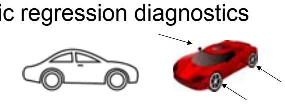
Model fit

HL Chapter 5 – part 2

Questions

- Does the model fit overall?
  - ➔ Summary goodness-of-fit tests √

- Are there any individual observations that don't fit?
  - ➔ Logistic regression diagnostics



Logistic regression diagnostics

Leverage
- How different from the other covariate pattern is this covariate pattern?

Change in Pearson chi-square or Deviance
- How much do Pearson X^2 and Deviance test statistics decrease if this covariate pattern is deleted
- I.e., is there any evidence of improved model fit if this covariate pattern is deleted?

Change in coefficients
- How much does deleting this covariate pattern affect the model coefficients?

Leverage

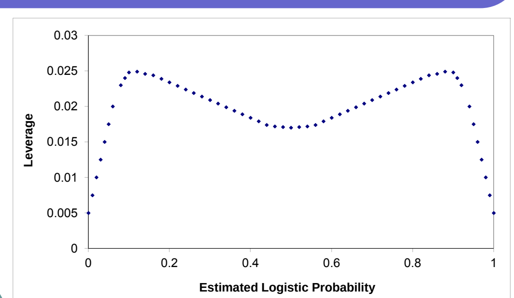
h

- How different from the other covariate patterns is this covariate pattern?

Leverage depends on $\hat{\pi}$

- 0.1< $\hat{\pi}$ <0.9: The greater h the more "unusual" the covariate pattern
- $\hat{\pi}$ ≤0.1 or $\hat{\pi}$ ≥0.9: h will be small even if the covariate pattern is "unusual"
- Look for deviations from the expected pattern (observations outside the "cloud")

Leverage – expected pattern (sort of)



Change in Pearson chi-square or Deviance

ΔX^2 or ΔD

- How much smaller are the Pearson X^2 and Deviance goodness-of-fit test statistics if this covariate pattern is deleted?
- I.e., does model fit improve if the covariate pattern is deleted?

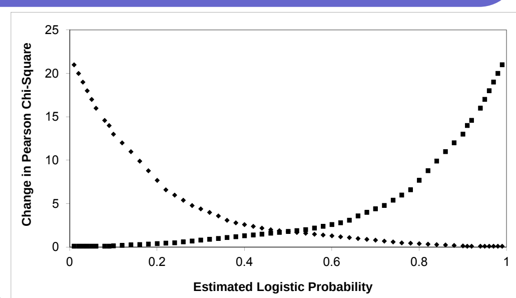
ΔX^2 and ΔD depend on $\hat{\pi}$

- Excellent fit most likely for
  - $\hat{\pi}$ <0.1 ($\hat{\pi}$ very close to y=0)
  - $\hat{\pi}$ >0.9 ($\hat{\pi}$ very close to y=1)
- Terrible fit most likely for
  - $\hat{\pi}$ <0.1 ($\hat{\pi}$ very different from y=1)
  - $\hat{\pi}$ >0.9 ($\hat{\pi}$ very different from y=0)
- Expect smallest and greatest ΔX^2 and ΔD for $\hat{\pi}$ <0.1 and $\hat{\pi}$ >0.9

ΔX^2 and ΔD depend on $\hat{\pi}$

- 0.3< $\hat{\pi}$ <0.7: ΔX^2 & ΔD likely to be fairly small
- Look for deviations from the expected pattern (observations outside the "cloud"), e.g.,
  - High ΔX^2 & ΔD for $\hat{\pi}$ > 0.3 or $\hat{\pi}$ <0.7
  - Very high ΔX^2 & ΔD for $\hat{\pi}$ < 0.1 or for $\hat{\pi}$ > 0.9

ΔX^2 and ΔD - Expected pattern (sort of)



Change in coefficients

$\Delta\hat{\beta}$

- How much does deleting this covariate pattern affect the model coefficients?
- $\Delta\hat{\beta}$ combines leverage and ΔX^2

$\Delta\hat{\beta}$ depends on $\hat{\pi}$

- Expect greatest $\Delta\hat{\beta}$ values when neither leverage nor ΔX^2 are very small
- I.e., $\Delta\hat{\beta}$ is expected to be highest for $\hat{\pi}$ between 0.1 and 0.3 and for $\hat{\pi}$ between 0.7 and 0.9
- Look for deviations from the expected pattern (observations outside the "cloud")

Change in specific coefficients

Note:

- $\Delta\hat{\beta}$ assesses the effect on all coefficients
- If $\Delta\hat{\beta}$ is large, it may be worth checking which coefficients are most affected by the deletion of the covariate pattern

If the model doesn't fit …

- Try rebuilding the model
  - Continuous covariates may have been modeled in the wrong scale
  - Important risk factors, confounders or effect modifiers may have been missed
- Standard logistic regression model may not work for small or large $\hat{\pi}$
  - Try model with the extra parameters that allow for the tails to vary (from Stukel test)

If the model doesn't fit …

- Logistic regression model may not work
  - Try a different regression model
- If nothing works, one or more crucial covariates may not have been measured
  - ☹
- However, lack of fit does not necessarily mean that the OR estimates are way off or that the model won't predict well

Example – GLOW500 data set

Create covariate patterns

```
proc sort data=glow500;
   by priorfrac momfrac armassist raterisk2 height age;
run;

proc means n sum noprint data=glow500;
     by priorfrac momfrac armassist raterisk2 height age;
     var fracture; output out=jdat n=m_j sum=y_j;
run;
```

Number of observations in covariate pattern j

Number of observations with outcome in covariate pattern j

Run proc logistic using m_ j and y_ j
Create graphs, save diagnostics

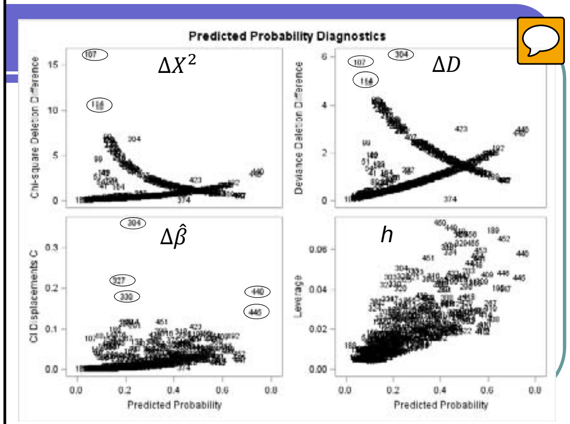
Plot h, ΔX^2, ΔD and $\Delta\hat{\beta}$vs. pihat

```
proc logistic descending data=jdat plots(only label)=(phat);

model y_j/m_j =priorfrac momfrac armassist raterisk2 height
             age priorfrac*age momfrac*armassist;

output out=diag
       h=h difchisq=difchisq difdev=difdev c=db p=pihat;
run;
```

Save h, ΔX^2, ΔD and $\Delta\hat{\beta}$



Evaluate outliers

```
data diag; set diag;j=_N_;run;

proc print data=diag noobs;

where j in (107,304,327,330,440,445);

var j priorfrac momfrac armassist raterisk2 height age
      y_j m_j pihat h difchisq difdev db;
run;
```

Add covariate pattern numbers to data set

Outlier table

j	PRIORFRAC	MOMFRAC	ARMASSIST	RATERISK2	HEIGHT	AGE	y_j	m_j	pihat	h	difchisq	difdev	db
107	0	0	0	0	166	57	1	1	0.059	0.005	16.099	5.744	0.075
304	0	1	0	0	168	65	2	2	0.238	0.051	6.738	6.080	0.359
327	0	1	1	0	175	75	1	1	0.175	0.042	4.934	3.698	0.217
330	0	1	1	1	162	60	1	1	0.208	0.043	3.973	3.308	0.177
440	1	1	0	0	142	70	0	1	0.747	0.057	3.131	2.928	0.191
445	1	1	0	1	153	63	0	1	0.735	0.046	2.913	2.793	0.139

Why is 107 an outliers?

- Low risk profile
- Therefore, low pihat
- But FRACTURE=1 (y_j and m_j are 1)
- Covariate pattern may be unusual and not represented well by the model
- None of the covariate values are unreasonable

Why is 304 an outlier?

- Fairly low risk profile
- Therefore, lowish pihat
- But both observations in the covariate pattern have FRACTURE=1 (y_j and m_j are 2)
- Covariate pattern may be unusual and not represented well by the model
- None of the covariate values are unreasonable

Why are 327 and 330 outliers?

- PRIORFRAC seems to greatly influence pihat
- Highish risk profile but PRIORFRAC=0
- Therefore, lowish pihat
- But both covariate patterns have FRACTURE=1 (y_j and m_j are 1 in both)
- Covariate pattern may be unusual and not represented well by the model
- None of the covariate values are unreasonable

Why are 440 and 445 outliers?

- PRIORFRAC seems to greatly influence pihat
- High(ish) risk profile and PRIORFRAC=1
- Therefore, highish pihat
- But both covariate patterns have FRACTURE=0 (y_j = 0 and m_j = 1 in both)
- Covariate pattern may be unusual and not represented well by the model
- None of the covariate values are unreasonable

Determine the effect of outliers on the model

- Delete outlier (one at a time or in groups)
- Rerun logistic regression model
- Compare ORs and p-values to those in the model based on all data
- Decide what to do with the outlier(s)

Model without covariate pattern 107

Compare to proc logistic for final model in Chapter4_4Results.pdf

```
proc logistic descending data=diag;
 where j ne 107;

 model y_j/m_j = priorfrac momfrac armassist raterisk2 height
        age priorfrac*age momfrac*armassist;

 contrast 'raterisk greater vs. same/less' raterisk2 1/estimate=exp;
 contrast 'height increase of 10 cm' height 10/estimate=exp;
```

Model w/o covariate pattern 107, cont.

```
contrast 'prior fracture yes vs. no at age 55' priorfrac 1 priorfrac*age 55
/estimate=exp;
contrast 'prior fracture yes vs. no at age 60' priorfrac 1 priorfrac*age 60
/estimate=exp;
< Etc.>
contrast 'prior fracture yes vs. no at age 90' priorfrac 1 priorfrac*age 90
/estimate=exp;
```

Model w/o covariate pattern 107, cont.

```
contrast 'age+10 at priorfrac=1' age 10 priorfrac*age 10/estimate=exp;
contrast 'age+10 at priorfrac=0' age 10 priorfrac*age 0/estimate=exp;

contrast 'momfrac yes vs. no at armassist=1' momfrac 1
                        momfrac*armassist 1 /estimate=exp;
contrast 'momfrac yes vs. no at armassist=0' momfrac 1
                        momfrac*armassist 0 /estimate=exp;
 contrast 'armassist yes vs. no at momfrac=1' armassist 1
                        momfrac*armassist 1/estimate=exp;
 contrast 'armassist yes vs. no at momfrac=0' armassist 1
                        momfrac*armassist 0/estimate=exp;
run;
```

ORs before and after deletion of outliers

Contrast	OR0	OR107	OR304	OR327	OR330	OR440	OR445
raterisk greater vs. same/less	1.60	1.63	1.65	1.64	1.57	1.56	1.63
height increase of 10 cm	0.63	0.62	0.61	0.60	0.63	0.59	0.61
prior fracture yes vs. no at age 55	4.82	5.09	4.95	4.69	5.00	5.07	5.27
prior fracture yes vs. no at age 60	3.65	3.80	3.74	3.60	3.77	3.83	3.93
prior fracture yes vs. no at age 65	2.77	2.84	2.83	2.76	2.84	2.89	2.93
prior fracture yes vs. no at age 70	2.10	2.12	2.14	2.12	2.14	2.18	2.18
prior fracture yes vs. no at age 75	1.59	1.59	1.61	1.63	1.62	1.64	1.62
prior fracture yes vs. no at age 80	1.21	1.19	1.22	1.25	1.22	1.24	1.21
prior fracture yes vs. no at age 85	0.92	0.89	0.92	0.96	0.92	0.94	0.90
prior fracture yes vs. no at age 90	0.70	0.66	0.70	0.74	0.69	0.71	0.67
age+10 at priorfrac=1	1.02	1.02	1.02	1.03	1.03	1.01	0.99
age+10 at priorfrac=0	1.77	1.83	1.78	1.75	1.82	1.77	1.78
momfrac yes vs. no at armassist=1	0.97	0.96	0.96	0.81	0.82	0.97	0.97
momfrac yes vs. no at armassist=0	3.48	3.62	2.87	3.52	3.51	3.91	3.85
armassist yes vs. no at momfrac=1	0.53	0.52	0.64	0.44	0.44	0.48	0.48
armassist yes vs. no at momfrac=0	1.90	1.95	1.90	1.92	1.89	1.92	1.90

P-values before and after deletion of outliers

Contrast	p0	p107	p304	p327	p330	p440	p445
raterisk greater vs. same/less	0.05	0.04	0.04	0.04	0.06	0.06	0.04
height increase of 10 cm	0.01	0.01	0.01	0.01	0.01	0.01	0.01
prior fracture yes vs. no at age 55	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prior fracture yes vs. no at age 60	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prior fracture yes vs. no at age 65	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prior fracture yes vs. no at age 70	0.00	0.00	0.00	0.00	0.00	0.00	0.00
prior fracture yes vs. no at age 75	0.07	0.07	0.06	0.06	0.06	0.05	0.06
prior fracture yes vs. no at age 80	0.55	0.59	0.54	0.49	0.54	0.50	0.55
prior fracture yes vs. no at age 85	0.84	0.77	0.85	0.92	0.84	0.88	0.80
prior fracture yes vs. no at age 90	0.49	0.44	0.49	0.56	0.49	0.51	0.45
age+10 at priorfrac=1	0.92	0.92	0.93	0.87	0.87	0.97	0.96
age+10 at priorfrac=0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
momfrac yes vs. no at armassist=1	0.94	0.93	0.93	0.67	0.69	0.95	0.96
momfrac yes vs. no at armassist=0	0.00	0.00	0.01	0.00	0.00	0.00	0.00
armassist yes vs. no at momfrac=1	0.27	0.26	0.44	0.17	0.17	0.20	0.21
armassist yes vs. no at momfrac=0	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Decision

- May want to rerun Stukel test (see Chapter 5, part 1) after deleting outliers to see if the tails assumption is now satisfied
- But…deleting outliers results in little change in ORs and p-values
- Would probably keep all outliers in the data set

What if your model had poor overall goodness-of-fit?

- May want to rerun goodness-of-fit tests (see Chapter 5, part 1) after deleting outliers to see if the model fits better
- Not necessary in this example

Assessment of fit via external validation

- A model always performs better on the developmental data set

Idea

- If the data set is large enough, exclude a subsample
- Or, if possible, collect additional data
- Build the model on the original data
- Test model fit on the subsample or the additional data
- Treat the coefficients as fixed constants rather than estimated values
- Or: Try bootstrapping as described by Austin et al.