

Biost 518 / Biost 515

Applied Biostatistics II / Biostatistics II



Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington

Lecture 10:
Effect Modification; Diagnostics

February 25, 2015

Lecture Outline



- Effect modification
 - Modeling effect modification
 - Examples
 - Administrative duties modifying the association between salary and sex
 - SEP by height, age, sex
- Introduction to Case Diagnostics
 - Leverage
 - Influence
 - Outliers

Effect Modification



Effect Modifier



- The association between Response and POI differs in strata defined by effect modifier
 - Statistical term: “Interaction”
- Depends on the measurement of effect
 - Summary measure
 - Mean, geometric mean, median, proportion, odds, hazard, etc.
 - Comparison across groups
 - Difference, ratio

Analysis of Effect Modification



- When the scientific question involves effect modification, analyses must be within each stratum separately
- If we want to estimate degree of effect modification or test for its existence, a regression model will typically include
 - Predictor of interest
 - Effect modifier
 - A covariate modeling the interaction (usually product)

Model for Effect Modification



- Typical model for effect modification
- Include “main effects” (can be bad not to)
 - X (or predictors that involve only X)
 - W (or predictors that involve only W)
- Include “interactions”
 - Predictor(s) derived from both X and W

$$\begin{aligned}g[\theta | X_i, W_i] &= \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times (XW)_i \\ &= \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i\end{aligned}$$

Interpretation of Parameters



$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Usual approach a bit more difficult
 - We can try using the idea of “comparison of θ across groups differing by 1 unit in corresponding predictor but agreeing in other modeled predictors”
 - However, terms involving two scientific variables makes this approach difficult

Intercept



$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of intercept straightforward
 - β_0 corresponds to $X=0, W=0$
 - May not be scientifically meaningful

Slopes for Main Effects



$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of main effects
 - β_X corresponds to 1 unit difference in X holding W and $(X \times W)$ constant
 - So 1 unit difference in X when $W=0$
 - May not be scientifically meaningful
 - β_W corresponds to 1 unit difference in W holding X and $(X \times W)$ constant
 - So 1 unit difference in W when $X=0$
 - May not be scientifically meaningful

Slope for interaction

$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of interaction difficult
 - β_{XW} corresponds to 1 unit difference in $(X \times W)$ holding X and W constant
 - Impossible, so we need another way to interpret this slope parameter

Consider Scientific Predictors



$$\begin{aligned}
 g[\theta | X_i, w] &= \beta_0 + \beta_X \times X_i + \beta_W \times w + \beta_{XW} \times X_i \times w \\
 &= (\beta_0 + \beta_W \times w) + (\beta_X + \beta_{XW} \times w) \times X_i
 \end{aligned}$$

In stratum with $W = w$

Intercept : $(\beta_0 + \beta_W \times w)$ corresponds to $X_i = 0$

Slope : $(\beta_X + \beta_{XW} \times w)$ compares groups differing by 1 unit in X

β_{XW} is difference in X slope per 1 unit difference in W

Consider Scientific Predictors



$$\begin{aligned}
 g[\theta | x, W_i] &= \beta_0 + \beta_X \times x + \beta_W \times W_i + \beta_{XW} \times x \times W_i \\
 &= (\beta_0 + \beta_X \times x) + (\beta_W + \beta_{XW} \times x) \times W_i
 \end{aligned}$$

In stratum with $X = x$

Intercept : $(\beta_0 + \beta_X \times x)$ corresponds to $W_i = 0$

Slope : $(\beta_W + \beta_{XW} \times x)$ compares groups differing
by 1 unit in W

β_{XW} is difference in W slope per 1 unit
difference in X

Symmetry of Effect Modification



- Note that if X modifies the association between Y and W, then W modifies the association between Y and X
- Aside: Confounding need not be symmetric
 - W can confound the association between Y and X, but X not confound the association between Y and W
 - W and X associated in the sample
 - Y and X not associated after adjusting for W
 - Y and W associated after adjusting for X

Inference for Effect Modification



$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- No effect modification if $\beta_{XW} = 0$
 - Hence, inference about existence of effect modification tests that $\beta_{XW} = 0$
 - We can perform such inference using standard regression output for the corresponding slope parameter

Inference for Main Effect Slope



$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Interpretation of $\beta_X = 0$
 - Same intercept in all strata defined by W
 - Generally a very uninteresting question
 - We rarely make inference on main effect slopes by themselves

Inference About Effect of X



$$g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_W \times W_i + \beta_{XW} \times X_i \times W_i$$

- Response parameter not associated with X if $\beta_X = 0$ AND $\beta_{XW} = 0$
 - We will need to construct special tests that both parameters are simultaneously 0
 - The t tests given in regression output consider only one slope parameter at a time

Stata: Testing Multiple Slopes



- Stata has easy method for performing test that multiple parameters are simultaneously 0
 - Perform any regression command
 - Then use `test var1 var2 ...`
 - Provides P value of the hypothesis test based on most recently executed regression command of any type of regression

Ex: Salary by Sex and Admin



- Does sex modify the association between mean salary and administrative duties?
- With two binary variables, modeling interaction by product is the obvious choice

Ex: Stata output

```

.....
. g admfem= admin * female
. regress salary female admin admfem if year==95, robust

```

Linear regression

```

Number of obs =    1597
F(   3, 1593) =  125.26
Prob > F      =   0.0000
R-squared     =   0.1615
Root MSE     =  1866.9

```

		Robust				[95% Conf Intrvl]	
salary	Coef	StdErr	t	P> t			
female	-1226.2	95.371	-12.86	0.000	-1413.3		-1039.2
admin	1951.4	176.39	11.06	0.000	1605.4		2297.4
admfem	-461.91	341.68	-1.35	0.177	-1132.1		208.28
_cons	6506.6	61.823	105.25	0.000	6385.3		6627.9

Automatic Creation of Dummy Variables



- Stata's prefix command can automatically create dummy variables and interactions
- To create dummy variables
 - `regress yvar i.x1var i.x2var ctsvar`
- To create dummy variables and interactions (use ##)
 - `regress yvar i.x1var##i.x2var ctsvar`
- To create dummy variables and interaction with a continuous variable (use c. prefix)
 - `regress yvar i.x1var i.x2var##c.ctsvar`
- Three way interactions
 - `regress yvar i.x1var##i.x2var##c.ctsvar`
- (Other options will create the interactions without main effects, but I urge you not to do that)

20

Ex: Stata output

```
. regress salary i.female##i.admin if year==95, robust
```

Linear regression

Number of obs = 1597
 F(3, 1593) = 125.26
 Prob > F = 0.0000
 R-squared = 0.1615
 Root MSE = 1866.9

salary	Robust					
	Coef	StdErr	t	P> t	[95% Conf	Intvl]
1.female	-1226.2	95.371	-12.86	0.000	-1413.3	-1039.2
1.admin	1951.4	176.39	11.06	0.000	1605.4	2297.4
female#admin						
1 1	-461.91	341.68	-1.35	0.177	-1132.1	208.28
_cons	6506.6	61.823	105.25	0.000	6385.3	6627.9

21

Ex: Descriptive Statistics



- Note that with two binary variables, the regression parameters agree exactly with the corresponding group sample means
 - A saturated model: Four distinct groups sampled, four regression parameters

```
. table admin female if year==95, co(mean salary)
```

admin	female	
	Male	Female
Nonadmin	6506.607	5280.373
Admin	8457.985	6769.844

Ex: Inference About Eff Mod

- Does sex modify association between mean salary and administrative duties?
 - Estimate that the “administrative supplement” averages \$462 less for women than men
 - 95% CI: \$1132 less to \$208 more
 - Not statistically significant: $P = 0.177$

	Robust					
salary	Coef	StdErr	t	P> t	[95% Conf	Intrvl]
female	-1226.2	95.371	-12.86	0.000	-1413.3	-1039.2
admin	1951.4	176.39	11.06	0.000	1605.4	2297.4
adm fem	-461.91	341.68	-1.35	0.177	-1132.1	208.28
_cons	6506.6	61.823	105.25	0.000	6385.3	6627.9

Ex: Inference About Sex Assoc



- Is sex associated with mean salary?
 - Need to test that slope parameters for `female` and `adm_fem` are simultaneously 0

```
. test female adm_fem
```

```
( 1) female = 0
```

```
( 2) adm_fem = 0
```

```
F( 2, 1593) = 95.90  
Prob > F = 0.0000
```


Ex: Inference for Admin Assoc



- Are administrative duties associated with mean salary?
 - Need to test that slope parameters for `admin` and `adm fem` are simultaneously 0

```
. test admin adm fem
```

```
( 1) admin = 0
```

```
( 2) adm fem = 0
```

```
F( 2, 1593) = 74.15  
Prob > F = 0.0000
```

Continuous Predictors



- Modeling interactions with continuous predictors is conceptually more complicated
- Is a multiplicative interaction at all a reasonable model for the data?
- Nonetheless, this is the most common way we detect interactions
 - I would caution against using the model as predictions without carefully examining the data
 - And this sort of *post hoc* assessment has its problems

Example: SEP “Normal Ranges”



- We want to find “normal ranges” for somatosensory evoked potential (SEP)
 - Time that it takes a signal to reach your brain from your ankle
- Method of analysis
 - Not recommended: “Prediction intervals” assuming same distribution (Gaussian) within each group
 - Recommended: 2.5th and 97.5th quantile within groups
- As a first step, we want to consider important predictors of nerve conduction times
- If any variables such as sex, age, height, race, etc. are important predictors of nerve conduction times, then it would make most sense to obtain “normal ranges” within such groups

SEP: Important Predictors



- Scientifically, we might expect that height, age, and sex are related to the nerve conduction time
 - Nerve length should matter, and height is a surrogate for nerve length
 - Age might affect nerve conduction times: People slow down with age
 - Sex: Men are SO fragile

SEP: Height – Age Interaction?



- Prior to looking at the data, we can also consider the possibility that interactions between these variables might be important
- Height - age interaction?
 - Do we expect the difference in conduction times between 6 foot tall and 5 foot tall 20 year olds to be the same as the difference in conduction times between 6 foot tall and 5 foot tall 80 year olds?

SEP: Height – Age Interaction Rationale



- We might suspect such an interaction due to the fact that height may not be as good a surrogate for nerve length in older people
 - With age, some people tend to shrink due to osteoporosis and compression of intervertebral discs
 - It is not clear that nerve length would be altered in such a process
- Thus, in young people, differences in height probably are a better measure of nerve length than in old people
 - Tall old people probably have been tall always
 - Short old people will include some who were much taller when they were young

SEP: Height – Age - Sex Interaction?

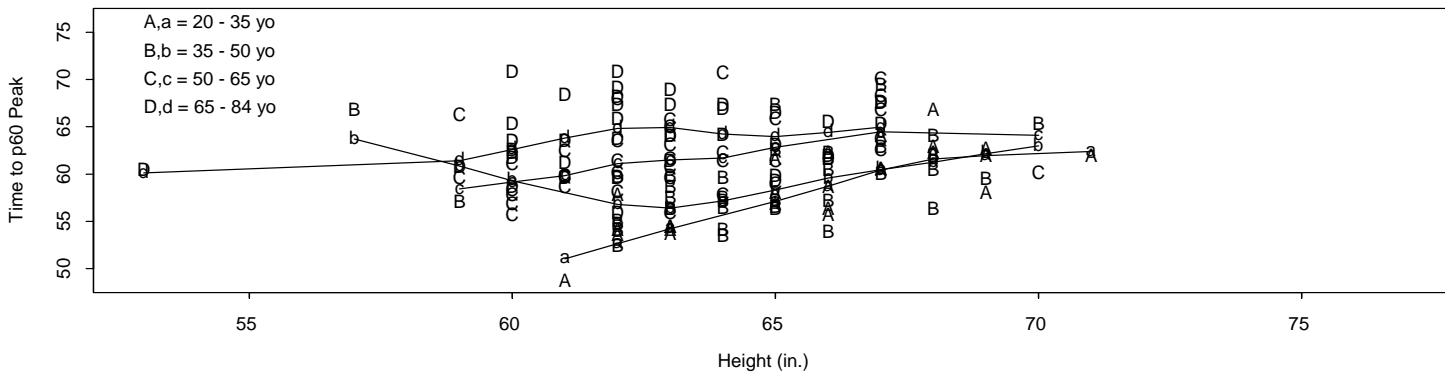


- We can also consider the possibility of three way interactions between height, age, and sex
- Osteoporosis affects women far more than men
 - Hence, we might expect the height - age interaction to be greatest in women and not so important in men
- A two way interaction between height and age that is different between men and women defines a three way interaction between height, age, and sex

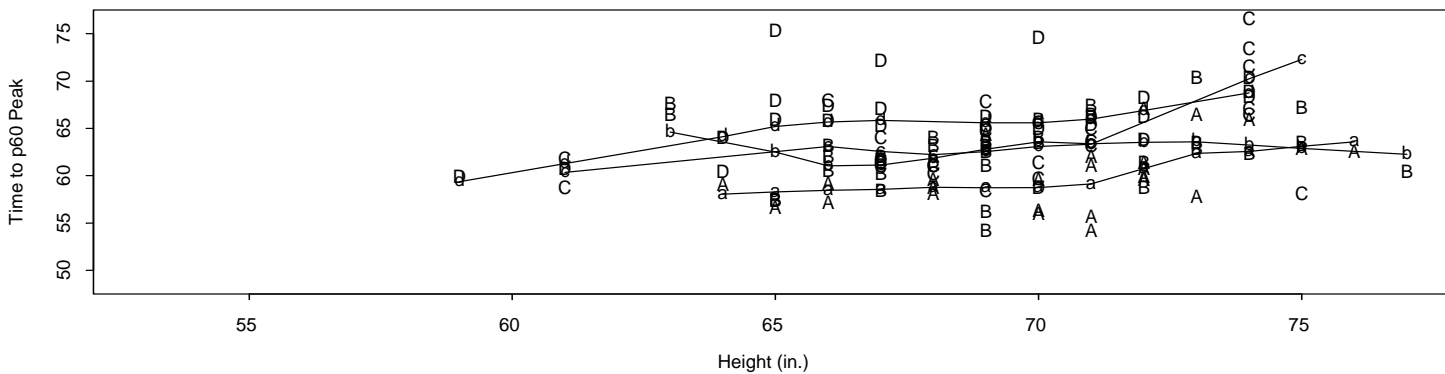
Stratified Scatterplots

- Scatterplots of p60 by height for each sex, lowess in age strata

Average Time to p60 Peak: Females



Average Time to p60 Peak: Males



SEP: Definition of Model



- Defining a regression model with interactions
 - We must create variables to model the three way interaction term
- Furthermore, it is a VERY GOOD idea to include all “main effects” and “lower order interactions” in the model as well
 - “main effects”: the individual variables which contribute to the interaction
 - “lower order terms”: all interactions that involve some combination of the variables which contribute to the 3-way interaction

SEP: Modeling Interactions



- Most often, we lack sufficient information to be able to guess what the true form of an interaction might be
- The most popular approach is thus to consider multiplicative interactions
- Create a new variable by merely multiplying the two (or more) interacting predictors

SEP: Creating New Interaction Terms



- Thus for this problem we could create variables
 - $HA = \text{Height} * \text{Age}$
 - $HM = \text{Height} * \text{Male}$
 - $AM = \text{Age} * \text{Male}$
 - $HAM = \text{Height} * \text{Age} * \text{Male}$

SEP: Interpretation of Parameters



- In the presence of higher order terms (powers, interactions) interpretation of parameters is not easy
- We can no longer use “the change associated with a 1 unit difference in predictor holding other variables constant”
- It is generally impossible to hold other variables constant when changing a covariate involved in an interaction
- (If not impossible, it is certainly uninteresting)

SEP: Interpretation Using Lines in Strata



- Interpretation of the model in terms of the SEP height relationship within age-sex strata

$$E(p60 | Ht, Age, Male) = \beta_0 + \beta_H Ht + \beta_A Age + \beta_M Male \\ + \beta_{HA} HA + \beta_{HM} HM + \beta_{AM} AM + \beta_{HAM} HAM$$

p60 - Height relationship for Age = a :

Sex	Intercept	Slope
F	$(\beta_0 + \beta_A a)$	$(\beta_H + \beta_{HA} a)$
M	$(\beta_0 + \beta_M + (\beta_A + \beta_{AM}) a)$	$(\beta_H + \beta_{HM} + (\beta_{HA} + \beta_{HAM}) a)$

SEP: Inclusion of Lower Order Terms



- From the above, we see the importance of including the main effects and lower order terms
- E.g., leaving out the height - sex interaction is tantamount to claiming that the p60 - height relationship among newborns is the same for the two sexes
 - (It might be, but the chance that our lines would predict the truth is very slight-- we are trying to approximate relationships in other age ranges)

SEP: Regression Output



```
. g p60=(p60R + p60L)/2
. regress p60 height age male ha hm am ham, robust
```

Linear regression

```
Number of obs =      250
F( 7, 242) =      18.76
Prob > F      =      0.0000
R-squared     =      0.3920
Root MSE     =      3.6125
```

		Robust					
p60	Coef.	StdErr	t	P> t	[95% Conf	Intrvl]	
height	1.3803	.43090	3.20	0.002	.53147	2.2291	
age	1.1294	.44597	2.53	0.012	.25094	2.0079	
male	74.958	34.860	2.15	0.033	6.2908	143.62	
ha	-.01500	.00684	-2.19	0.029	-.02848	-.00152	
hm	-1.1270	.52083	-2.16	0.031	-2.1529	-.10107	
am	-1.1629	.57238	-2.03	0.043	-2.2904	-.03537	
ham	.01750	.00859	2.04	0.043	.00057	.03443	
_cons	-36.443	28.304	-1.29	0.199	-92.196	19.311	

39

SEP: Using Stata's i. Prefix



```
. regress p60 i.male##c.age##c.height, robust
```

Linear regression

```
Number of obs =    250
F( 7, 242) =    18.76
Prob > F      =    0.0000
R-squared     =    0.3920
Root MSE     =    3.6125
```

p60	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
1.male	74.95773	34.85959	2.15	0.033	6.290783	143.6247
age	1.129423	.445972	2.53	0.012	.2509409	2.007906
male#c.age						
1	-1.162866	.5723847	-2.03	0.043	-2.290358	-.0353737
height	1.380275	.4309057	3.20	0.002	.5314702	2.229079
male#c.height						
1	-1.127006	.5208285	-2.16	0.031	-2.152942	-.1010699
c.age#c.height	-.0149985	.0068449	-2.19	0.029	-.0284817	-.0015154
male#c.age# c.height						
1	.0175005	.0085949	2.04	0.043	.00057	.0344309
_cons	-36.44286	28.30387	-1.29	0.199	-92.19624	19.31053

40

SEP: An Aside - Subgroup Analysis

- If I restrict analysis to females, estimates are the same in this quasi “saturated” model
 - (Restricting by age or height would differ due to “borrowing information” across groups)
- Inference differs due to the estimate of the residual standard error

```
. regress p60 height age ha if male==0, robust
```

Linear regression

```
Number of obs =    137
F( 3, 133) =    24.21
Prob > F      =    0.0000
R-squared     =    0.3810
Root MSE     =    3.6006
```

p60	Coef.	Robust StdErr	t	P> t	[95% Conf	Intrvl]
height	1.3803	.43028	3.21	0.002	.52919	2.2314
age	1.1294	.44533	2.54	0.012	.24858	2.0103
ha	-.01500	.0068	-2.19	0.030	-.02852	-.00148
_cons	-36.443	28.263	-1.29	0.199	-92.346	19.460

41

SEP: Interpreting Estimates

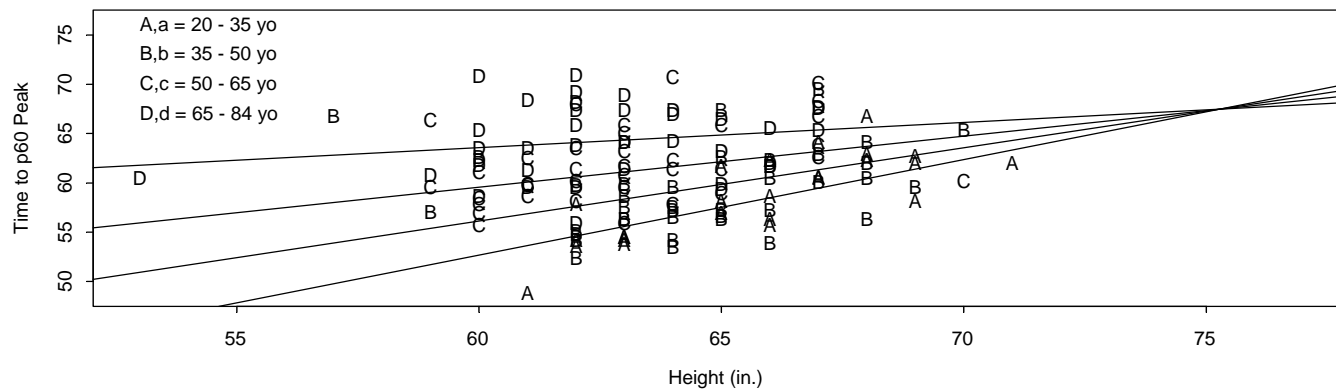


- Figuring out what all these estimates mean is nearly impossible
- I find it easiest to graph the predicted values
- I did this by graphing the fitted values for midpoints of strata defined by age

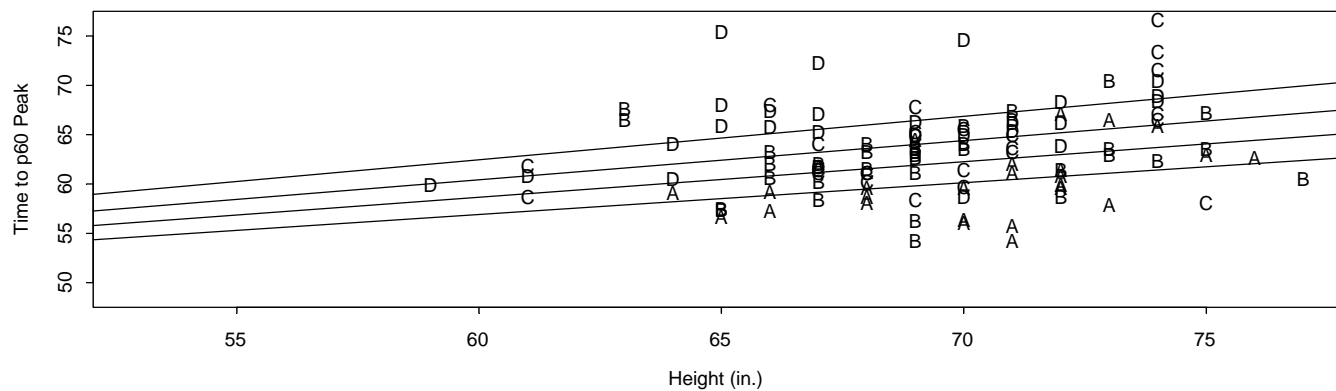
SEP: Lines Predicted By Model



Average Time to p60 Peak: Females



Average Time to p60 Peak: Males



SEP: Detecting 3-way Interaction



- From the inference, we find a statistically significant three way interaction
 - $P = .047$
- This would argue that I should make predictions based on a model including the 3-way interaction
 - But...

SEP: Influence of Individual Cases



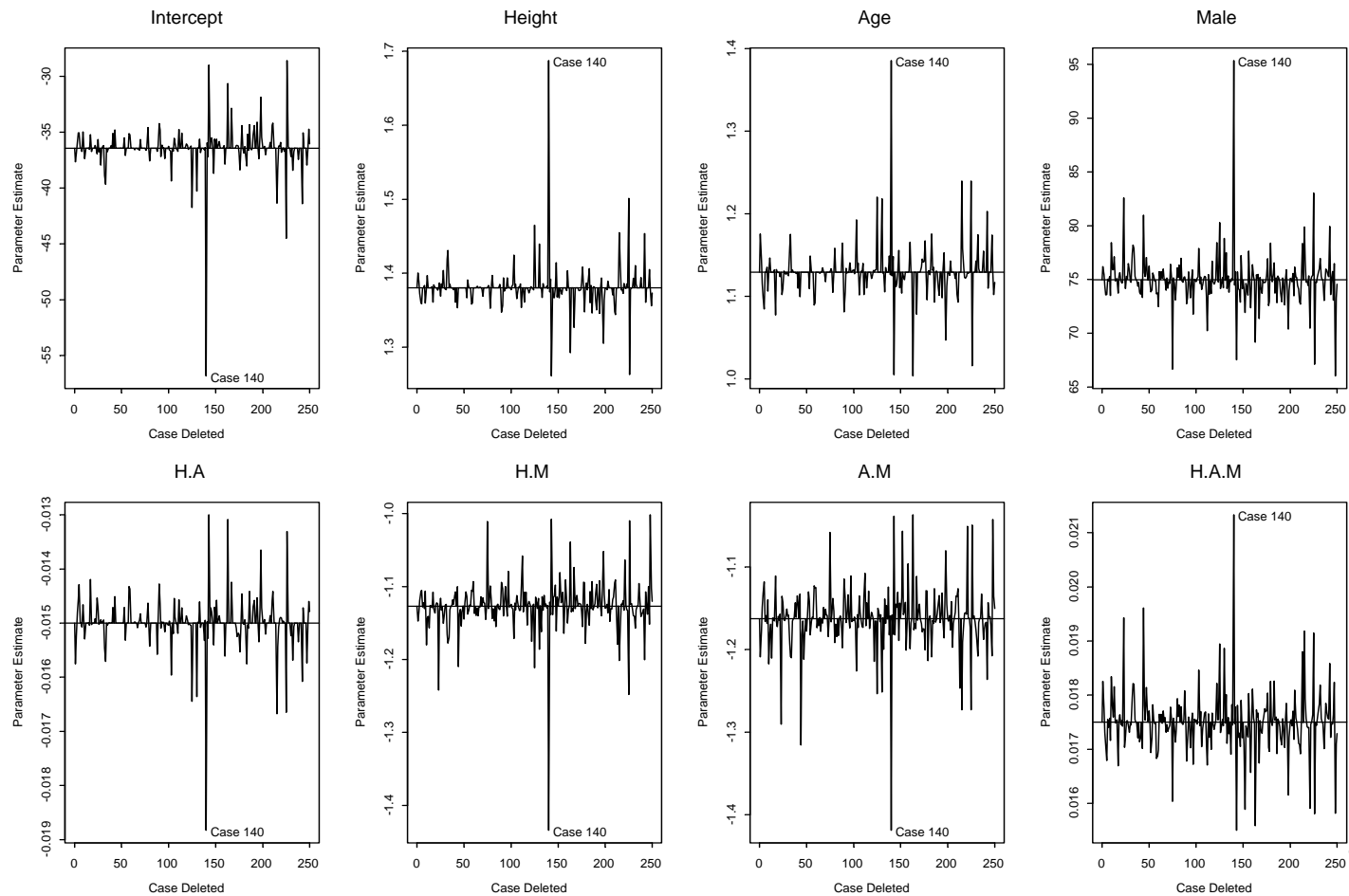
- I always worry that interactions might be significant only because of a single “outlier”
- If that were the case, I might choose not to include the interaction (but I always include the case)
- Looking ahead: I can “diagnose” such a problem by assessing the influence of each case

Example: SEP “Normal Ranges”



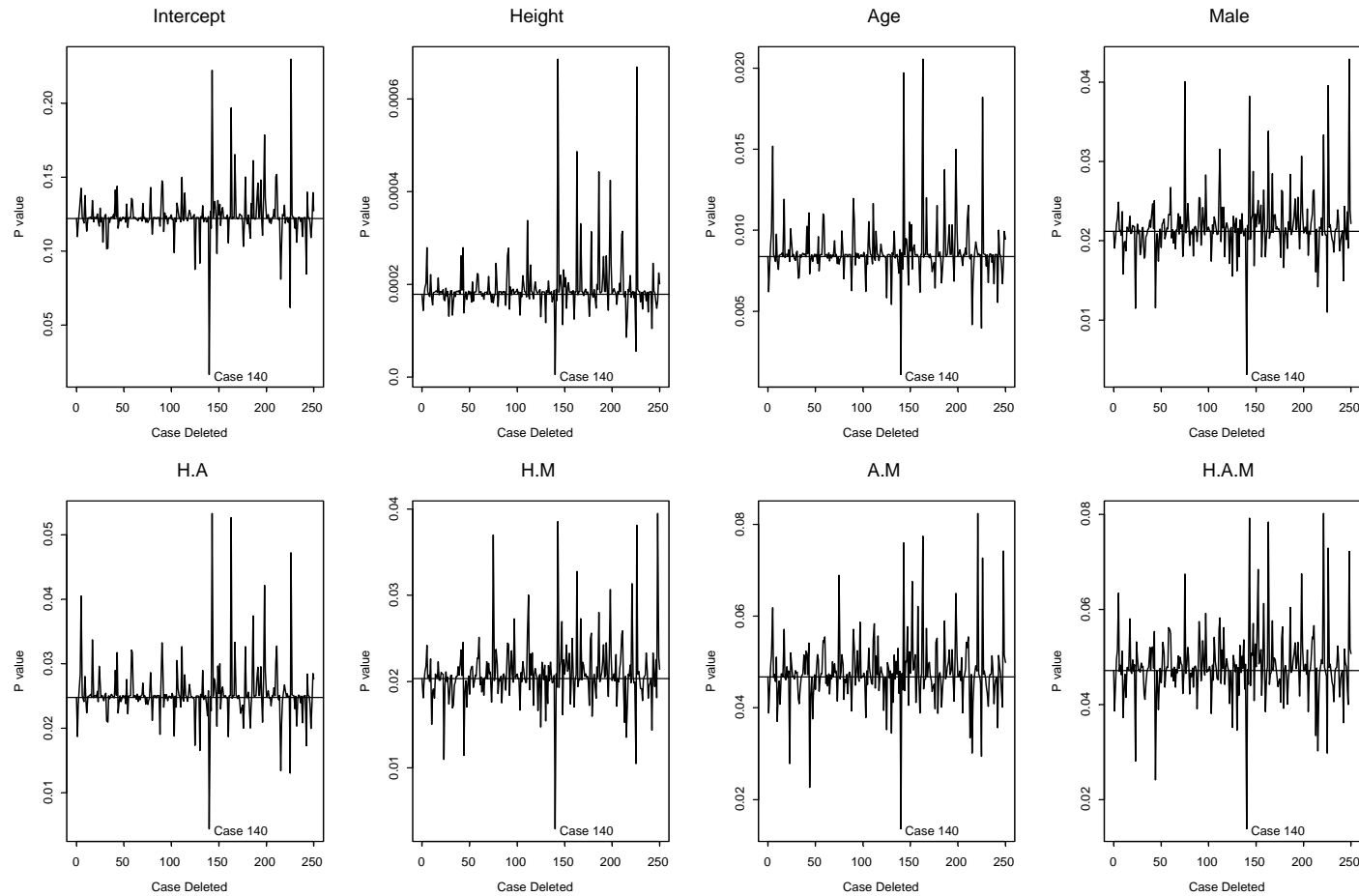
- I am now interested in ensuring that the evidence for an interaction is not based solely on a single person's observation
- Hence, I consider 250 different regressions in which I leave out each case in turn
- I plot the slope estimates and P values for each variable as a function of which case I left out
 - Case 0 corresponds to using the full data set

Influence on Estimates



47

Influence on P values



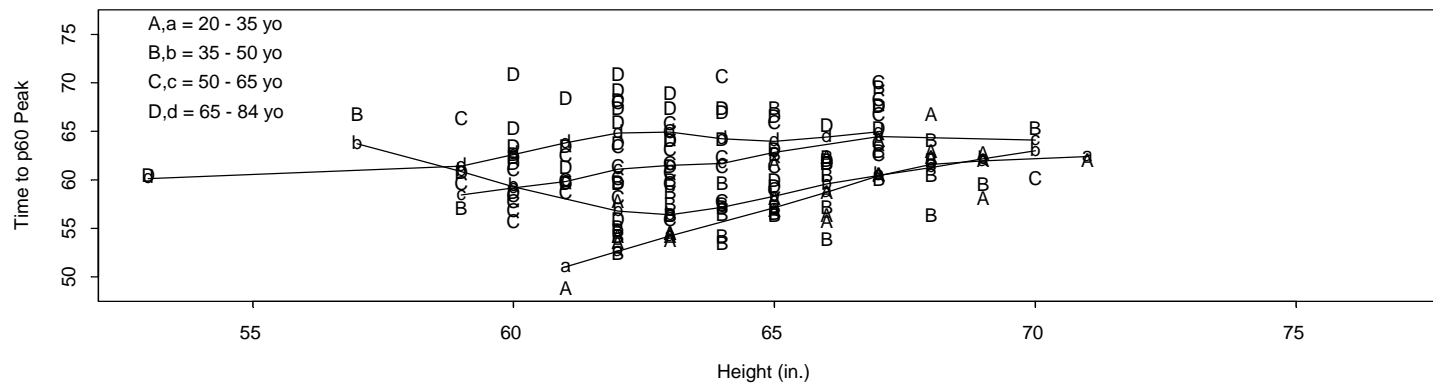
Example: SEP “Normal Ranges”



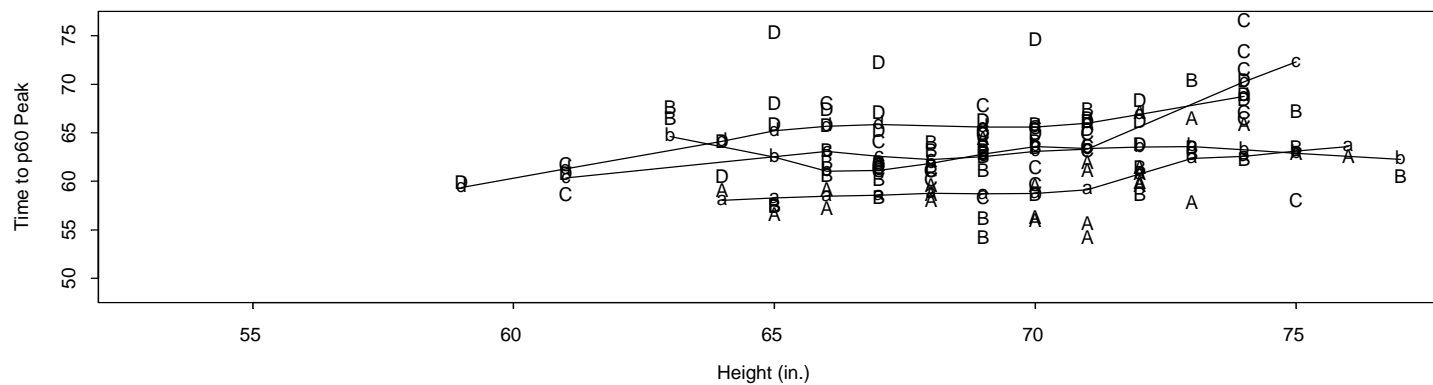
- Contrary to what I was afraid of, the only influential case actually lessened the evidence of an interaction
- When Case 140 is removed from the data, the evidence for an interaction is a larger estimate and a lower P value
- We can examine the scatterplot to see why Case 140 might be so influential

Stratified Scatterplots

Average Time to p60 Peak: Females



Average Time to p60 Peak: Males



50

Example: SEP “Normal Ranges”



- So now what do I do with Case 140?
- From the influence diagnostics, I now feel comfortable with the fact that the data really do suggest a three way interaction
- Personally, I do NOT remove the case from the dataset when making my prediction intervals
 - I do not know why Case 140 is so unusual
 - It is possible that people like her are actually more prevalent in the population than my sample would suggest
 - My best guess is that she represents 0.4% of the population, so leave her in