Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 19, part 5: Analysis of FEV in Children

Jinko Graham

Forced Expiratory Volume (FEV) and Height in Children

- Let's look at a different dataset than the one from the low-birthweight babies to illustrate another sequence of steps in a MLR analysis.
- This dataset was collected to investigate the association between FEV and height in children aged 6-22 years.
- The data were retrieved from

http://www.statsci.org/data/general/fev.html

- The description of the data from the website is as follows:

FEV is an index of pulmonary function that measures the volume of air expelled after one second of constant effort. The data contains determinations of FEV on 654 children ages 6-22 years who were seen in the Childhood Respiratory Desease Study in 1980 in East Boston, Massachusetts. The data are part of a larger study to follow the change in pulmonary function over time in children.

Variables and Reference

- Variables are:
  - FEV: in litres
  - Height: in inches
  - Sex: Male or Female
  - Smoker:
    - Non = not currently a smoker,
    - Current = currently a smoker

- Reference is:
  - Tager, I. B., Weiss, S. T., Rosner, B., and Speizer, F. E. (1979). Effect of parental cigarette smoking on pulmonary function in children. *American Journal of Epidemiology*, **110**: 15-26.

Exploratory Analysis

- We read in the data and summarize.

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/fev.csv")
fev <- read.csv(uu)
head(fev,n=3)
```

```
##     FEV Height    Sex Smoker
## 1 1.708   57.0 Female    Non
## 2 1.724   67.5 Female    Non
## 3 1.720   54.5 Female    Non
```

```
summary(fev)
```

```
##       FEV             Height          Sex          Smoker
##  Min.   :0.791   Min.   :46.00   Female:318   Current: 65
##  1st Qu.:1.981   1st Qu.:57.00   Male  :336   Non    :589
##  Median :2.547   Median :61.50
##  Mean   :2.637   Mean   :61.14
##  3rd Qu.:3.119   3rd Qu.:65.50
##  Max.   :5.793   Max.   :74.00
```

Data Processing

- To simplify the investigation we exclude smokers.
- We start with the original dataframe `fev` and modify it by applying the functions `filter()` and `select()` from the `dplyr` package.

```
library(dplyr)
fev <-
  fev %>% filter(Smoker=="Non") %>% select(-Smoker)
```

- On the right-hand side of the assignment operator <-, note the forward pipe %>%.
- Recall that the forward pipe pushes the output from one function to the input of another.

```
fev <-
  fev %>% filter(Smoker=="Non") %>% select(-Smoker)
```

- We start by pushing the original dataframe `fev` to the `filter()` function as input.
  - The `filter()` function extracts a subset of rows (subjects) from a dataframe.
  - The logical condition `Smoker=="Non"` identifies the subset we are extracting.
- The result of the push to `filter()` is a reduced dataframe containing only the non-smokers.
- This reduced dataframe of non-smokers is then piped as input to the `select()` function.

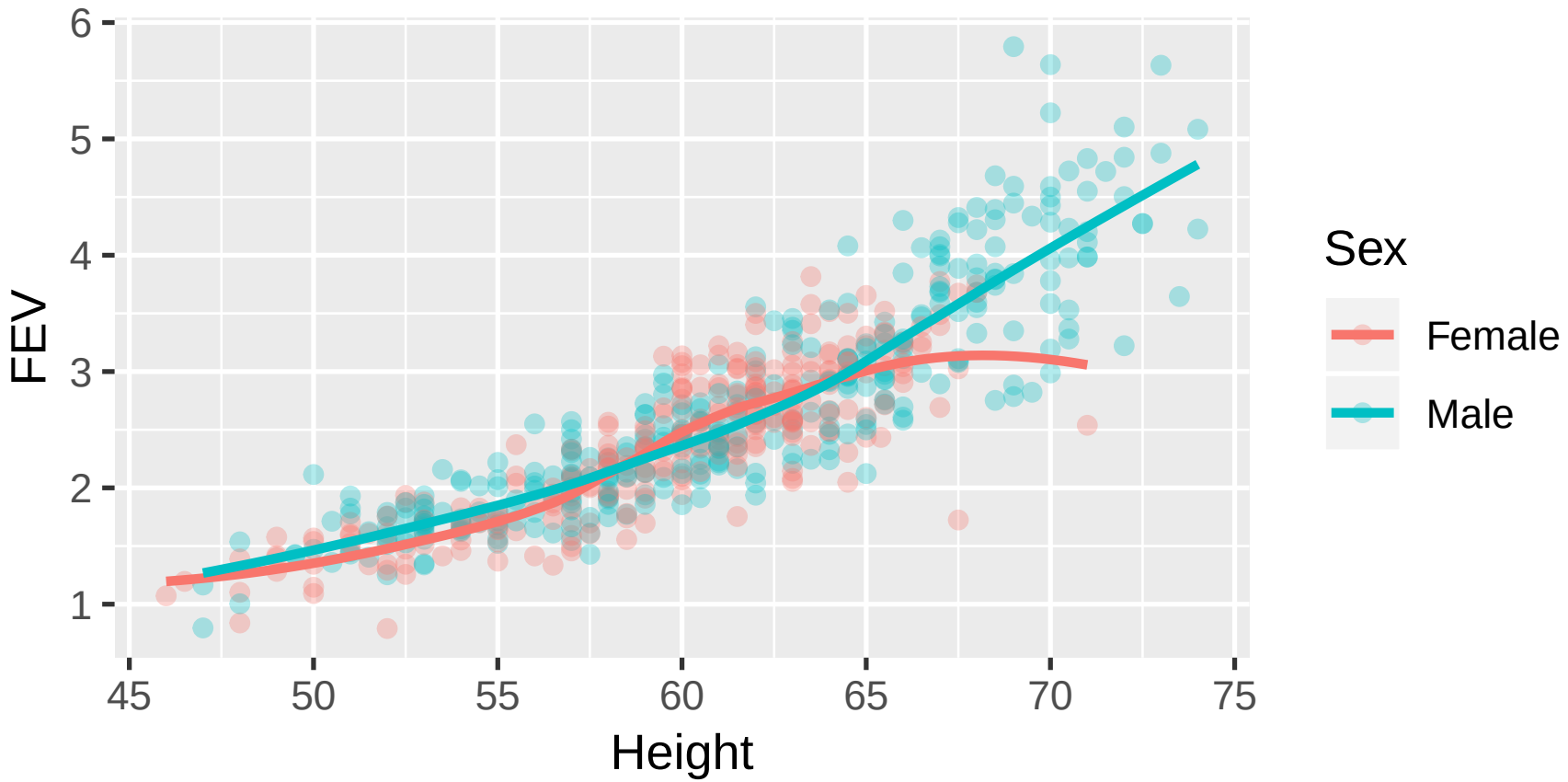
```
fev <-
  fev %>% filter(Smoker=="Non") %>% select(-Smoker)
```

- `select()` is used to select a subset of columns of the dataframe.
- The minus argument `-Smoker` means to keep all columns in the dataframe except for the column for the variable `Smoker`.
- The result of the final push to `select()` is a reduced dataframe containing only the non-smokers and with the column for the variable `Smoker` removed.
- This final dataframe is assigned to the object `fev` with the assignment operator `<-`, thereby changing the value of `fev` from what it was originally.

Exploratory Plot

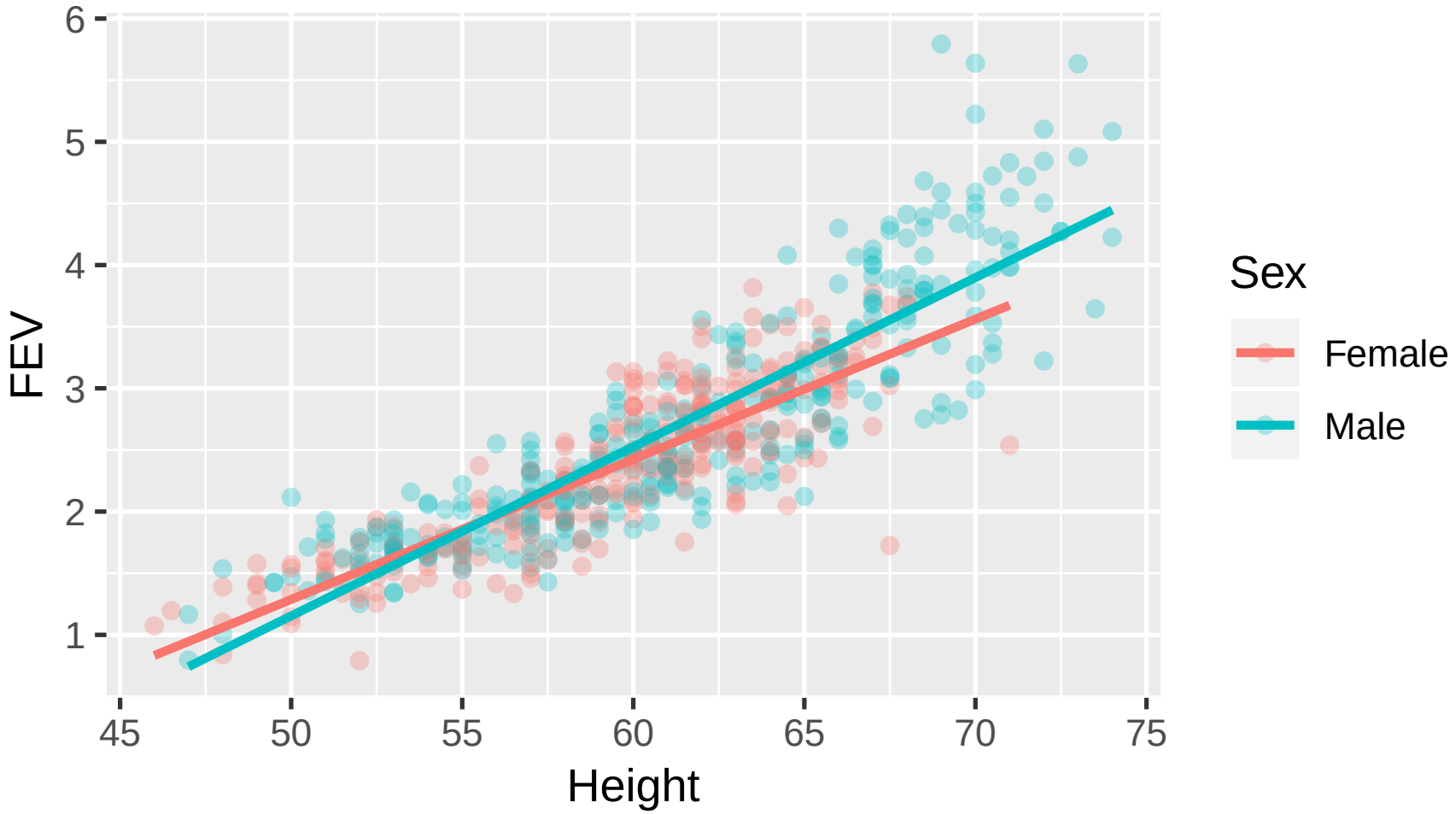
- Suppose we are interested in whether FEV is influenced by height and we think sex might modify the relationship. A scatterplot to visualize these relationships is as follows.

```
library(ggplot2)
ggplot(fev,aes(x=Height,y=FEV,color=Sex)) +
  geom_point(alpha=.3) + geom_smooth(se=FALSE)
```



- The `alpha` argument in the `geom_point()` component can take values between 0 and 1; values closer to 0 make the points more transparent so that the lines are easier to see.
- The lines drawn by the `geom_smooth()` component show a moving average of the points, without shading to indicate their standard errors (`se=FALSE`).
- `FEV` appears to be positively associated with `Height` within each `Sex`.
  - The relationship within males looks reasonably linear.
  - The relationship within females less so.
- However, let's try fitting linear relationships in both ...

```
ggplot(fev,aes(x=Height,y=FEV,color=Sex)) +
  geom_point(alpha=0.3) + geom_smooth(method="lm",se=FALSE)
```



- The Sex-specific slopes for Height look slightly different, but is this difference significant?

Regression Modelling

- To make inference about the difference between the slopes for `Height` we must fit a model that includes statistical interaction between `Height` and `Sex`.

```
fit <- lm(FEV ~ Height+Sex+Height:Sex,data=fev)
summary(fit)$coefficients
```

```
##                    Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    -4.39833385 0.308932467 -14.237202 1.044004e-39
## Height          0.11370695 0.005166719  22.007571 1.166264e-78
## SexMale        -1.30999234 0.384616766  -3.405968 7.045153e-04
## Height:SexMale  0.02352965 0.006357862   3.700874 2.351494e-04
```

- The sex variable is labelled `SexMale` in the output.
  - The `lm()` function converts factors with 2 levels to 0/1 numeric variables for regression.
  - `SexMale` tells us that `lm()` has coded the value `Male` of the factor `Sex` as 1 (and the value `Female` as 0).

Hypothesis Test for Interaction

- Looking at the row of output for the interaction term `Height:SexMale`, we see a tiny *p*value.

```
summary(fit)$coefficients
```

```
##                    Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    -4.39833385 0.308932467 -14.237202 1.044004e-39
## Height          0.11370695 0.005166719  22.007571 1.166264e-78
## SexMale        -1.30999234 0.384616766  -3.405968 7.045153e-04
## Height:SexMale  0.02352965 0.006357862   3.700874 2.351494e-04
```

- The test of statistical interaction rejects the null hypothesis of no interaction at any of the standard significance levels.
- We conclude that there is strong evidence that sex modifies the relationship between FEV and height.

Interpretation

```
summary(fit)$coefficients
```

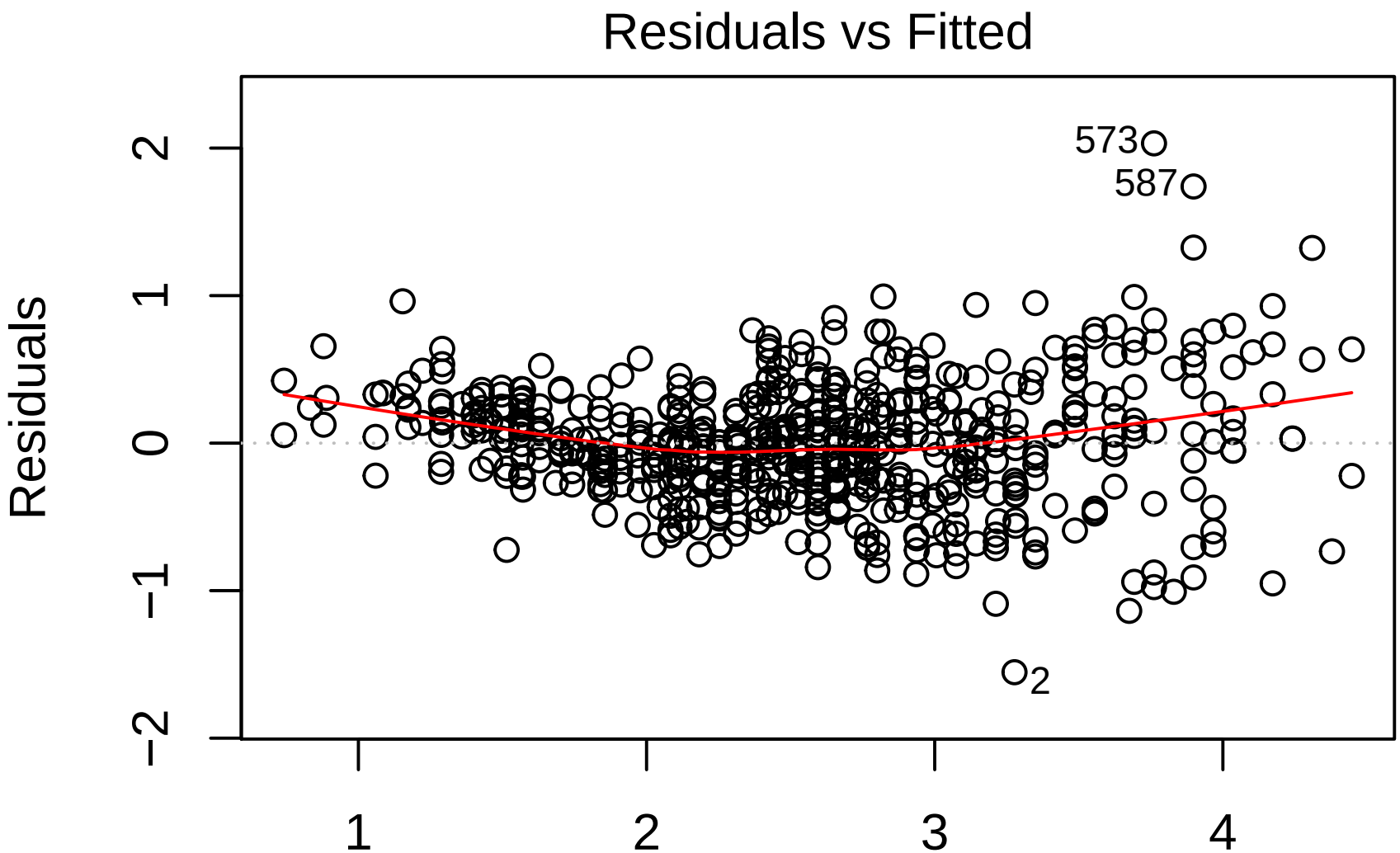
```
##                   Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    -4.39833385 0.308932467 -14.237202 1.044004e-39
## Height          0.11370695 0.005166719  22.007571 1.166264e-78
## SexMale        -1.30999234 0.384616766  -3.405968 7.045153e-04
## Height:SexMale  0.02352965 0.006357862   3.700874 2.351494e-04
```

- We estimate that a one inch increase in female height is associated with a 0.114 litre increase in FEV.
- We estimate that a one inch increase in male height is associated with a $0.114+0.024 = 0.138$ litre increase in FEV.

Residual Diagnostics

- Our first diagnostic plot is of residuals *vs.* fitted values.

```
plot(fit,which=1)
```



Linear predictor

- The red trend line is slightly curved and suggests a possible missed trend in the linear predictor.
  - In our exploratory scatterplot of FEV *vs.*. height, stratified by sex, recall that the relationship between FEV and height in females looked slightly non-linear.
  - This may explain the slight curve in the red line.
- Non-linear regression is beyond the scope of this course.
  - Please see Stat 452 (Statistical Learning and Prediction) if interested.

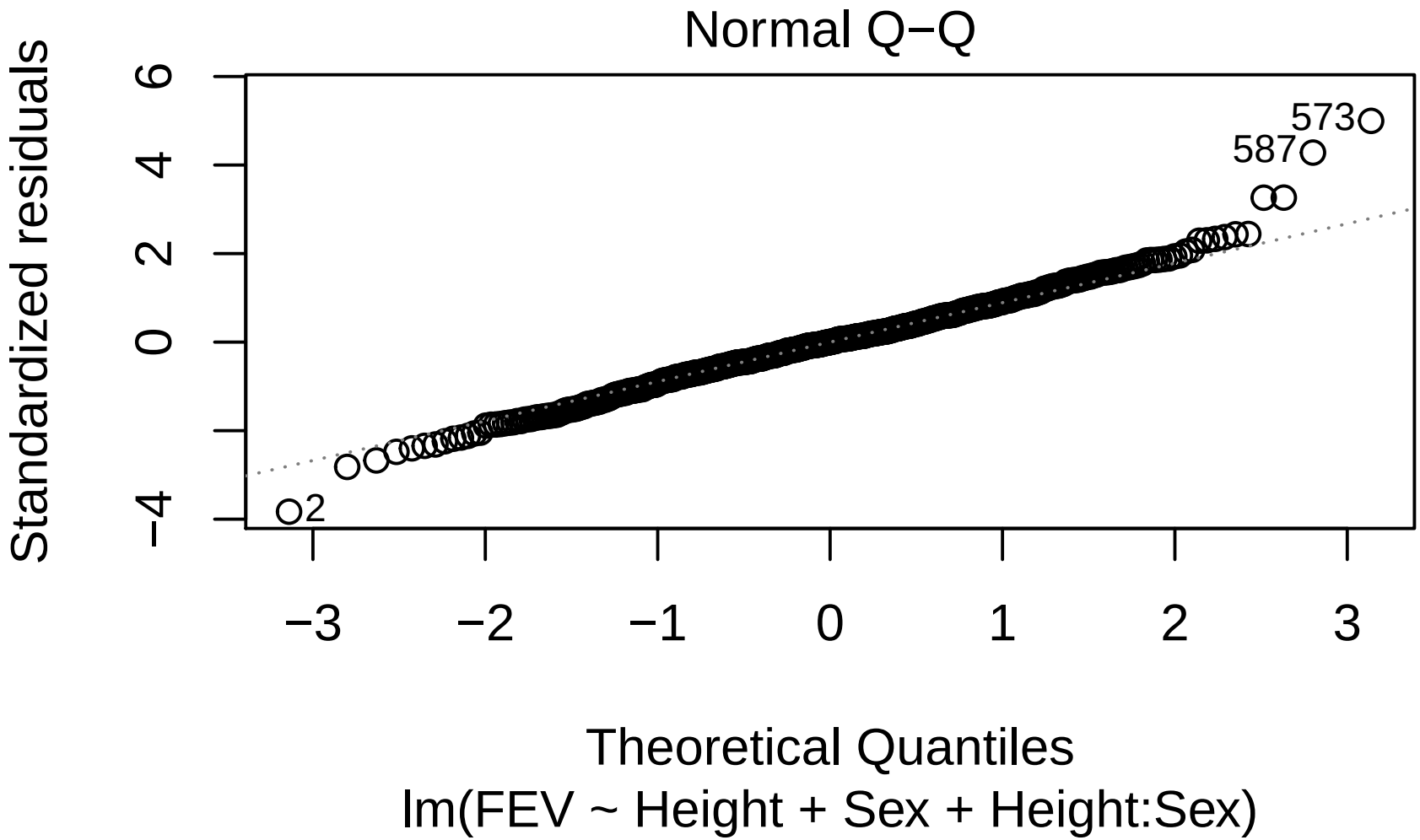
Constant error SD

- The plot of residuals *vs.* fitted values gives an impression of a funnel with a narrow end at the left (for small fitted values) and a wide end at the right (for large fitted values).
- One approach to correct non-constant error SD is to **transform** the response variable.
  - However, transformations are beyond the scope of this course.
  - Please see Stat 452 if interested.

Q-Q Plot

- The 2nd residual diagnostic is the Q-Q plot to check the assumption of normal errors.

```
plot(fit,which=2)
```



- Except for a few outliers (see below), this plot suggests that the distribution of standardized residuals is fairly close to a normal distribution.
- The cases which depart from the straight line at the left and right of the plot are in the left and right tails of the empirical distribution of the standardized residuals.
  - The tails are subject to more random fluctuation than the centre
  - Hence, we don't necessarily interpret departures from the straight line in the tails as suggesting non-normal errors.
  - However, large departures for a few points could suggest they are outliers.

Obvious Outliers

- The Q-Q plot has five cases whose standardized residuals are greater than 3 in absolute value (four positive and one negative).
- We can look at these extreme residuals by printing the head and tail ends of the sorted values.

```
head(sort(rstandard(fit)))
```

```
##         2       453       507       539       563       331
## -3.830125 -2.819514 -2.673469 -2.478851 -2.400295 -2.343346
```

```
tail(sort(rstandard(fit)))
```

```
##       358       324       435       580       587       573
## 2.435521 2.442184 3.262644 3.264111 4.281478 4.997344
```

- According to our ± 3 rule, cases 2, 435, 580, 587 and 573 are obvious outliers.

Remove Outliers?

- People often ask if they should discard the obvious outliers from the dataset and interpret the model that has been fitted without them.
- Not advisable: Try to follow up on the outliers first.
  - They could be data-recording errors, and are worth double-checking if you have access to the original data source.
  - Should remove them and say why if you find that they are data-recording errors.
  - If they are not data-recording errors, they may reveal something interesting about the natural phenomenon that you are investigating.
- Unfortunately, we don't have access to the data source and so can't check these data.
  - Should discuss these outliers in your scientific report.
  - Point them out and indicate why they are unusual. Say you weren't able to follow them up.
  - Can discuss the results of the analysis with the outliers included and then discuss what happens when the outliers are taken out.