5.2043    5.34 0.000000091
Hired:sex          0.9829     1.5909    0.62     0.53671
Interviewed:sex    1.3891     1.5520    0.90     0.37077
Hired:gre         -0.0467     0.0109   -4.27 0.000019394
Interviewed:gre   -0.0320     0.0107   -3.00     0.00274
Hired:years       -5.0333     0.9902   -5.08 0.000000371
Interviewed:years -3.6000     0.9723   -3.70     0.00021
Hired:pubs         6.4368     1.3141    4.90 0.000000968
Interviewed:pubs   4.4414     1.2953    3.43     0.00061
```



Odds Ratios			
Predictor	Odds Ratio	Lower 95%	Upper 95%
Interviewed: Sex	4.011	0.192	84.013
Interviewed: GRE	0.968	0.948	0.989
Interviewed: Years	0.027	0.004	0.184
Interviewed: Pubs	84.893	6.703	1075.197
Hired: Sex	2.672	0.118	60.409
Hired: GRE	0.954	0.934	0.975
Hired: Years	0.007	0.001	0.045
Hired: Pubs	624.389	47.515	8204.987

Each publication multiplies the odds of getting an interview versus not getting an interview by 84.9. Each publication multiplies the odds of getting hired versus not getting an interview by 624.4.



Odds Ratios			
Predictor	Odds Ratio	Lower 95%	Upper 95%
Interviewed: Sex	4.011	0.192	84.013
Interviewed: GRE	0.968	0.948	0.989
Interviewed: Years	0.027	0.004	0.184
Interviewed: Pubs	84.893	6.703	1075.197
Hired: Sex	2.672	0.118	60.409
Hired: GRE	0.954	0.934	0.975
Hired: Years	0.007	0.001	0.045
Hired: Pubs	624.389	47.515	8204.987

Each year to complete multiplies the odds of getting an interview versus not getting an interview by .027. This represents a 97.3% decrease in the odds of getting interviewed.

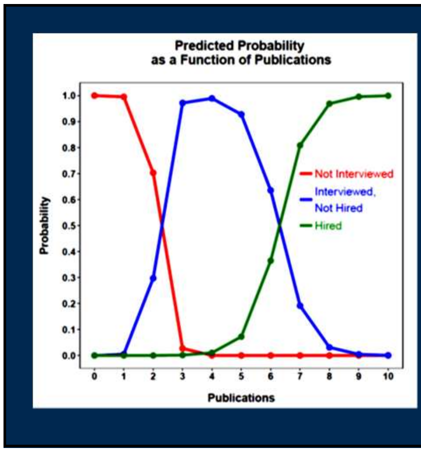
Logits for each model are calculated in the usual fashion. Probabilities are a little more complicated.

For outcomes, m=2 to M, the probability of outcome, m:

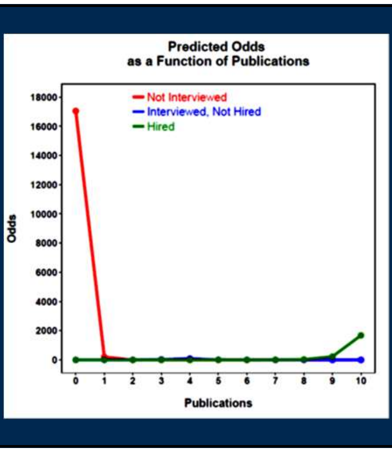
$$\frac{e^{L_m}}{1+\sum_{j=2}^{M} e^{L_j}}$$

For the reference category, m = 1, the probability is:

$$\frac{1}{1+\sum_{j=2}^{M} e^{L_j}}$$



These probabilities assume other predictors are constant at their means.



If the outcome categories have an order, that information can be used in the analysis. There are several varieties of ordinal logistic models, the most common being the proportional odds model.

This means that predictors are constrained to have a constant influence across ordinal transitions.

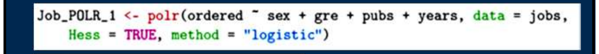
The logistic regression attempts to find the linear model that distinguishes the outcome groups defined by successive divisions along the ordinal scale.

1 = Not interviewed
2 = Interviewed but not hired
3 = Hired

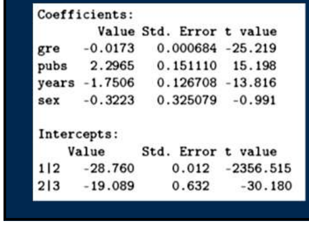
What is the probability that someone will not be interviewed?

What is the probability that someone will either not be interviewed or be interviewed but not hired?

If the assumptions hold, most notably the *proportional odds assumption*, then the ordinal regression is simpler and easier to estimate and interpret than a multinomial regression for the same data.



```
Job_POLR_1 <- polr(ordered ~ sex + gre + pubs + years, data = jobs,
    Hess = TRUE, method = "logistic")
```

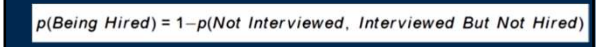


```
Coefficients:
        Value Std. Error t value
gre   -0.0173   0.000684 -25.219
pubs   2.2965   0.151110  15.198
years -1.7506   0.126708 -13.816
sex   -0.3223   0.325079  -0.991

Intercepts:
    Value     Std. Error t value
1|2  -28.760     0.012   -2356.515
2|3  -19.089     0.632     -30.180
```

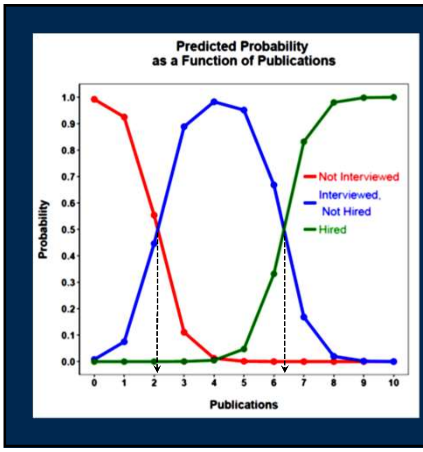
Each coefficient refers to a binary distinction along the ordinal outcome scale. The same coefficients apply to all distinctions. Note different intercepts.

The logits are calculated for a particular cut-point. They can then be converted to probabilities, but these are cumulative probabilities for the outcomes at or below the given cut-point.

The probabilities of a particular outcome can be calculated through subtraction:

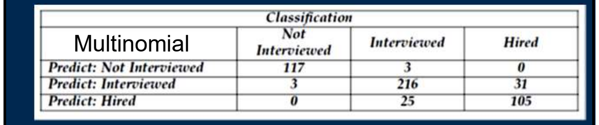


$$p(Being\ Hired) = 1 - p(Not\ Interviewed,\ Interviewed\ But\ Not\ Hired)$$

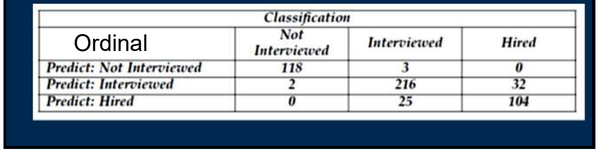


Transition points can be identified that show where outcomes become more probable than those preceding it on the ordinal scale.

Despite the more restrictive model, classification is still impressive and comparable to the multinomial model in which separate coefficients were estimated for each binary comparison.



	Classification		
Multinomial	*Not Interviewed*	*Interviewed*	*Hired*
Predict: Not Interviewed	*117*	*3*	*0*
Predict: Interviewed	*3*	*216*	*31*
Predict: Hired	*0*	*25*	*105*



	Classification		
Ordinal	*Not Interviewed*	*Interviewed*	*Hired*
Predict: Not Interviewed	*118*	*3*	*0*
Predict: Interviewed	*2*	*216*	*32*
Predict: Hired	*0*	*25*	*104*

Next time . . .

MANOVA