Outline

Florida 2000 Presidential Election

- For each of the 67 counties in Florida:
  Total votes for Bush, Gore, Nader, and Buchanan

```
> fl = read.csv("florida.dat"); head(fl)
    Bush   Gore Nader Buchanan
1  34062  47300  3215      262
2   5610   2392    53       73
3  38637  18850   828      248
4   5413   3072    84       65
5 115185  97318  4470      570
6 177279 386518  7099      789
```

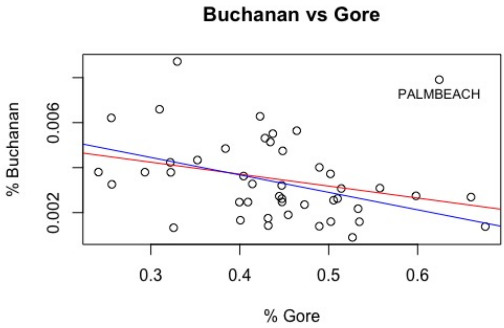
- Proportion of voters who support Buchanan in each county.

```
> flp = fl/apply(fl, 1, sum); round(head(flp),4)
    Bush   Gore  Nader Buchanan
1 0.4015 0.5575 0.0379   0.0031
2 0.6902 0.2943 0.0065   0.0090
3 0.6598 0.3219 0.0141   0.0042
4 0.6269 0.3558 0.0097   0.0075
5 0.5295 0.4474 0.0205   0.0026
6 0.3101 0.6761 0.0124   0.0014
> round(summary(flp[,4]), 4)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0009  0.0026  0.0040  0.0047  0.0056  0.0174
```

Florida 2000 Presidential Election

```
> fl2 = read.table("florida2000.txt",header=T);  # 67 counties
> head(fl2)
    County   Gore   Bush Buchanan Nader Total_Votes Reg_Reform Reg_Rep Reg_Ind Reg_Grn
1  ALACHUA  47300  34062      262  3215       84839         91   34319    1639     198
2    BAKER   2392   5610       73    53        8128          4    1684      58       0
3      BAY  18850  38637      248   828       58563         55   34286       0       8
4 BRADFORD   3072   5413       65    84        8634          3    2832      96       4
5  BREVARD  97318 115185      570  4470      217543        148  131427    6815      98
6  BROWARD 386518 177279      789  7099      571685        332  266829     125     179
  Reg_Dem Total_Reg
1   64135    120876
2   10261     12352
3   44209     92749
4    9639     13547
5  107840    283680
6  456789    887764
```

Florida 2000 Presidential Election



- Obs 50 is Palm Beach County with 656,694 registered voters in the 2000 election. Palm Beach is where a butterfly ballot was used.

Florida 2000 Presidential Election

- Objective: predict the Palm Beach vote from a general linear model in which Palm Beach is omitted from the fit.
- Let N_i be the total number of votes cast and Y_i be the number of votes for Buchanan in county i.
- Let π_i be the proportion of voters who support Buchanan in county i.
- Then the Binomial distribution would suggest that the variance of Y_i is approximately $N_i\pi_i(1-\pi_i)$.
- Quasi-likelihood approach to overdispersion based on models of the form

$$Var(Y_i) = \sigma^2 N_i\pi_i(1-\pi_i)$$

- For π_i, we assume that

$$\pi_i = \frac{1}{1+\exp(-\boldsymbol{X}_i'\beta)}$$

where $\boldsymbol{X}_i = (1, x_{i1}, x_{i2})'$ are predictor variables: the proportion voting for Bush and Nader.

Florida 2000 Presidential Election

- Palm Beach (case 50) is excluded for the purpose of model fitting and prediction.
- Consider only counties with more than 10,000 actual voters in the 2000 presidential election.

```
> index = apply(fl, 1, sum) > 10000;
> index2 = index;
> index2[50] = FALSE;
> yi = fl[index2,4]
> xi1 = flp[index2,1]; xi2 = flp[index2,3]
> Ni = apply(fl, 1, sum);
> ni = Ni[index2]-yi;
> glmfit.no50 = glm(cbind(yi, ni)~xi1+xi2, family=quasibinomial("logit"));
> summary(glmfit.no50)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.9313     0.4317 -18.371  < 2e-16 ***
xi1           2.8367     0.6845   4.144 0.000167 ***
xi2          26.9536    10.3612   2.601 0.012858 *

(Dispersion parameter for quasibinomial family taken to be 50.94362)

Null deviance: 3115.9  on 43  degrees of freedom
Residual deviance: 1958.0  on 41  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Florida 2000 Presidential Election

- Using the resulting model to predict π_{50} leads to a point estimate 0.00139063, and a 95% confidence interval (0.001044842,0.001850643).
- With $N_{50} = 430762$, this leads to a point estimate 599 and 95% confidence interval (450,797) for the mean vote $N_{50}\pi_{50}$.
- The interval is described as confidence interval rather than prediction interval, because it does not take account of the variability of Y_{50} given π_{50}.
- Since quasi-likelihood models do not specify the full distribution of Y_{50}, a precise resoluation is hard to achieve.

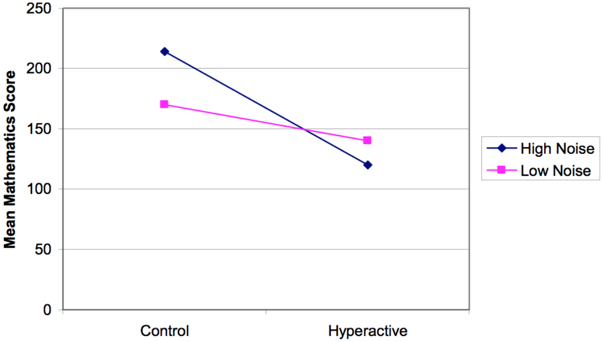
Outline

Interaction

- Interaction is a three-variable concept. One of these is the response variable (Y) and the other two are explanatory variables (X_1 and X_2).
- There is an interaction between X_1 and X_2 if the impact of an increase in X_2 on Y depends on the level of X_1.
- To incorporate interaction in multiple regression model, we add the explanatory variable $X_1 X_2$ (or $(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)$). There is evidence of an interaction if the coefficient on $X_1 X_2$ is significant (p-value < 0.05).

Example

- An experiment to study how noise affects the performance of children.
- Study sample consists of hyperactive children and a control group of children who were not hyperactive.
- The children solved problems under both high-noise and low-noise conditions.
- Here are the mean scores:



Interactions between categorical predictors

- Let Y=mathematics score, X_1 = type of child, X_2 = type of noise.
- There is an interaction between type of child and type of noise.

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i} + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

$$X_{1,i} = \begin{cases} 1 & \text{if } i \text{ is a hyperactive child} \\ 0 & \text{if } i \text{ is a control child} \end{cases} \quad X_{2,i} = \begin{cases} 1 & \text{if } i \text{ is in high noise} \\ 0 & \text{if } i \text{ is in low noise} \end{cases}$$

$$X_{1,i} X_{2,i} = \begin{cases} 1 & \text{if } i \text{ is a hyperactive child in high noise} \\ 0 & \text{otherwise} \end{cases}$$

What is the interpretation of $\beta_0, \beta_1, \beta_2, \beta_3$?

$$\begin{aligned} \beta_3 = \quad & \mathbb{E}(Y|\text{hyperactive, high noise}) + \mathbb{E}(Y|\text{control, low noise}) \\ & - \mathbb{E}(Y|\text{hyperactive, low noise}) - \mathbb{E}(Y|\text{control, high noise}) \end{aligned}$$

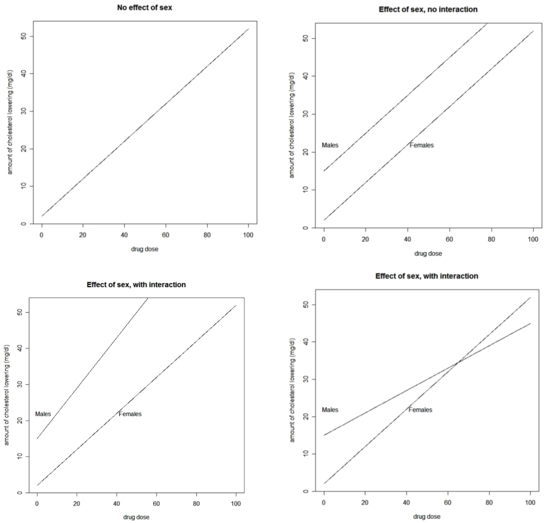
Interaction between categorical and continuous predictors

- An interaction occurs when an exploratory variable has a different effect on the outcome depending on the values of another exploratory variable.
- Let Y represents the output (amount of cholesterol lowering), β_1 represents the effect of the drug, β_2 the effect of gender, and β_3 the interaction effect

$$\mathbb{E}(Y) = \beta_0 + \beta_1 \text{Dose} + \beta_2 \mathbf{1}_{female} + \beta_3 \text{Dose} * \mathbf{1}_{\text{female}}$$

- How to depict each of the following scenario?
  1. No gender effect
  2. Gender has an main effect, but no interaction
  3. Gender has an effect with interaction.

Cont.


No effect of sex
amount of cholesterol lowering (mg/dl)
drug dose
Effect of sex, no interaction
Males
Females
amount of cholesterol lowering (mg/dl)
drug dose
Effect of sex, with interaction
Males
Females
amount of cholesterol lowering (mg/dl)
drug dose
Effect of sex, with interaction
Males
Females
amount of cholesterol lowering (mg/dl)
drug dose

Interactions between two continuous predictors

- The number of car accidents on a highway (Y) is related to the number of vehicles that travel over it (X_1) and the speed at which they are traveling (X_2).
- Data covering the last few years were provided.
- Possible model with interactions

$$Y_t \sim Poi(\lambda_t) \quad \text{independently},$$
$$\log(\lambda_t) = \beta_0 + \beta_1 \text{cars}_t + \beta_2 \text{speed}_t + \beta_3 (\text{cars}_t - 60) * (\text{speed}_t - 9.9)$$

Parameter	Estimate	Std Error	Z-ratio	Prob
Intercept	-0.85	7.31	-0.12	0.90
Cars	0.41	0.13	3.05	0.003
Speed	0.06	0.11	0.54	0.58
(Speed-60) * (Cars-9.9)	1.07	0.07	12.26	< 0.0001

- Increases in speed have a worse impact on the number of accidents when there are a large number of cars than when there are a small number of cars on the road.
- What is $\mathbb{E}$(Cars = 8, Speed 66) $-$ $\mathbb{E}$(Cars = 8, Speed 65)?
- How about $\mathbb{E}$(Cars = 11, Speed 66) $-$ $\mathbb{E}$(Cars = 11, Speed 65)?

Trick that sometime helps

- Subtract the mean from each independent variable, and use these so-called "centered" variables to create the interaction variables.
- This will not change the correlations among the non-interaction terms, but may reduce correlations for interaction terms.
- We have looked only at "first order" interactions, and only at interactions between two variables at a time. However, second order interactions, or interactions between three or more variables are also possible. E.g.

$$\mathbb{E}(Y) = \beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_1 X_2+\beta_5 X_1 X_3+\beta_6 X_1 X_2 X_3$$

- ANOVA analysis similarly extends to two-way (or higher-way) interactions.

Revisit the CEB dataset

- Number of Children Ever Born (CEB) to Women of Indian Race By Marital Duration, Type of Place of Residence and Educational Level
- Each cell shows the mean, variance, and sample size.

Marr.	Suva				Urban				Rural			
Dur.	N	LP	UP	S+	N	LP	UP	S+	N	LP	UP	S+
0–4	0.50	1.14	0.90	0.73	1.17	0.85	1.05	0.69	0.97	0.96	0.97	0.74
	1.14	0.73	0.67	0.48	1.06	1.59	0.73	0.54	0.88	0.81	0.80	0.59
	8	21	42	51	12	27	39	51	62	102	107	47
5–9	3.10	2.67	2.04	1.73	4.54	2.65	2.68	2.29	2.44	2.71	2.47	2.24
	1.66	0.99	1.87	0.68	3.44	1.51	0.97	0.81	1.93	1.36	1.30	1.19
	10	30	24	22	13	37	44	21	70	117	81	21
10–14	4.08	3.67	2.90	2.00	4.17	3.33	3.62	3.33	4.14	4.14	3.94	3.33
	1.72	2.31	1.57	1.82	2.97	2.99	1.96	1.52	3.52	3.31	3.28	2.50
	12	27	20	12	18	43	29	15	88	132	50	9
15–19	4.21	4.94	3.15	2.75	4.70	5.36	4.60	3.80	5.06	5.59	4.50	2.00
	2.03	1.46	0.81	0.92	7.40	2.97	3.83	0.70	4.91	3.23	3.29	–
	14	31	13	4	23	42	20	5	114	86	30	1
20–24	5.62	5.06	3.92	2.60	5.36	5.88	5.00	5.33	6.46	6.34	5.74	2.50
	4.15	4.64	4.08	4.30	7.19	4.44	4.33	0.33	8.20	5.72	5.20	0.50
	21	18	12	5	22	25	13	3	117	68	23	2
25–29	6.60	6.74	5.38	2.00	6.52	7.51	7.54	–	7.48	7.81	5.80	–
	12.40	11.66	4.27	–	11.45	10.53	12.60	–	11.34	7.57	7.07	–
	47	27	8	1	46	45	13	–	195	59	10	–

Additive model

TABLE 4.4: Estimates for Additive Log-Linear Model of Children Ever Born by Marital Duration, Type of Place of Residence and Educational Level

Parameter		Estimate	Std. Error	z-ratio
Constant		−0.1173	0.0549	−2.14
Duration	0–4	–		
	5–9	0.9977	0.0528	18.91
	10–14	1.3705	0.0511	26.83
	15–19	1.6142	0.0512	31.52
	20–24	1.7855	0.0512	34.86
	25–29	1.9768	0.0500	39.50
Residence	Suva	–		
	Urban	0.1123	0.0325	3.46
	Rural	0.1512	0.0283	5.34
Education	None	–		
	Lower	0.0231	0.0227	1.02
	Upper	-0.1017	0.0310	-3.28
	Sec+	-0.3096	0.0552	-5.61

ANOVA model

TABLE 4.3: Deviances for Poisson Log-linear Models Fitted to the Data on CEB by Marriage Duration, Residence and Education

Model	Deviance	d.f.
Null	3731.52	69
One-factor Models		
Duration	165.84	64
Residence	3659.23	67
Education	2661.00	66
Two-factor Models		
$D+R$	120.68	62
$D+E$	100.01	61
DR	108.84	52
DE	84.46	46
Three-factor Models		
$D+R+E$	70.65	59
$D+RE$	59.89	53
$E+DR$	57.06	49
$R+DE$	54.91	44
$DR+RE$	44.27	43
$DE+RE$	44.60	38
$DR+DE$	42.72	34
$DR+DE+RE$	30.95	28

Interpretation

- Null model has a deviance of 3732 on 69 degrees of freedom, which does not pass the goodness-of-test. $\Rightarrow$ reject the hypothesis that "the expect number of children is the same for all these groups".
- Introducing marital duration leads to substantial reduction of 3566 at only 5 d.f. $\Rightarrow$ significant effect of "duration" on the number of children
- The additive model D+R+E has a deviance of 70.65 on 59 d.f. The associated P-value under a χ^2 distribution is 0.14, so the model provides a good description of the data.
- Education effect:
  - compare model E to model Null (1071 on 3 d.f.)
  - compare model D+E to model D (65.8 on 3 d.f.)
  - compare model D+R+E to model D+R. (50.1 on 3 d.f.)
  - part of education effect may be attributed to the fact that more educated women tend to live in Suva or in other urban areas (collinearity between E and R).

Interaction effect

- Does education make more of a difference in rural areas than in urban areas?
- Compare $D + R + E$ to $D + RE$. $\Rightarrow$ reduces the deviance by 10.8 at the expense of 6 d.f. $\Rightarrow$ not significant, with a P-value of 0.096.
- Does education effect increase with marital duration?
- Compare $D + R + E$ to $R + DE$ $\Rightarrow$ reduces the deviance by 15.7 at the expense of 15 d.f. $\Rightarrow$ hardly a bargain.