Lec 02: Linear Regression Analysis

MATH 456 - Spring 2016

Navbar: [Home] [Schedule] [Data] [Week 3 Overview] [HW Info] [Google Group] [Google Drive]

Assigned Reading

Afifi: Chapters 6-7

Simple Regression and Correlation (*Afifi Ch 6*)

Aims

- Describe the relationship between an independent variable X and a continuous dependent variable Y as a straight line. The textbook discusses two cases:
  - Fixed-X: values of X are preselected by investigator
  - Variable-X: have random sample of (X, Y) values
  - Calculations are the same,
- Draw inferences regarding this relationship
- Predict value of Y for a given value of X

Mathmatical Model

- The mean of Y values at any given X is $\beta_0 + \beta_1 X$
- The variance of Y values at any X is σ^2 (same for all X)
- Y values are normally distributed at any given X (need for inference)

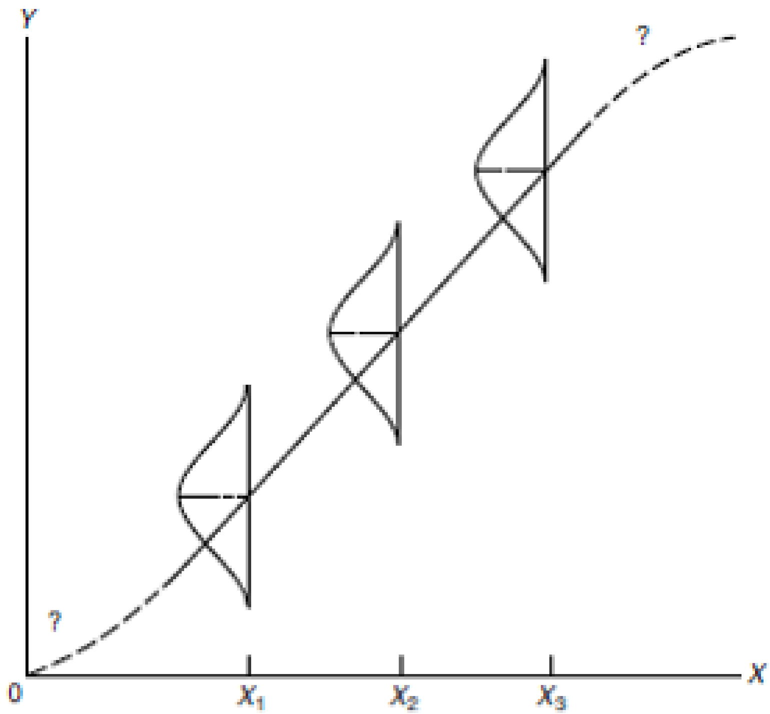


Figure 6.2: *Simple Linear Regression Model for Fixed X's*

Parameter Estimates

- Estimate the slope β_1 and intercept β_0 using least-squares methods.
- The residual mean squared error (RMSE) is an estimate of the variance s^2
- Typically interested in inference on β_1
  - Assume no relationship between X and Y ($H_0 : \beta_1 = 0$) until there is reason to believe there is one ($H_0 : \beta_1 \neq 0$)

Interval estimation

- Everything is estimated with some degree of error
- Confidence intervals for the mean of Y
- Prediction intervals for an individual Y

Which one is wider? Why?

Corelation Coefficient

- The correlation coefficient ρ measures the strength of association between X and Y in the *population*.
- $\sigma^2 = VAR(Y|X)$ is the variance of Y for a specific X.
- $\sigma_y^2 = VAR(Y)$ is the variance of Y for all X's.

$$\sigma^2 = \sigma_y^2(1 - \rho^2)$$

$$\rho^2 = \frac{\sigma_y^2 - \sigma^2}{\sigma_y^2}$$

- ρ^2 = reduction in variance of Y associated with knowledge of X/original variance of Y
- **Coefficient of Determiniation**: $100\rho^2$ = % of variance of Y associated with X or explained by X
- Caution: association vs. causation.

Example: Lung Function

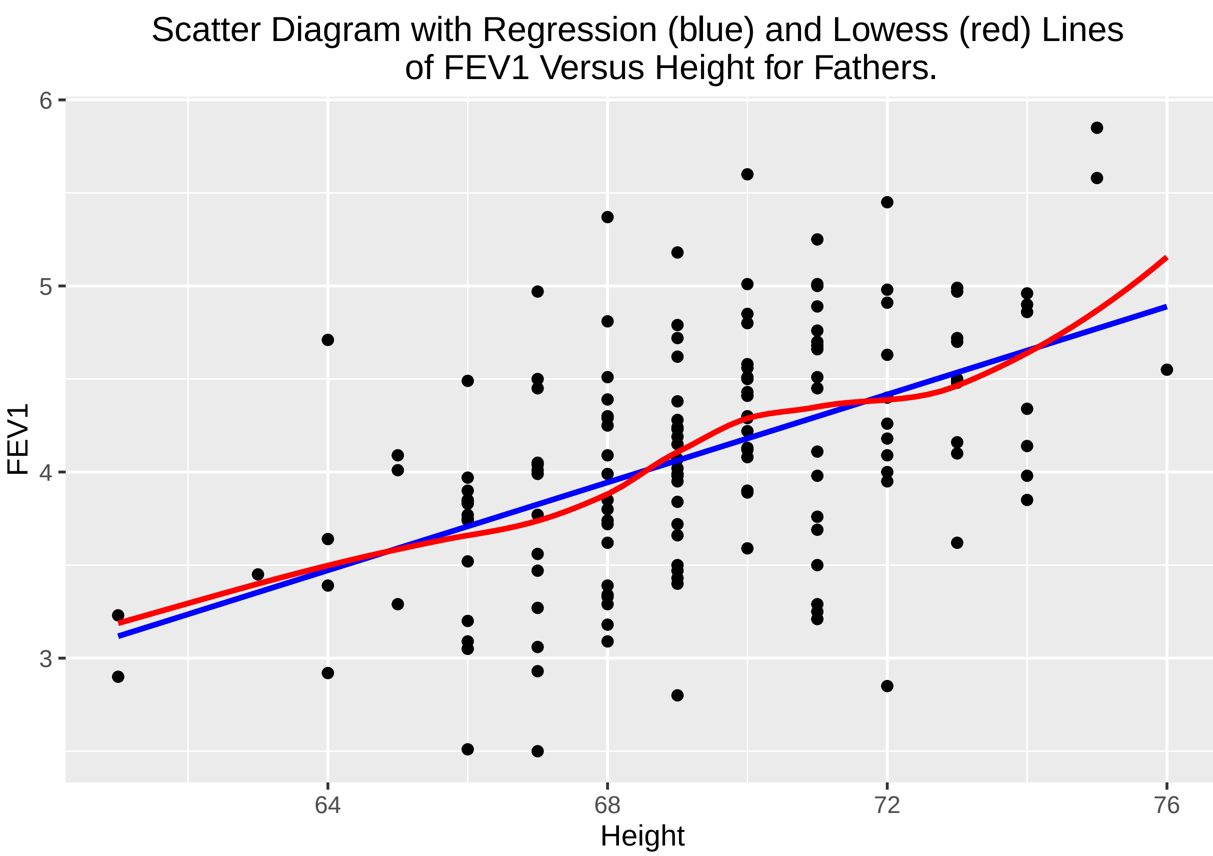
Note: The data management markdown file for this data set can be found here. you will need to download the raw data, and this data management file, update the file locations and run it on your machine to create the analysis data set.

Section 6.3 1. Read in the analysis data set for the lung function.

```
fev <- read.delim("C:/GitHub/MATH456/data/Lung_020716.txt", sep="\t", header=TRUE)
```

2. Create a scatterplot of FEV versus height for fathers. Add a blue linear regression line and red lowess line. Add appropriate plot titles and axes labels. R Cookbook reference

```
library(ggplot2)
qplot(y=FFEV1, x=FHEIGHT, geom="point", data=fev, xlab="Height", ylab="FEV1",
      main="Scatter Diagram with Regression (blue) and Lowess (red) Lines
      of FEV1 Versus Height for Fathers.") +
      geom_smooth(method="lm", se=FALSE, col="blue") +
      geom_smooth(se=FALSE, col="red")
```



There does appear to be a tendency for taller men to have higher FEV1.

3. Fit a linear model and report the regression parameter estimates.

```
model <- lm(FFEV1 ~ FHEIGHT, data=fev)
summary(model)
```

```
##
## Call:
## lm(formula = FFEV1 ~ FHEIGHT, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56688 -0.35290  0.04365  0.34149  1.42555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.08670    1.15198  -3.548 0.000521 ***
## FHEIGHT      0.11811    0.01662   7.106 4.68e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5638 on 148 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2494
## F-statistic:  50.5 on 1 and 148 DF,  p-value: 4.677e-11
```

The least squares equation is $Y = -4.087 + 0.118X$.

4. Test for a significant relationship between height and FEV. Include a p-value and a confidence interval for the parameter estimate in your conclusion.

```
confint(model)
```

```
##                   2.5 %     97.5 %
## (Intercept) -6.36315502 -1.8102499
## FHEIGHT      0.08526328  0.1509472
```

For ever inch taller a father is, his FEV1 measurement significantly increases by .12 (95%CI: .09, .15, p<.0001). The correlation between FEV1 and height is $\sqrt{.2544} = 0.5$.

5. Predict the FEV for a 6' and 6'4" male.

```
newdata <- data.frame(FHEIGHT=c(72, 76))
predict(model, newdata, interval="confidence")
```

```
##        fit      lwr      upr
## 1 4.416875 4.288918 4.544832
## 2 4.889296 4.649977 5.128615
```

```r
predict(model, newdata, interval="prediction")
```

```
##        fit      lwr      upr
## 1 4.416875 3.295410 5.538340
## 2 4.889296 3.749741 6.028851
```

Why couldn't I create a prediction for a 5' male (60")? Hint, look at the range of height values below.

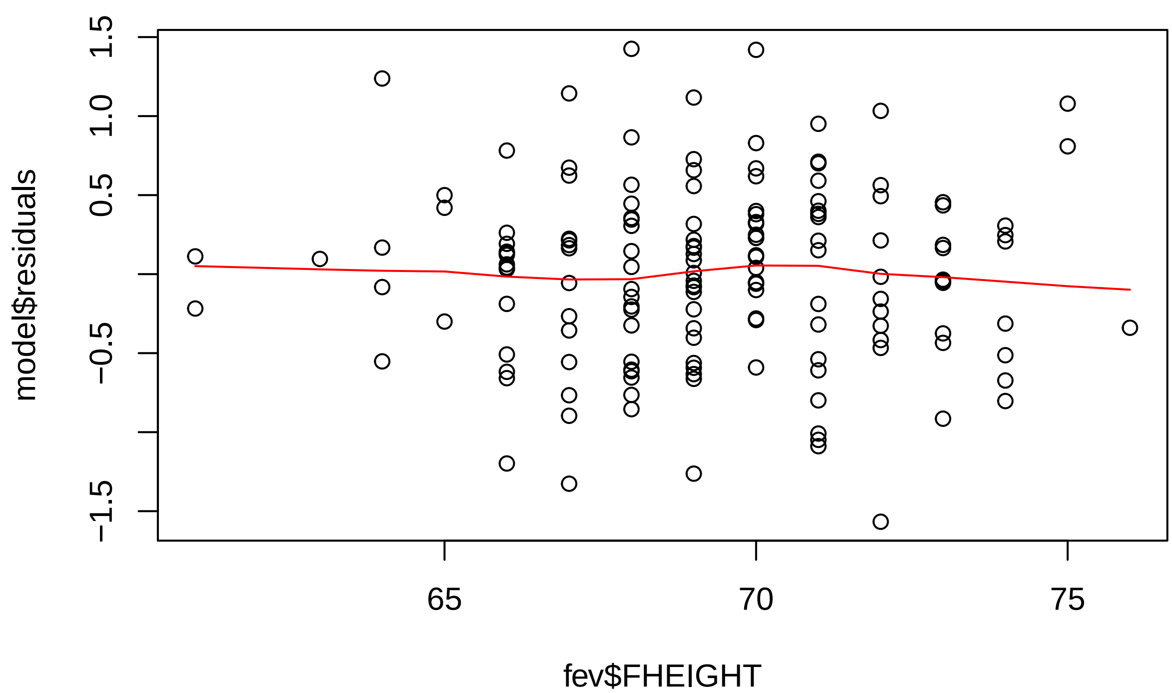
```r
summary(fev$FHEIGHT)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61.00   67.25   69.00   69.26   71.00   76.00
```

Assumptions of Linear Regression

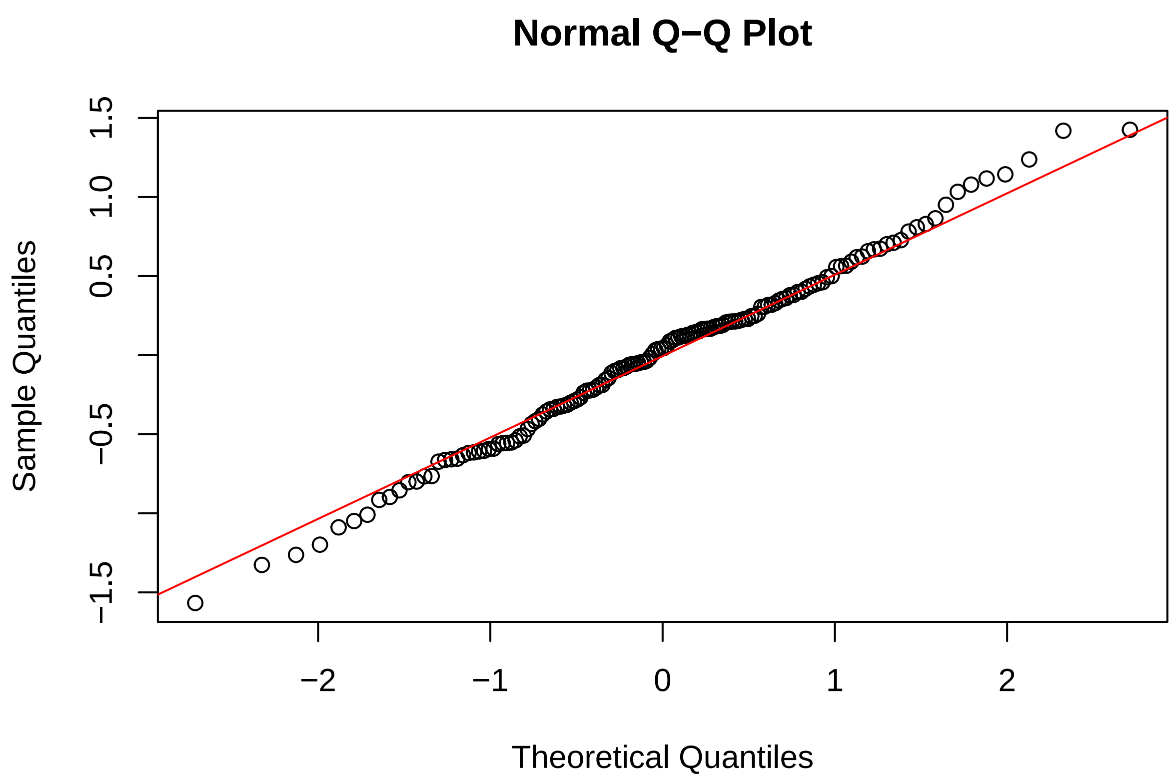
- Homogeneity of variance (same σ^2)
  - Not extremely serious
  - Can use transformations to achieve it
  - Graphical assessment: Plot the residuals against the x variable, add a lowess line. This assumption is upheld if there is no relationship/trend between the residuals and the predictor.

```r
plot(model$residuals ~ fev$FHEIGHT)
lines(lowess(model$residuals ~ fev$FHEIGHT), col="red")
```



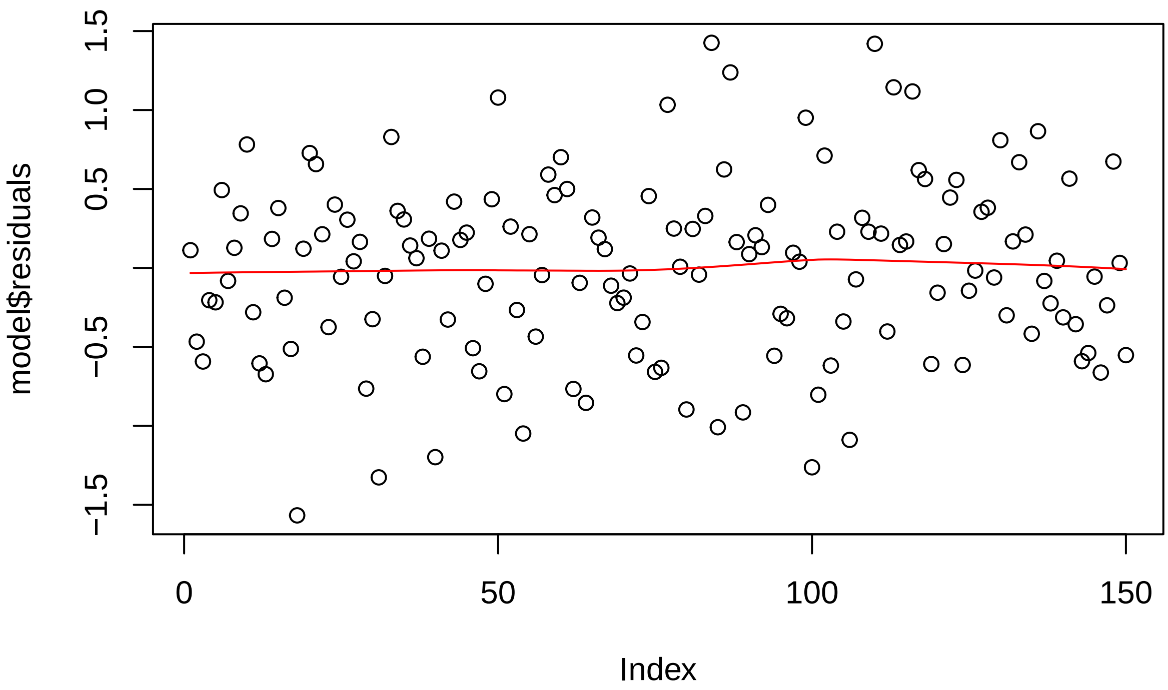
- Normal residuals
    - Slight departures OK
    - Can use transformations to achieve it
    - Graphical assessment: normal qqplot of the model residuals.

```
qqnorm(model$residuals)
qqline(model$residuals, col="red")
```



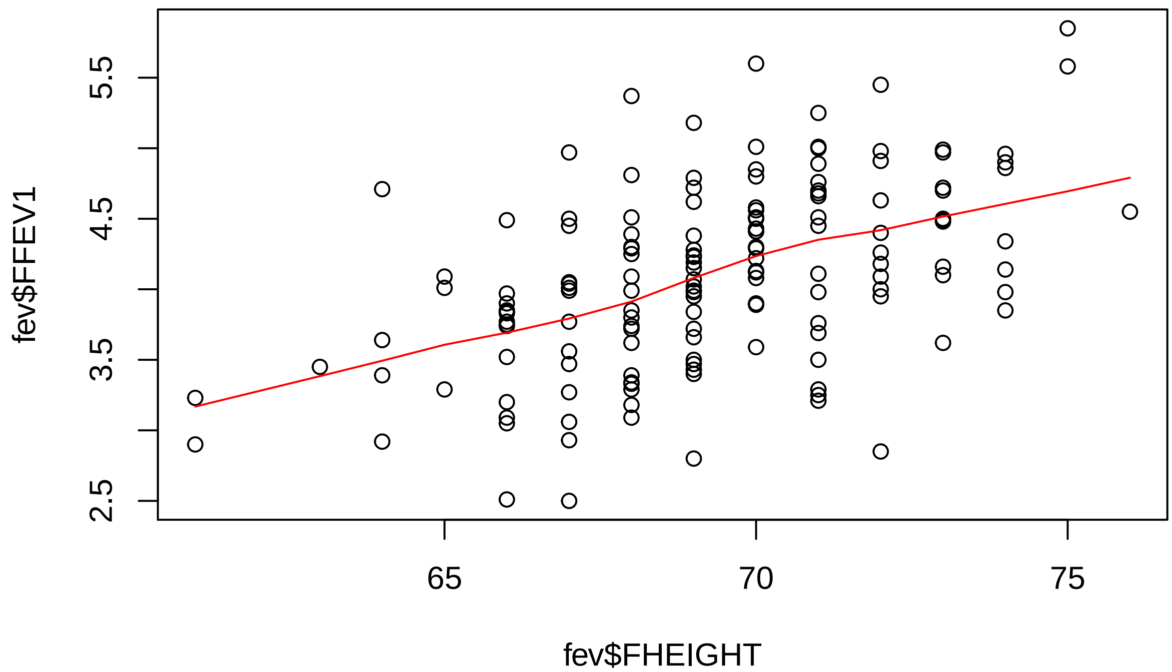
- Randomness / Independence
    - Very serious
    - Can use hierarchical models for clustered samples
    - Graphical assessment: Plot the residuals against the index (row number), add a lowess line. This assumption **may** upheld if there is no relationship/trend seen.

```
plot(model$residuals)
lines(lowess(model$residuals ~ 1:NROW(fev)), col="red")
```



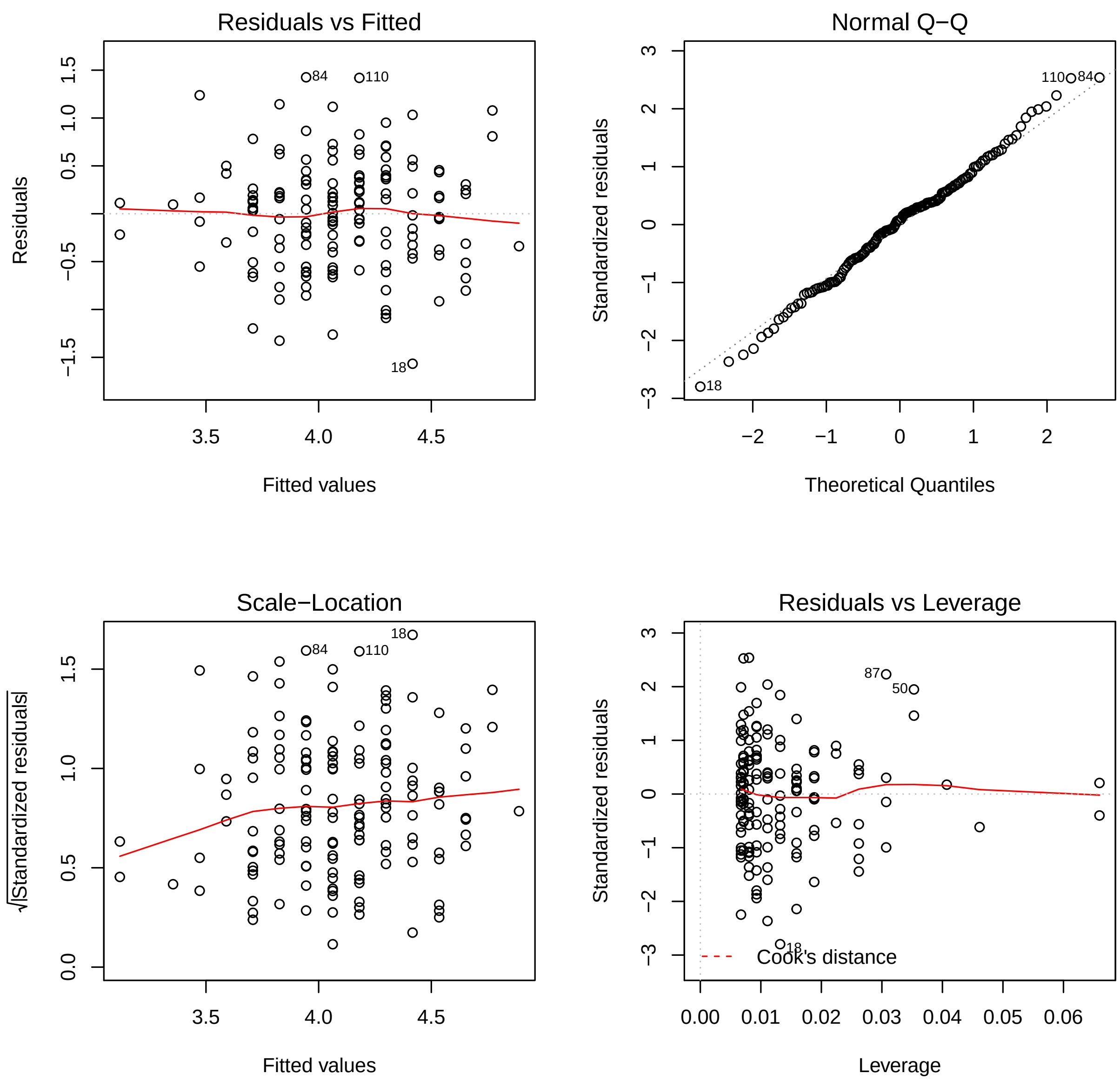
- Linear relationship
  - Slight departures OK
  - Can use transformations to achieve it
  - Graphical assessment: Simple scatterplot of y vs x. Looking for linearity in the relationship. Should be done prior to any analysis.

```
plot(fev$FFEV1 ~ fev$FHEIGHT)
lines(lowess(fev$FFEV1 ~ fev$FHEIGH), col="red")
```



Some of these plots can be displayed by simply plotting the model output. The advantage of this is that the observations that are potential outliers are labeled with their row number.

```
par(mfrow=c(2,2))
plot(model)
```



What to watch out for

- Representative sample
- Range of prediction should match observed range of X in sample
- Use of nominal or ordinal, rather than interval or ratio data
- Errors-in-variables
- Correlation does not imply causation
- Violation of assumptions
- Influential points
- Appropriate model

The book goes into more detail about influential points, and how outliers can have different affects on the model results depending on if they are an outlier in Y vs an outlier in X (or both).

On Your Own: Afifi Problem 6.2 (modified). Write your responses in a new Markdown file named *userid_ch6.rmd*. Don't forget to use the *analysis* data set for this problem.

1. Perform a regression analysis of weight on height for fathers
    a. Determine the correlation coefficient and the regression equation.

    b. Interpret the coefficients in context of the problem.

2. Test that the coefficients are significantly different from zero.
3. Assess model fit graphically. Comment on any outliers.
4. Repeat 1 -3 for mothers.

5. Discuss why the correlation for fathers appears higher than that for mothers.