BIOS 312: MODERN REGRESSION ANALYSIS

James C (Chris) Slaughter
Department of Biostatistics
Vanderbilt University School of Medicine
james.c.slaughter@vanderbilt.edu

biostat.mc.vanderbilt.edu/CourseBios312

Updated April 3, 2012

Contents

Chapter 15

Regression Diagnostics

15.1 Overview

- There are many available techniques for diagnosing both the appropriateness of the chosen model and the impact of individual observations
  - Model diagnostics
  - Case diagnostics
- Checking your model is an important part of any analysis
  - Usually conducted at the end of the analysis to verify your approach and determine if a different approach gives different scientific conclusions
  - However, need to avoid data-driven model building approaches or will suffer from inflated Type 1 error rate
  - My philosophy: Carefully consider the scientific question, propose a reasonable model to answer question, then check modeling assumptions
    - Always clearly report the process you used

- Three levels of statistical analyst
  - Novice analyst
    - Understands how to fit appropriate regression models and interpret the output
    - Fits one regression model
    - Doesn't necessarily understand all of the underlying assumptions of the chosen model
    - May or may not fit an appropriate model, and doesn't know how to conduct case diagnostics
  - Intermediate/Student
    - Understands the assumptions inherent in chosen model
    - Can conduct numerous regression diagnostics
    - Fits many different models to answer scientific question
    - Often relies heavily on diagnostics or other data-driven techniques to arrive at a final model
      - Will report the best model, but not all of the multiple-testing involved in arriving at the final model
      - Statistical inference is much more questionable than what is reported
  - Experienced analyst
    - Is aware of the danger inherent with data driven approaches
      - Biased parameter estimates (often overstating the effect size)
      - Smaller variance estimates (increase Type I error rate)

- Responsible for many of the “findings” in published research that are never able to be reproduced again

* Knows that model assumptions can’t be tested with p-values
  - Such tests are asking you to interpret large p-values as “equal”, which is wrong

  - Like with confounding, tests at best help to diagnoses problem with the proposed model

* Conducts an analysis more like the novice statistician than the student
  - Thinks about the scientific problem and how best to answer the proposed research questions using a few (one?) statistical models

  - Knows that only the pre-specified models can be reliably trusted to be testing scientific hypotheses

  - Considers the underlying assumptions of the model and whether or not those assumptions would be appropriate for the particular problem

* More likely to pre-specify a model that makes fewer underlying assumptions

* After fitting the pre-specified model, will conduct exploratory analyses and model diagnostics
  - Suggest areas for future research

  - Does using a different model alter my scientific conclusions?

  - If conclusions remain the same, more confident that the pre-specified model correctly answer the scientific question

    - If conclusions are different, need to understand the reason for the differences. Diagnostics can help explain differences.

- Diagnostics for regression models include considering the appropriateness of the proposed model and the impact of individual data points on our scientific conclusions

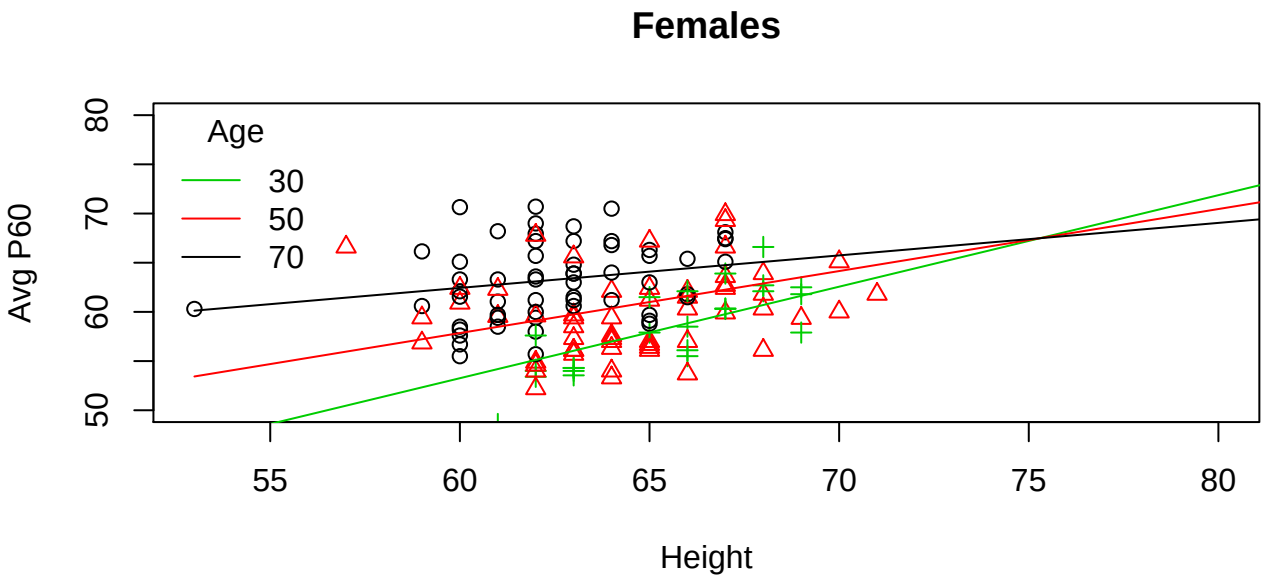
- Model diagnostics
  - Assessing distributional assumptions
  - Assessing model fit

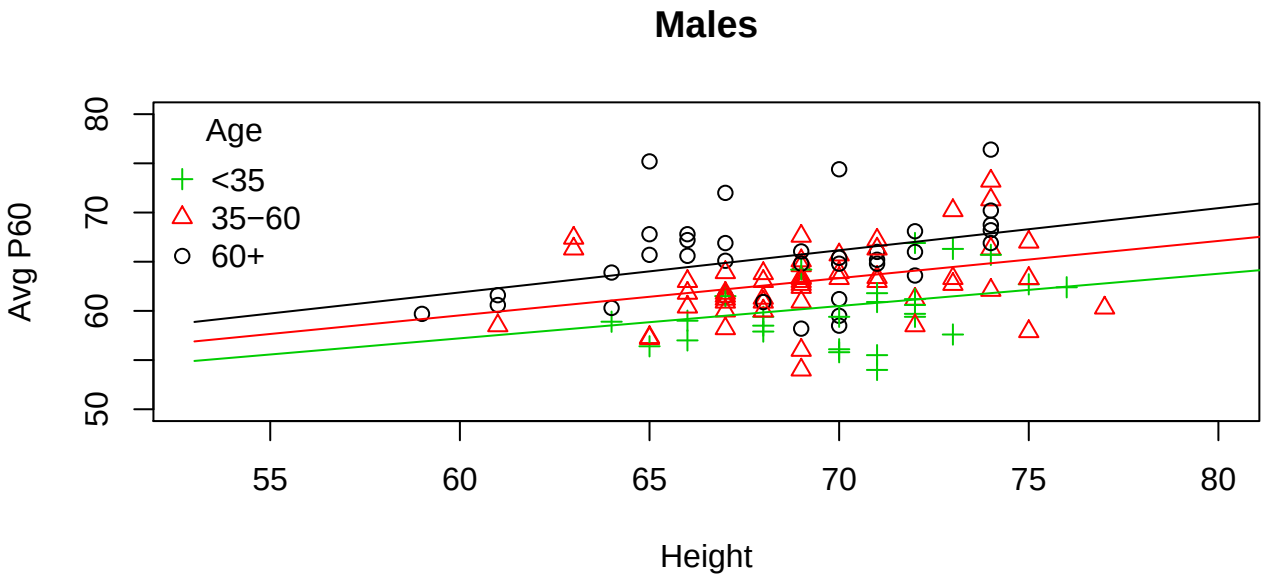
- Case diagnostics
  - Leverage
  - Influence
  - Outliers

15.1.1 Example: Normal ranges for SEP

- We want to find normal ranges for somatosensory evoked potential (SEP)
  - p60: Average time (in milliseconds) to detection of the second positive SEP following stimulation of the posterior right and left tibial nerve
  - Consider predictors age, gender, height
  - Pre-specify that we want to fit a regression model with main effects, all two-way interactions, and the three-way interaction
    $E[p60|Ht, Age, Male] = \beta_0 + \beta_H Ht + \beta_A Age + \beta_M Male + \beta_{HA} HA + \beta_{HM} HM + \beta_{AM} AM + \beta_{HAM} HAM$

- Modeling assumptions
  - Specifying a model with so many interactions will likely model the mean well, as long as the effects are linear over our range of observed covariates
    - Could be non-linear for different ranges of age, height
    - Could be better modeled using regression splines


Females
Age
30
50
70
Avg P60
Height
50
60
70
80
55
60
65
70
75
80


Males
Age
<35
35−60
60+
Avg P60
Height
50
60
70
80
55
60
65
70
75
80

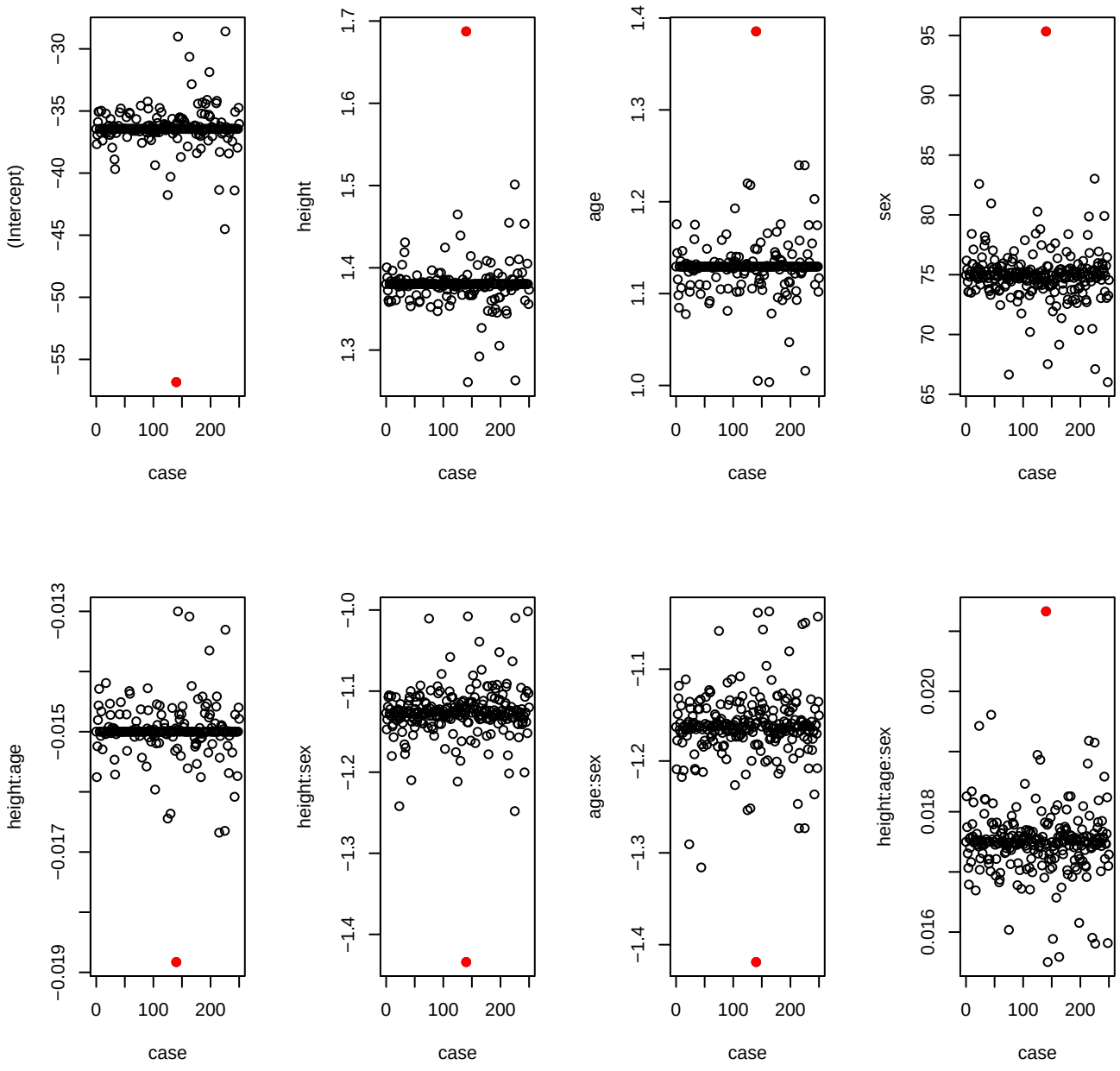
- We found a statistically significant three way interaction (p = 0.047)

- This would argue for making prediction based on a model that include the 3-way interaction

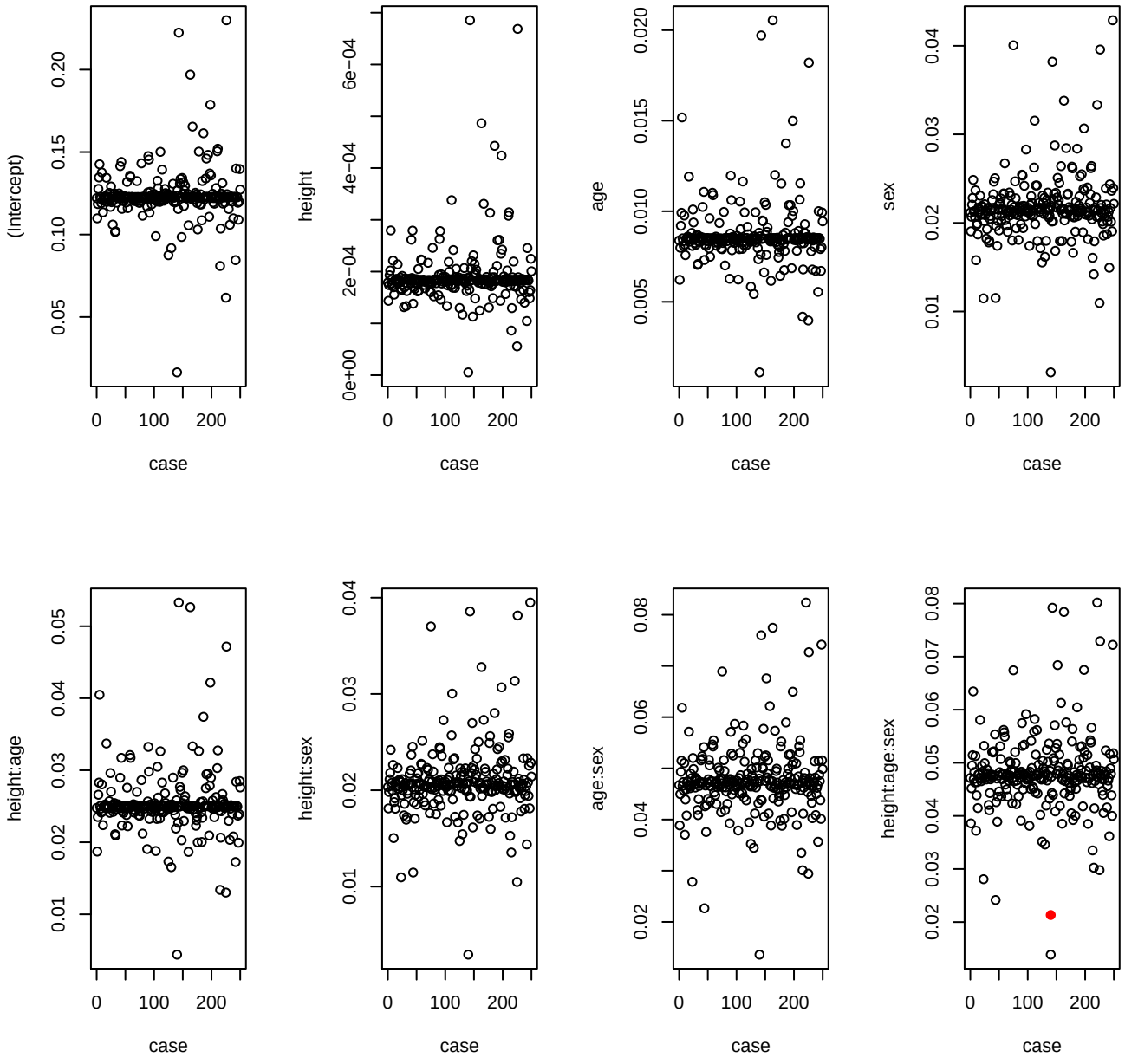
- However, interactions might be significant only because of a single outlier
  - If that were the case, I might choose not to include the interaction
  - But, I would include the influential data point
  - We will look at the results of a “diangosis” of influential observations now, and cover in more detail later

- In particular, I am interested in ensuring that the evidence for an interaction is not based solely on a single person’s observation
  - Hence, I consider 250 different regression in which I leave out each subject in turn
  - I plot the slope estimates and p-values for each variable as a function of which case I left out
  - For comparison, case 0 corresponds to using the full dataset

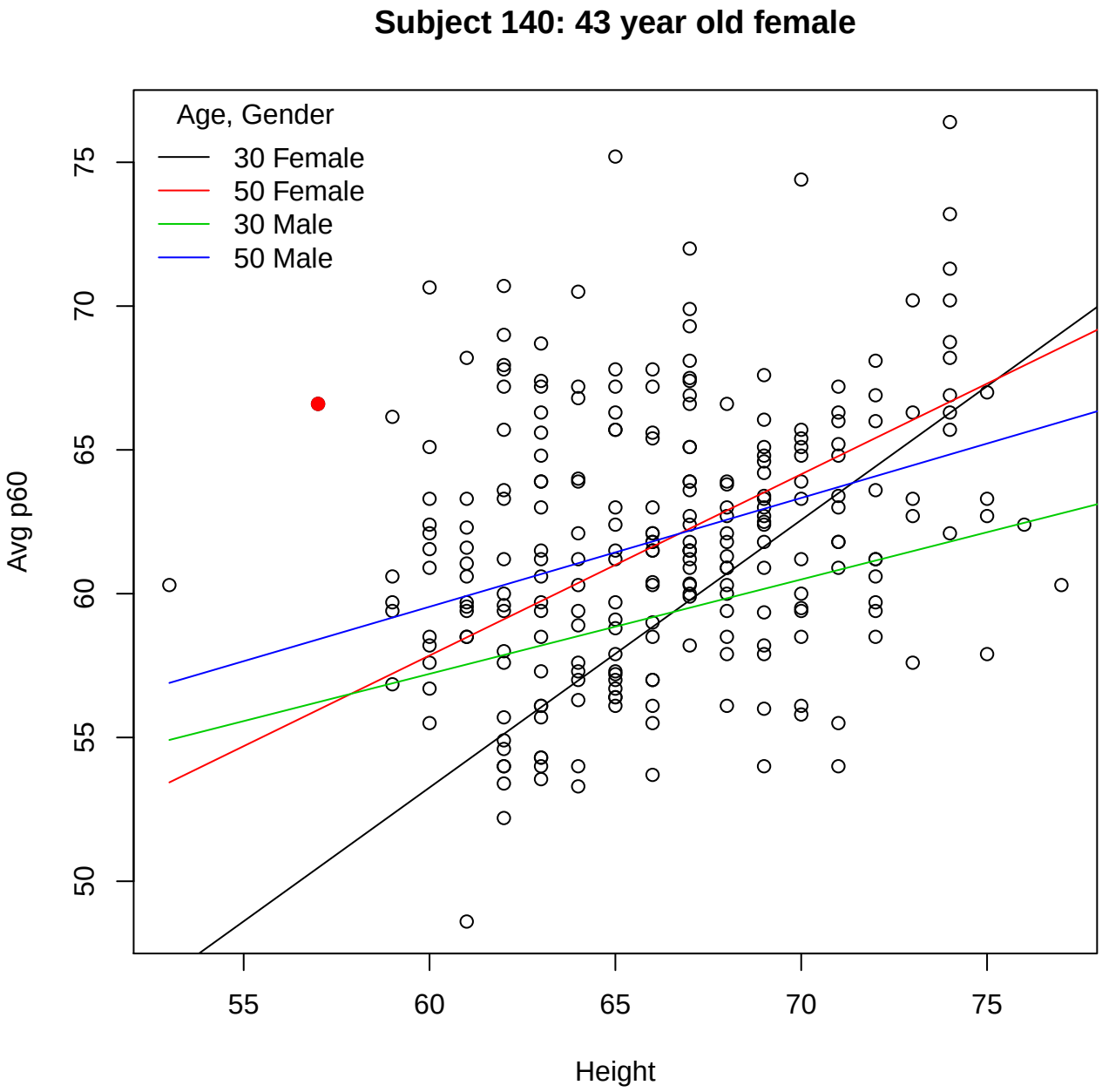
- Changes in coefficients (βs)



- Changes in p-values for βs



- Contrary to my fear, the only influential observation actually lessened the evidence of an interaction
  - When observation 140 is removed from the data, the evidence of an interaction is a larger estimate and lower p-value
  - We can examine the scatterplot to see why subject 140 might be so influential
  - Subject 140 is a 43 year old, 57 inch female with an average p60 of 66.6


Subject 140: 43 year old female
Age, Gender
30 Female
50 Female
30 Male
50 Male
Avg p60
Height
50
55
60
65
70
75
55
60
65
70
75

- So, what do I do with observation 140?
  - From the influence diagnostics, I am still comfortable that the data suggest a 3-way interaction
  - Personally, I do not remove observation 140 when making prediction intervals
    - I do not know why observation 140 is unusual
    - It is possible that people like 140 are actually more prevalent in the population than my sample would suggest
    - My best guess is observation 140 represents only 0.4% of the population, but would still leave her in the analysis
  - Removing subject 140 could bias parameter estimates, so I would rarely remove observations based on diagnostics

15.2 Regression Model Assumptions

- General regression model notation

$$g(\theta_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} \tag{15.1}$$

$g()$	Link function used for modeling
β_0	Intercept
β_p	Slope for predictor X_p

  - The link function is often either none (for modeling means) or log (for modeling geometric means, odds, hazards)

15.2.1 Maximal number of assumptions

- At most, there are 5 necessary assumptions

  - Independence between identified clusters
  - Sufficient sample sizes for asymptotic distributions to be a good approximation
  - Variance appropriate to the model
  - Regression model accurately describes summary measure across groups
  - Shape of distribution the same in each group

- Note that for some regression models (e.g. logistic, Poisson, PH) there is an implied relationship between the mean and variance
  - If we have correctly modeled the mean, we have correctly modeled the variance
  - Model with more parameters (e.g. complex dose-response models) will do a better job of modeling the mean, and thus a better job of getting the variance correct

- Depending on our scientific questions, we may need to check all or some of the above assumptions

15.2.2 Assumptions needs to detect linear trends in $g(\theta)$

- Independence between identified clusters
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model (relaxed if using robust standard errors)

15.2.3 Assumptions needs to estimate $g(\theta)$ in groups

- Independence between identified clusters
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model (relaxed if using robust standard errors)
- Regression model accurately describes summary measure across groups

15.2.4 Assumptions needs to predict range of new Y in groups

- Independence between identified clusters
- Sufficient sample sizes for asymptotic distributions to be a good approximation
- Variance appropriate to the model (NOT relaxed if using robust standard errors)
- Regression model accurately describes summary measure across groups
- Shape of distribution the same in each group

15.2.5 Role of Diagnostics

- Determine if the regression model fits the data well
  - Model diagnostics

  - Evaluate independence, link function, transformation of predictors, interactions, assumptions about variance

- Determine if individual cases are difference from the bulk of the data
  - Case diagnostics
  - Evaluate leverage, influence, outliers

- Caveats
  - Diagnostic methods are always approximate and rarely conclusive on their own
  - Using diagnostics alters the analysis plan (and thus the scientific question) and should always lessen our confidence in the statistical evidence
  - Unfortunately, we do not have good methods for quantifying the lessened confidence in the p-values and confidence intervals
  - Diagnostics only tell you about data in your sample, not the data in the population which you didn't sample
    * Data in the population is more concerning. Does my model adequately describe the population?
    * Only the sampling scheme can tell you if you have adequately sampled from the population. This question is more fundamental than the analysis you performed.

15.3 Model Diagnostics

15.3.1 Assessing Independence

- Requires consideration of the study design

- Are there variables in the dataset that identify clusters?

- Thing to look for
  - Correlations in time
  - Correlations in location
  - Correlations within families, hospitals, etc.
  - Correlations within subjects

- Note that we are interested in correlations after adjustment for predictors

15.3.2 Assessing Asymptotic Distribution

- We rely on an approximate Normal distribution for regression parameters (βs)

- Generally, Normality will hold in large sample sizes
  - Large depends on the distribution of the data
  - As a rule, heavy tails of response distribution requires larger sample size because it increases the tendency to have outliers

- Rules of thumb
  - Linear regression is quite robust for tests of zero slope when $n > 50$ (Lumley et al.)
  - Logistic, Poisson, PH asymptotics depend on the number of events observed

15.3.3 Assessing Appropriate Variance

- In classic linear regression, constant variance can be assessed by
  - Stratified estimates of the variances
    - Calculate variance of the outcome with groups as defined by the predictor of interest
    - Problem: Heterogeneity of the means within strata can look like variability of response variable
    - Might work well if have large number of observations in each level of grouping variable (e.g. categorical dose)
  - Scatter plots
    - Response versus predictor
    - Residuals versus predictors
    - Residuals versus fitted values
- Types of residuals
  - "residuals": $Y - \hat{Y}$
  - "standardized residuals": residuals standardized to have variance of 1
  - "studentized", "deleted", "studentized-deleted"
- Estimation of residuals from software
  - In Stata, following any regression command
    - `predict varname, resid`
    - `predict varname, rstu`

  - In R, after saving any model fit
    * `resid(model1)`
    * `rstudent(model1)`

- Assumption in linear regression are primarily about the errors, which we can examine by looking at residuals
  - De-trends the data by subtracting the expected value (residuals have mean 0)
  - Can examine the impact of multiple covariates simultaneously

- Residuals in logistic, PH, Poisson regression
  - Can also calculate residuals for these models, but the interpretations are more complex
  - Details left for specific courses on categorical and time to event data analysis

- Common misconceptions when using classic linear regression
  - Myth: Need the maximally number of assumptions (as given) earlier to hold to coduct *any* statistical inference
    * Truth: Far fewer assumptions are needed for testing first order trends, and the maximal number of assumptions are only needed for making individual predictions
  - Myth: Ignoring covariates, the outcome Y must follow an approximately Normal distribution (i.e. the marginal distribution of Y is approximately Normal)
    * Truth: Conditional on covariates, the outcome Y must follow an approximately Normal distribution (i.e. the conditional distribution of Y given X is approximately Normal)

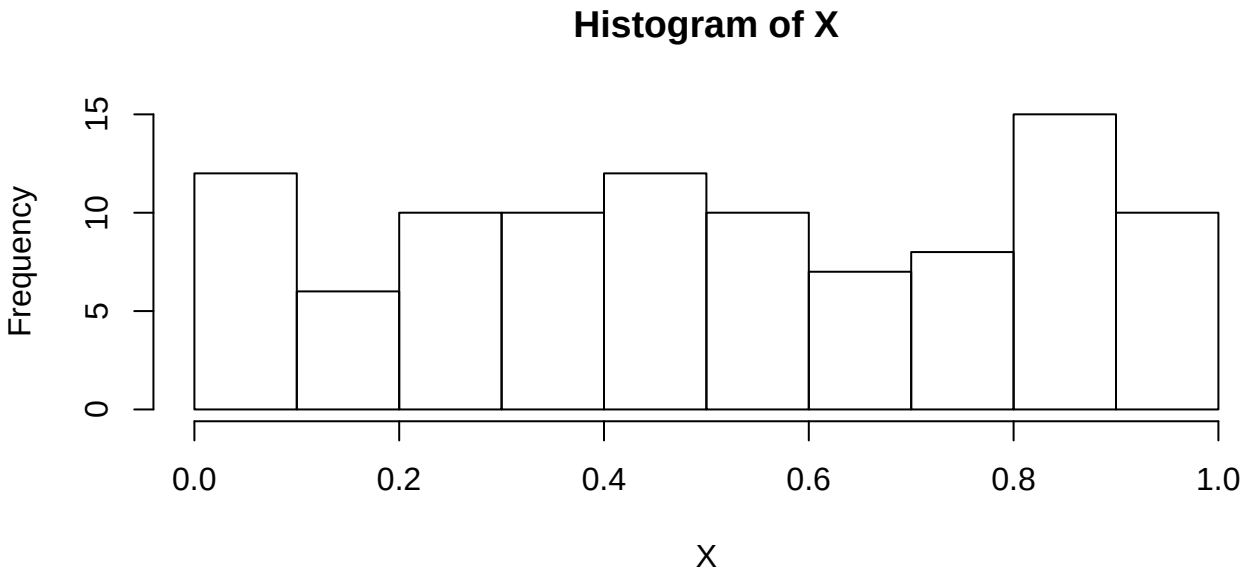
    - Implication: A histogram of Y is not useful for evaluating if the classic linear model is appropriate or not; must look at the residuals

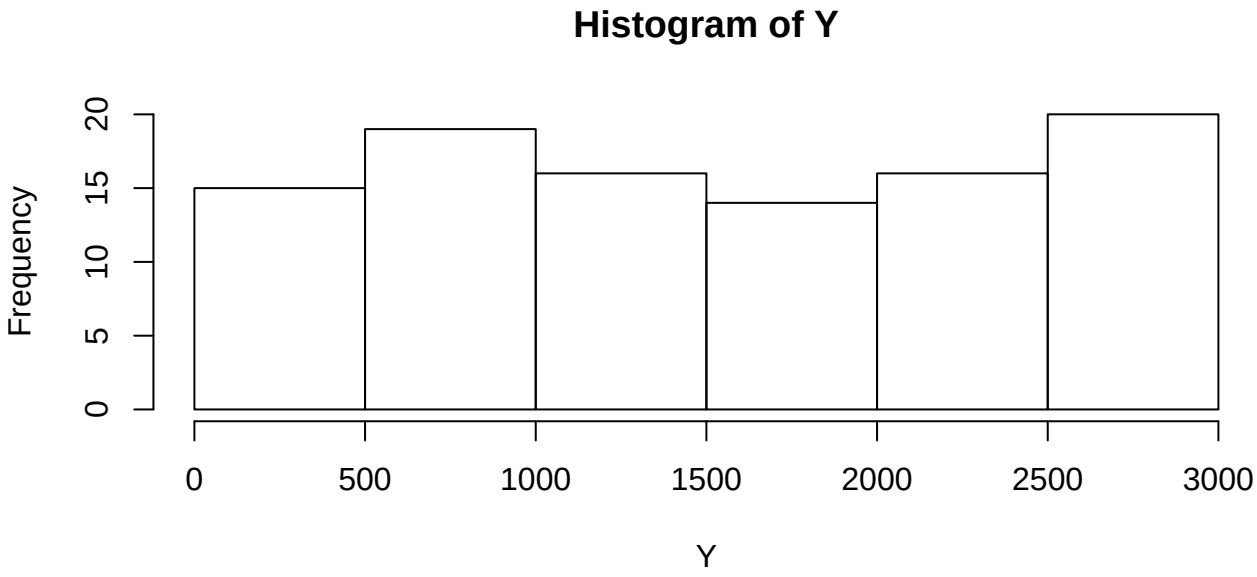
  - Myth: The covariates must follow an approximately Normal distribution
    - Truth: For the linear model to be correct, it does not matter what distribution X follows. However,
      - If want to detect non-linear association, will have greatest power if X is approximately uniform
      - If want to assume a linear trend, will have greatest power if half of the observations are at the maximum values of X and half are at the minimum values of X

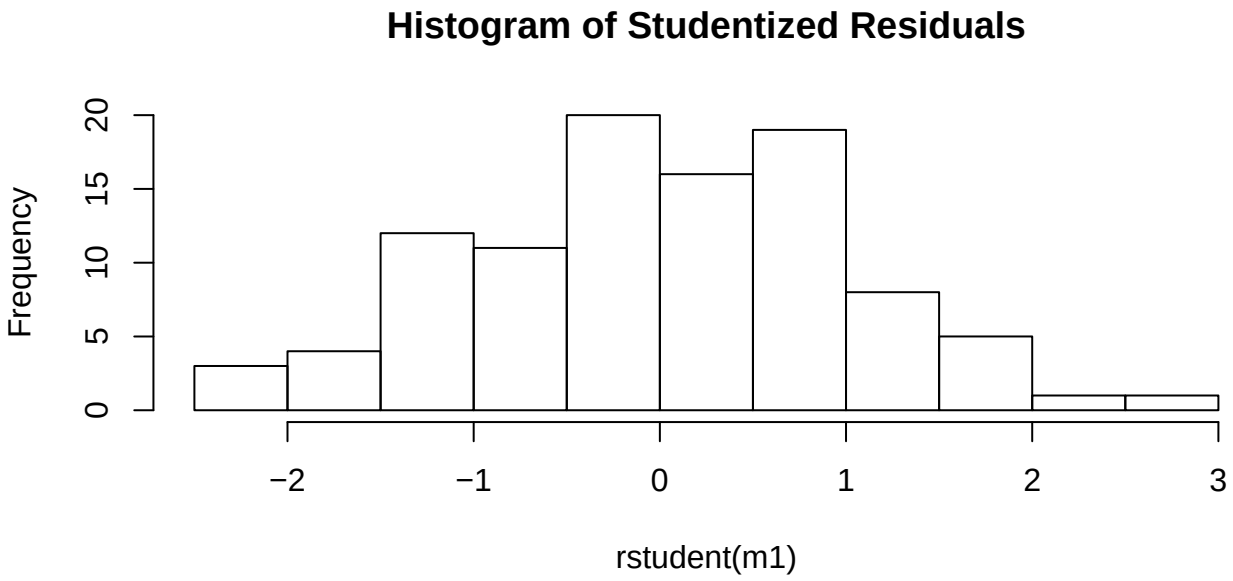
- To assess Normality, examine Normality of the residuals
  - Histogram of the residuals
  - QQ plot
    - Graph ordered residuals versus what we would expect from a Normal distribution having the same mean and variance
    - Truly Normal data will approximately fall on a straight line
  - (Statistical tests of Normality are virtually worthless)

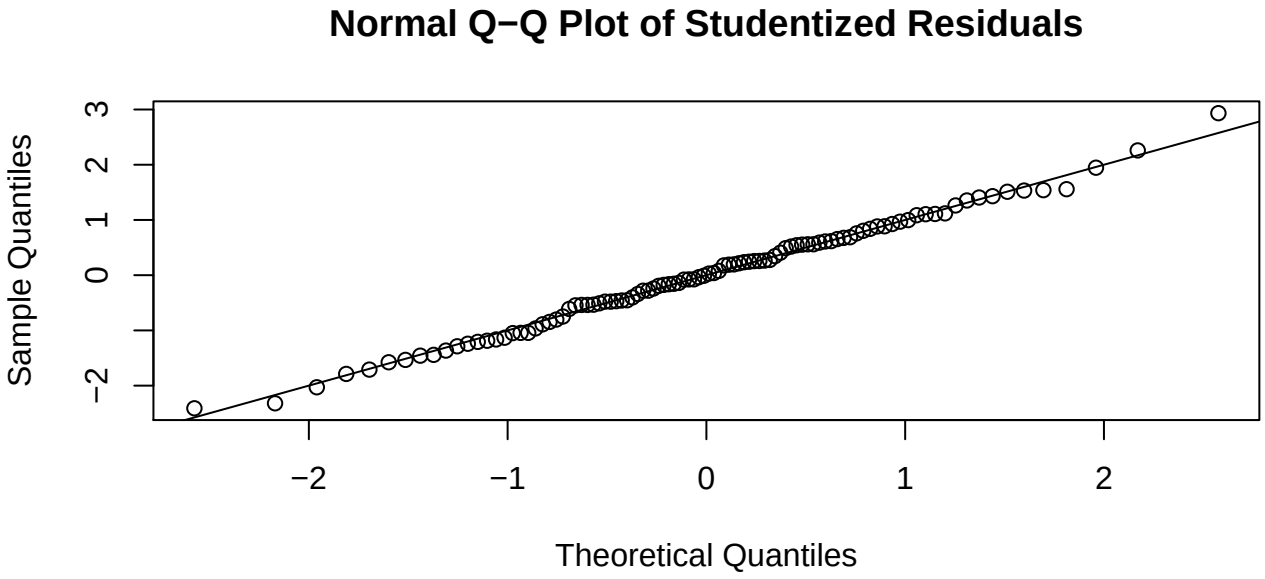
- The following plots show
  - The distribution of Y (clearly not Normal)
  - The distribution of X (clearly not Normal)
  - The distribution of the residuals from a simple linear regression of Y on X (approximately Normal)

– Here, the linear model (even assuming homoscedasticity) would be appropriate






Histogram of Studentized Residuals
Frequency
0
5
10
15
20
−2
−1
0
1
2
3
rstudent(m1)


Normal Q−Q Plot of Studentized Residuals
Sample Quantiles
−2
0
1
2
3
−2
−1
0
1
2
Theoretical Quantiles

15.3.4 Assessing Model Fit

- In this course, we only consider models that are linear in the predictors
  - The summary of the response distribution is predicted to vary in some way according to a linear function of the modeled predictors
  - We can still model transformations of the predictors (e.g. log, splines, polynomials), but the βs are linear
  - We assess model fit by considering linearity on the appropriate scale for each type model
    - Linear regression: Linearity of means
    - Logistic regression: Linearity of log odds
    - Poisson regression: Linearity of log rates
    - PH regression: Linearity of log hazards
  - Example: log odds across strata defined by bilirubin

Strata	bili	log odds	odds
1	0.0 to 1.0	-0.50	0.68
2	1.0 to 2.0	-0.25	0.78
3	2.0 to 3.0	-0.00	1.00
4	3.0 to 4.0	0.25	1.28
5	3.0 to 4.0	0.50	1.65
6	3.0 to 4.0	0.75	2.12

  - A logistic regression model that predicts that, on average, for every 1 unit increase in bilirubin, the log-odds increase by 0.25 units would be a good model fit
  - Evaluation on the odds scale is not possible

15.4 Case Diagnostics

- When using regression models to explore associations between variables, we are always very interested in whether there are individual cases that behave differently from the bulk of the data

- Some cases may be poorly described by the model ("outliers")

- Some cases may be overly influential in fitting the regression model
  - Influential cases affect estimates

  - Highly leveraged cases also affect statistical significance

- Outliers can be judged by evaluating if the observed response is far from that predicted by the model
  - Residual will be large (in absolute value) relative to other residuals

  - Well developed for linear regression, assuming Normally distributed errors

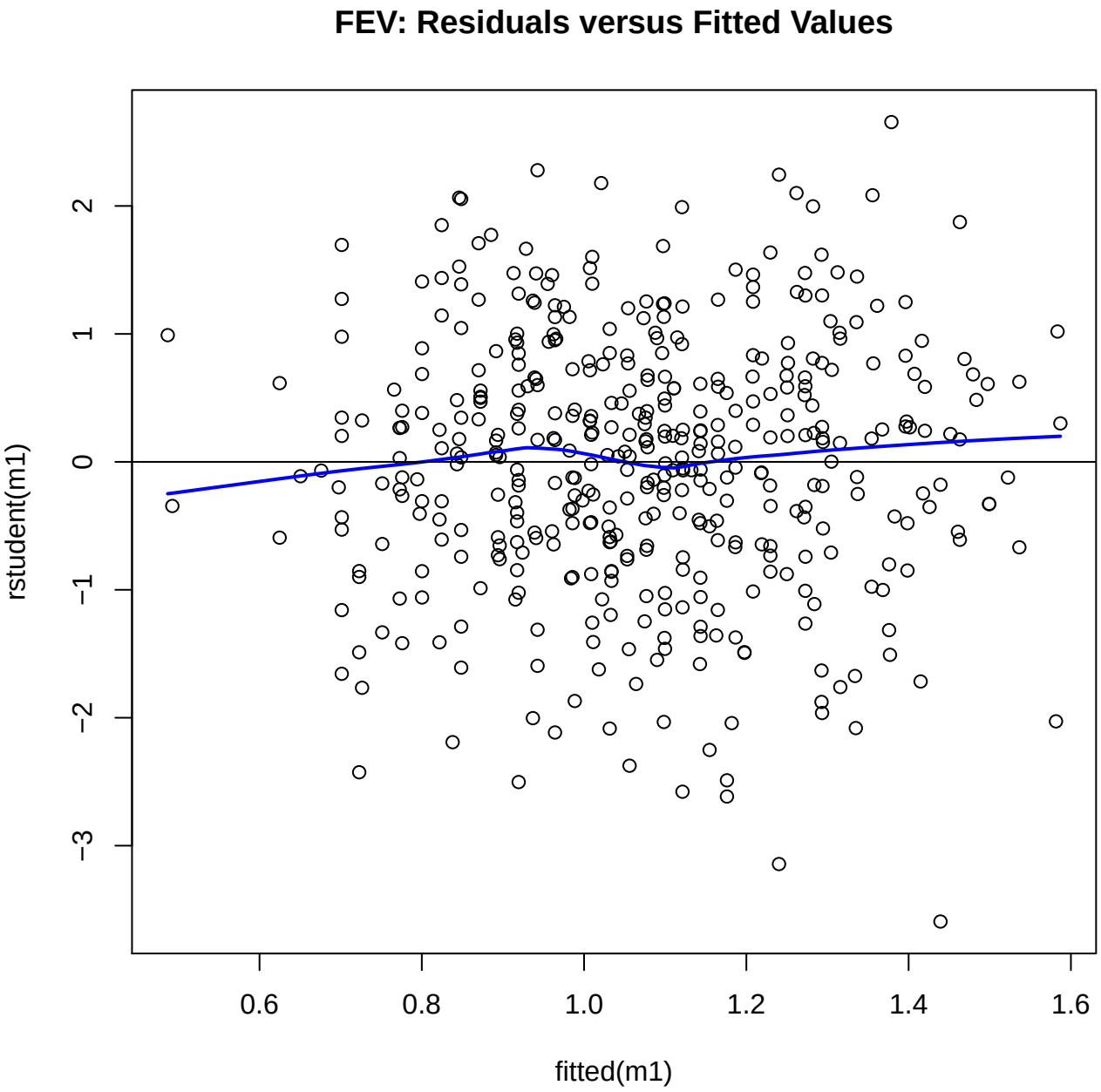
  - Determine ho many standard deviations a single case is from its group mean relative to the sample size of the dataset
    * Often easier to considered studentized (standardize) residuals because they have a variance of 1

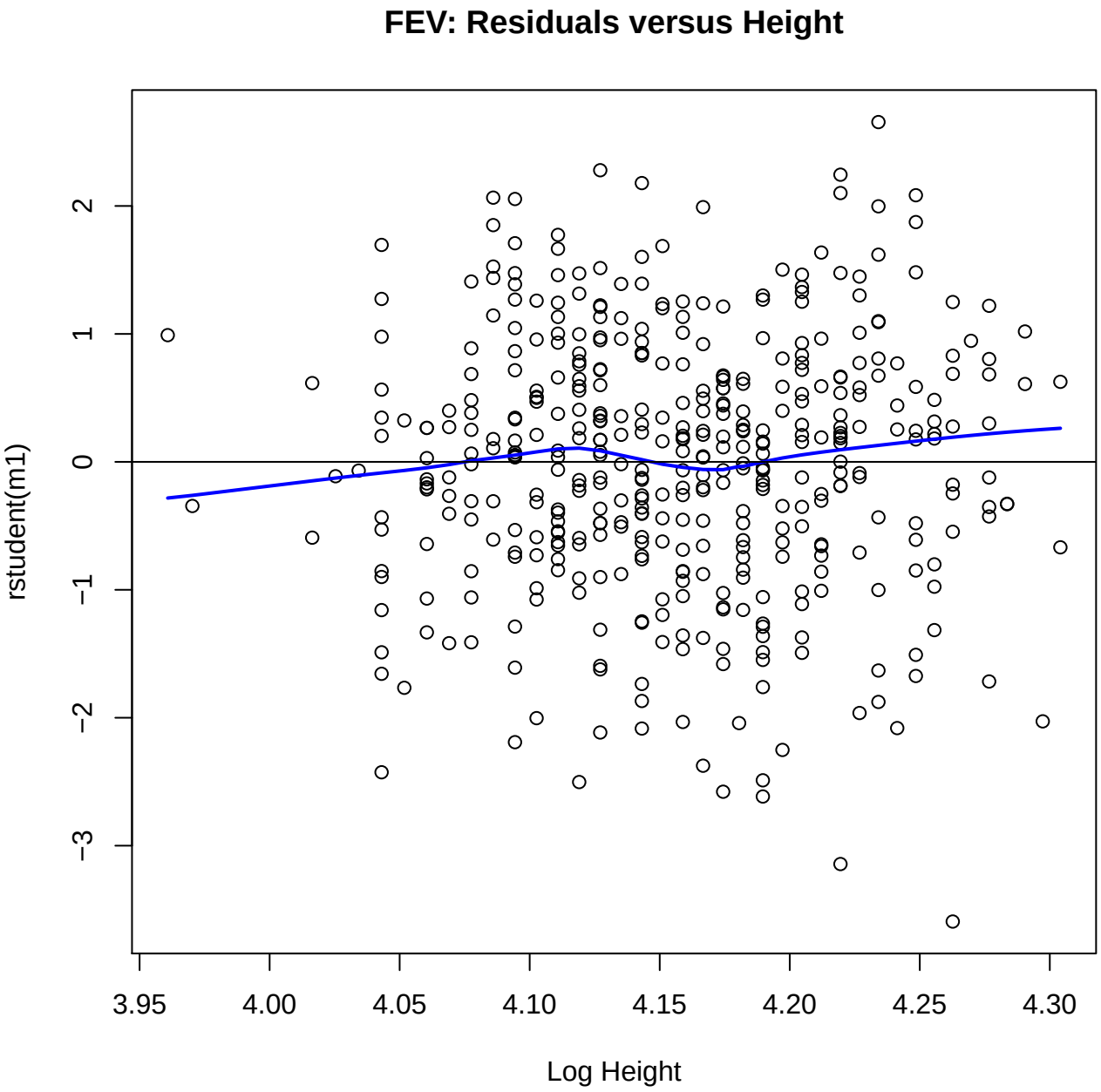
15.4.1 Example: FEV and Smoking

- Consider the following multiple regression model

```
. regress logfev smoker age loght if age >=9

      Source |       SS       df       MS              Number of obs =     439
-------------+------------------------------           F(  3,   435) =  294.73
       Model |   18.352492     3  6.11749734           Prob > F      =  0.0000
    Residual |  9.02895108   435  .020756209           R-squared     =  0.6703
-------------+------------------------------           Adj R-squared =  0.6680
       Total |  27.3814431   438   .06251471           Root MSE      =  .14407

------------------------------------------------------------------------------
      logfev |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |  -.0535896   .0209462    -2.56   0.011     -.094758   -.0124213
         age |   .0215295   .0038187     5.64   0.000      .014024    .0290349
       loght |   2.869658   .1300579    22.06   0.000     2.614038    3.125278
       _cons |  -11.09461   .5201256   -21.33   0.000    -12.11688   -10.07234
------------------------------------------------------------------------------

. predict resid if age>=9, rstud
. predict resid2 if age>=9, rstand
```

- Note that because the regression model restrict to subjects at least 9 years old, our prediction must specify the same restriction
  - Otherwise, Stata would use the model to predict for all ages


FEV: Residuals versus Fitted Values
rstudent(m1)
2
1
0
−1
−2
−3
0.6
0.8
1.0
1.2
1.4
1.6
fitted(m1)

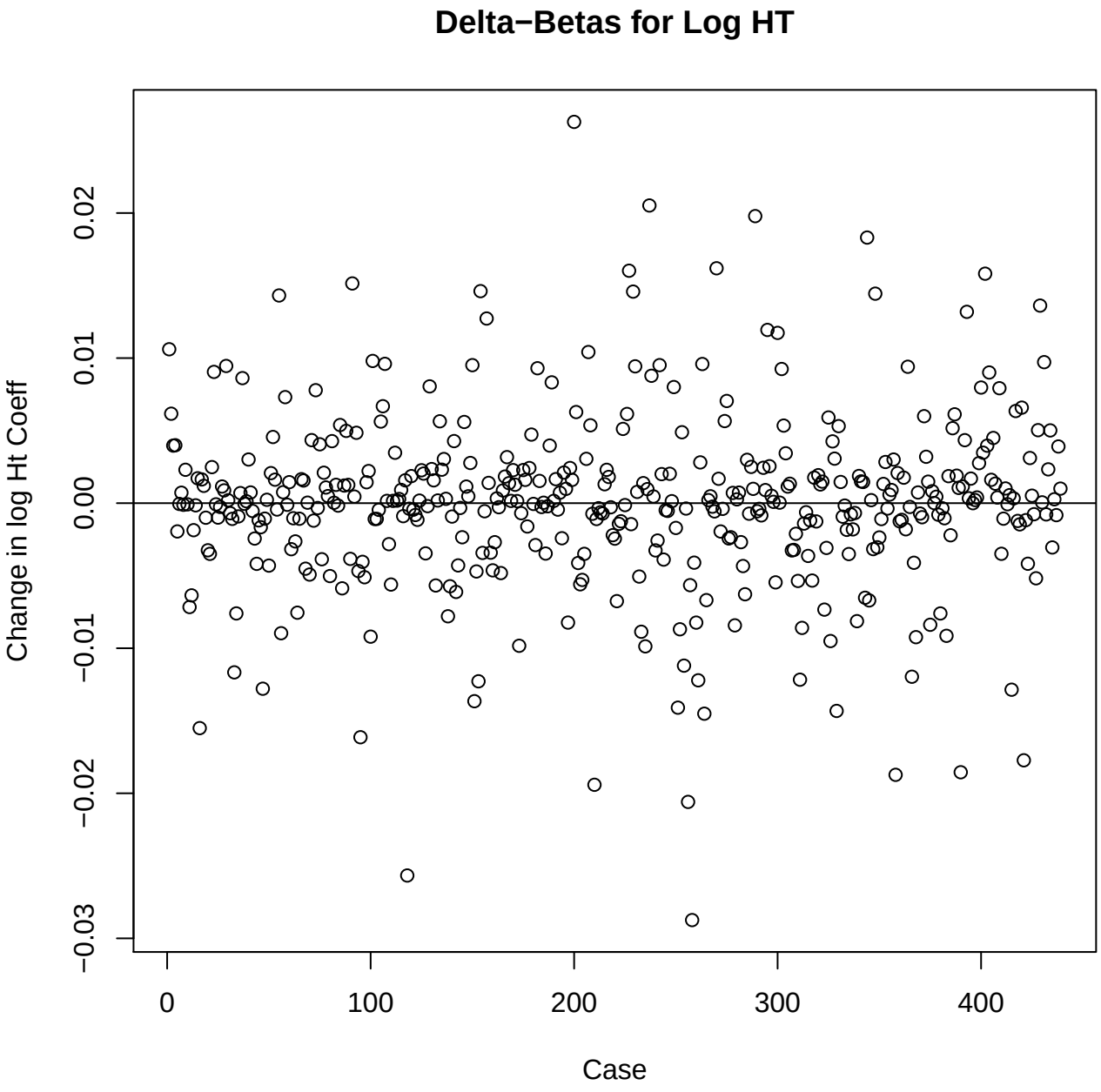

FEV: Residuals versus Height
rstudent(m1)
2
1
0
−1
−2
−3
3.95
4.00
4.05
4.10
4.15
4.20
4.25
4.30
Log Height

- The largest residual is 2.66 standard deviations away from the mean
  - Is this observations abnormally large?
  - Based on a t distribution with 435 degrees of freedom, we would expect 1.2% of the residuals to be this large or larger if Normality holds
  - But, we only observed 1 out of 439 points this large or larger(0.6%)
- The two smallest residuals were -3.593 and -3.143 standard deviations away from the mean
  - Using the t with 435 df, we expect 0.068% and 0.30% of residuals to be this small or smaller if Normality holds
  - Scientifically we might expect more small residuals
    - FEV is highly dependent on effort of subject
    - If a subject did not force out maximal amount of air in 1 second, would lead to low FEV
    - Alternatively, could be due to unmeasured variables (asthma, illness, etc.)

Detecting Influential Cases

- Conceptually, finding influential cases is easy and can be done in any regression analysis
  - Leave one case out of the analysis, and evaluate if the coefficients change enough to alter the scientific conclusions
  - Of course, there can be influential pairs (or triplets, etc.) of observations, but searching through all possibilities quickly becomes very computationally demanding

- Calculating change in coefficients can be accomplished without fitting n different models in linear and logistic regression
  - Stata (linear): `dfbeta` will create delta-beta for each variable
  - Stata (logistic): `predict varname, dbeta`
  - R: `dfbeta(m1)`


Delta−Betas for Log HT
Change in log Ht Coeff
0.02
0.01
0.00
−0.01
−0.02
−0.03
0
100
200
300
400
Case

- With the full dataset, our coefficient for log height was 2.87 with a standard error of (0.13)

- Removing a single data point changed the coefficient by at my 0.03 units, which is not scientifically (or statistically) relevant

- Could keep going...
  - Delta-betas for other coefficients
  - Look at the effect of removing observations on p-values
  - Other diagnostic measures

- Want to avoid data-driven decisions in the model building process to control the experiment-wise error rate

- Experiment-wise error rates ($\alpha = 0.05$ at each decision)

- The more comparisons we make, the more likely we are to make a Type 1 error

Number of Comparisons	Worst Case Scenario	Independent Errors
1	0.0500	0.0500
2	0.1000	0.0975
3	0.1500	0.1426
5	0.2500	0.2262
10	0.5000	0.4013
20	1.0000	0.6415
50	1.0000	0.9231