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

R Demo for Chapter 18, part 3: Prediction intervals, r^2 and Residual Plots

Jinko Graham

Load the data, fit the regression model, get CIs

- We'll continue working with the data on low-birthweight babies.

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu) #load the data
lfit <- lm(headcirc ~ gestage,data=lbwt) #fit the regression model
lCI <- predict(lfit,interval="confidence") #get the 95% CI
#Now make a dataframe of the results
lbwtCIs<-data.frame(gestage=lbwt$gestage, lCI )
head(lbwtCIs)
```

```
##   gestage      fit      lwr      upr
## 1      29 26.53581 26.21989 26.85172
## 2      31 28.09591 27.68437 28.50745
## 3      33 29.65602 29.05247 30.25956
## 4      31 28.09591 27.68437 28.50745
## 5      30 27.31586 26.97102 27.66070
## 6      25 23.41559 22.83534 23.99584
```

- Next, we'll compare these CIs to *prediction intervals* ....

95% PIs for example data

```
lPI <- predict(lfit,interval="prediction")
lbwtPIs <- data.frame(gestage=lbwt$gestage, lPI)
head(lbwtPIs,n=3)
```

```
##   gestage      fit      lwr      upr
## 1      29 26.53581 23.36391 29.70770
## 2      31 28.09591 24.91307 31.27875
## 3      33 29.65602 26.44271 32.86933
```

- Predicted values $\hat{y}$ and lower and upper limits of PI are in the columns `fit`, `lwr` and `upper`, respectively.

PIs are wider than CIs

- E.G. Compare for gestational age 29 weeks:

```
lbwtCIs[1,] #95% CI for gestational age 29 weeks
```

```
##   gestage      fit      lwr      upr
## 1      29 26.53581 26.21989 26.85172
```

```
lbwtPIs[1,] #95% PI for gestational age 29 weeks
```

```
##   gestage      fit      lwr     upr
## 1      29 26.53581 23.36391 29.7077
```

- The CI **estimates a parameter**, namely the population-mean head circumference of a baby of gestational age 29 weeks.
  - The CI is an interval ***estimate*** of a ***parameter***.
- The PI **predicts a future random value**, namely the head circumference of a new baby of gestational age 29 weeks.
  - The PI is an interval ***prediction*** of a future ***random value***.

r^2 in simple linear regression

- For the regression of head circumference on gestational age, we can get r^2, the coefficient of determination as follows:

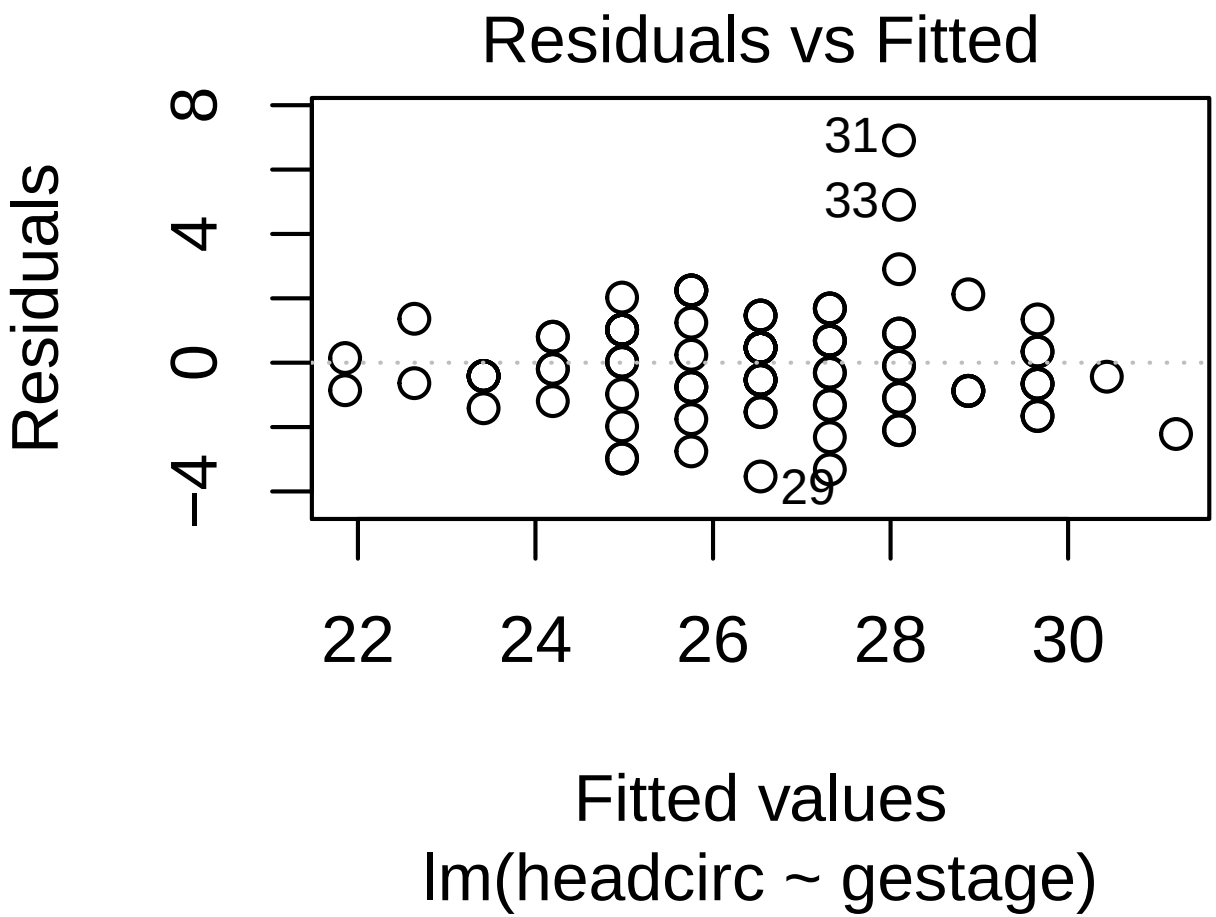
```
with(summary(lfit), r.squared)
```

```
## [1] 0.6094799
```

Plots of residuals vs. fitted values

- In the low-birthweight babies data, the regression of head circumference on gestational age gives:

```
plot(lfit,which=1, add.smooth=FALSE)
```



- R's `plot()` function will plot the regression-modelling output in `lfit`.
- Recall that we got `lfit` by applying the `lm()` function to the dataframe `lbwt` in `lfit <- lm(headcirc ~ gestage,data=lbwt)`
- For regression-model objects such as `lfit` the `plot()` function has 6 different diagnostic plots, specified by the `which` argument.
  - The first (`which=1`) is the plot of the residuals *vs* fitted values.
- The resulting plot labels the three most extreme residuals (i.e. residuals farthest from zero) by their case number in the `lbwt` dataframe: 29, 31 and 33.
- These cases can be viewed in the context of the dataframe:

```
lbwt[c(29,31,33),]
```

```
##    headcirc length gestage birthwt momage toxemia
## 29       23     33      29     560     29       0
## 31       35     36      31     900     23       0
## 33       33     39      31    1440     26       0
```