Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Practical Aspects of Normal Linear Models

Mauricio Sadinle
Department of Biostatistics
University of Washington
msadinle@uw.edu

Practical Aspects of Normal Linear Models

- Data example
- Functions to fit linear models in `R`
- Diagnostics for detecting violations of the model's assumptions

Assumptions of the Normal Linear Model

So far, we have derived results to perform inference under the normal linear model:

$$\mathbf{Y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).$$

For these inferences to be valid we need the assumptions of the model to hold:

- Homoscedasticity: $\text{var}(\mathbf{Y} \mid \mathbf{X}) = \sigma^2\mathbf{I}_n$
  - The errors (and thus the Y_i's at each $\mathbf{x}_i$) have constant variance
- Normality: $\mathbf{Y} \mid \mathbf{X} \sim N[\mu(\mathbf{X}), \sigma^2\mathbf{I}_n]$
  - The errors (and thus the Y_i's at each $\mathbf{x}_i$) are normally distributed
- Correct specification of the linear model: $\text{E}(\mathbf{Y} \mid \mathbf{X}) = \mu(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$
  - There is a linear relationship between Y_i and $\mathbf{x}_i$ given by $\text{E}(Y_i \mid \mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta}$

Are these assumptions reasonable given our data?

Assumptions of the Normal Linear Model

Other considerations:

- Are there single points that determine our conclusions? Outliers?
- Did we skip other possible covariates that should be included in the model?
- Are the errors independent? (see 571)

Diagnostics

- Goal: to assess the adequacy of assumptions underlying a confirmatory analysis, or to be used for model exploration

- Not to be viewed as a way of avoiding careful initial thought about the model, especially in a confirmatory analysis

- Inference (confirmatory analysis) requires the model to *not* have been chosen on the basis of the current data set, otherwise, inferences are invalid!

Diagnostics

- In a frequentist analysis, the operating characteristics seen in this class (e.g. coverage of confidence intervals, error rates of hypothesis tests) are based upon repeated sampling under the same fixed model

Diagnostics

- In a frequentist analysis, the operating characteristics seen in this class (e.g. coverage of confidence intervals, error rates of hypothesis tests) are based upon repeated sampling under the same fixed model

- Inferences under a model selected using the data will lead to understatements of variability or overconfidence in your results

Diagnostics

- In a frequentist analysis, the operating characteristics seen in this class (e.g. coverage of confidence intervals, error rates of hypothesis tests) are based upon repeated sampling under the same fixed model

- Inferences under a model selected using the data will lead to understatements of variability or overconfidence in your results

- Incidentally, *post-selection* inference is currently a very hot topic (lot's of new work, none of it covered here)

Example: Exploratory Analysis

- FEV: forced expiratory volume. FEV1: amount of air you can force from your lungs in one second.

- Data from 654 children and youths ages 3–19 in East Boston, 1980. (Childhood Respiratory Disease Study).

- For more information visit: `http://www.statsci.org/data/general/fev.html`

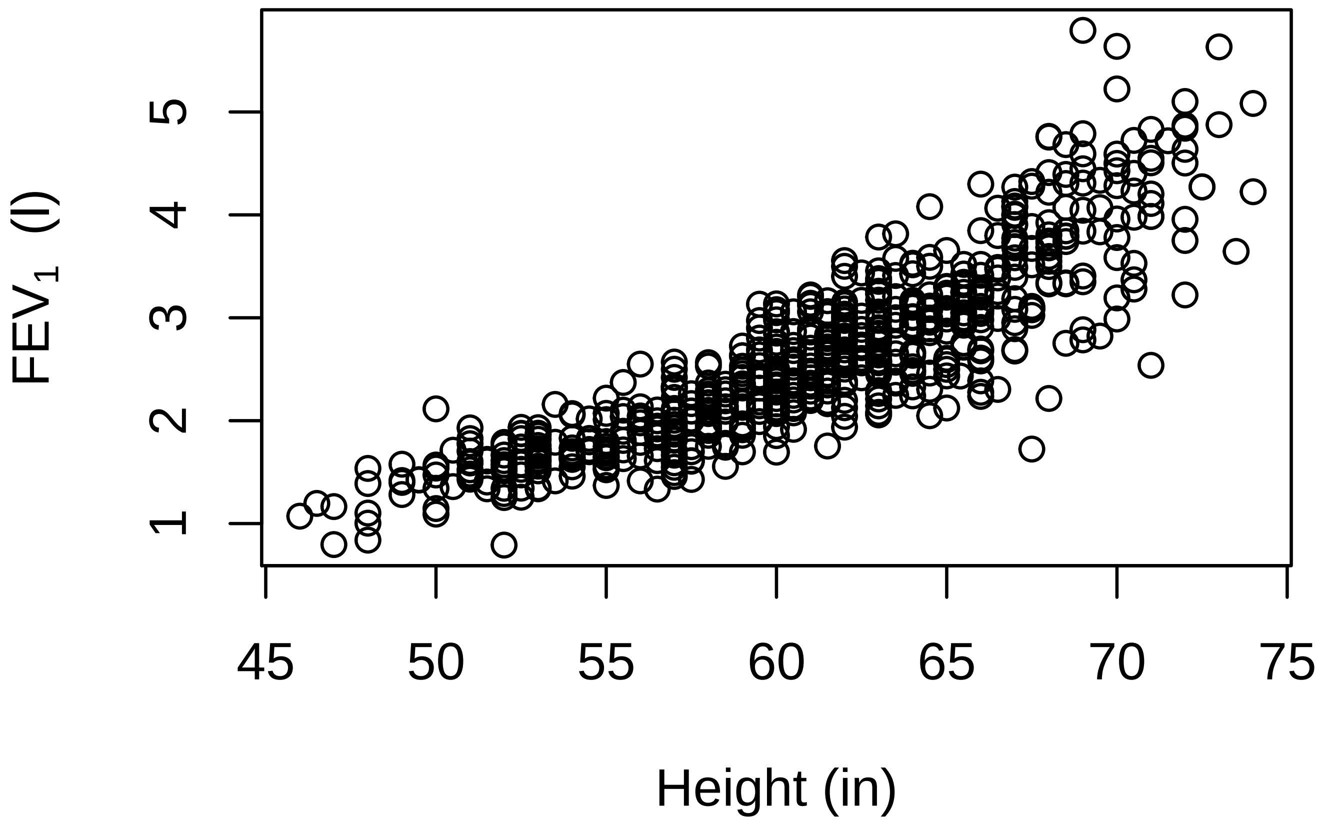
```
> url <- "http://www.statsci.org/data/general/fev.txt"
> data  <- read.table(file = url, header = T, sep="\t", stringsAsFactors = F)
> data$Sex <- factor(data$Sex, levels=c("Male","Female"), labels=c(0,1))
> data$Smoker <- factor(data$Smoker, levels=c("Non","Current"), labels=c(0,1))
> head(data)
```

```
    ID Age   FEV Height Sex Smoker
1  301   9 1.708   57.0   1      0
2  451   8 1.724   67.5   1      0
3  501   7 1.720   54.5   1      0
4  642   9 1.558   53.0   0      0
5  901   9 1.895   57.0   0      0
6 1701   8 2.336   61.0   1      0
```

Example: Exploratory Analysis

Let's first consider an *exploratory analysis*:

Can FEV1 be predicted from something more easily measured, such as height?



Example: Exploratory Analysis

Let's start assuming $\quad \text{E(FEV1} \mid \text{Height)} = \beta_0 + \beta_1 \text{Height}$

(to illustrate how to detect problems with a model; this model is not going to be the greatest)

The `R` function `lm` fits linear models:

```
> model <- lm(FEV ~ Height, data=data)
```

Example: Exploratory Analysis

Let's start assuming $\quad \text{E}(\text{FEV1} \mid \text{Height}) = \beta_0 + \beta_1 \text{Height}$

(to illustrate how to detect problems with a model; this model is not going to be the greatest)

The `R` function `lm` fits linear models:

```
> model <- lm(FEV ~ Height, data=data)
```

Exact t-test assuming normal i.i.d. errors

```
> coef(summary(model))
```

```
              Estimate  Std. Error   t value      Pr(>|t|)
(Intercept) -5.4326788 0.181459887 -29.93873 1.453077e-124
Height       0.1319756 0.002954958  44.66241 1.574556e-200
```

Example: Exploratory Analysis

Let's start assuming $\mathrm{E}(\mathrm{FEV1} \mid \mathrm{Height}) = \beta_0 + \beta_1 \mathrm{Height}$

(to illustrate how to detect problems with a model; this model is not going to be the greatest)

The `R` function `lm` fits linear models:

```
> model <- lm(FEV ~ Height, data=data)
```

Exact t-test assuming normal i.i.d. errors

```
> coef(summary(model))

              Estimate  Std. Error   t value      Pr(>|t|)
(Intercept) -5.4326788 0.181459887 -29.93873 1.453077e-124
Height       0.1319756 0.002954958  44.66241 1.574556e-200
```

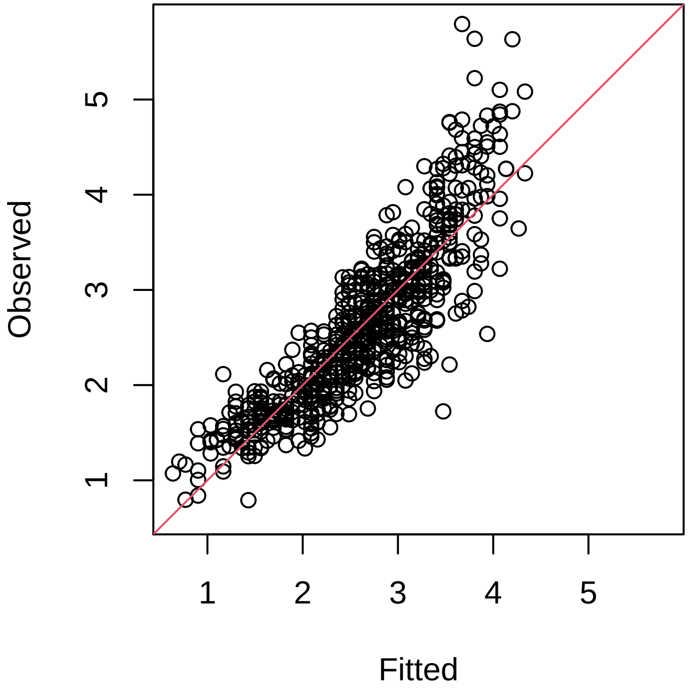
Confidence intervals

```
> confint(model, level = 0.95)

                 2.5 %    97.5 %
(Intercept) -5.7889951 -5.076363
Height       0.1261732  0.137778
```

Example: Exploratory Analysis

Does the model have a decent fit? An observed vs. fitted plot:

```
> limits <- range(data$FEV, fitted(model))
> plot(fitted(model), data$FEV, xlab="Fitted", ylab="Observed", xlim=limits, ylim=limits)
> abline(a=0, b=1, col=2)
```



We want all points to lie around the diagonal. Not a great model, as expected.

Diagnostics: Residuals

Goal: Identify points that are not well fit by the model

- Raw residuals: $\boldsymbol{e} = (e_1, \ldots, e_n)^T$, with $e_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}$
- 'Hat' matrix: $\mathbf{P} = \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- Note that we had used **P** before to denote **H**, but the literature on diagnostics prefers **H** for 'hat' matrix, since **H** 'puts a hat' on **Y**:

$$\mathbf{H}\mathbf{Y} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

Diagnostics: Residuals

Goal: Identify points that are not well fit by the model

- Raw residuals: $\boldsymbol{e} = (e_1, \dots, e_n)^T$, with $e_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}$
- 'Hat' matrix: $\mathbf{P} = \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- Note that we had used $\mathbf{P}$ before to denote $\mathbf{H}$, but the literature on diagnostics prefers $\mathbf{H}$ for 'hat' matrix, since $\mathbf{H}$ 'puts a hat' on $\mathbf{Y}$:

$$\mathbf{HY} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{Y}}$$

- Under i.i.d. errors $\text{var}(\boldsymbol{e} \mid \mathbf{X}) = \text{var}[(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \mid \mathbf{X}] = \sigma^2(\mathbf{I}_n - \mathbf{H})$, so letting h_i be the ith diagonal element of $\mathbf{H}$,

$$\text{var}(e_i \mid \mathbf{X}) = \sigma^2(1 - h_i) \quad \text{and} \quad \frac{e_i}{\widehat{\sigma}\sqrt{(1-h_i)}}$$

  has mean zero and variance 1 in large samples if the model is correct.

- These residuals are called both 'standardized' (`MASS R` package, Wakefield) and 'internally studentized' (Seber and Lee).

Diagnostics: Residuals

- The value of $\widehat{\sigma}^2$ can be influenced by poorly fit points.

- May prefer to use 'externally studentized' (Seber and Lee) or 'studentized' (`MASS` `R` package) residuals

$$\frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{(1-h_i)}}$$

  that replace $\widehat{\sigma}^2$ with $\widehat{\sigma}^2_{(i)}$: the estimate of σ^2 based on all observations but the ith (Why is this better?)

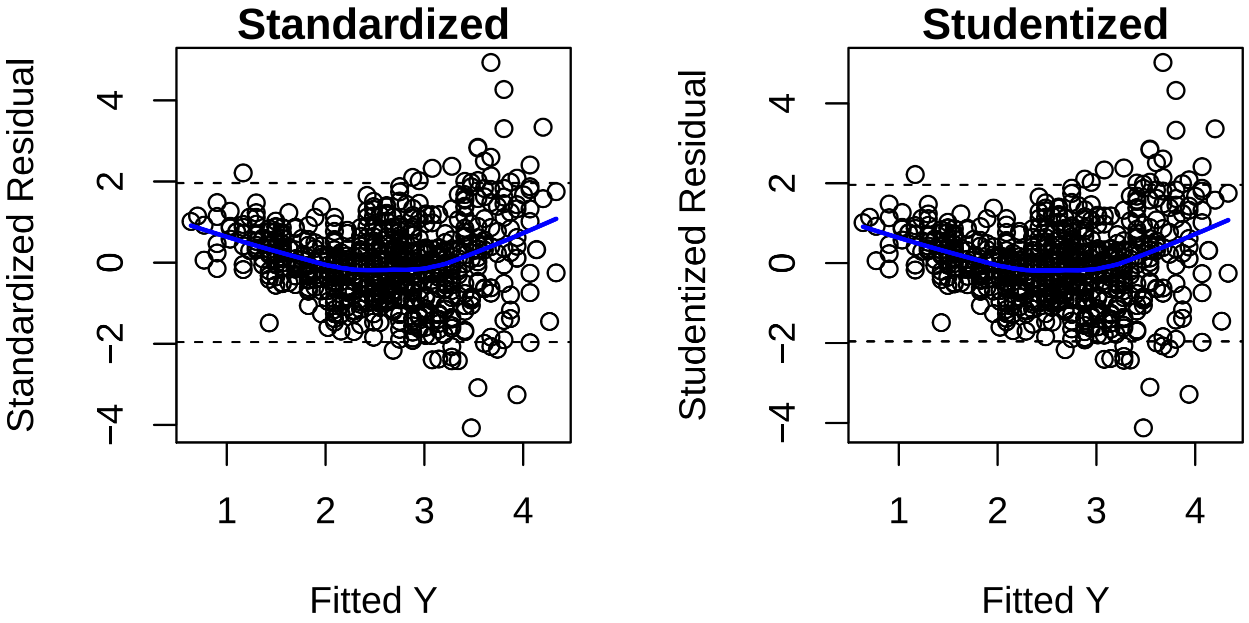
- Under normality of **Y**, can show that these have a t_{n-k-2} distribution conditional on **X**. (Seber and Lee, Chapter 10.)

Example

Continuing with our example

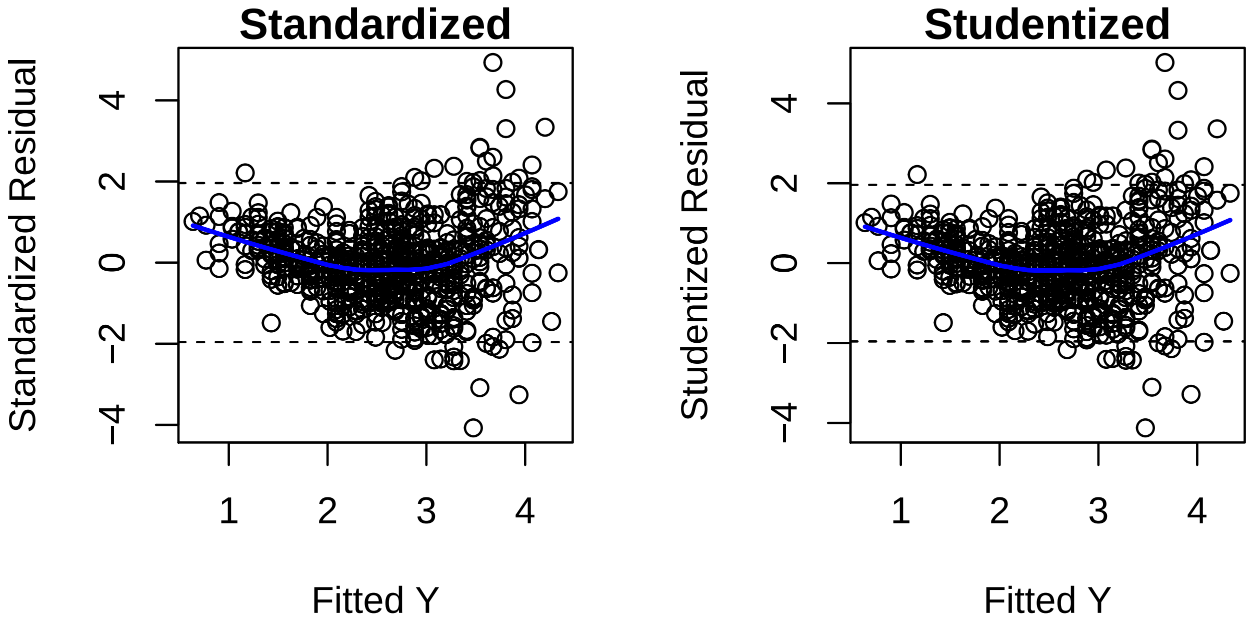
```
> library(MASS)
> estd <- stdres(model) # Standardized residuals
> estud <- studres(model) # Studentized residuals
> yhat <- predict(model)

> par(mfrow = c(1,2))
> plot(yhat, estd, ylab = "Standardized Residual", xlab = "Fitted Y")
> lines(lowess(yhat, estd), lwd = 2, col = "blue")
> abline(h = 1.96, lty = 2)
> abline(h = -1.96, lty = 2)
> title(main = "Standardized")
> plot(yhat, estud, ylab = "Studentized Residual", xlab = "Fitted Y")
> lines(lowess(yhat, estud), lwd = 2, col = "blue")
> abline(h = 1.96, lty = 2)
> abline(h = -1.96, lty = 2)
> title(main = "Studentized")
```

Example



Linearity? Constant variance? Outliers?

Example



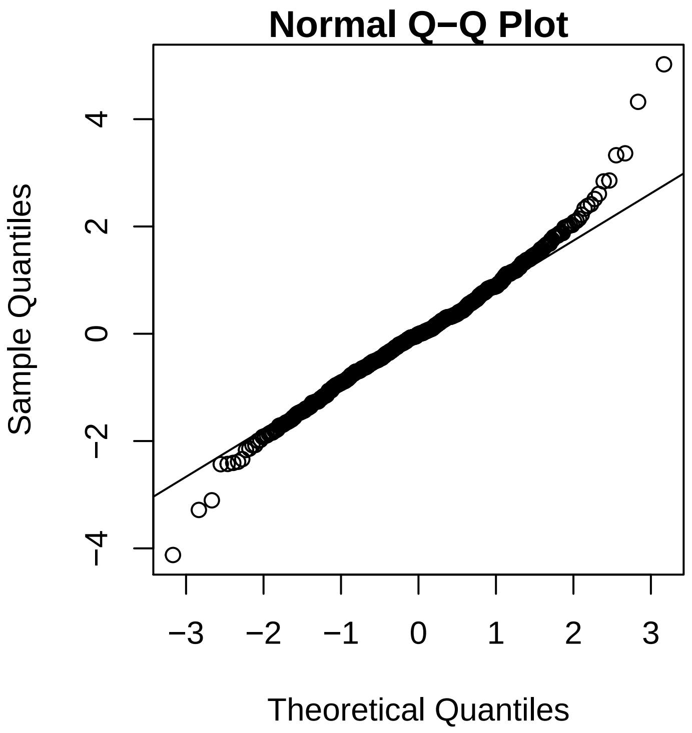
Linearity? Constant variance? Outliers?

- Homoscedasticity assumption is not justified
- Mean model (regression function) is misspecified
- How about normality?

Example

If we still want to check normality of errors, we may check QQ plot of some type of the residuals: this compares the quantiles of a dataset with those of the standard normal.

```
> qqnorm(estud)
> qqline(estud)
```



In this case we have evidence of heavier tails than under the standard normal.

Example

Also, there are lots of tests of normality available:

- Kolmogorov-Smirnov
- Lilliefors
- Shapiro-Wilk
- Anderson-Darling
- Cramer-von Mises
- D'Agostino
- Anscombe-Glynn
- D'Agostino-Pearson
- Jarque-Bera
- Martinez-Iglewicz

Deletion Diagnostics

- To measure the influence of the ith observation on the coefficient estimates, it can be useful to compute
$$\Delta\boldsymbol{\beta}_{(i)} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}$$
where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the OLS estimate of $\boldsymbol{\beta}$ based on all observations but the ith.

- Measures how much higher or lower each element of $\widehat{\boldsymbol{\beta}}$ becomes when the ith observation is added to the data.

- Observations for which this difference is 'large' for an element β_j of $\boldsymbol{\beta}$ have a high influence on the estimation of β_j.

- These diagnostics are a good practice regardless of whether your inferences are fully parametric.

Deletion Diagnostics

- Usually only care about $\Delta\beta$'s for coefficients of interest, or linear combinations of them.

- Can identify important data errors, or influential single observations that should be reported.

- Computation:

$$\Delta\boldsymbol{\beta}_{(i)} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T e_i}{1 - h_i}$$

  (details in Seber & Lee, Theorem 10.1).

- This is implemented in `R` as `dfbeta`

Deletion Diagnostics

- We can also examine influence on the test statistic or p-value, but this does depend on the assumptions of the model being used

- Easy to compute difference in test statistic if the sample size is not large

$$\Delta t_{j(i)} = \frac{\widehat{\beta}_j}{\text{se}(\widehat{\beta}_j)} - \frac{\widehat{\beta}_{j(i)}}{\text{se}(\widehat{\beta}_{j(i)})}$$

  and the associated $\Delta p_{j(i)}$ (the change in p-values)

- For homoscedastic linear models

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{(i)}) = \widehat{\sigma}^2_{(i)}(\mathbf{X}^T_{(i)}\mathbf{X}_{(i)})^{-1},$$

  with $\mathbf{X}_{(i)}$ obtained from removing the ith row from $\mathbf{X}$, and

$$\widehat{\sigma}^2_{(i)} = \frac{1}{n-k-2}\left[(n-k-1)\widehat{\sigma}^2 - \frac{e_i^2}{1-h_i}\right]$$

  (see SL, p. 268)

- se$(\widehat{\beta}_{j(i)})$ is the square root of the jth diagonal entry of $\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}_{(i)})$

Deletion Diagnostics

- `R` implements an approximation of $\Delta t_{j(i)}$ as `dfbetas`, given by

$$\Delta\beta_{j(i)}/\sqrt{\widehat{\sigma}^2_{(i)}(\mathbf{X}^T\mathbf{X})^{-1}_j}$$

with $(\mathbf{X}^T\mathbf{X})^{-1}_j$ being the jth diagonal entry of $(\mathbf{X}^T\mathbf{X})^{-1}$

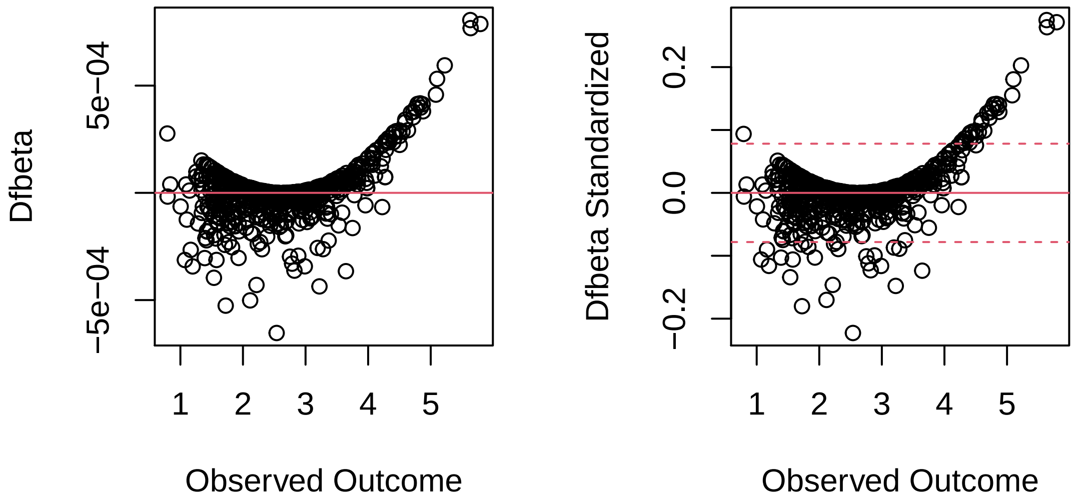
By looking at delta-betas, we can see whether the estimates and inferences are unduly influenced by a single observation.

- If nothing weird, lucky you!
- If yes, examine data points for validity
  - if not valid, omit or correct
  - if valid, make scientific judgement of which $\widehat{\beta}$ is primary <u>and report both</u>.

Example

Continuing with our example

```
> deltabetas <- dfbeta(model)
> deltabetas_st <- dfbetas(model)

> par(mfrow=c(1,2))
> plot(data$FEV, deltabetas[,2], ylab = "Dfbeta", xlab = "Observed Outcome")
> abline(h=0, col=2)
> plot(data$FEV, deltabetas_st[,2], ylab = "Dfbeta Standardized", xlab = "Observed Outcome")
> abline(h=0, col=2);
> abline(h=2/sqrt(nrow(data)), lty=2, col=2); abline(h=-2/sqrt(nrow(data)), lty=2, col=2)
```



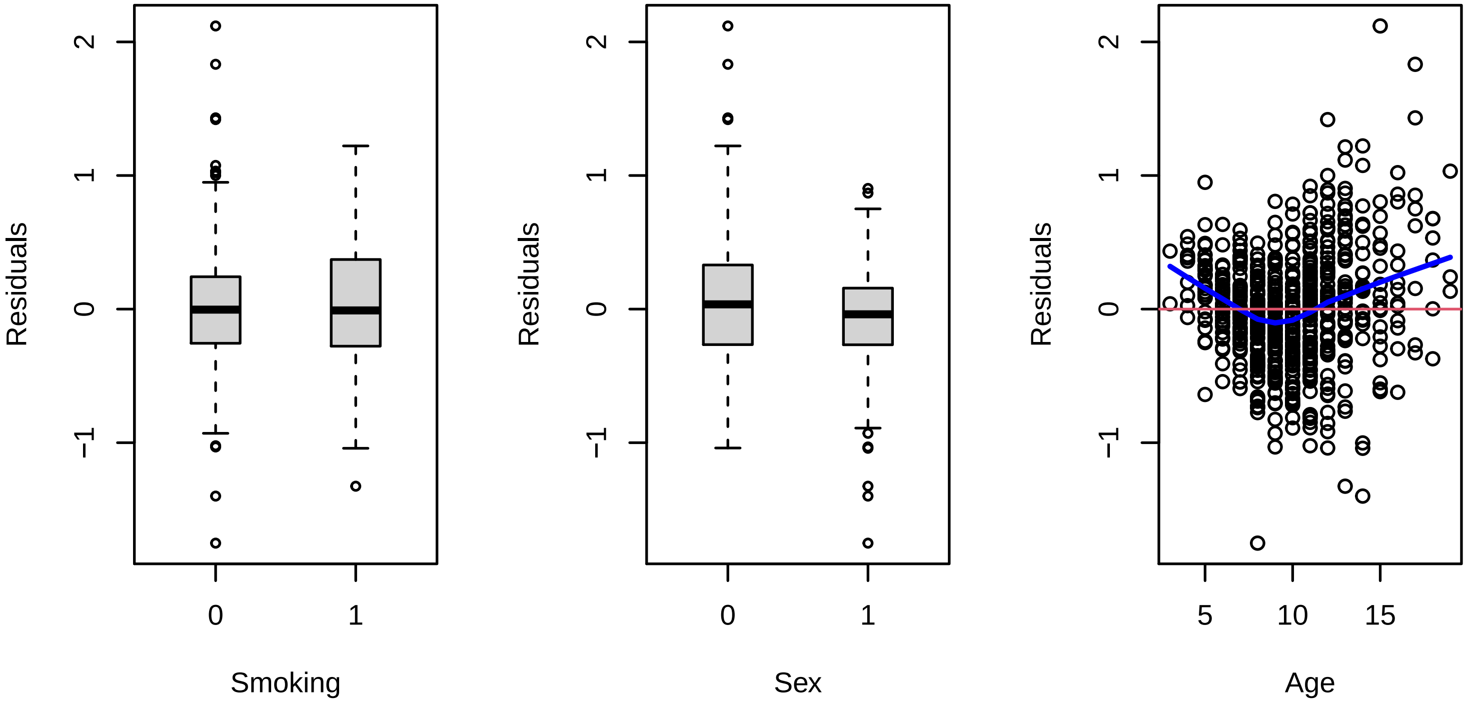
The standardized delta-betas are deemed to be of concern if larger than $2/\sqrt{n}$ in absolute value, but this guidance implicitly relies on the distribution of the test statistics to be approx. normal

Example

How about other variables? Recommended to plot residuals vs. other variables

```
> par(mfrow = c(1,3))
> plot(model$residuals~data$Smoker, xlab = "Smoking", ylab = "Residuals",
+  col = 'lightgray', boxwex = .35)
> plot(model$residuals~data$Sex, xlab = "Sex", ylab = "Residuals",
+  col = 'lightgray', boxwex = .35)
> plot(data$Age, model$residuals, xlab = "Age", ylab = "Residuals",)
> lines(lowess(data$Age, model$residuals), lwd = 2, col = "blue")
> abline(h=0, col=2)
```

Example



No surprises here, this model is terrible!

How do Ideal Diagnostics Look Like?

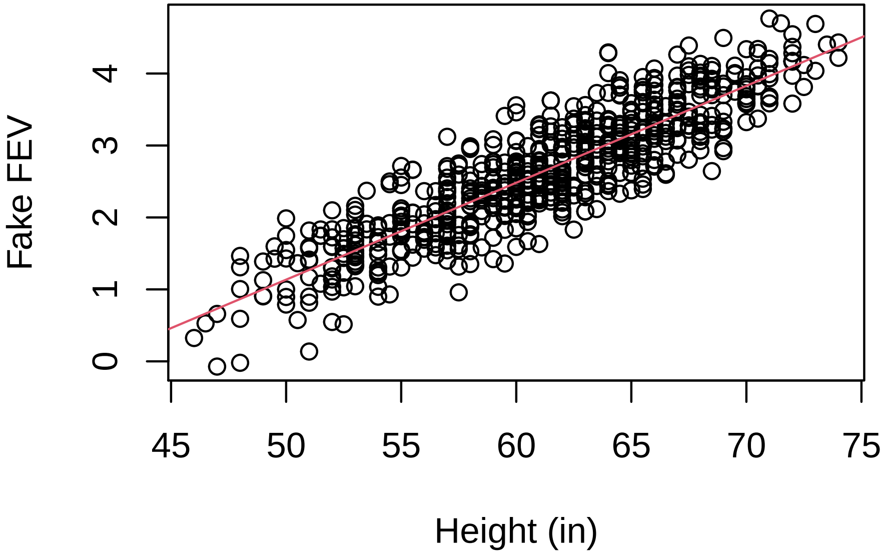
- The assumptions of the normal linear model do not hold in the previous example, but how would the diagnostics look like if the model was correct?

- Since we have a fully parametric model, we can simulate fake data from it!

How do Ideal Diagnostics Look Like?

Fake data generated from the model:

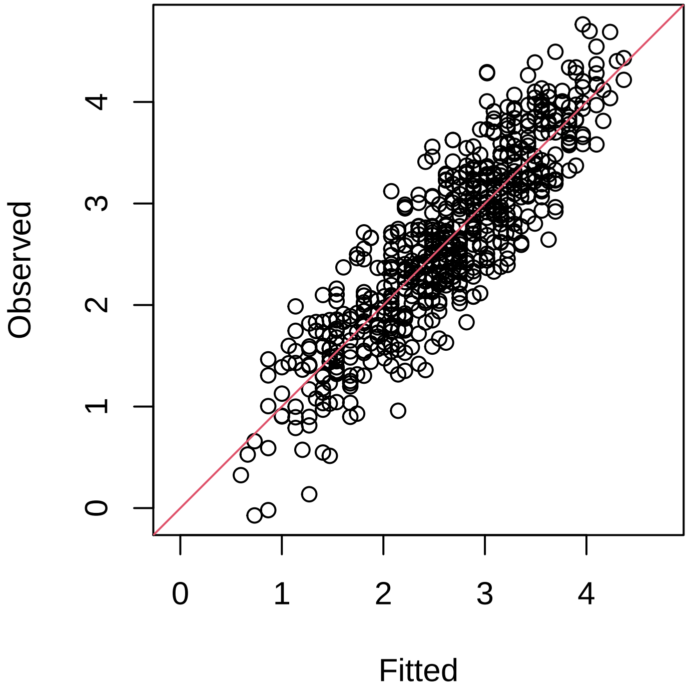
```
> n_samp <- nrow(data)
> sd_error <- summary(model)$sigma # estimate of the error std. dev.
> set.seed(32)
> Y_ideal <- fitted(model) + rnorm(n_samp, 0, sd_error) # fake responses according to fitted model
> model_ideal <- lm(Y_ideal ~ data$Height) # re-fit model using fake response
> plot(Y_ideal ~ data$Height, xlab = "Height (in)", ylab = "Fake FEV")
> abline(coef(model_ideal), col=2)
```



How do Ideal Diagnostics Look Like?

Ideal fitted vs observed plot:

```
> limits_fake <- range(fitted(model_ideal), Y_ideal)
> plot(fitted(model_ideal), Y_ideal, xlab="Fitted", ylab="Observed",
+          xlim=limits_fake, ylim=limits_fake)
> abline(a=0, b=1, col=2)
```

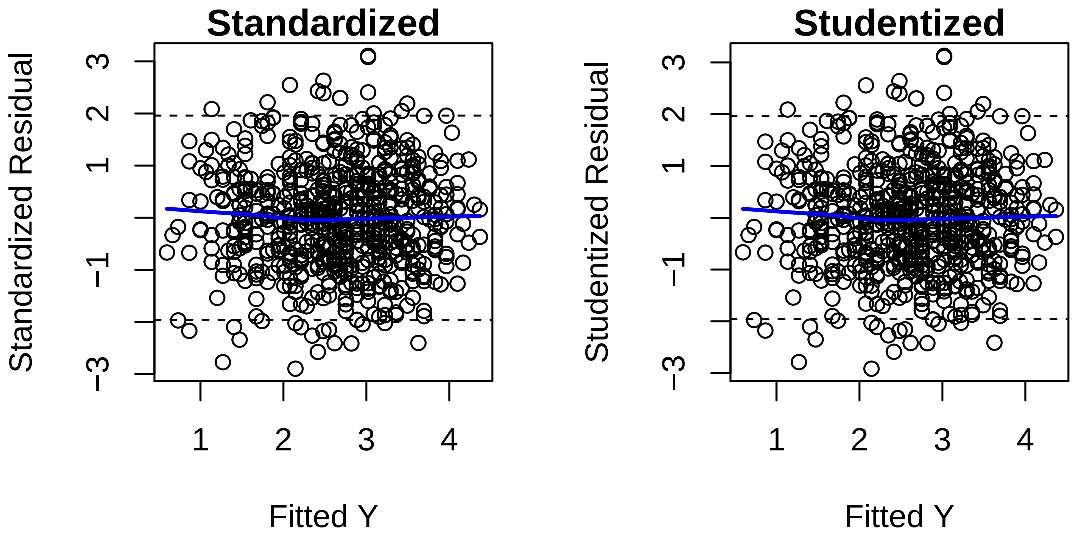


(even if the model is correct, these plots could make you think there's something wrong)

How do Ideal Diagnostics Look Like?

Residual plots:

```
> estd_ideal <- stdres(model_ideal); estud_ideal <- studres(model_ideal)
> yhat_ideal <- fitted(model_ideal); par(mfrow = c(1,2))
> plot(yhat_ideal, estd_ideal, ylab = "Standardized Residual", xlab = "Fitted Y")
> lines(lowess(yhat_ideal, estd_ideal), lwd = 2, col = "blue")
> abline(h = 1.96, lty = 2); abline(h = -1.96, lty = 2); title(main = "Standardized")
> plot(yhat_ideal, estud_ideal, ylab = "Studentized Residual", xlab = "Fitted Y")
> lines(lowess(yhat_ideal, estud_ideal), lwd = 2, col = "blue")
> abline(h = 1.96, lty = 2); abline(h = -1.96, lty = 2); title(main = "Studentized")
```

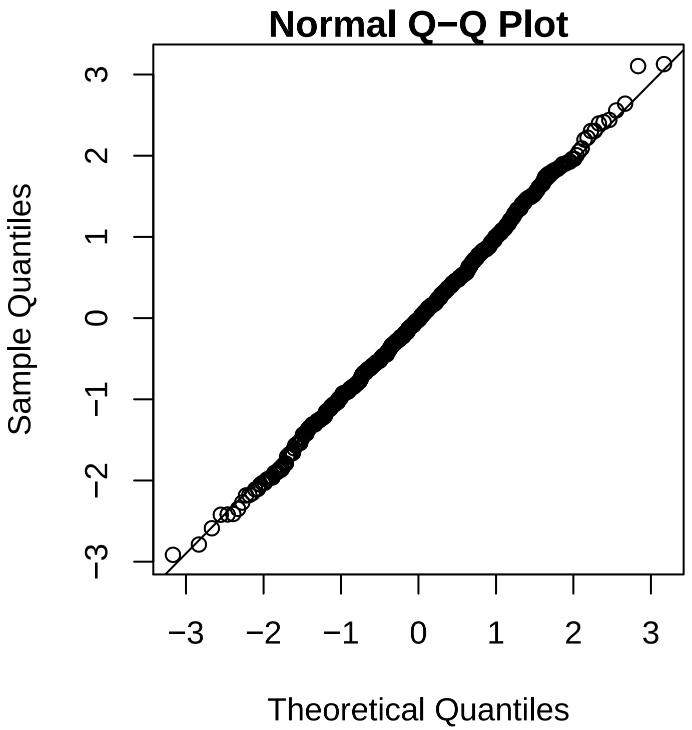


(residuals don't have to all be between -2 and 2, trend line doesn't have to be exactly flat)

How do Ideal Diagnostics Look Like?

Q-Q plot:

```
> qqnorm(estud_ideal)
> qqline(estud_ideal)
```

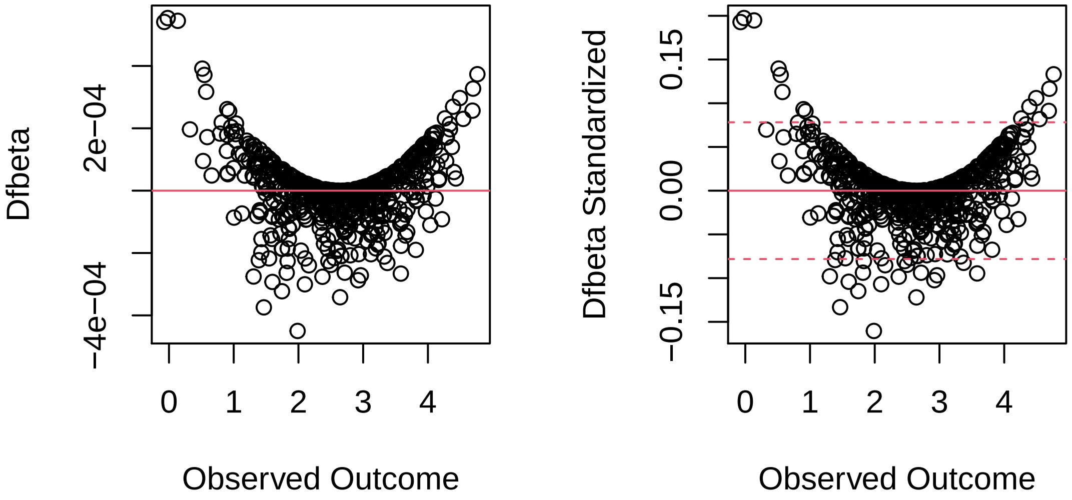


(even if the model is correct, upper/lower quantile pairs are more affected by variability)

How do Ideal Diagnostics Look Like?

Dfbetas:

```
> deltabetas_ideal <- dfbeta(model_ideal); deltabetas_st_ideal <- dfbetas(model_ideal)
> par(mfrow=c(1,2))
> plot(Y_ideal, deltabetas_ideal[,2], ylab = "Dfbeta", xlab = "Observed Outcome")
> abline(h=0, col=2)
> plot(Y_ideal, deltabetas_st_ideal[,2], ylab = "Dfbeta Standardized", xlab = "Observed Outcome")
> abline(h=0, col=2)
> abline(h=2/sqrt(nrow(data)), lty=2, col=2); abline(h=-2/sqrt(nrow(data)), lty=2, col=2)
```



(the threshold $2/\sqrt{n}$ for detecting influential points is simply a rough guide; better to compare dfbetas across observations: is there one that is orders of magnitude larger than the rest?)

Recap on Model Exploration

Traditionally, *iterative model building* follows these steps:

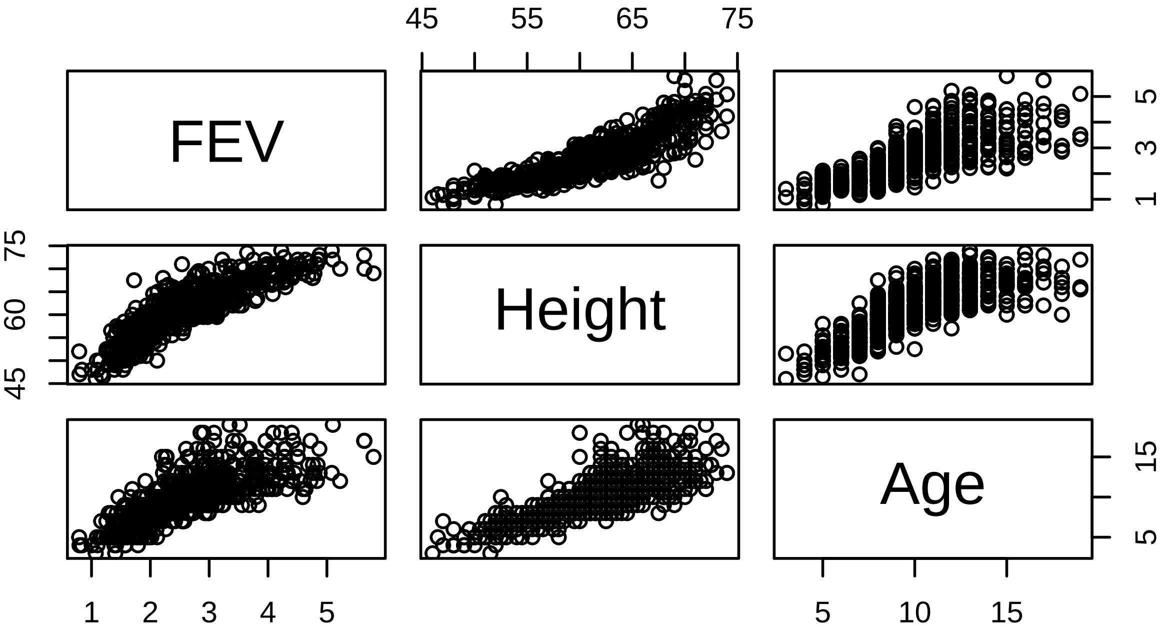
- Start with a model, check its assumptions against the data using diagnostics
- If something is off, make adjustments to the model, and check assumptions again
- Repeat until finding a satisfactory fit

But how to choose the initial model?: Unless there is a strong scientific motivation, simple data exploration can inform the initial model

Model Exploration

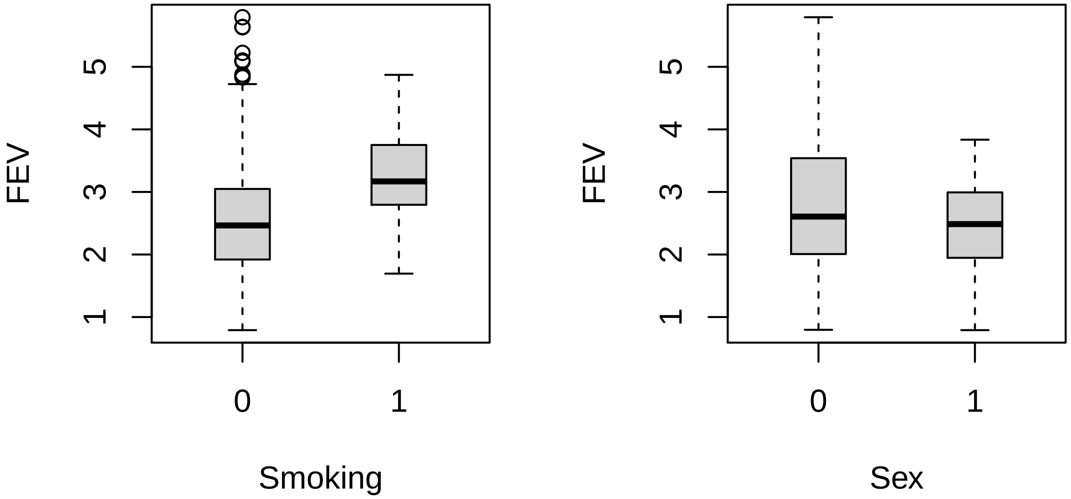
Model exploration usually starts with simple plots

```
> pairs(data[,c(3,4,2)])
```



Model Exploration

```
> par(mfrow=c(1,2))
> plot(FEV ~ Smoker, data=data, xlab = "Smoking", ylab = "FEV",
+  col = 'lightgray', boxwex = .35)
> plot(FEV ~ Sex, data=data, xlab = "Sex", ylab = "FEV",
+  col = 'lightgray', boxwex = .35)
```



Model Exploration

You may try a simple model first, say:

$$\text{E}(\text{FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Height}^2 + \beta_3 \text{Age} + \beta_4 I(\text{Sex}{=}\text{Female}) + \beta_5 I(\text{Smoker}{=}\text{Yes})$$

Model Exploration

You may try a simple model first, say:

$$\text{E}(\text{FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Height}^2 + \beta_3 \text{Age} + \beta_4 I(\text{Sex=Female}) + \beta_5 I(\text{Smoker=Yes})$$

Interpretation of individual parameters:

- β_3: a child from this population who is one-year older than another one with the same height, sex and smoking status, is expected to have β_3 liters higher FEV (or $-\beta_3$ liters lower FEV).

Model Exploration

You may try a simple model first, say:

$$\text{E(FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1\text{Height} + \beta_2\text{Height}^2 + \beta_3\text{Age} + \beta_4 I(\text{Sex=Female}) + \beta_5 I(\text{Smoker=Yes})$$

Interpretation of individual parameters:

- β_3: a child from this population who is one-year older than another one with the same height, sex and smoking status, is expected to have β_3 liters higher FEV (or $-\beta_3$ liters lower FEV).

  Question: how about “for every one-year increase in a child’s age we expect β_3 liters increase in FEV”?

Model Exploration

You may try a simple model first, say:

$$\text{E(FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1\text{Height} + \beta_2\text{Height}^2 + \beta_3\text{Age} + \beta_4 I(\text{Sex=Female}) + \beta_5 I(\text{Smoker=Yes})$$

Interpretation of individual parameters:

- β_3: a child from this population who is one-year older than another one with the same height, sex and smoking status, is expected to have β_3 liters higher FEV (or $-\beta_3$ liters lower FEV).

  Question: how about "for every one-year increase in a child's age we expect β_3 liters increase in FEV"?

- β_4: a girl from this population is expected to have β_4 liters higher FEV compared with a boy from this population with the same height, age and smoking status. (or $-\beta_4$ liters lower FEV). Analogous way of interpreting β_5.

Model Exploration

You may try a simple model first, say:

$$\text{E(FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1\text{Height} + \beta_2\text{Height}^2 + \beta_3\text{Age} + \beta_4 I(\text{Sex=Female}) + \beta_5 I(\text{Smoker=Yes})$$

Interpretation of individual parameters:

- β_3: a child from this population who is one-year older than another one with the same height, sex and smoking status, is expected to have β_3 liters higher FEV (or $-\beta_3$ liters lower FEV).

  Question: how about "for every one-year increase in a child's age we expect β_3 liters increase in FEV"?

- β_4: a girl from this population is expected to have β_4 liters higher FEV compared with a boy from this population with the same height, age and smoking status. (or $-\beta_4$ liters lower FEV). Analogous way of interpreting β_5.

- β_0: nonsensical: expected FEV for a non-smoker boy who is zero years old and zero inches tall. Better approach: instead of Age and Height, use Age$-c_1$ and Height$-c_2$, where c_1 and c_2 are meaningful.

Model Exploration

You may try a simple model first, say:

$$\text{E(FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1 \text{Height} + \beta_2 \text{Height}^2 + \beta_3 \text{Age} + \beta_4 I(\text{Sex=Female}) + \beta_5 I(\text{Smoker=Yes})$$

Interpretation of individual parameters:

- β_3: a child from this population who is one-year older than another one with the same height, sex and smoking status, is expected to have β_3 liters higher FEV (or $-\beta_3$ liters lower FEV).

  Question: how about "for every one-year increase in a child's age we expect β_3 liters increase in FEV"?

- β_4: a girl from this population is expected to have β_4 liters higher FEV compared with a boy from this population with the same height, age and smoking status. (or $-\beta_4$ liters lower FEV). Analogous way of interpreting β_5.

- β_0: nonsensical: expected FEV for a non-smoker boy who is zero years old and zero inches tall. Better approach: instead of Age and Height, use Age$-c_1$ and Height$-c_2$, where c_1 and c_2 are meaningful.

- β_1, β_2: not as easy, could say "β_1Height $+$ β_2Height^2 adjusts for height"

Model Exploration

You may try a simple model first, say:

$$\text{E(FEV1} \mid \mathbf{X}) = \beta_0 + \beta_1\text{Height} + \beta_2\text{Height}^2 + \beta_3\text{Age} + \beta_4 I(\text{Sex=Female}) + \beta_5 I(\text{Smoker=Yes})$$

Interpretation of individual parameters:

- β_3: a child from this population who is one-year older than another one with the same height, sex and smoking status, is expected to have β_3 liters higher FEV (or $-\beta_3$ liters lower FEV).

  Question: how about "for every one-year increase in a child's age we expect β_3 liters increase in FEV"?

- β_4: a girl from this population is expected to have β_4 liters higher FEV compared with a boy from this population with the same height, age and smoking status. (or $-\beta_4$ liters lower FEV). Analogous way of interpreting β_5.

- β_0: nonsensical: expected FEV for a non-smoker boy who is zero years old and zero inches tall. Better approach: instead of Age and Height, use Age$-c_1$ and Height$-c_2$, where c_1 and c_2 are meaningful.

- β_1, β_2: not as easy, could say "β_1Height $+$ β_2Height^2 adjusts for height"

Does this model make scientific sense?

Model Exploration

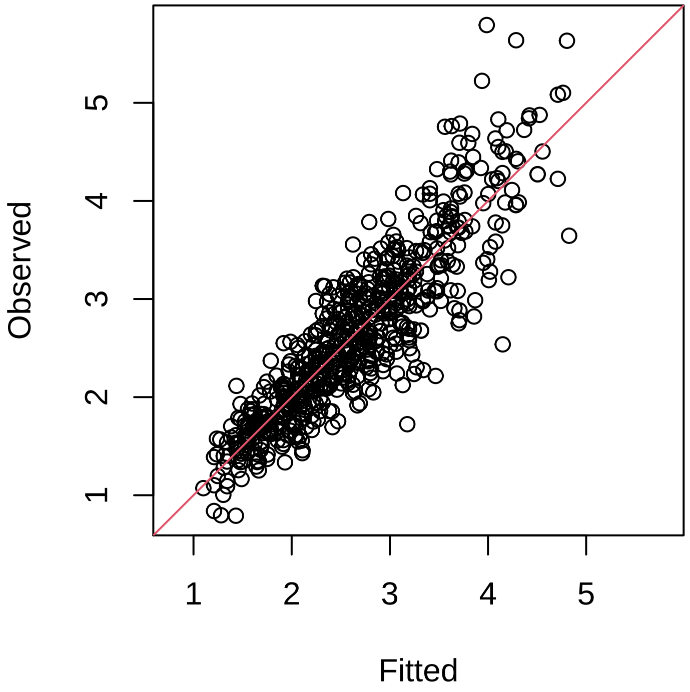
```
> data$HeightC <- data$Height - 60
> data$AgeC <- data$Age - 10
> data$HeightCSq <- data$HeightC^2
> model2 <- lm(FEV ~ HeightC + HeightCSq + Age + Sex + Smoker, data=data)

> round(coef(summary(model2)),2)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.79       0.09   20.21     0.00
HeightC         0.10       0.00   21.99     0.00
HeightCSq       0.00       0.00    7.65     0.00
Age             0.07       0.01    7.63     0.00
Sex1           -0.09       0.03   -2.88     0.00
Smoker1        -0.13       0.06   -2.33     0.02
```

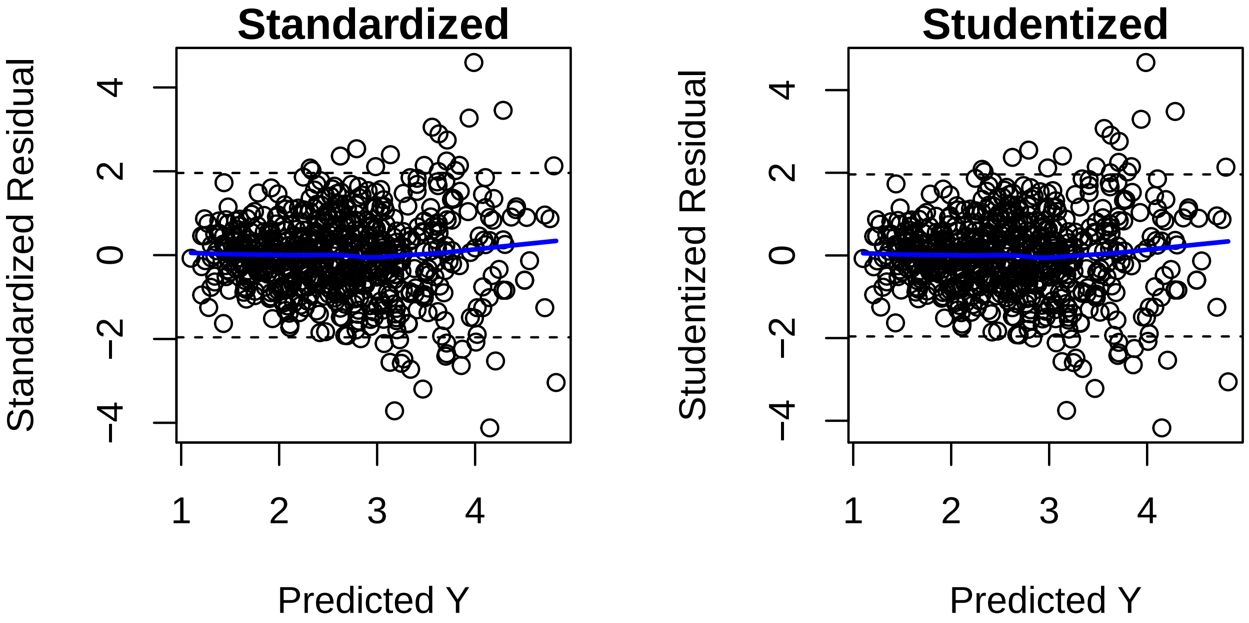
Model Exploration

Does the model have a decent fit?

```
> limits2 <- range(data$FEV, fitted(model2))
> plot(fitted(model2), data$FEV, xlab="Fitted", ylab="Observed", xlim=limits2, ylim=limits2)
> abline(a=0, b=1, col=2)
```



Model Exploration

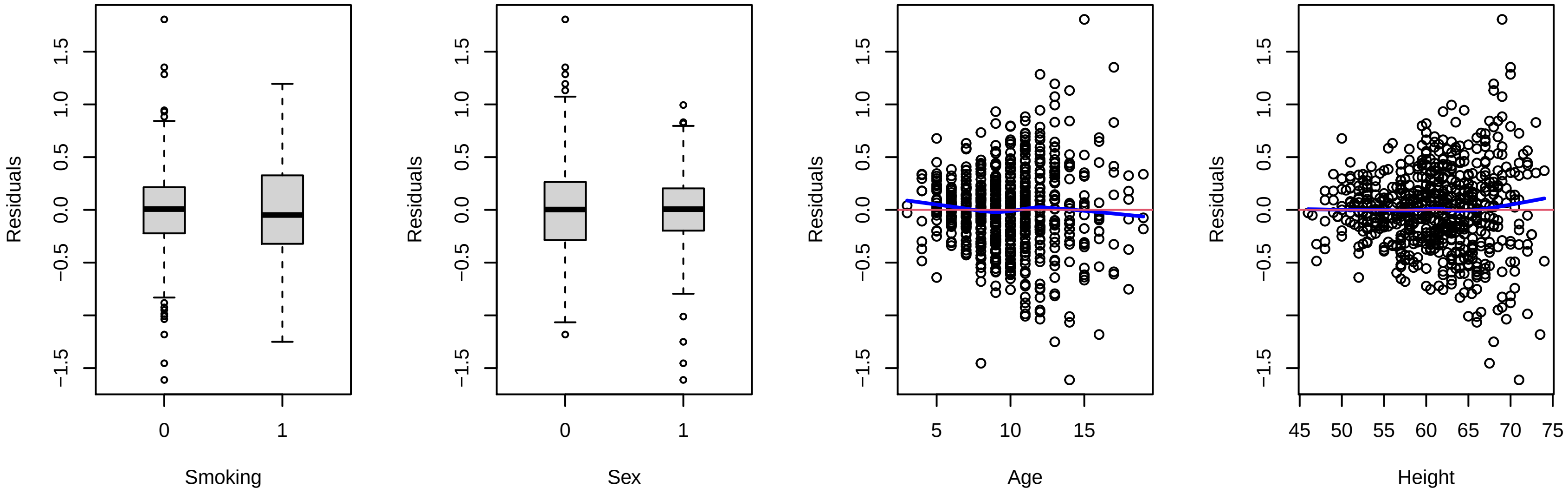


Lack of homoscedasticity seems to be the main issue here, even if we liked the mean model

Model Exploration

Residuals vs covariates

```
> par(mfrow = c(1,4))
> plot(model2$residuals ~ data$Smoker, xlab = "Smoking", ylab = "Residuals",
+  col = 'lightgray', boxwex = .35)
> plot(model2$residuals ~ data$Sex, xlab = "Sex", ylab = "Residuals",
+  col = 'lightgray', boxwex = .35)
> plot(data$Age, model2$residuals, xlab = "Age", ylab = "Residuals",)
> lines(lowess(data$Age, model2$residuals), lwd = 2, col = "blue")
> abline(h=0, col=2)
> plot(data$Height, model2$residuals, xlab = "Height", ylab = "Residuals",)
> lines(lowess(data$Height, model2$residuals), lwd = 2, col = "blue")
> abline(h=0, col=2)
```



Comments on Model Building

- We could continue iterating, using these plots and tests to decide if we should include/remove variables, add interactions or higher order terms
- There are model selection methods that automate such strategies, termed *stepwise regression*
  - Forward selection: start with no variables, at each step test the addition of a new variable, end when no variable is worth adding
  - Backward elimination: start with all variables, at each step test the deletion of each variable, end when no further variable can be deleted
  - Bidirectional elimination: a combination of the above, testing at each step for variables to be included or excluded
  - The 'tests' might be done using F-tests, t-tests, adjusted R^2, AIC, BIC, etc. See the functions `add1`, `drop1`, `step`, `stepAIC` in `R`

Comments on Model Building

- We could continue iterating, using these plots and tests to decide if we should include/remove variables, add interactions or higher order terms

- There are model selection methods that automate such strategies, termed *stepwise regression*

  - Forward selection: start with no variables, at each step test the addition of a new variable, end when no variable is worth adding

  - Backward elimination: start with all variables, at each step test the deletion of each variable, end when no further variable can be deleted

  - Bidirectional elimination: a combination of the above, testing at each step for variables to be included or excluded

  - The 'tests' might be done using F-tests, t-tests, adjusted R^2, AIC, BIC, etc. See the functions `add1, drop1, step, stepAIC` in `R`

- Criticism

  - The tests that you are running are all based on the same data: you are using the same data to test dozens, hundreds of hypothesis
  - Leads to lots of 'false discoveries', data-dredging
  - Selected models might not be scientifically meaningful

Comments on Model Building

- Someone wants to study a phenomenon/association and has access to some data
- You are aware of the above issues, so you want to avoid them as much as possible (don't torture your data; they'll tell you whatever you want).
- What do you do?
  - Read about the problem
  - Determine which variables would impact the response and should control for
  - Conduct exploratory analyses (plots for associations, summary statistics, etc)
  - Think carefully about the model/assumptions you'll use
  - Specify a model that makes *scientific* sense
  - You make it clear that this is still an *exploratory* analysis, and might recommend a replication study for a *confirmatory* analysis
  - Try to avoid fooling yourself!

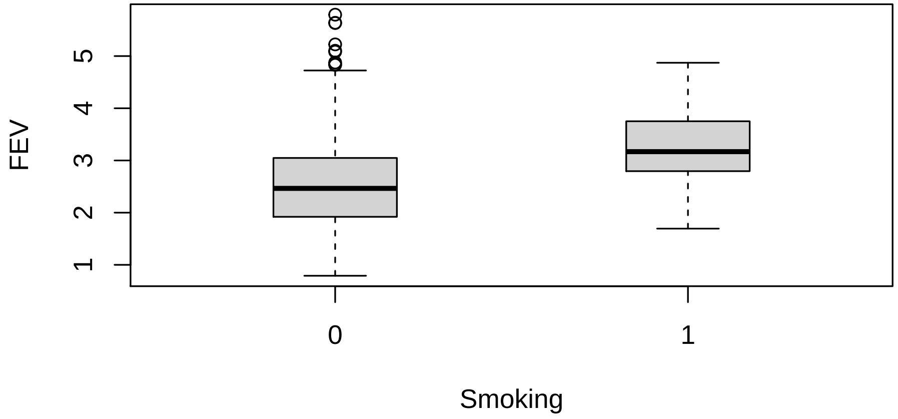
Example: Context-Based Initial Model Formulation

We previously used the FEV data to illustrate some practical issues of diagnostics and model building.

A more sensible way in which these data could appear is as follows:

- A researcher approaches you because she is interested in assessing children's pulmonary function in the absence or presence of smoking cigarettes
- A preliminary analysis shows a relationship, but not in the direction we expected!

```
> plot(FEV~Smoker, data=data, xlab = "Smoking", ylab = "FEV",
+  col = 'lightgray', boxwex = .35)
```



Example: Context-Based Initial Model Formulation

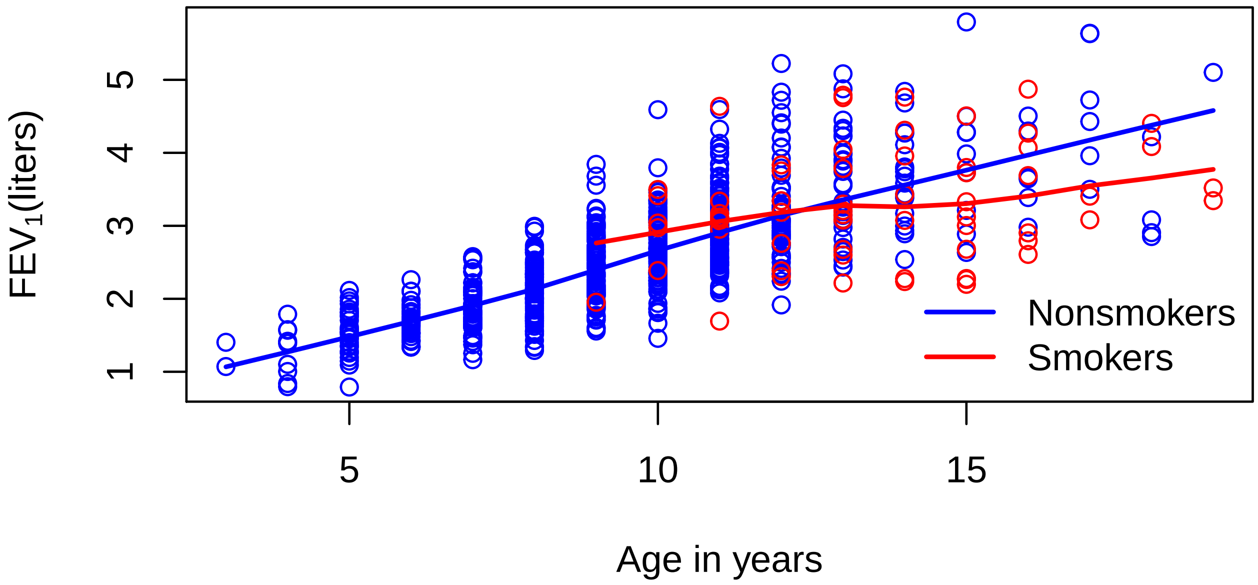
- You think hard about it: what factors could be 'confounding' this relationship and what variables are generally associated with FEV?
  - Men and women have different body-types, so sex should be included
  - Who smokes? Little kids don't, hopefully, so we should account for age
  - Taller people should have higher pulmonary function
  - While age and height should be related with FEV, the relationship might be different depending on smoking status and sex
- You formulate a concrete scientific question: after accounting for other factors, is there still association between smoking status and FEV?
- You do some exploratory analysis (we already did some in this class), formulate a model that helps answer the question

Example: Context-Based Initial Model Formulation

Let's look at the relationship of FEV with age for smokers and non-smokers.

```
> plot(FEV~Age, data=data, xlab = "Age in years",
+         ylab = expression(paste(FEV[1], "(liters)")), type = "n")
> with(data,
+         {
+         points(Age[Smoker == 0], FEV[Smoker == 0],  col = "blue");
+         points(Age[Smoker == 1], FEV[Smoker == 1],  col = "red");
+         lines(lowess(Age[Smoker == 0], FEV[Smoker == 0]), lwd = 2, col = "blue");
+         lines(lowess(Age[Smoker == 1], FEV[Smoker == 1]), lwd = 2, col = "red");
+         }
+         )
> legend("bottomright", col = c("blue", "red"), lwd = 2,
+   legend = c("Nonsmokers", "Smokers"), bty = "n")
```

Example: Context-Based Initial Model Formulation

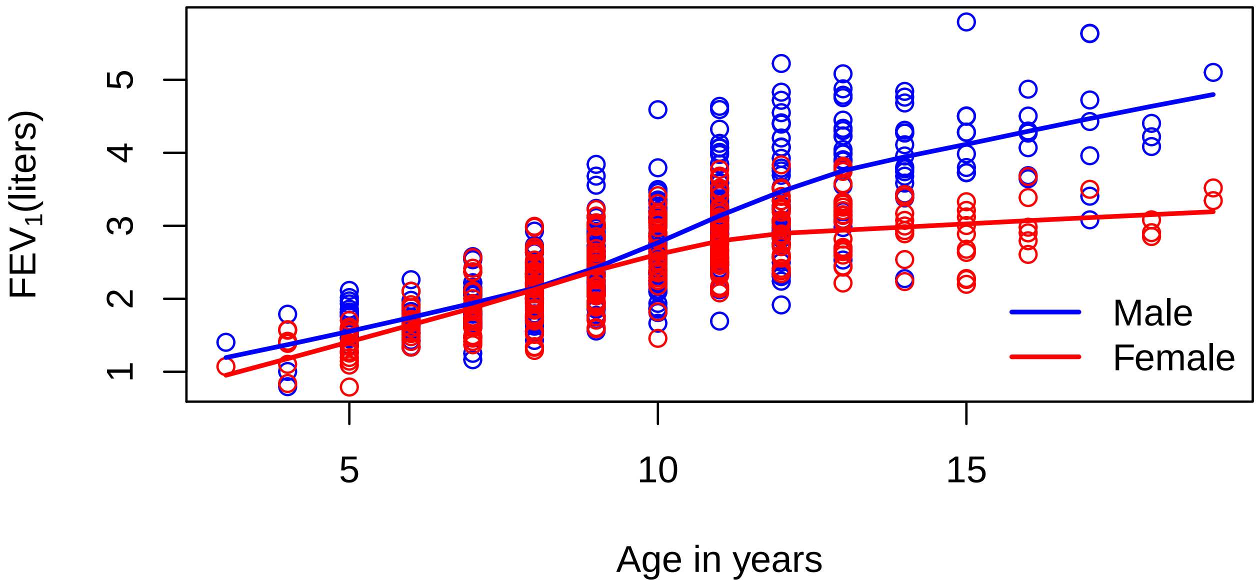


Example: Context-Based Initial Model Formulation

Now the relationship of FEV with age for males and females.

```
> plot(FEV~Age, data=data, xlab = "Age in years",
+         ylab = expression(paste(FEV[1], "(liters)")), type = "n")
> with(data,
+         {
+         points(Age[Sex == 0], FEV[Sex == 0],  col = "blue");
+         points(Age[Sex == 1], FEV[Sex == 1],  col = "red");
+         lines(lowess(Age[Sex == 0], FEV[Sex == 0]), lwd = 2, col = "blue");
+         lines(lowess(Age[Sex == 1], FEV[Sex == 1]), lwd = 2, col = "red");
+         }
+         )
> legend("bottomright", col = c("blue", "red"), lwd = 2,
+   legend = c("Male", "Female"), bty = "n")
```

Example: Context-Based Initial Model Formulation

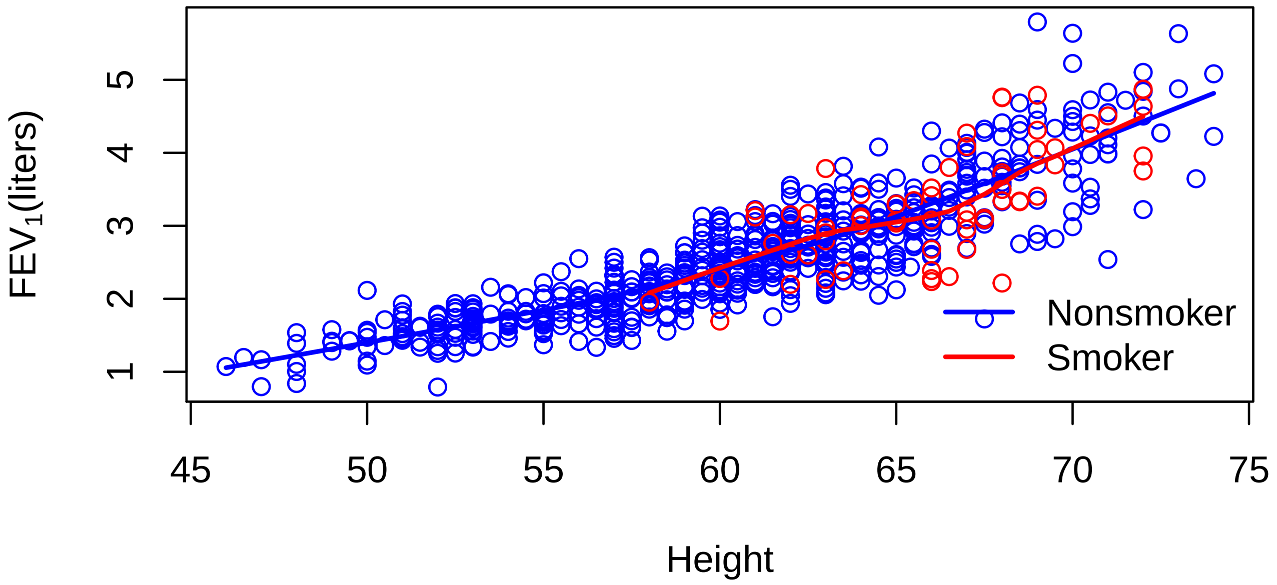


Example: Context-Based Initial Model Formulation

Now, let's look at the relationship of FEV with height for smokers and nonsmokers.

```
> plot(FEV~Height, data=data, xlab = "Height",
+         ylab = expression(paste(FEV[1], "(liters)")), type = "n")
> with(data,
+         {
+         points(Height[Smoker == 0], FEV[Smoker == 0],  col = "blue");
+         points(Height[Smoker == 1], FEV[Smoker == 1],  col = "red");
+         lines(lowess(Height[Smoker == 0], FEV[Smoker == 0]), lwd = 2, col = "blue");
+         lines(lowess(Height[Smoker == 1], FEV[Smoker == 1]), lwd = 2, col = "red");
+         }
+         )
> legend("bottomright", col = c("blue", "red"), lwd = 2,
+   legend = c("Nonsmoker", "Smoker"), bty = "n")
```

Example: Context-Based Initial Model Formulation

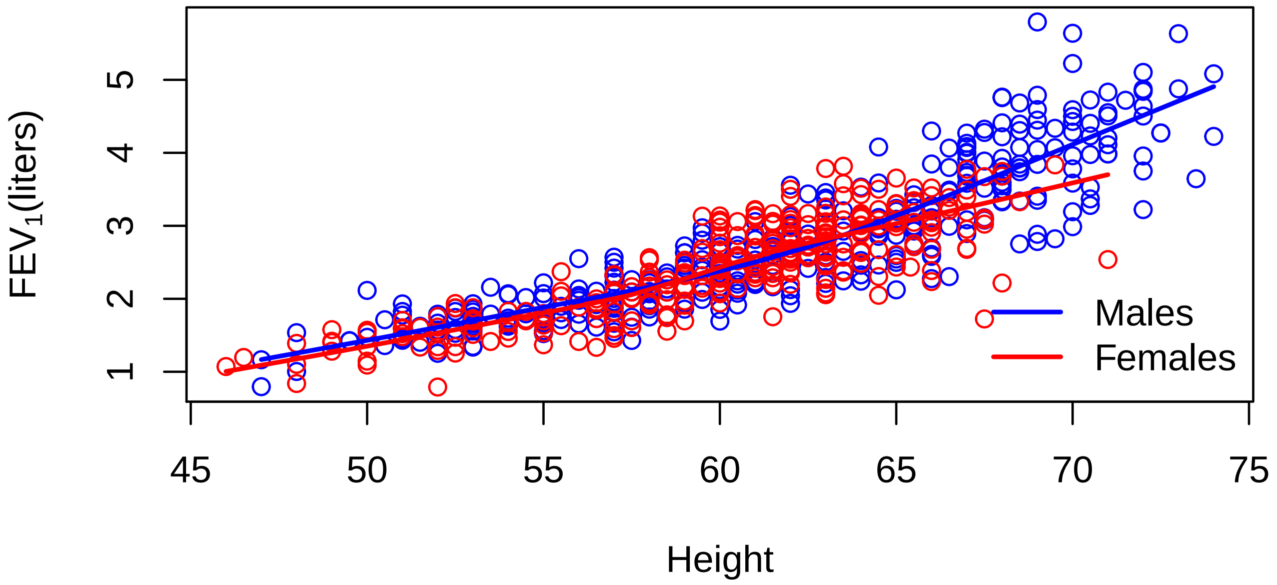


Example: Context-Based Initial Model Formulation

Now, let's look at the relationship of FEV with height for males and females.

```
> plot(FEV~Height, data=data, xlab = "Height",
+         ylab = expression(paste(FEV[1], "(liters)")), type = "n")
> with(data,
+         {
+         points(Height[Sex == 0], FEV[Sex == 0],  col = "blue");
+         points(Height[Sex == 1], FEV[Sex == 1],  col = "red");
+         lines(lowess(Height[Sex == 0], FEV[Sex == 0]), lwd = 2, col = "blue");
+         lines(lowess(Height[Sex == 1], FEV[Sex == 1]), lwd = 2, col = "red");
+         }
+         )
> legend("bottomright", col = c("blue", "red"), lwd = 2,
+   legend = c("Males", "Females"), bty = "n")
```

Example: Context-Based Initial Model Formulation



Example: Context-Based Initial Model Formulation

Based on the previous exploratory analysis, a model that could make sense would be

$$\begin{aligned}\mathrm{E}(\mathsf{FEV1} \mid \mathbf{X}) = & \beta_0 + \beta_1 \mathsf{Height} + \beta_2 \mathsf{Height}^2 + \beta_3 \mathsf{Age} + \\ & (\beta_4 + \beta_5 \mathsf{Height} + \beta_6 \mathsf{Height}^2 + \beta_7 \mathsf{Age}) I(\mathsf{Sex}{=}\mathsf{Female}) + \\ & (\beta_8 + \beta_9 \mathsf{Height} + \beta_{10} \mathsf{Height}^2 + \beta_{11} \mathsf{Age})) I(\mathsf{Smoker}{=}\mathsf{Yes})\end{aligned}$$

```
> model3 <- lm(
+         FEV ~ (HeightC+HeightCSq+AgeC)*(Sex+Smoker),
+         data=data)
```

If there is no association between FEV and smoking after accounting for other variables, under this model we would have that

$$\beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0,$$

which gives us a null hypothesis to test

Example: Context-Based Initial Model Formulation

```
> summary(model3)

Call:
lm(formula = FEV ~ (HeightC + HeightCSq + AgeC) * (Sex + Smoker),
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32240 -0.23285  0.00171  0.24423  1.72654

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.4580448  0.0326854  75.203  < 2e-16 ***
HeightC            0.0966865  0.0063384  15.254  < 2e-16 ***
HeightCSq          0.0037413  0.0005387   6.945 9.30e-12 ***
AgeC               0.0870387  0.0137570   6.327 4.69e-10 ***
Sex1              -0.0011669  0.0431706  -0.027   0.9784
Smoker1           -0.0487881  0.1459980  -0.334   0.7384
HeightC:Sex1      -0.0141184  0.0098161  -1.438   0.1508
HeightC:Smoker1    0.0270083  0.0491119   0.550   0.5826
HeightCSq:Sex1    -0.0037595  0.0009502  -3.956 8.46e-05 ***
HeightCSq:Smoker1 -0.0013206  0.0040106  -0.329   0.7421
AgeC:Sex1         -0.0192941  0.0177341  -1.088   0.2770
AgeC:Smoker1      -0.0426888  0.0238088  -1.793   0.0734 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Context-Based Initial Model Formulation

```
Residual standard error: 0.3886 on 642 degrees of freedom
Multiple R-squared:  0.8025,	Adjusted R-squared:  0.7991
F-statistic: 237.1 on 11 and 642 DF,  p-value: < 2.2e-16
```

- Residual standard error: $\widehat{\sigma} = 0.3886$ computed dividing the RSS by 642 (number of observations minus number of beta parameters)

Example: Context-Based Initial Model Formulation

```
Residual standard error: 0.3886 on 642 degrees of freedom
Multiple R-squared:  0.8025,        Adjusted R-squared:  0.7991
F-statistic: 237.1 on 11 and 642 DF,  p-value: < 2.2e-16
```

- Residual standard error: $\widehat{\sigma} = 0.3886$ computed dividing the RSS by 642 (number of observations minus number of beta parameters)

- Multiple R-squared: $R^2 = 1 - (RSS/TSS)$, with $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total sum of squares

```
> 1 - sum(residuals(model3)^2) / sum((data$FEV - mean(data$FEV))^2)
```

```
[1] 0.8024861
```

Example: Context-Based Initial Model Formulation

```
Residual standard error: 0.3886 on 642 degrees of freedom
Multiple R-squared:  0.8025,       Adjusted R-squared:  0.7991
F-statistic: 237.1 on 11 and 642 DF,  p-value: < 2.2e-16
```

- Residual standard error: $\widehat{\sigma} = 0.3886$ computed dividing the RSS by 642 (number of observations minus number of beta parameters)

- Multiple R-squared: $R^2 = 1 - (RSS/TSS)$, with $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total sum of squares

  ```
  > 1 - sum(residuals(model3)^2) / sum((data$FEV - mean(data$FEV))^2)
  ```

  ```
  [1] 0.8024861
  ```

  note that $R^2 = 1 - (RSS/TSS) = 1 - (RSS/n)/(TSS/n)$ which we can see as an estimate of 1-(residual variance/total variance). Unfortunately R^2 has the bad property of always increasing with new variables (HW1).

Example: Context-Based Initial Model Formulation

```
Residual standard error: 0.3886 on 642 degrees of freedom
Multiple R-squared:  0.8025,	Adjusted R-squared:  0.7991 
F-statistic: 237.1 on 11 and 642 DF,  p-value: < 2.2e-16
```

- Residual standard error: $\widehat{\sigma} = 0.3886$ computed dividing the RSS by 642 (number of observations minus number of beta parameters)

- Multiple R-squared: $R^2 = 1 - (RSS/TSS)$, with $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total sum of squares

  ```
  > 1 - sum(residuals(model3)^2) / sum((data$FEV - mean(data$FEV))^2)
  ```

  ```
  [1] 0.8024861
  ```

  note that $R^2 = 1 - (RSS/TSS) = 1 - (RSS/n)/(TSS/n)$ which we can see as an estimate of 1-(residual variance/total variance). Unfortunately R^2 has the bad property of always increasing with new variables (HW1).

- Adjusted R-squared: instead of using RSS/n for the residual variance use $RSS/(n-k-1)$, and instead of TSS/n use $TSS/(n-1)$. This penalizes for the number of betas in the model

Example: Context-Based Initial Model Formulation

```
Residual standard error: 0.3886 on 642 degrees of freedom
Multiple R-squared:  0.8025,       Adjusted R-squared:  0.7991
F-statistic: 237.1 on 11 and 642 DF,  p-value: < 2.2e-16
```

- Residual standard error: $\hat{\sigma} = 0.3886$ computed dividing the RSS by 642 (number of observations minus number of beta parameters)

- Multiple R-squared: $R^2 = 1 - (RSS/TSS)$, with $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total sum of squares

  ```
  > 1 - sum(residuals(model3)^2) / sum((data$FEV - mean(data$FEV))^2)
  ```

  ```
  [1] 0.8024861
  ```

  note that $R^2 = 1 - (RSS/TSS) = 1 - (RSS/n)/(TSS/n)$ which we can see as an estimate of 1-(residual variance/total variance). Unfortunately R^2 has the bad property of always increasing with new variables (HW1).

- Adjusted R-squared: instead of using RSS/n for the residual variance use $RSS/(n-k-1)$, and instead of TSS/n use $TSS/(n-1)$. This penalizes for the number of betas in the model

- F-statistic: uses the F test seen before in this class for $H_0: \quad \beta_1 = \cdots = \beta_k = 0$

Example: Context-Based Initial Model Formulation

Is there any association between FEV and smoking after controlling for other variables in this model?

Let's test the hull hypothesis

$$H_0 : \quad \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$$

We can do this by comparing `model3` with a model that does not include the Smoker variable, say `model4`:

```
> model4 <- lm( FEV ~ (HeightC+HeightCSq+AgeC)*Sex, data=data)
```

Example: Context-Based Initial Model Formulation

We saw how to do this comparison based on an F test:

$$\frac{(RSS_{H_0} - RSS)/q}{RSS/(n-k-1)} \sim F_{q,n-k-1}$$

where q is the number of restrictions of the null hypothesis

In R, the function anova does it for us

```
> anova(model4, model3)

Analysis of Variance Table

Model 1: FEV ~ (HeightC + HeightCSq + AgeC) * Sex
Model 2: FEV ~ (HeightC + HeightCSq + AgeC) * (Sex + Smoker)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    646 97.811
2    642 96.963  4   0.84781 1.4033 0.2313
```

Example: Context-Based Initial Model Formulation

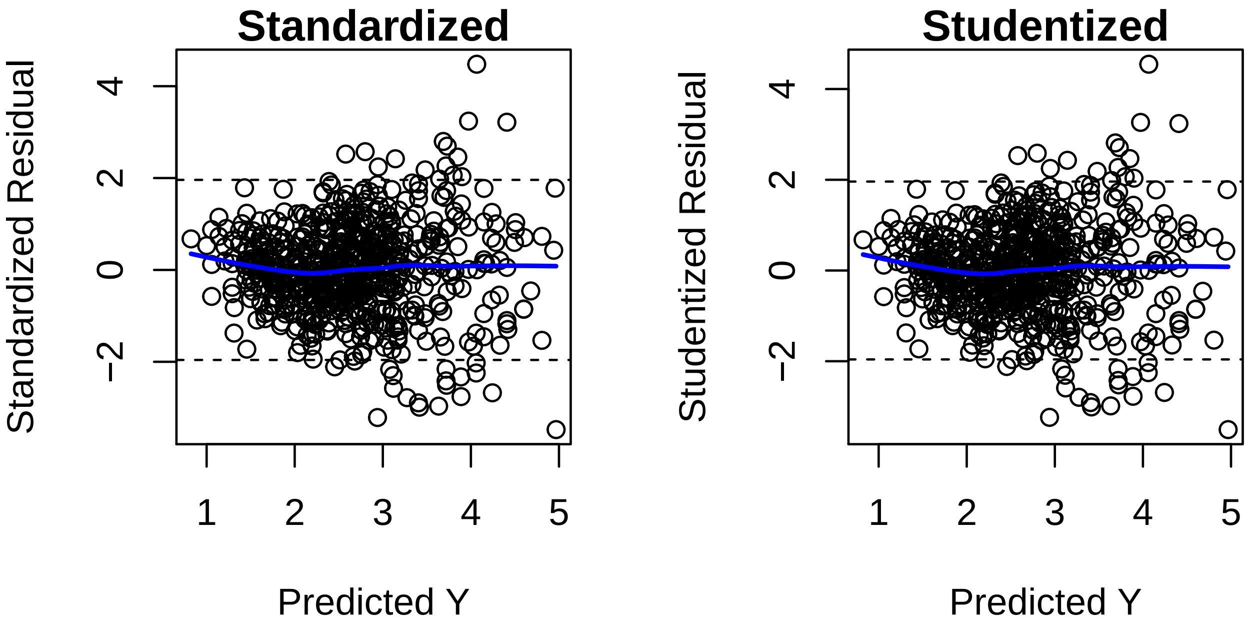
Or we can do the test “by hand”

```
> df_mod4 <- summary(model4)$df[2]
> df_mod3 <- summary(model3)$df[2]
> sd_error_mod4 <- summary(model4)$sigma
> sd_error_mod3 <- summary(model3)$sigma
> RSS_mod4 <- sd_error_mod4^2*df_mod4
> RSS_mod3 <- sd_error_mod3^2*df_mod3
> q <- df_mod4 - df_mod3 # 4, difference in number of params
> ( F_obs <- ((RSS_mod4 - RSS_mod3)/q)/sd_error_mod3^2 )

[1] 1.403349

> 1 - pf(F_obs, q, df_mod3)

[1] 0.231275
```

Example: Context-Based Initial Model Formulation

Does the model have a decent fit?

Unfortunately, we still have heteroscedasticity of the errors, so the reliability of these tests is questionable



Comments on Model Building

What about a *confirmatory analysis?*

- In a serious study there will be a protocol that specifies how the data will be analyzed even *before* the data are collected:
  - Which model will be used, covariates to control for
  - Which hypothesis will be tested
  - How it will be tested
  - What to do with outliers or missing values
  - ...

Comments on Model Building

What about a *confirmatory analysis?*

- In a serious study there will be a protocol that specifies how the data will be analyzed even *before* the data are collected:
  - Which model will be used, covariates to control for
  - Which hypothesis will be tested
  - How it will be tested
  - What to do with outliers or missing values
  - ...

- After you receive the data:
  - You will report the results obtained from the pre-specified analysis, and this will be the *primary analysis*
  - Even then, you will want to check some diagnostics and might discover that not everything went as planned, or some detail was overlooked in the protocol
  - What do you do?: secondary analyses and sensitivity analyses are often reported
  - What if the primary analysis results are not very robust??