Optional Exercises
with Solutions in SAS and R
Short Course on Modeling Ordinal Categorical Data

Bernhard Klingenberg

1. Consider the mental impairment data analyzed in the course, which are available at http://sites.williams.edu/bklingen/ordinal.

   a. Using your choice of software, fit the cumulative logit model discussed in the notes.

   b. Conduct a likelihood-ratio or Wald test about the life events effect, and interpret.

   c. Construct a confidence interval for a cumulative odds ratio to interpret the life events effect.

   d. Now fit the more general model that allows interaction between life events and SES in their effects on mental impairment. Interpret the nature of the interaction.

   e. Test whether the interaction effect is needed in the model. Interpret.

   f. Plot the estimated cumulative logits, cumulative probabilities and category probabilities against the life score for each SES category.

2. Table 1 from a study of the efficacy of seat-belt use in auto accidents has the response categories (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized but did not die, and (5) injured and died. Table 2 shows output for a cumulative logit model, using indicator variables for predictors, that allows the effect of seat-belt use to vary by location.

   a. Why are there four intercepts? Explain how they determine the estimated response distribution for males in urban areas wearing seat belts.

   b. Estimate and interpret the cumulative odds ratio that describes the effect of gender, given seat-belt use and location. (Since gender does not occur in an interaction term, it is valid to estimate this "main effect.") Construct a 95% confidence interval for the effect, and interpret.

   c. Find the estimated cumulative odds ratio between the response and seat-belt use for those in rural locations and for those in urban locations, given gender. Based on this, explain how the effect of seat-belt use varies by location, and explain how to interpret the interaction estimate.

3. Analyze the family income and happiness data mentioned in the notes, treating family income as quantitative with scores (3, 2, 1).

   a. Fit a cumulative logit model and interpret the income effect estimate.

b. Now treat income as a qualitative factor instead of a quantitative predictor with scores. Interpret the effects. Analyze whether a significantly improved fit results from treating income as a qualitative factor.

c. Plot the models in 3a and 3b on the logit scale. If possible, include the sample cumulative logits in your plot to check the fit of the model. Also plot the fitted cumulative and category probabilities for the model in part a.

d. Fit a model that allows non-proportional odds (treating income as quantitative) and plot it. Check if the proportional odds assumption is reasonable.

e. Test goodness of fit for the proportional odds model when treating income as quantitative. (This shows some lack of fit, but note the large sample size. Note how the plots in 3c show a good fit to the sample logits.)

f. Fit an adjacent-categories logit model analog of the model in (a), and interpret the income effect estimate. Compare the fitted values to those for the cumulative logit model in (a), and note how similar they are. These two models describe similar behavior (e.g., stochastically ordered distributions, varying in location rather than dispersion) and fit well in similar situations.

4. Fit a continuation-ratio logit model to the happiness and income data on p. 11 of the notes. Interpret results. (You might note that you'll get different results if you reverse the order of response categories to which you apply the continuation-ratio logits.)

5. Fitting the cumulative probit model (using R) to the happiness and income data, using scores (3, 2, 1) for the income levels, gives results

```
Coefficients:
           Value Std. Error  t value
income 0.3634578 0.02988844 12.16048

Intercepts:
    Value   Std. Error t value
1|2 -0.4681  0.0607    -7.7115
2|3  1.2067  0.0632    19.0893

Residual Deviance: 5487.385
AIC: 5493.385
```

a. See if you replicate the output and interpret the effect estimate of income (i.e., 0.363).

b. The corresponding cumulative logit model gives results

```
Coefficients:
           Value Std. Error  t value
income 0.6310666  0.0523753 12.04894

Intercepts:
    Value   Std. Error t value
1|2 -0.7613  0.1057    -7.1992
2|3  2.0461  0.1122    18.2428
```

Table 1: **Data for Exercise on Degree of Injury in Auto Accident**

			Response on Injury Outcome				
Gender	Location	Seat Belt	1	2	3	4	5
Female	Urban	No	7,287	175	720	91	10
		Yes	11,587	126	577	48	8
	Rural	No	3,246	73	710	159	31
		Yes	6,134	94	564	82	17
Male	Urban	No	10,381	136	566	96	14
		Yes	10,969	83	259	37	1
	Rural	No	6,123	141	710	188	45
		Yes	6,693	74	353	74	12

Table 2: **Output for Exercise on Auto Accident Injuries**

Parameter			DF	Estimate	Std Error
Intercept1			1	3.3074	0.0351
Intercept2			1	3.4818	0.0355
Intercept3			1	5.3494	0.0470
Intercept4			1	7.2563	0.0914
gender	female		1	−0.5463	0.0272
gender	male		0	0.0000	0.0000
location	rural		1	−0.6988	0.0424
location	urban		0	0.0000	0.0000
seatbelt	no		1	−0.7602	0.0393
seatbelt	yes		0	0.0000	0.0000
location*seatbelt	rural	no	1	−0.1244	0.0548
location*seatbelt	rural	yes	0	0.0000	0.0000
location*seatbelt	urban	no	0	0.0000	0.0000
location*seatbelt	urban	yes	0	0.0000	0.0000

```
Residual Deviance: 5487.699
AIC: 5493.699
```

Interpret the income effect (0.631), and compare substantive results to those for the cumulative probit model.

c. Plot the fitted cumulative probabilities in terms of income for the logit and probit model.

Solutions to Exercises using SAS

Exercise 1: Cumulative Logit model for mental impairment data:

SAS code (using proc genmod):

1a)

```
data mental;
input impair ses life;
datalines;
1 1 1
1 1 9
1 1 4
1 1 3
1 0 2
...
4 0 8
4 0 9
;
proc genmod data=mental;
model impair = life ses / dist=multinomial link=clogit type3 aggregate;
run;
```

Selected Output:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Log Likelihood		-49.5489	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	-0.2819	0.6423	-1.5615	0.9839	0.19	0.6607
Intercept2	1	1.2128	0.6607	-0.0507	2.5656	3.37	0.0664
Intercept3	1	2.2094	0.7210	0.8590	3.7123	9.39	0.0022
life	1	-0.3189	0.1210	-0.5718	-0.0920	6.95	0.0084
ses	1	1.1112	0.6109	-0.0641	2.3471	3.31	0.0689

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
life	1	7.78	0.0053
ses	1	3.43	0.0641

1b) Likelihood Ratio (LR) test for H0: $\beta_1 = 0$ vs H1: $\beta_1 \neq 0$ (coefficient for life): LR statistic = 7.78 on df = 1: P-value = 0.0053. The data provide evidence (P-value = 0.0053) of a significant effect of the number of life events on the cumulative log-odds of mental impairment. (Wald Statistic in Chi-square form: (-0.3189/0.1210)^2 = 6.95, df = 1: P-value = 0.0084.)

Interpretation of effect: For both low and high SES adults, the odds of a mental impairment score less than or equal to *j* (instead of greater than *j*) decrease by a factor of exp(-0.3189) = 0.73 for every unit increase in the life event score. This is true for all *j* (proportional odds assumption). For instance, when *j* = 1 = "well", the estimated odds of feeling well

decrease by a factor of 0.73 for every unit increase in life events. When j = 2, the odds of feeling well or showing mild symptoms of mental impairment (versus moderate symptoms or mental impairment) decrease by a factor of 0.73 for every unit increase in the life event score.

1c) 95% Likelihood Ratio confidence interval for β_1: [-0.572;-0.092] (from `lrci` option). exp([-0.572;-0.092]) = [0.564;0.912].
We are 95% confident that the odds of mental impairment below any level j decrease by a factor of at least 0.91 and at most 0.56 for every unit increase in the life event score. (For Wald interval, leave out option `lrci`.)

1d)

```
proc genmod data=mental;
model impair = life ses life*ses / dist=multinomial link=clogit lrci type3;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
life	1	-0.4204	0.1864	-0.8379	-0.0828	5.09	0.0241
ses	1	0.3709	1.1361	-1.8745	2.6318	0.11	0.7441
life*ses	1	0.1813	0.2383	-0.2761	0.6778	0.58	0.4468

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
life*ses	1	0.59	0.4411

Interpretation: The decrease in the estimated cumulative odds of mental impairment below any level is stronger for those adults with low socioeconomic status (ses=0) than for those with high socioeconomic status (ses=1). In particular, for adults with low SES, the estimated cumulative odds decrease by a factor of exp(-0.4202) = 0.66 for every unit increase in the life events score, compared to a decrease of exp(-0.4204 + 0.1813) = 0.79 for those of high SES.

1e) However, the coefficient for the interaction term (estimate = 0.1813, SE = 0.238) is not significant. Likelihood ratio test statistic = 0.59, df = 1: P-value = 0.441. The decrease in the estimated odds can be considered the same for both high and low SES.

Exercise 2: Auto Accidents Injuries:

SAS code (using proc genmod):

```
data accidents;
input gender$ location$ seatbelt$ y1-y5;
resp=1; count=y1; output;
resp=2; count=y2; output;
resp=3; count=y3; output;
resp=4; count=y4; output;
resp=5; count=y5; output;
drop y1-y5;
datalines;
```

```
female urban no 7287 175 720 91 10
female urban yes 11587 126 577 48 8
female rural no 3246 73 710 159 31
female rural yes 6134 94 564 82 17
male urban no 10381 136 566 96 14
male urban yes 10969 83 259 37 1
male rural no 6123 141 710 188 45
male rural yes 6693 74 353 74 12
;
proc genmod data=accidents;
class gender location seatbelt;
model resp = gender location seatbelt location*seatbelt/ dist=multinomial link=clogit lrci
type3;
freq count;
run;
```

Selected Output:

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1			1	3.3074	0.0351	3.2392	3.3767	8894.46	<.0001
Intercept2			1	3.4818	0.0355	3.4127	3.5520	9603.94	<.0001
Intercept3			1	5.3494	0.0470	5.2580	5.4420	12978.6	<.0001
Intercept4			1	7.2563	0.0914	7.0811	7.4398	6296.51	<.0001
gender	**female**		**1**	**-0.5463**	**0.0272**	**-0.5997**	**-0.4929**	401.92	<.0001
gender	male		0	0.0000	0.0000	0.0000	0.0000	.	.
location	rural		1	-0.6988	0.0424	-0.7819	-0.6159	272.15	<.0001
location	urban		0	0.0000	0.0000	0.0000	0.0000	.	.
seatbelt	no		1	-0.7602	0.0393	-0.8374	-0.6833	374.14	<.0001
seatbelt	yes		0	0.0000	0.0000	0.0000	0.0000	.	.
location*seatbelt	rural	no	1	-0.1244	0.0548	-0.2318	-0.0170	5.16	0.0232
location*seatbelt	rural	yes	0	0.0000	0.0000	0.0000	0.0000	.	.
location*seatbelt	urban	no	0	0.0000	0.0000	0.0000	0.0000	.	.
location*seatbelt	urban	yes	0	0.0000	0.0000	0.0000	0.0000	.	.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
gender	1	406.33	<.0001
location	1	763.96	<.0001
seatbelt	1	917.46	<.0001
location*seatbelt	1	5.15	0.0232

2a) The response variable is ordinal with 5 categories. Therefore, we model the odds of 4 cumulative probabilities, where each model has its own intercept parameter ($\alpha_1 < \alpha_2 < \alpha_3 < \alpha_4$).
For males in urban areas wearing seat belts, all dummy variables equal 0 and the estimated cumulative probabilities are exp(3.3074)/[1 + exp(3.3074)] = 0.965, exp(3.4818)/[1 + exp(3.4818)] = 0.970, exp(5.3494)/[1 + exp(5.3494)] = 0.995, exp(7.2563)/[1 + exp(7.2563)] = 0.9993, and 1.0. The corresponding response probabilities are 0.965, 0.005, 0.025, 0.004, and 0.0007.

2b) Cumulative log-odds for female drivers – Cumulative log-odds for male drivers = β_1 (for any location and seatbelt use).
For given location and seatbelt use, the estimated cumulative odds of the extent of injury being less than or equal to j for females are exp(-0.5463) = 0.58 times the ones for males. I.e., for given location and seatbelt use, the estimated cumulative odds of the extent of injury below any level are 0.58 times smaller for males than for females. For instance, if j = 2, the estimated cumulative odds of not being injured or being injured but not transported by emergency medical services are 0.58 times smaller for males than for females. Females, more so than males, tend to fall on the lower end of the response scale. The P-value for a likelihood ratio test for this parameter is less than 0.0001.
A 95% profile likelihood confidence interval for the difference in the estimated cumulative log-odds for females versus males is given by [-0.5997;-0.4929]. exp([-0.5997;-0.4929]) = [0.549; 0.611]. Hence, the estimated cumulative odds for females are at most 0.549 and at least 0.611 times smaller for females compared to males.

2c) For any gender and *rural* location: Cumulative log odds for those using seat belt – Cumulative log-odds for those not using seat belt = $-(\beta_3 + \beta_4)$. For those in rural locations, the estimated cumulative odds of the extent of the injury being less than or equal to j are $\exp\{-(-0.76022-0.1244)\}$ = 2.42 times larger for those using seatbelts than those not using a seat belt. Hence, the probability of an extent of injury less than or equal to j are larger for those using seat belts, i.e., they are more likely to fall on the low end of the response scale, where the severity of the injury is not so dramatic.
For any gender and *urban* location: estimated cumulative log odds for those using seat belt – estimated cumulative log-odds for those not using seat belt = $-\beta_3$. For those in urban locations, the estimated cumulative odds of the extent of the injury being less than or equal to j are $\exp\{-(-0.76022)\}$ = 2.14 times larger for those using a seatbelt than those not using a seat belt. Hence, the probability of an extent of injury less than or equal to j are larger for those using seat belts, i.e., they are more likely to fall on the low end of the response scale, where the severity of the injury is not so dramatic.

Note that the estimated cumulative odds ratio in urban locations is smaller by a factor of exp(-0.1244) = 0.88 compared to rural locations. I.e., the effect of seat belt use on the odds of the extent of the injury is more pronounced in rural compared to urban locations. This effect is statistically significant (likelihood ratio P-value = 0.0232) but in practice the two estimates of 2.42 for rural and 2.14 for urban locations are very similar.

Exercise 3: Happiness and Family Income:

SAS code (using proc genmod):

```
data happy;
input Income Happiness$ count;
datalines;
3 Very 272
3 Pretty 294
3 Not 49
2 Very 454
2 Pretty 835
2 Not 131
1 Very 185
1 Pretty 527
1 Not 208
;
proc genmod data=happy descending;
model Happiness = Income / dist=multinomial link=clogit lrci type3;
freq count;
run;
```

Response Profile

Ordered Value	Happiness	Total Frequency
1	Very	911
2	Pretty	1656
3	Not	388

PROC GENMOD is modeling the probabilities of levels of Happiness having LOWER Ordered Values in the response profile table.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	-2.0461	0.1122	-2.2672	-1.8275	332.80	<.0001
Intercept2	1	0.7613	0.1057	0.5546	0.9692	51.83	<.0001
Income	1	0.6311	0.0524	0.5287	0.7341	145.18	<.0001

3a) Treating income as quantitative with scores (1=below average, 2=average, 3=above average), the estimated coefficient equals 0.631. It is important to note the way SAS orders the ordinal levels of happiness in the response profile. By specifying the descending option, we requested the ordering that ranges from very happy to not happy. (By default, by the way we named the categories for happiness, the ordinal response would have been ordered from not happy to very happy.)
The estimated cumulative odds of happiness below any level are multiplied by exp(0.631)=1.88 for every unit increase in the family income. Since happiness is ordered from very happy to not happy, the probability of happiness below any level increases with increasing family income. That is, responses are more likely to fall at the low end (happy end) of the scale for increasing income.

3b)

```
proc genmod data=happy descending;
class Income;
model Happiness = Income / dist=multinomial link=clogit lrci type3;
freq count;
run;
```

Response Profile

Ordered Value	Happiness	Total Frequency
1	Very	911
2	Pretty	1656
3	Not	388

PROC GENMOD is modeling the probabilities of levels of Happiness having LOWER Ordered Values in the response profile table.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1		1	-0.2521	0.0790	-0.4072	-0.0972	10.17	0.0014
Intercept2		1	2.5643	0.0950	2.3793	2.7518	728.38	<.0001
Income	1	1	-1.2369	0.1055	-1.4445	-1.0307	137.37	<.0001
Income	2	1	-0.4501	0.0941	-0.6348	-0.2658	22.88	<.0001
Income	3	0	0.0000	0.0000	0.0000	0.0000	.	.

For families with below average income (Income=1), the estimated odds of being very or pretty happy (instead of not happy) are exp(2.5643 - 1.2369) = 3.77, while they are exp(2.5643 – 0.4501) = 8.28 for families with average income and exp(2.5643) = 12.99 for families with above average income. I.e., the odds of being very or pretty happy are exp(1.2369)=3.44 times higher for families with above average income than those with below average income.

For a comparison, the estimated odds for the model assuming a linear trend (on the logit scale) as in part a are exp(0.7613 + 0.6311) = 4.02 for below average income, exp(0.7613 + 2*0.6311) = 7.56 for average income and exp(0.7613 + 3*0.6311) = 14.22 for above average income, and the odds of being very or pretty happy are exp(2*0.6311)=3.53 times higher for families with above average income than those with below average income. These estimates are rather similar to those from the model treating income as a factor.

Formally, since the model in part a is a special case of the model in part b (namely, when beta1 - 2*beta2 + beta3 = 0, where beta1, beta2 and beta3 are the coefficients for Income in the model in part b) we can compare the models via a liklihood ratio test. I we add the SAS command

```
contrast "LR test" Income 1 -2 1 /E;
```

to the code above, we get

Coefficients for Contrast LR test

Label	Row	Intercept1	Intercept2	Prm1	Prm2	Prm3
LR test	1	0	0	1	-2	1

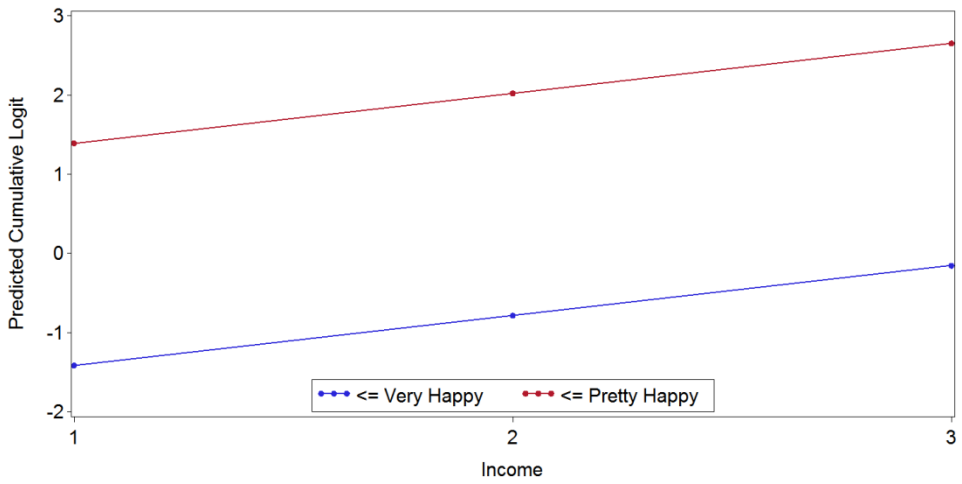
Contrast Results

Contrast	DF	Chi-Square	Pr > ChiSq	Type
LR test	1	5.34	0.0208	LR

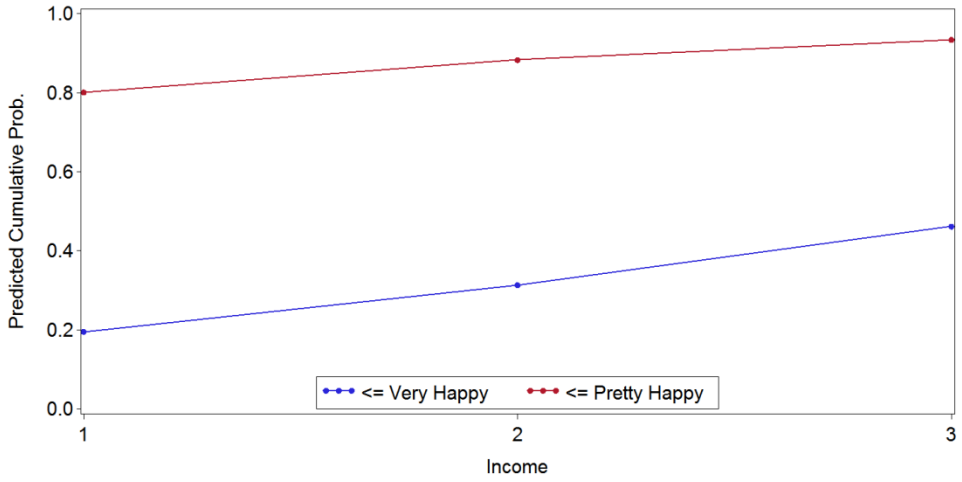
The likelihood ratio test statistic equals 5.34, yielding a P-value of 0.0208, showing that the model treating the effect of income as linear on the logit scale is not adequate.

3c) For plotting, it is easier to use proc logistic to create a dataset with the fitted cumulative and category probabilities:

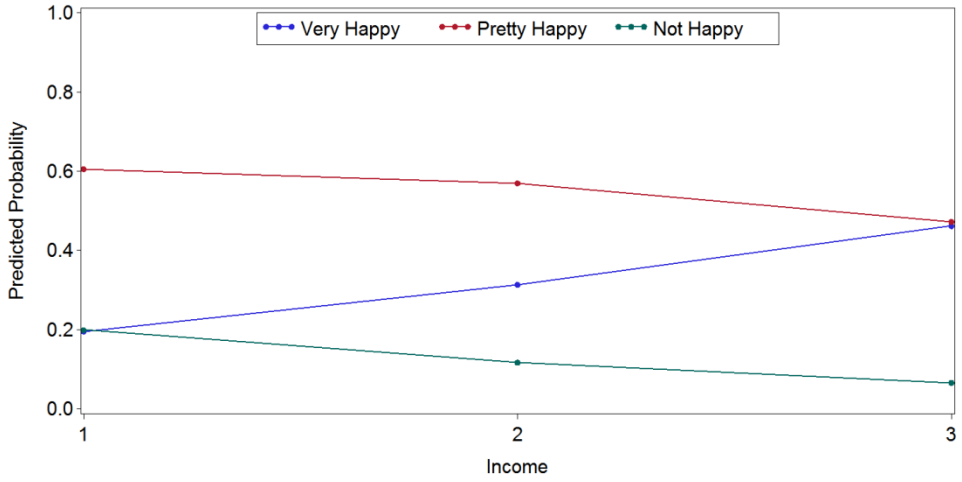
```
proc logistic data=happy descending;
model Happiness = Income / aggregate scale=none;
freq count;
output out=prediction PREDPROBS=I; *requests fitted category probabilities;
run;
/* setting graphing parameters */
goptions htext=2;
axis1 label=("Income") order = (1 to 3 by 1) minor = none;
axis2 label=(angle=90 "Predicted Cumulative Logit") order = (-2 to 3 by 1) minor=none;
legend1 label=none value=('<= Very Happy' '<= Pretty Happy')
       position=(bottom center inside) mode=share cborder=black;
symbol interpol=join value=dot width=2;
proc gplot data = probs;
plot (logit1 logit2)*Income /overlay haxis=axis1 vaxis=axis2 legend=legend1;
run;
quit;
```




```
/* fitted cumulative probabilities */
axis3 label=(angle=90 "Predicted Cumulative Prob.") order = (0 to 1 by .2) minor=none;
proc gplot data = probs;
plot (CP_Very CP_Pretty)*Income /overlay haxis=axis1 vaxis=axis3 legend=legend1;
run;
quit;
```




```
/* fitted category probabilities */
axis4 label=(angle=90 "Predicted Probability") order = (0 to 1 by .2) minor=none;
legend2 label=none value=('Very Happy' 'Pretty Happy' 'Not Happy')
       position=(top center inside) mode=share cborder=black;
proc gplot data = probs;
plot (IP_Very IP_Pretty IP_Not)*Income /overlay haxis=axis1 vaxis=axis4 legend=legend2;
run;
quit;
```



3d) The output of the proc logistic call above shows the score test for the proportional odds assumption:

```
proc logistic data=happy descending;
model Happiness = Income / aggregate scale=none;
freq count;
run;
```

Response Profile

Ordered Value	Happiness	Total Frequency
1	Very	911
2	Pretty	1656
3	Not	388

Probabilities modeled are cumulated over the lower Ordered Values.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
3.2595	1	0.0710

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept Very	1	-2.0461	0.1113	337.9827	<.0001
Intercept Pretty	1	0.7613	0.1049	52.6391	<.0001
Income	1	0.6311	0.0520	147.2925	<.0001

3e) Test goodness of fit for the proportional odds model when treating income as quantitative: At each of the 3 income levels, we have a multinomial response with 3 categories, hence, there are 3*(3-1) = 6 multinomial probabilities. The model specifies these in terms of 3 parameters. A goodness of fit test compares the fitted cell counts based on the model to the observed cell counts through a statistic such as the likelihood ratio statistic or the Pearson statistic and has df=6-3 = 3. From the proc logistic output:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	16.1867	3	5.3956	0.0010
Pearson	15.8487	3	5.2829	0.0012

3f) Fit adjacent category logit model. Here, each row shows the multinomial counts in the three categories.

```
/* Fit Adjacent Category Model with proportional odds */
data happy1;
input Income y1 y2 y3;
datalines;
3 272 294 49
2 454 835 131
1 185 527 208
;
proc nlmixed data=happy1;
eta1 = alpha1+alpha2+2*beta*Income;
eta2 = alpha2+beta*Income;
p3 = 1/(1+exp(eta1)+exp(eta2));
p1 = exp(eta1)*p3;
p2 = exp(eta2)*p3;
ll = y1*log(p1) + y2*log(p2) + y3*log(p3);
model y1 ~ general(ll);
run;
```

The NLMIXED Procedure

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha1	-1.6116	0.09813	3	-16.42	0.0005	0.05	-1.9239	-1.2993	1.527E-6
alpha2	0.5648	0.08866	3	6.37	0.0078	0.05	0.2827	0.8470	-6.99E-7
beta	0.5130	0.04331	3	11.84	0.0013	0.05	0.3751	0.6508	3.255E-6

Note: When there is a continuous predictor or when you have subject-level data, one option is to create a multivariate binary response for each subject. E.g., with three response categories

```
data subj;
input subj Income y1 y2 y3;
datalines;
1 1 1 0 0
2 1 0 1 0
3 1 0 0 1
4 3 0 1 0
...
;
```

would indicate that the first subject is in income category 1 and made response 'very happy', that the second subject is in income category 1 and made response 'pretty happy', that the third subject is in income category 1 and made response 'not happy' and that the fourth subject is in income category 3 and made response 'pretty happy'.

Interpretation of effect: The estimated odds of response very happy instead of pretty happy, and the estimated odds of response pretty happy instead of not too happy increase by a factor of exp(0.513) = 1.67 for every category increase in average family income.

One can get predicted category probabilities from the adjacent category model by including estimate statements in the nlmixed call:

```
/* Predicted category probabilities */
estimate "P(Y = very happy|Income=3)" exp(alpha1+alpha2+2*beta*3)/
(1+exp(alpha1+alpha2+2*beta*3)+exp(alpha2+beta*3));
estimate "P(Y = very happy|Income=2)" exp(alpha1+alpha2+2*beta*2)/
(1+exp(alpha1+alpha2+2*beta*2)+exp(alpha2+beta*2));
estimate "P(Y = very happy|Income=1)" exp(alpha1+alpha2+2*beta)/
(1+exp(alpha1+alpha2+2*beta)+exp(alpha2+beta));

estimate "P(Y = pretty happy|Income=3)" exp(alpha2+beta*3)/
(1+exp(alpha1+alpha2+2*beta*3)+exp(alpha2+beta*3));
estimate "P(Y = pretty happy|Income=2)" exp(alpha2+beta*2)/
(1+exp(alpha1+alpha2+2*beta*2)+exp(alpha2+beta*2));
estimate "P(Y = pretty happy|Income=1)" exp(alpha2+beta)/
(1+exp(alpha1+alpha2+2*beta)+exp(alpha2+beta));

estimate "P(Y = pretty happy|Income=3)" 1/
(1+exp(alpha1+alpha2+2*beta*3)+exp(alpha2+beta*3));
estimate "P(Y = not happy|Income=2)" 1/
(1+exp(alpha1+alpha2+2*beta*2)+exp(alpha2+beta*2));
estimate "P(Y = not happy|Income=1)" 1/
(1+exp(alpha1+alpha2+2*beta)+exp(alpha2+beta));
```

Additional Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower
P(Y = very happy|Income=3)	**0.4532**	0.01559	3	29.08	<.0001	0.05	0.4036
P(Y = very happy|Income=2)	**0.3162**	0.008785	3	36.00	<.0001	0.05	0.2883
P(Y = very happy|Income=1)	**0.1992**	0.01052	3	18.93	0.0003	0.05	0.1657
P(Y = pretty happy|Income=3)	**0.4874**	0.01218	3	40.00	<.0001	0.05	0.4486
P(Y = pretty happy|Income=2)	**0.5680**	0.009284	3	61.18	<.0001	0.05	0.5385
P(Y = pretty happy|Income=1)	**0.5975**	0.009713	3	61.51	<.0001	0.05	0.5666
P(Y = pretty happy|Income=3)	**0.05946**	0.005523	3	10.76	0.0017	0.05	0.04188
P(Y = not happy|Income=2)	**0.1157**	0.005932	3	19.51	0.0003	0.05	0.09686
P(Y = not happy|Income=1)	**0.2033**	0.01069	3	19.01	0.0003	0.05	0.1693

Exercise 4: Continuation-Ratio Model for Happiness and Family Income:

SAS code (using proc nlmixed):

```
/* Fit Continuation-Ratio Model with proportional odds */
proc nlmixed data=happy1;
eta1 = alpha1 + beta*Income;
eta2 = alpha2 + beta*Income;
p1 = exp(eta1)/(1+exp(eta1));
p2 = exp(eta2)/((1+exp(eta1))*(1+exp(eta2)));
p3 = 1-p1-p2;
ll = y1*log(p1) + y2*log(p2) + y3*log(p3);
model y1 ~ general(ll);
/* Predicted category probabilities */
predict p1 out=pred1;
run;
```

The NLMIXED Procedure

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper	Gradient
alpha1	-1.9332	0.1043	3	-18.54	0.0003	0.05	-2.2651	-1.6014	-2.5E-7
beta	0.5766	0.04794	3	12.03	0.0012	0.05	0.4240	0.7291	0.000204
alpha2	0.4583	0.09708	3	4.72	0.0180	0.05	0.1493	0.7673	0.000072

```
proc print data=pred1 label;
label Pred = 'P(Y = very happy | Income)';
var Income Pred;
run;
```

Obs	Income	P(Y = very happy | Income)
1	3	0.44928
2	2	0.31429
3	1	0.20478

Interpretation of effect: The estimated odds of response very happy instead of pretty or not happy and the estimated odds of response pretty happy instead of not happy increase by a factor of exp(0.5766) = 1.78 for every category increase in family income. (Note that this effect is in between the effect for the cumulative logit model, exp(0.631)=1.88 and the effect of the adjacent category model exp(0.513) = 1.67, as these models contrast different (cumulative) probabilities.

SAS code (using proc genmod):

To use proc genmod, we have to create separate 3x2 tables: Income x (very happy vs. pretty and not happy) and Income x (very and pretty happy vs. not happy). This is done using a stratum, as then proc genmod uses a different intercept for each stratum, but a common effect parameter for income:

```
/* Fit Continuation-Ratio Model with proportional odds */
data happy2;
input stratum Income successes failures;
n=successes+failures;
datalines;
1 3 272 343
1 2 454 966
1 1 185 735
2 3 294 49
2 2 835 131
2 1 527 208
;
proc genmod data=happy2;
class stratum;
model successes/n = stratum Income /noint dist=binomial link=logit;
run;
```

The GENMOD Procedure

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		0	0.0000	0.0000	0.0000	0.0000	.	.
stratum	1	1	-1.9332	0.1043	-2.1376	-1.7289	343.71	<.0001
stratum	2	1	0.4583	0.0971	0.2680	0.6486	22.28	<.0001
Income		1	0.5766	0.0479	0.4826	0.6705	144.63	<.0001

Exercise 5: Probit Model

SAS code (using proc genmod):

```
/* Fit Cumulative Probit Model */
proc genmod data=happy descending;
 model Happiness = Income / dist=multinomial link=cprobit lrci type3;
 freq count;
run;
```

Response Profile

Ordered Value	Happiness	Total Frequency
1	Very	911
2	Pretty	1656
3	Not	388

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	-1.2072	0.0632	-1.3314	-1.0835	364.69	<.0001
Intercept2	1	0.4678	0.0607	0.3489	0.5868	59.38	<.0001
Income	1	0.3637	0.0299	0.3051	0.4223	148.04	<.0001

5a) Interpretation of Effect: The estimate of 0.36 implies that the fitted regression model for the underlying latent variable on happiness (ranging from greater happiness to unhappiness) has slope -0.36: For every category increase in income, the mean of an underlying latent variable for happiness decreases (i.e., moves towards more happiness) by 0.36 standard deviations (of the latent normal distribution).

5b) The estimate for the cumulative logit model was equal to 0.6311 (see 3a above), hence the fitted regression model for the underlying (logistic) latent variable on happiness has slope -0.63: For every category increase in income, the mean of an underlying latent variable for happiness decreases (towards unhappiness) by 0.63 standard deviations (of the latent logistic distribution, which has standard deviation π/sqrt(3) = 1.81). Since the standard deviation of the standard normal cdf is 1/1.81=0.55 times the standard deviation of the standard logistic cdf, this corresponds to a decrease of 0.63*0.55 = 0.35 on the standard normal scale. It follows that the cumulative probit and logit models lead to very similar estimates of the effect of income on the latent variable. This can also be seen by comparing fitted category probabilities (see below).

5c) see R code

Solutions to Exercises using R

Exercise 1: Cumulative Logit model for mental impairment data:

R code (using package "VGAM"):
(This package uses same parameterization as SAS, i.e., linear predictor = $\alpha_j + \beta x$.)

```
> mental <- read.table("mental.dat", header=TRUE)
> head(mental)
  impair ses life
1      1   1    1
2      1   1    9
3      1   1    4
4      1   1    3
5      1   0    2
6      1   1    0
> require(VGAM)
> fit <- vglm(impair ~ life + ses, family=cumulative(parallel=TRUE), data=mental)
> summary(fit)

Selected Output:
Coefficients:
             Estimate Std. Error  z value
(Intercept):1 -0.28176    0.62304 -0.45223
(Intercept):2  1.21291    0.65119  1.86260
(Intercept):3  2.20947    0.71719  3.08075
life          -0.31888    0.11944 -2.66973
ses            1.11112    0.61427  1.80884
```

1b)

```
> maxl=logLik(fit)
> maxl
[1] -49.54895
> fit0 <- vglm(impair ~ ses, family=cumulative(parallel=TRUE), data=mental)
> maxl0 <- logLik(fit0)
> maxl0
[1] -53.43718
> LR.stat <- -2*(maxl0 - maxl)
> LR.stat
[1] 7.776457
> 1 - pchisq(LR.stat,df=1)
[1] 0.005293151
```

Likelihood ratio statistic: 7.78; P-value for Likelihood ratio test: 0.0053

1c) Cannot compute profile likelihood interval directly (package ordinal below can compute profile likelihood intervals). Wald interval from output above: -0.3188 +/- 1.96*0.119.

1d)

```
> fit1 <- vglm(impair ~ life + ses + life*ses, family=cumulative(parallel=TRUE),
data=mental)
> summary(fit1)
Coefficients:
              Estimate Std. Error  z value
(Intercept):1  0.098131    0.81107  0.12099
(Intercept):2  1.592521    0.83729  1.90199
(Intercept):3  2.606616    0.90980  2.86504
life          -0.420448    0.19034 -2.20893
ses            0.370876    1.13027  0.32813
life:ses       0.181294    0.23613  0.76777
```

Estimated coefficient for interaction effect: -0.1813

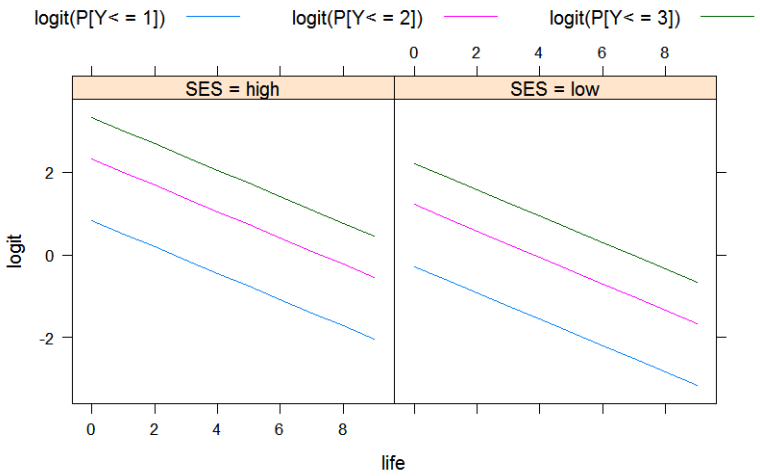
1e)

```
> fit1 <- vglm(impair ~ life + ses + life*ses, family=cumulative(parallel=TRUE),
data=mental)
> LR.stat <- -2*(maxl - logLik(fit1))
> LR.stat
[1] 0.5934586
> 1 - pchisq(LR.stat,df=1)
[1] 0.4410848
```

LR-stat = 0.594; P-value = 0.441

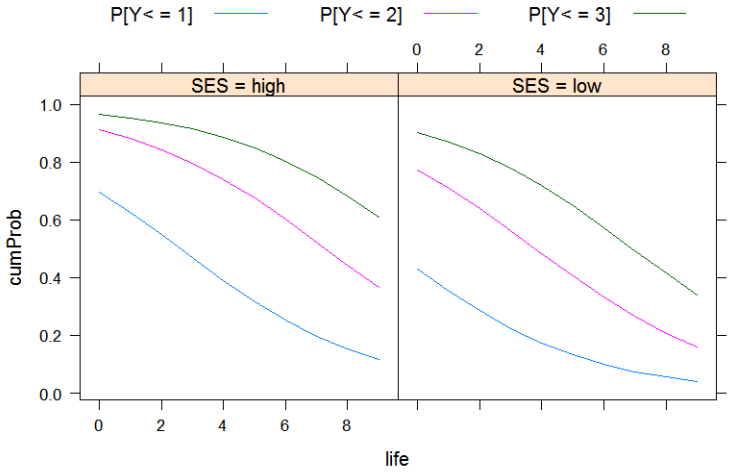
1f) Plot the estimated cumulative logits, cumulative probabilities and category probabilities against the life score for each SES category.

```
> ### fitted logits
> life1 <- seq(0,9,1)
> fit.logit.ses0 <- predict(fit,newdata=data.frame(life=life1,ses=0))
> fit.logit.ses1 <- predict(fit,newdata=data.frame(life=life1,ses=1))
> name <- colnames(fit.logit.ses1)
> plot.data <- data.frame(life=rep(life1,3),ses=rep(c("SES = low","SES = high"),each=3*10),
type=rep(name,each=10), logit=c(fit.logit.ses0,fit.logit.ses1))
> head(plot.data)
  life       ses             type      logit
1    0 SES = low logit(P[Y< = 1]) -0.2817575
2    1 SES = low logit(P[Y< = 1]) -0.6006407
3    2 SES = low logit(P[Y< = 1]) -0.9195239
4    3 SES = low logit(P[Y< = 1]) -1.2384071
5    4 SES = low logit(P[Y< = 1]) -1.5572904
6    5 SES = low logit(P[Y< = 1]) -1.8761736
> xyplot(logit~life|ses, group=type, data=plot.data, type="l", auto.key =
list(points=FALSE, lines=TRUE, columns=3))
```

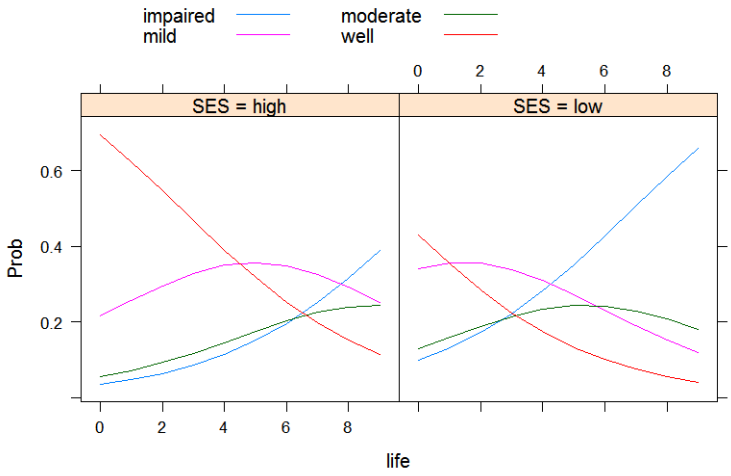



```
> ### fitted cumulative probabilities
> fit.cumprob.ses0 <- predict(fit,newdata=data.frame(life=life1,ses=0), untransform=TRUE)
> fit.cumprob.ses1 <- predict(fit,newdata=data.frame(life=life1,ses=1), untransform=TRUE)
```

```
> name <- colnames(fit.cumprob.ses1)
> plot.data <- data.frame(life=rep(life1,3),ses=rep(c("SES = low","SES = high"),each=3*10),
type=rep(name,each=10), cumProb=c(fit.cumprob.ses0,fit.cumprob.ses1))
> head(plot.data)
  life       ses      type   cumProb
1    0 SES = low P[Y< = 1] 0.4300230
2    1 SES = low P[Y< = 1] 0.3541971
3    2 SES = low P[Y< = 1] 0.2850549
4    3 SES = low P[Y< = 1] 0.2247134
5    4 SES = low P[Y< = 1] 0.1740358
6    5 SES = low P[Y< = 1] 0.1328290
> xyplot(cumProb~life|ses, group=type, data=plot.data, type="l", auto.key =
list(points=FALSE, lines=TRUE, columns=3))
```




```
> fit.prob.ses0 <- predict(fit,newdata=data.frame(life=life1,ses=0), type="response")
> fit.prob.ses1 <- predict(fit,newdata=data.frame(life=life1,ses=1), type="response")
> name <- colnames(fit.prob.ses1)
> plot.data <- data.frame(life=rep(life1,4),ses=rep(c("SES = low","SES = high"),each=4*10),
type=rep(name,each=10), Prob=c(fit.prob.ses0,fit.prob.ses1))
> head(plot.data)
  life       ses type      Prob
1    0 SES = low well 0.4300230
2    1 SES = low well 0.3541971
3    2 SES = low well 0.2850549
4    3 SES = low well 0.2247134
5    4 SES = low well 0.1740358
6    5 SES = low well 0.1328290
> xyplot(Prob~life|ses, group=type, data=plot.data, type="l", auto.key = list(points=FALSE,
lines=TRUE, columns=3))
```



R code (using package "ordinal"):

Attention, package "ordinal" uses the latent variable coding, i.e., linear predictor = $\alpha_j - \beta x$!

1a)

```
> mental <- read.table("mental.dat", header=TRUE)
> mental$impair <- factor(mental$impair, labels=c("well","mild","moderate","impaired"),
ordered=TRUE)
> head(mental)
  impair ses life
1   well   1    1
2   well   1    9
3   well   1    4
4   well   1    3
5   well   0    2
6   well   1    0
> require(ordinal)
> fit <- clm(impair ~ life + ses, data=mental)
> summary(fit)
 link  threshold nobs logLik AIC    niter max.grad cond.H
 logit flexible  40   -49.55 109.10 4(0)  3.17e-08 3.6e+02

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
life   0.3189     0.1210   2.635   0.0084 **
ses   -1.1112     0.6109  -1.819   0.0689 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Threshold coefficients:
                 Estimate Std. Error z value
well|mild         -0.2819     0.6423  -0.439
mild|moderate      1.2128     0.6607   1.836
moderate|impaired  2.2094     0.7210   3.064
```

1b)

```
> fit0 <- clm(impair ~ ses, data=mental) # model without life effect
> # or:
> # fit0 <- update(fit, ~ -life)
> anova(fit,fit0)
```

```
Likelihood ratio tests of cumulative link models:

     formula:            link: threshold:
fit0 impair ~ ses        logit flexible
fit  impair ~ life + ses logit flexible

     no.par    AIC  logLik LR.stat df Pr(>Chisq)
fit0      4 114.87 -53.437
fit       5 109.10 -49.549  7.7765  1   0.005293 **
---
```

Likelihood ratio statistic: 7.77; P-value for Likelihood ratio test: 0.0053.

1c)

```
> confint(fit)
           2.5 %     97.5 %
life  0.09203351 0.57184548
ses  -2.34711898 0.06410755
```

95% Profile Likelihood Confidence for β_1: [0.092;0.572] (Remember, β_1 here is $-\beta_1$ from SAS output)

1d)

```
> fit1 <- clm(impair ~ life + ses + ses*life, data=mental)
> summary(fit1)
 link  threshold nobs logLik AIC    niter max.grad cond.H
 logit flexible  40   -49.25 110.50 4(0)  2.30e-08 1.2e+03

Coefficients:
         Estimate Std. Error z value Pr(>|z|)
ses       -0.3709     1.1361  -0.326   0.7441
life       0.4204     0.1864   2.255   0.0241 *
ses:life  -0.1813     0.2383  -0.761   0.4468
```

Estimated coefficient for interaction effect: -0.1813

1e)

```
> anova(fit1,fit)
     no.par   AIC  logLik LR.stat df Pr(>Chisq)
fit       5 109.1 -49.549
fit1      6 110.5 -49.252  0.5935  1     0.4411
```

LR-stat = 0.594; P-value = 0.441

Exercise 2: Auto Accidents Injuries:

R code (using package "VGAM"):

```
> accident <- read.table("accident.dat", header=TRUE)
> accident
  gender location seatbelt    y1  y2  y3  y4 y5
1 female    urban       no  7287 175 720  91 10
2 female    urban      yes 11587 126 577  48  8
3 female    rural       no  3246  73 710 159 31
4 female    rural      yes  6134  94 564  82 17
5   male    urban       no 10381 136 566  96 14
6   male    urban      yes 10969  83 259  37  1
7   male    rural       no  6123 141 710 188 45
8   male    rural      yes  6693  74 353  74 12
> require(VGAM)
```

```
> fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ gender + location + seatbelt + location*seatbelt,
data=accident, family=cumulative(parallel=TRUE))
> summary(fit)
Coefficients:
                          Estimate Std. Error z value
(Intercept):1              1.17775   0.028486 41.3442
(Intercept):2              1.35217   0.028837 46.8895
(Intercept):3              3.21971   0.041260 78.0344
(Intercept):4              5.12666   0.088571 57.8820
gendermale                 0.54625   0.027211 20.0748
locationurban              0.82326   0.034756 23.6871
seatbeltyes                0.88457   0.038363 23.0577
locationurban:seatbeltyes -0.12442   0.054765 -2.2718
```

2a) R uses a different coding for the dummy variables than SAS. (By default, SAS treats the last category as the reference category, while R uses the first.) For males in urban areas wearing seat belts, the estimated cumulative logits and probabilities are:

logit[P(resp = not injured)] = 1.1777 + 0.5462 + 0.8233 + 0.8846 - 0.1244 = 3.3074
P(resp = not injured) = exp(3.3074)/[1+exp(3.3074)] = 0.965

logit[P(resp ≤ not transported)] = 1.3522 + 0.5462 + 0.8233 + 0.8846 - 0.1244 = 3.4818
P(resp ≤ not transported) = exp(3.4818)/[1+exp(3.4818)] = 0.970
...

2b) With the parameterization in R: Cumulative log-odds for female drivers – Cumulative log-odds for male drivers = $-\beta_1$ (for any location and seatbelt use), estimated as –0.546. Hence, the estimated cumulative log-odds ratio is equal to exp(-0.546) = 0.58. Cannot compute profile likelihood interval directly (package ordinal below can compute profile likelihood intervals). Wald interval for β_1 from output above: 0.5462 +/- 1.96* 0.0272 = [0.49;0.60]. Wald interval for exp($-\beta_1$) : exp{ -[0.49;0.60] } = [0.55;0.61].

2c) For any gender and *rural* location: Cumulative log odds for those using seat belt – Cumulative log-odds for those not using seat belt = β_3, estimated as 0.8846. Estimated cumulative odds ratio = exp(0.8846) = 2.42.

For any gender and *urban* location: Cumulative log odds for those using seat belt – Cumulative log-odds for those not using seat belt = $\beta_3+\beta_4$,estimated as 0.8846 + (–0.1244) = 0.7602. Estimated cumulative odds ratio = exp(0.7602) = 2.14.

R code (using package "ordinal"):
(Package "ordinal" uses the latent variable coding, i.e, linear predictor = $\alpha_j - \beta x$!)

```
> require("reshape2")
> accident.long <- melt(accident,1:3)
> colnames(accident.long)[4:5] <- c("resp","count")
> accident.long$resp = factor(accident.long$resp, labels=c("not injured","not
      transported","not hospitalized","hospitalized", "died"), ordered=TRUE)
> head(accident.long)  gender location seatbelt             resp count
1 female    urban       no      not injured  7287
2 female    urban       no  not transported   175
3 female    urban       no not hospitalized   720
4 female    urban       no     hospitalized    91
5 female    urban       no             died    10
6 female    urban      yes      not injured 11587
> require(ordinal)
> fit <- clm(resp ~ gender + location + seatbelt + location*seatbelt, weights= count,
data=accident.long)
```

```
> summary(fit)
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
gendermale                -0.54625    0.02725 -20.048   <2e-16 ***
locationurban             -0.82326    0.03483 -23.637   <2e-16 ***
seatbeltyes               -0.88457    0.03848 -22.985   <2e-16 ***
locationurban:seatbeltyes  0.12442    0.05479   2.271   0.0232 *
---
Threshold coefficients:
                                  Estimate Std. Error z value
not injured|not transported       1.17775    0.02865   41.10
not transported|not hospitalized  1.35217    0.02901   46.61
not hospitalized|hospitalized     3.21971    0.04137   77.82
hospitalized|died                 5.12666    0.08862   57.85
```

2a) R uses a different coding for the dummy variables than SAS: For males in urban areas wearing seat belts, the estimated cumulative logits and probabilities are:

logit[P(resp = not injured)] = 1.1777 – (-0.5462) – (-0.8233) – (-0.8846) – 0.1244 = 3.3074
P(resp = not injured) = exp(3.3074)/[1+exp(3.3074)] = 0.965

logit[P(resp ≤ not transported)] = 1.3522 – (-0.5462) – (-0.8233) – (-0.8846) – 0.1244 = 3.4818
P(resp ≤ not transported) = exp(3.4818)/[1+exp(3.4818)] = 0.970
…

The corresponding response probabilities are:

```
> cbind(accident1,fitted(fit))
   gender location seatbelt            resp count  fitted(fit)
1  female    urban       no      not injured  7287 0.8809034672
2  female    urban       no  not transported   175 0.0171185026
...
26   male    urban      yes      not injured 10969 0.9646825779
27   male    urban      yes  not transported    83 0.0054842229
28   male    urban      yes not hospitalized   259 0.0251045854
29   male    urban      yes     hospitalized    37 0.0040234217
30   male    urban      yes             died     1 0.0007051921
...
40   male    rural      yes             died    12 0.0014174344
```

2b) With the parameterization in R and the ordinal package: Cumulative log-odds for female drivers – Cumulative log-odds for male drivers = β_1 (for any location and seatbelt use), estimated as -0.546. Hence, the estimated cumulative log-odds ratio is equal to exp(-0.546) = 0.58.

2c) For any gender and *rural* location: Cumulative log odds for those using seat belt – Cumulative log-odds for those not using seat belt = $-\beta_3$ (remember, package "ordinal" uses $\alpha_j - \beta x$), estimated as –(–0.8846). Estimated cumulative odds ratio = exp(0.8846) = 2.42.

For any gender and *urban* location: Cumulative log odds for those using seat belt – Cumulative log-odds for those not using seat belt = $-\beta_3 - \beta_4$, estimated as –(–0.8846) –0.1244 = 0.7602. Estimated cumulative odds ratio = exp(0.7602) = 2.14.

Exercise 3: Happiness and Family Income:

R code (using package "VGAM"):

```
> happy <- read.table("happiness.dat", header=TRUE)
> happy
  income very pretty not
1      3  272    294  49
2      2  454    835 131
3      1  185    527 208
> ## VGAM Package
> require(VGAM)
> fit <- vglm(cbind(very,pretty,not) ~ income, data=happy, family =
cumulative(parallel=TRUE))
> summary(fit)
Coefficients:
             Estimate Std. Error z value
(Intercept):1 -2.04610   0.111294 -18.385
(Intercept):2  0.76130   0.104935   7.255
income         0.63107   0.051997  12.137

Residual deviance: 16.18668 on 3 degrees of freedom

Log-likelihood: -28.22221 on 3 degrees of freedom
```

3a) Treating income as quantitative with scores (1=below average, 2=average, 3=above average), the estimated coefficient for income equals 0.631.

3b)

```
> fit1 <- vglm(cbind(very,pretty,not) ~ factor(income), data=happy, family =
cumulative(parallel=TRUE))
> summary(fit1)

Coefficients:
                Estimate Std. Error  z value
(Intercept):1   -1.48905   0.073324 -20.3077
(Intercept):2    1.32736   0.071596  18.5396
factor(income)2  0.78681   0.085370   9.2164
factor(income)3  1.23694   0.104841  11.7982

Residual deviance: 10.84597 on 2 degrees of freedom

Log-likelihood: -25.55186 on 2 degrees of freedom
```

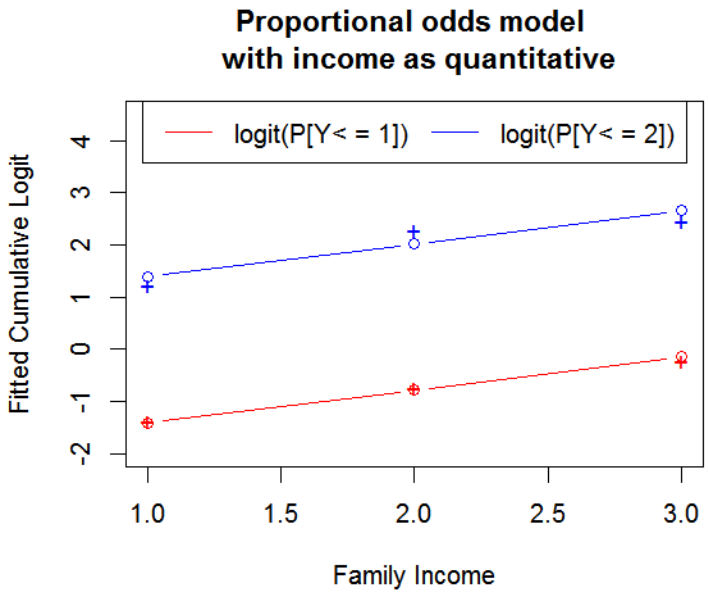
For families with below average income (income=1), the estimated odds of being very or pretty happy (instead of not happy) are exp(1.32736) = 3.77, while they are exp(1.32736 + 0.78681) = 8.28 for families with average income and exp(1.32736+1.23694) = 12.99 for families with above average income. I.e., the odds of being very or pretty happy are exp(1.23694)=3.44 times higher for families with above average income than those with below average income.

One can compare the two models (model in 3b is special case of model in 3a with beta1 - 2*beta2 + beta3 = 0) through a likelihood ratio test (= difference in the deviance):

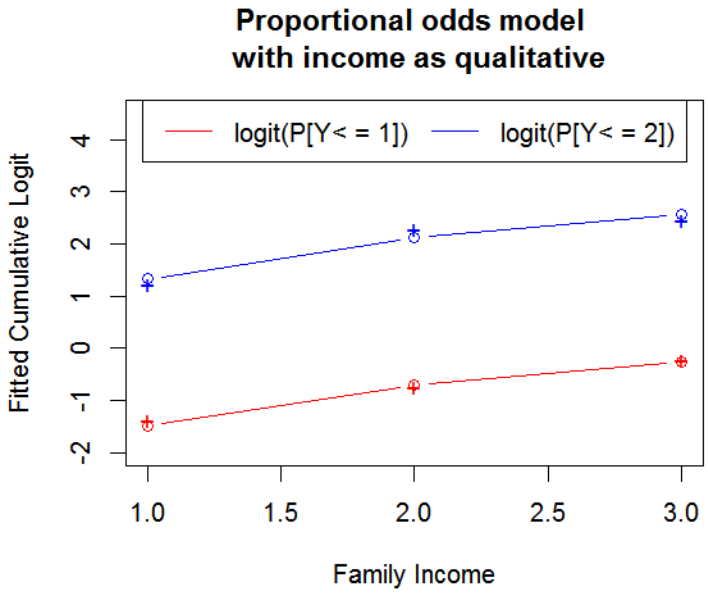
```
> LR <- -2*(logLik(fit)-logLik(fit1))
> LR
[1] 5.34071
> 1-pchisq(LR,df=1)
[1] 0.02083299
> deviance(fit1) - deviance(fit)
[1] -5.34071
```

3c) Plot the models in 3a and 3b on the logit scale. If possible, include the sample cumulative logits in your plot to check the fit of the model. Also plot the fitted cumulative and category probabilities for the model in part a.

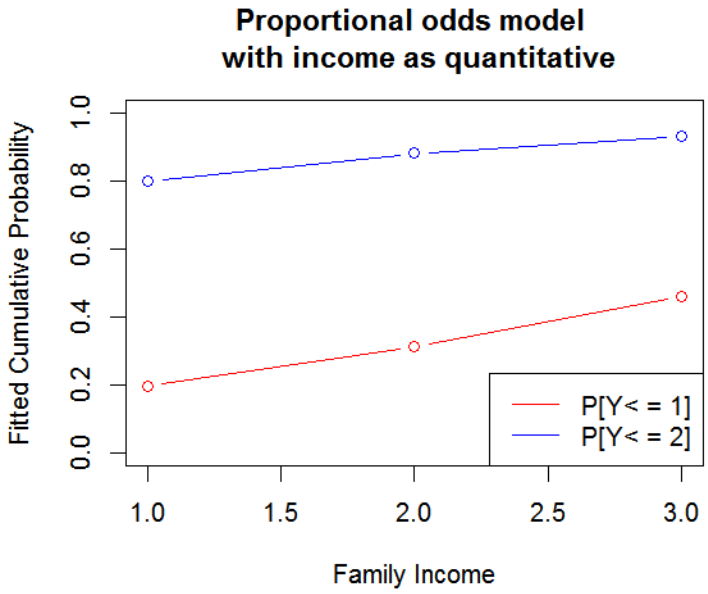
```
> ### fitted logits
> fit.logit <- predict(fit)
> name <- colnames(fit.logit)
> attach(happy)
> plot(fit.logit[,1]~income, type="b", col="red", ylim=c(-2,4.5), ylab="Fitted Cumulative
Logit", xlab="Family Income", main="Proportional odds model \n with income as
quantitative")
> lines(fit.logit[,2]~income, type="b", col="blue")
> ## add sample logits:
> n <- rowSums(happy[,2:4]) #total sample size
> sample.cumprob1 <- happy[,2]/n
> sample.cumprob2 <- rowSums(happy[,2:3])/n
> sample.logit1 <- logit(sample.cumprob1)
> sample.logit2 <- logit(sample.cumprob2)
> points(sample.logit1~income, pch="+", col="red")
> points(sample.logit2~income, pch="+", col="blue")
> legend("top", legend=name, lty=c(1,1), col=c("red","blue"), ncol=2, bty=n)
```




```
> fit1.logit <- predict(fit1)
> plot(fit1.logit[,1]~income, type="b", col="red", ylim=c(-2,4.5), ylab="Fitted Cumulative
Logit", xlab="Family Income", main="Proportional odds model \n with income as qualitative")
> lines(fit1.logit[,2]~income, type="b", col="blue")
> points(sample.logit1~income, pch="+", col="red")
> points(sample.logit2~income, pch="+", col="blue")
> legend("top", legend=name, lty=c(1,1), col=c("red","blue"), ncol=2, bty=n)
```

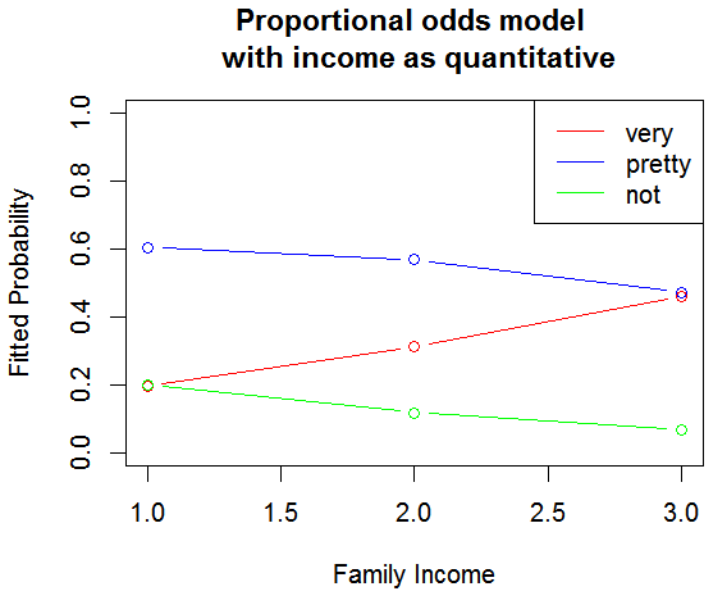



```
> ### fitted cumulative Probs
> fit.cumProb <- predict(fit, untransform=TRUE)
> name <- colnames(fit.cumProb)
> plot(fit.cumProb[,1]~income, type="b", col="red", ylim=c(0,1), ylab="Fitted Cumulative
Probability", xlab="Family Income", main="Proportional odds model \n with income as
quantitative")
> lines(fit.cumProb[,2]~income, type="b", col="blue")
> legend("bottomright", legend=name,lty=c(1,1),col=c("red","blue"), ncol=1, bty=n)
```




```
> ### fitted category Probs
> fit.prob <- predict(fit, type="response")
> name <- colnames(fit.prob)
> plot(fit.prob[,1]~income, type="b", col="red", ylim=c(0,1), ylab="Fitted Probability",
xlab="Family Income", main="Proportional odds model \n with income as quantitative")
> lines(fit.prob[,2]~income, type="b", col="blue")
```

```
> lines(fit.prob[,3]~income, type="b", col="green")
> legend("topright", legend=name,lty=c(1,1,1),col=c("red","blue","green"), ncol=1)
```

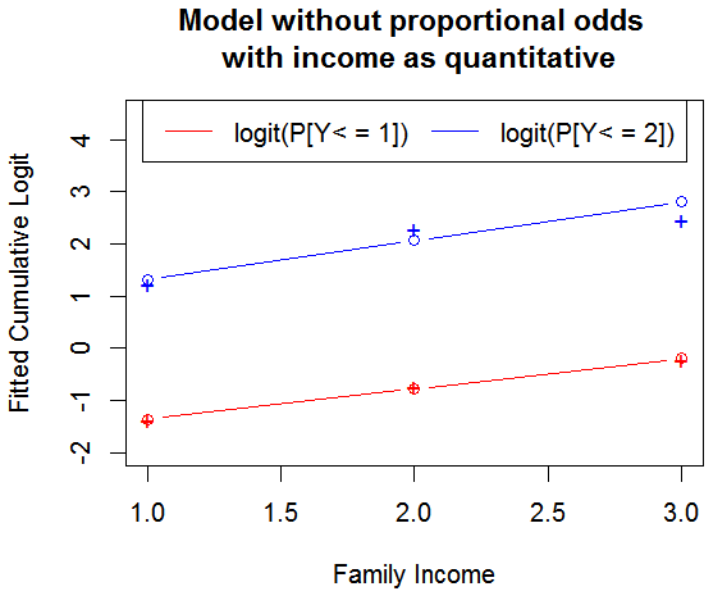


3d) Fit a model that allows non-proportional odds (treating income as quantitative) and plot it. Check if the proportional odds assumption is reasonable.

```
> fit2 <- vglm(cbind(very,pretty,not) ~ income, data=happy, family =
cumulative(parallel=FALSE))
> summary(fit2)

Coefficients:
              Estimate Std. Error  z value
(Intercept):1 -1.94760   0.122263 -15.9295
(Intercept):2  0.57208   0.147450   3.8798
income:1       0.58405   0.057518  10.1543
income:2       0.74798   0.083775   8.9285

> ### fitted logits
> fit2.logit <- predict(fit2)
> name <- colnames(fit2.logit)
> plot(fit2.logit[,1]~income, type="b", col="red", ylim=c(-2,4.5), ylab="Fitted Cumulative
Logit", xlab="Family Income", main="Model without proportional odds \n with income as
quantitative")
> lines(fit2.logit[,2]~income, type="b", col="blue")
> ## add sample logits:
> points(sample.logit1~income, pch="+", col="red")
> points(sample.logit2~income, pch="+", col="blue")
> legend("top", legend=name,lty=c(1,1),col=c("red","blue"), ncol=2, bty=n)
```



The lines appear are almost parallel, so the proportional odds assumption seems to be justified. To test the proportional odds assumption via a liklihood ratio test:

```
> LR <- -2*(logLik(fit) - logLik(fit2))
> LR
[1] 3.302767
> 1 - pchisq(LR,df=1)
[1] 0.0691633
```

3e) Test goodness of fit for the proportional odds model when treating income as quantitative:

```
> #Goodness of Fit
> obs <- happy[2:4]
> n <- rowSums(obs)
> exp <- apply(fitted(fit),2,function(col) col*n)
> X2 <- sum((obs-exp)^2/exp)
> G2 <- 2*sum(obs*log(obs/exp))
> X2
[1] 15.84874
> 1-pchisq(X2,df=6-3)
[1] 0.001217895
> G2
[1] 16.18668
> 1-pchisq(G2,df=6-3)
[1] 0.001038301
```

Note: The G2 statistic is the deviance and included in the output from summary(fit), see above.

3f) Adjacent Category Logit Model

```
> # by default, VGLM fits log(P[Y=j+1]/P[Y=j])
> # reverse=TRUE reverses this to log(P[Y=j]/P[Y=j+1])
> fit3 <- vglm(cbind(very,pretty,not) ~ income, data=happy, family = acat(reverse=TRUE,
parallel=TRUE))
> summary(fit3)
Coefficients:
             Estimate Std. Error  z value
(Intercept):1 -1.61160   0.098133 -16.4226
(Intercept):2  0.56484   0.088657   6.3711
income         0.51297   0.043311  11.8440

Names of linear predictors: log(P[Y = 1]/P[Y = 2]), log(P[Y = 2]/P[Y = 3])
```

Fitted category probabilities for cumulative logit model with proportional odds and adjacent category model are similar:

```
> cbind(income,fitted(fit))
  income      very    pretty        not
1      3 0.4618501 0.4724380 0.06571191
2      2 0.3134663 0.5697697 0.11676397
3      1 0.1954419 0.6055286 0.19902947

> cbind(income,fitted(fit3))
  income      very    pretty        not
1      3 0.4531901 0.4873543 0.05945556
2      2 0.3162396 0.5680182 0.11574214
3      1 0.1991606 0.5974905 0.20334891
```

Exercise 4: Continuation-Ratio Model for Happiness and Family Income:

R code (using package "VGAM"):

The continuation-ratio model is available via the family function 'family=sratio' in VGAM ('family=cratio' is also an option). For interpretation of effect see the SAS solutions.

```
> fit5 <- vglm(cbind(very,pretty,not) ~ income, data=happy, family=sratio(parallel=TRUE))
> summary(fit5)
Coefficients:
              Estimate Std. Error z value
(Intercept):1 -1.93324   0.104364 -18.524
(Intercept):2  0.45830   0.097179   4.716
income         0.57655   0.047988  12.015

Names of linear predictors: logit(P[Y = 1|Y> = 1]), logit(P[Y = 2|Y> = 2])

Residual deviance: 14.97402 on 3 degrees of freedom
```

Compare fitted category probabilities for continuation-ratio model and cumulative logit model:

```
> cbind(income,fitted(fit5))
  income      very    pretty        not
1      3 0.4492813 0.4951859 0.05553285
2      2 0.3142920 0.5716105 0.11409753
3      1 0.2047798 0.5867573 0.20846296
> cbind(income,fitted(fit))
  income      very    pretty        not
1      3 0.4618501 0.4724380 0.06571191
2      2 0.3134663 0.5697697 0.11676397
3      1 0.1954419 0.6055286 0.19902947
```

Exercise 5: Probit Model

R code (using package VGAM):

```
> fit.probit <- vglm(cbind(very,pretty,not) ~ income, data=happy,
      family=cumulative(link=probit, parallel=TRUE))
> summary(fit.probit)
Coefficients:
              Estimate Std. Error  z value
(Intercept):1 -1.20722   0.063263 -19.0826
(Intercept):2  0.46775   0.060703   7.7056
income         0.36365   0.029918  12.1552

Names of linear predictors: probit(P[Y< = 1]), probit(P[Y< = 2])

Residual deviance: 15.87322 on 3 degrees of freedom
```

5a) Estimated effect = 0.3636

5c) Plot the fitted cumulative probabilities for in terms of income for the logit and probit model.

```
> fit.cumProb.clogit <- predict(fit, untransform=TRUE)
> fit.cumProb.cprobit <- predict(fit.probit, untransform=TRUE)
> plot(fit.cumProb.clogit[,1]~income, type="b", lwd=2, col="blue", ylim=c(0,1),
      ylab="Fitted Cumulative Probability", xlab="Family Income", main="Comparison of
      cumulative \n logit and probit model")
> lines(fit.cumProb.clogit[,2]~income, type="b", lwd=2, col="red")
> lines(fit.cumProb.cprobit[,1]~income, type="b", lwd=2, col="blue")
> lines(fit.cumProb.cprobit[,2]~income, type="b", lwd=2, col="blue")
> legend("bottomright", legend=c("cum. logit","cum. probit"),lty=c(1,1),lwd=c(2,2),
      col=c("red","blue"), ncol=1, bty=n)
> ## add sample logits:
> n <- rowSums(happy[,2:4]) #total sample size
> sample.cumprob1 <- happy[,2]/n
> sample.cumprob2 <- rowSums(happy[,2:3])/n
> sample.logit1 <- logit(sample.cumprob1)
> sample.logit2 <- logit(sample.cumprob2)
> points(sample.logit1~income, pch="+", col="red")
> points(sample.logit2~income, pch="+", col="red")
```