CAAP Statistics - Lec18

R Session8

Aug 4, 2022

Review

- Line Fitting, Residuals and Correlation
  - $\hat{y} = b_0 + b_1 x$
  - Residuals = Observed(y) - Fitted($\hat{y}$)
- Fitting a line by Least Squares Regression
  - Idea: want to minimize the sum of squared residuals
- Types of outliers in Linear Regression
  - Outliers, High leverage point
  - Influential point
- Inference for Linear Regression
  - Hypothesis testing for the slope: $H_0 : \beta_1 = 0$

Learning Objectives

- Understand the output from `lm`
- Check the condition for linear regression
  - Independence
  - Linearity
  - Normality
  - Constant Variance

Load packages

```r
library(tidyverse)
library(openintro)
library(ggplot2)
#
install.packages("faraway")
library(faraway)
```

`Twins` Data

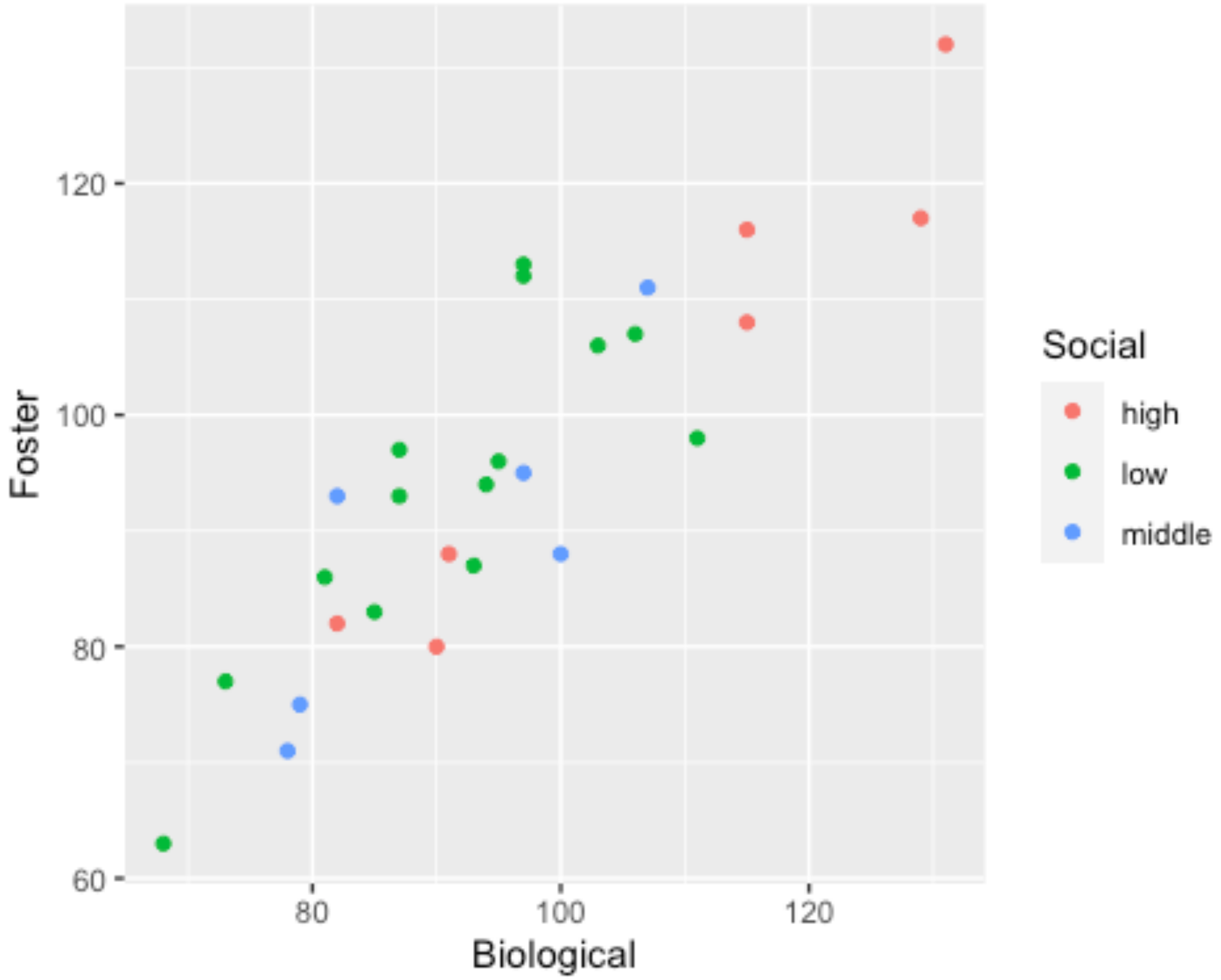
The data for this example come from a 1966 paper by Cyril Burt entitled "The genetic determination of differences in intelligence: A study of monozygotic twins reared apart". The data consist of IQ scores for identical twins, one raised by foster parents, the other by the natural parents. We also know the social class of natural parents (high, middle or low). **We are interested in predicting the IQ of the twin with foster parents from the IQ of the twin with the natural parents and the social class of natural parents.**

Source: [Practical Regression and Anova using R]

```
head(twins)
##   Foster Biological Social
## 1     82         82   high
## 2     80         90   high
## 3     88         91   high
## 4    108        115   high
## 5    116        115   high
## 6    117        129   high
```

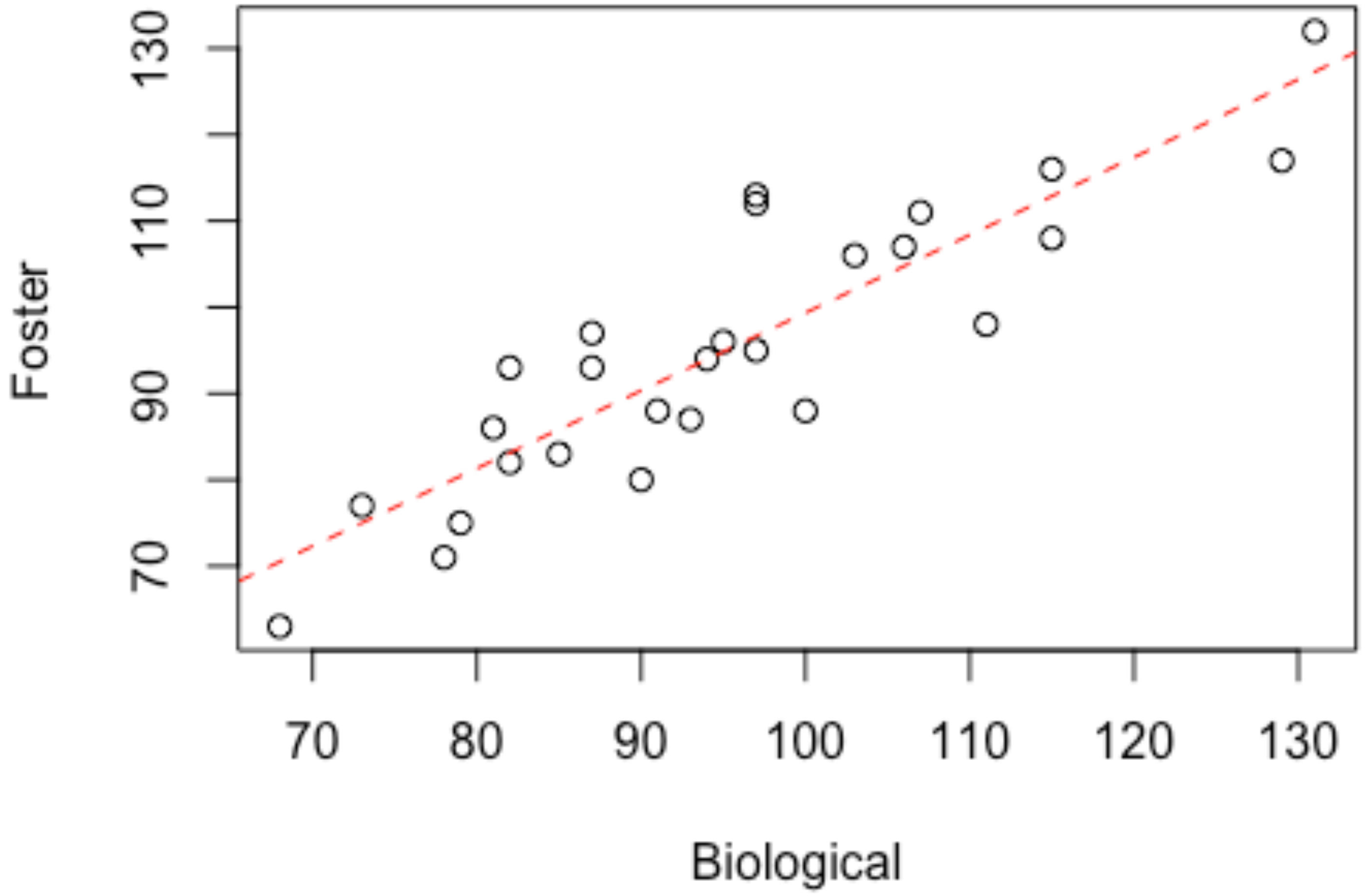
Preliminary Check

```
cor(twins$Biological, twins$Foster)
## [1] 0.8819877
twins %>%
  ggplot(aes(x=Biological, y = Foster))+
  geom_point(aes(color=Social))
```



Line Fitting

```
m1 = lm(Foster~Biological, data = twins)
plot(Foster~Biological, data = twins)
abline(m1$coefficients[1], m1$coefficients[2], col="red", lty = 2)
```

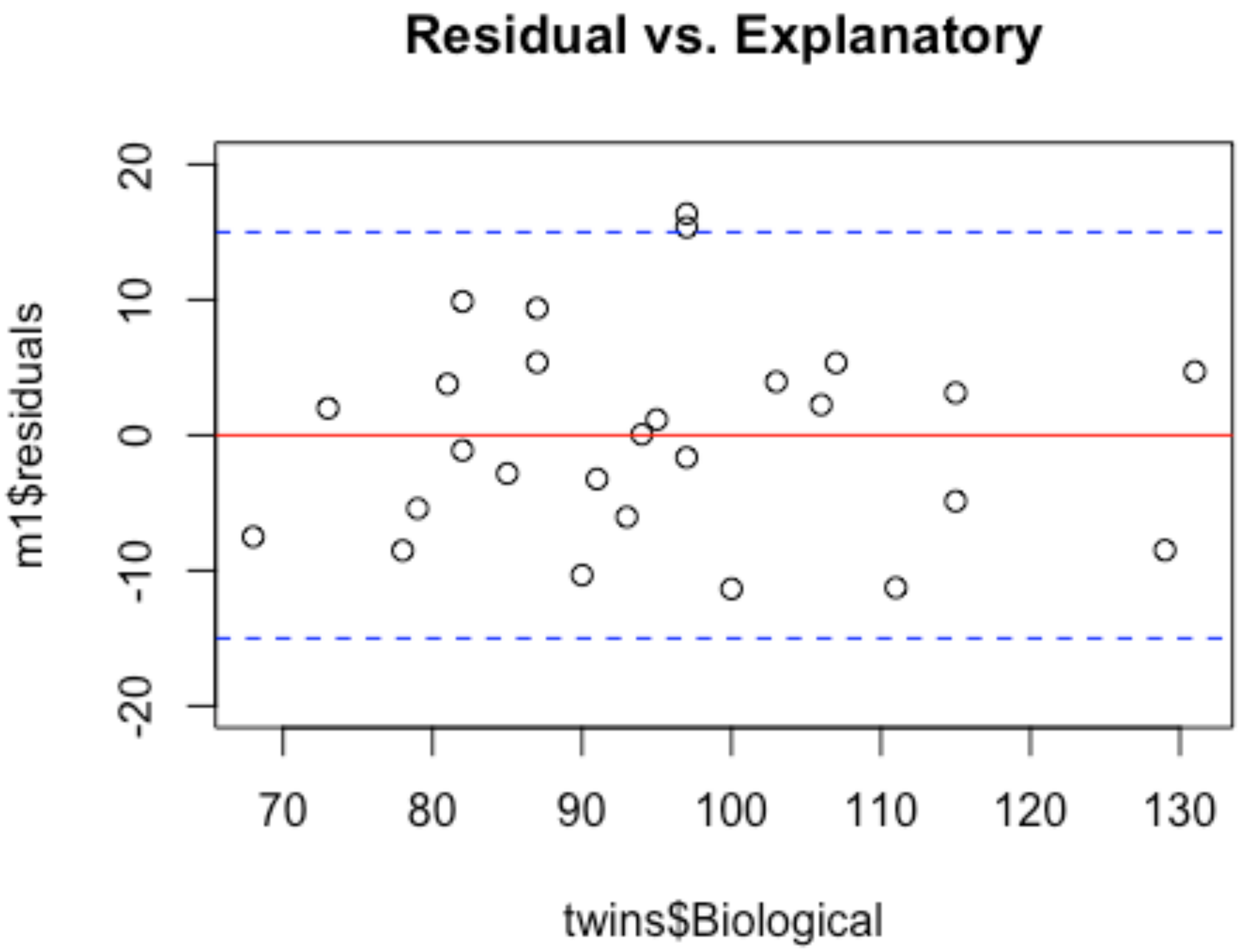


Line Fitting

```
summary(m1)
##
## Call:
## lm(formula = Foster ~ Biological, data = twins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3512  -5.7311   0.0574   4.3244  16.3531
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.20760    9.29990   0.990    0.332
## Biological   0.90144    0.09633   9.358  1.2e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.729 on 25 degrees of freedom
## Multiple R-squared:  0.7779, Adjusted R-squared:  0.769
## F-statistic: 87.56 on 1 and 25 DF,  p-value: 1.204e-09
```

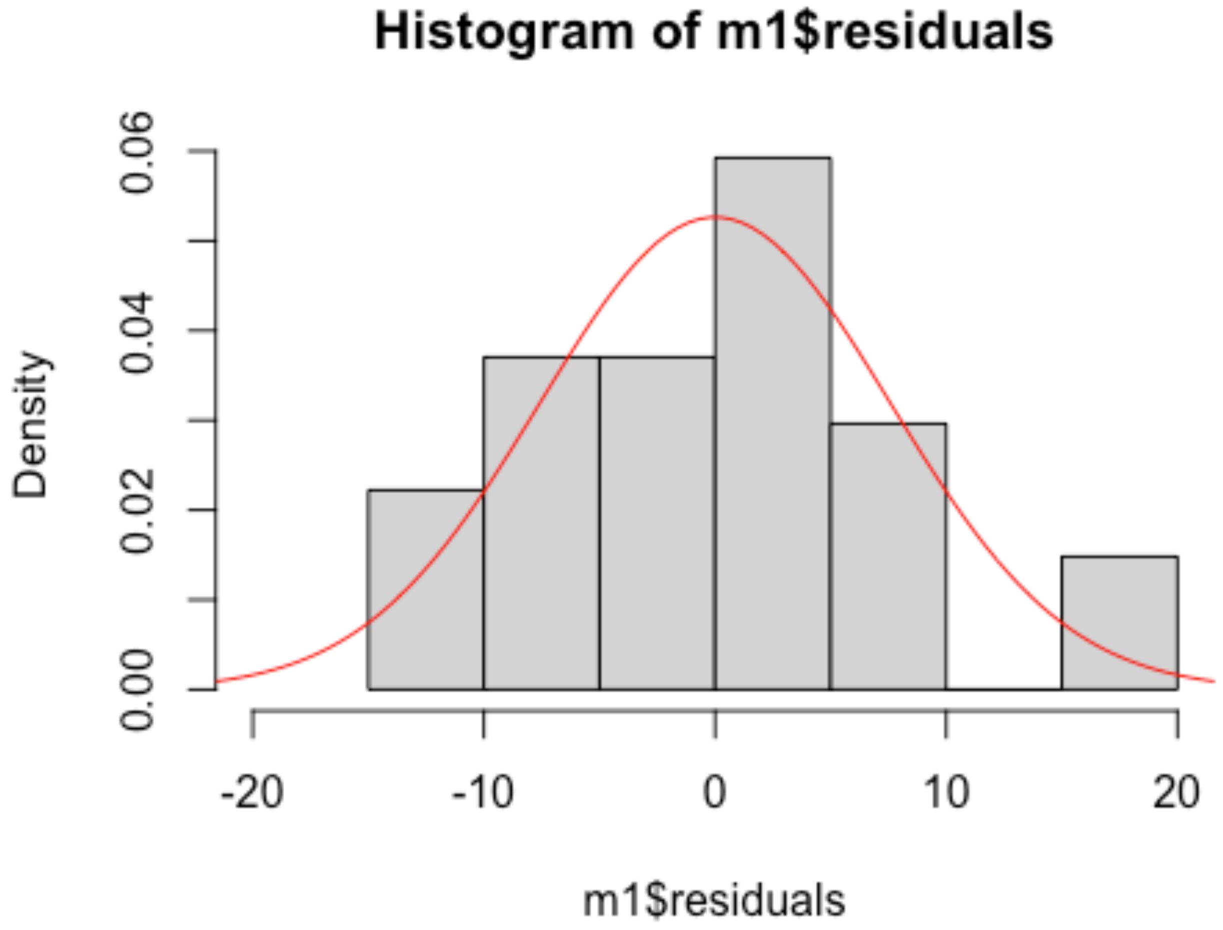
Residual Plots

```
plot(m1$residuals~twins$Biological, main = "Residual vs. Explanatory",
     ylim = c(-20,20))
abline(h = 0, col="red")
abline(h = 15, col="blue", lty =2)
abline(h = -15, col="blue", lty =2)
```



Assumption Check

```
mu = mean(m1$residuals)
sigma = sd(m1$residuals)
x = seq(-3,3, length = 100)
hist(m1$residuals, freq=FALSE, xlim = c(-20,20))
lines(x*sigma+mu, dnorm(x*sigma+mu, mean = mu, sd = sigma),
col="red")
```


Histogram of m1$residuals
Density
m1$residuals