Multiple Regression Parameterizations

Lab 7 Handout

Statistics 139

Topics

- Q1: Influential points
- Q2: Categorical predictors with multiple levels
- Q3: Nonlinearities and polynomials
- Q4: Interactions

Question 1: Influential points The `census_2010.csv` dataset has data on infant mortality and number of doctors for each of the 50 states including Washington, D.C.

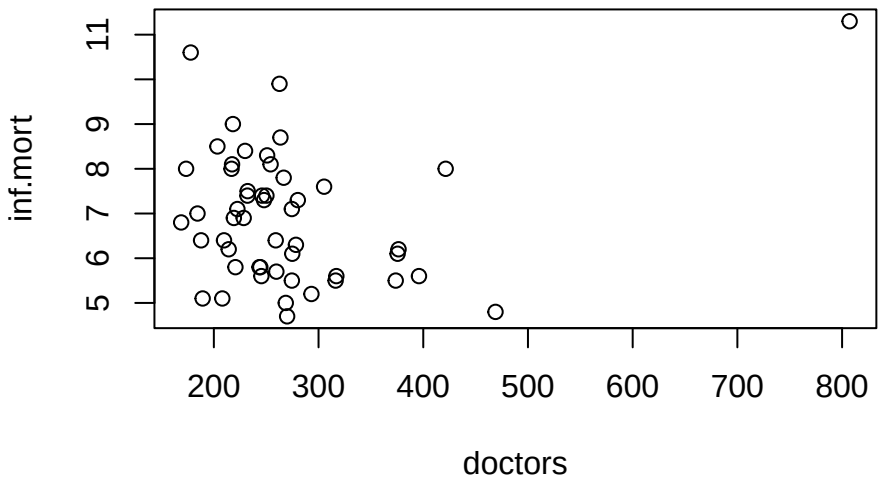
- Infant mortality (`inf.mort`) is measured as number of infant deaths in the first year of life per 1,000 births.
- Number of doctors (`doctors`) is recorded as number of doctors per 100,000 members of the population.

Suppose we are interested in modeling infant mortality rate from number of doctors.

a) Plot the data. Describe what you see—specifically with regards to unusual points? Identify this unusual point.

DC is a high leverage point because it has a lot of doctors compared to other states.

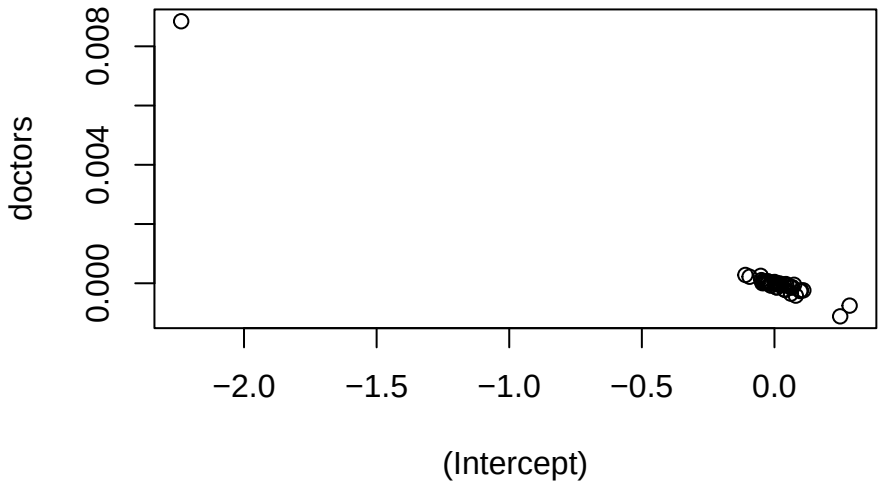
```
census <- read.csv("data/census_2010.csv")
plot(inf.mort~doctors,census)
```



\vspace{1cm}

b) Fit a model predicting infant mortality rate from number of doctors using the complete data, then fit the same model but excluding the influential observation. On a single scatterplot, illustrate the effect of the influential point on the estimated model coefficients.

```
lm.dc <- lm(inf.mort~doctors, census)
dfbeta <- as.data.frame(dfbeta(lm.dc))
plot(dfbeta)
```



\vspace{1cm}

c) What happens to the sampling distribution of $\hat{\beta}_1$ in the presence of an influential point? Apply a bootstrapping approach to the pairs of observations in the complete dataset and describe what you see.

d) From a model interpretation perspective, why might it be reasonable to exclude Washington, DC from an analysis of infant mortality and number of doctors based on this data?

Problem 2: Categorical predictors with multiple levels

The Prevention of REnal and Vascular END-stage Disease (PREVEND) study took place between 2003 and 2006 in the Netherlands. Clinical and demographic data for the 4,095 participants are in the `prevend.csv` data set.

Is RFFT score associated with educational attainment? Yhe variable `Education` indicates the highest level of education that an individual completed: primary school (0), lower secondary school (1), higher secondary school (2), or university (3).

a) Add a variable to the `prevend` data frame that recodes `Education` as a factor variable. The original numeric version of the variable will be used in part d).

b) Create a plot that shows the association between RFFT score and educational attainment. Describe what you see.

c) Apply the ANOVA procedure to explore whether RFFT score is associated with educational attainment. For the purposes of part d), do not apply a correction for multiple testing.

d) Fit a linear model that regresses RFFT score on education level.

i. Fit the model using the factor version of `Education`. Interpret the coefficients, including the intercept. How do the values of the coefficients and associated p-values relate to the output from part c)?

ii. Fit the model using the numeric version of `Education`. How does the interpretation of this model differ from the interpretation of the model in part i.? Which model is preferable?

iii. Check the assumptions for the model in part i. Briefly comment on whether the assumptions seem reasonably satisfied.

e) Is there evidence that mean RFFT score varies across levels of educational attainment? Perform a formal hypothesis test.

f) Let's consider two nested models for predicting RFFT score. The variables of interest are statin use (`Statin`), age (`Age`), and educational attainment (`Age`).

- Model 1: statin use, age
- Model 2: statin use, age, educational attainment

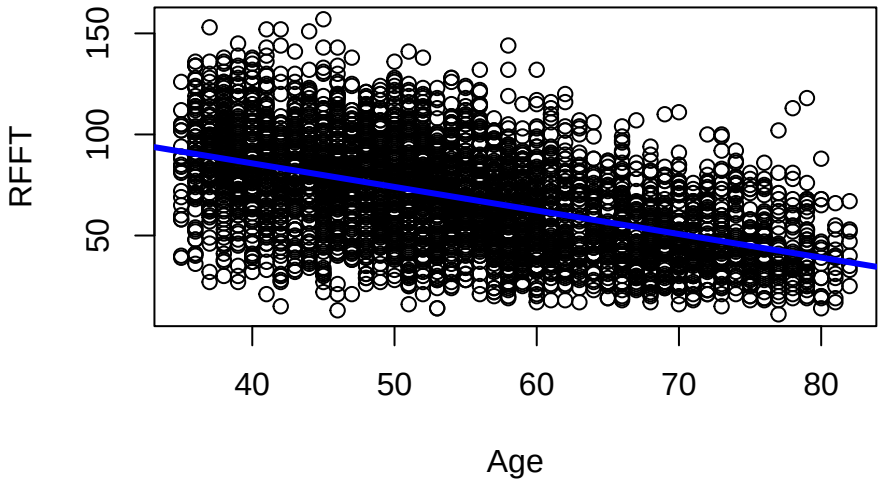
Formally compare the two models to assess whether educational attainment is a useful predictor.

Question 3: Non-linearities

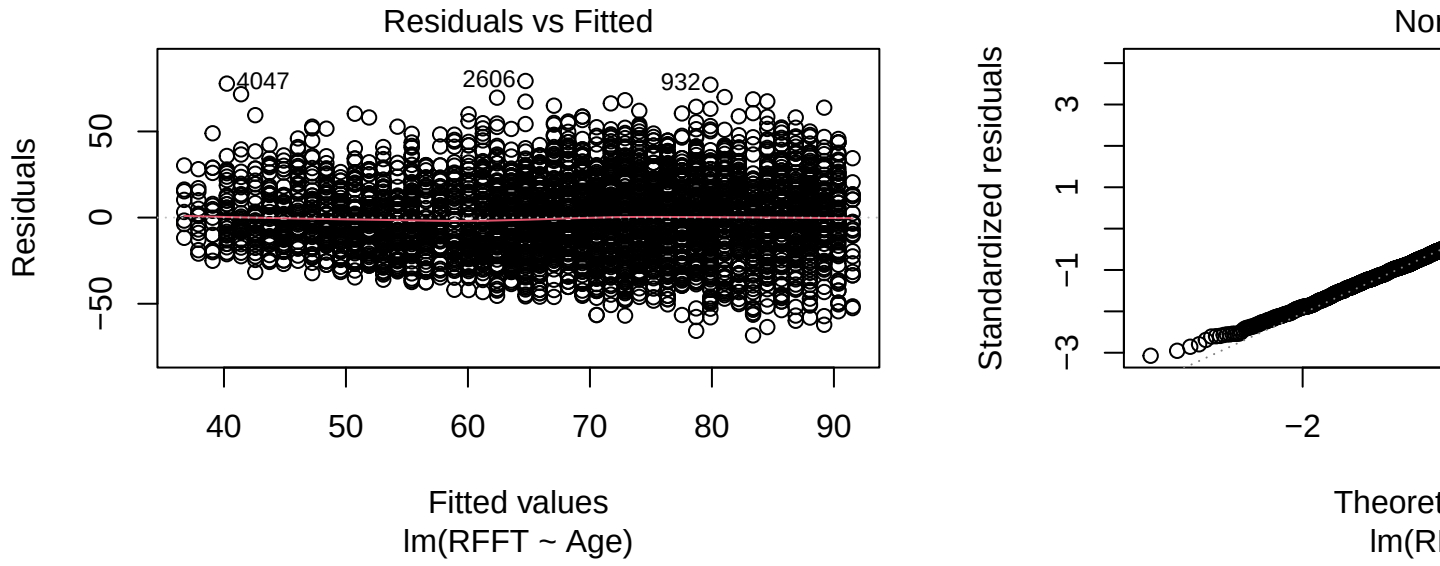
a) Fit a [linear] model to predict RFFT score from Age. Add the estimated line to the scatterplot and comment on the appropriateness of a simple linear model here.

```
prevend <- read.csv("data/prevend.csv")
lm.age <- lm(RFFT~Age, prevend)

plot(RFFT~Age, prevend)
abline(lm.age, col="blue", lwd=3)
```



```
plot(lm.age, which=c(1,2))
```



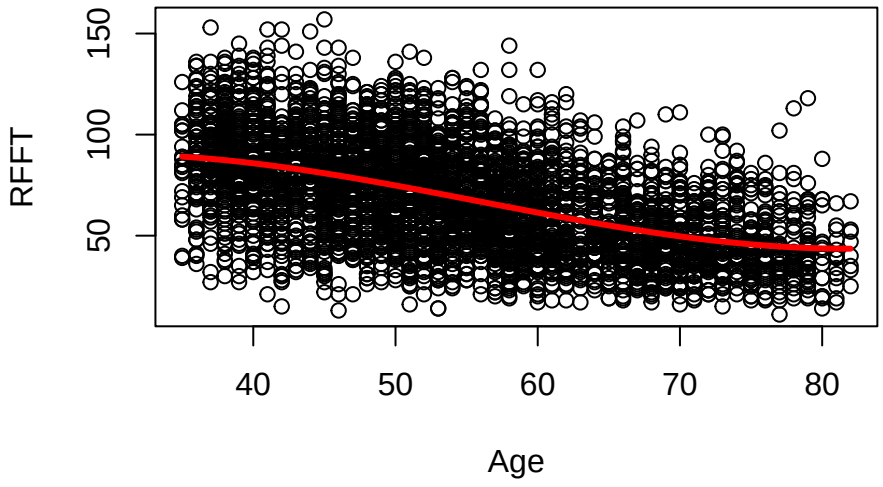
b) Fit a model to predict RFFT score from a cubic model (3rd-order polynomial function) of Age. Interpret the estimates of this model, create a visual to illustrate the relationship of RFFT score with Age based on this model, and formerly test with this model is preferred to handling age simply as a linear effect.

High order term is highly significant.

```
# fit model
lm.age.cubic = lm(RFFT~poly(Age, 3, raw=T), prevend)
lm.age.cubic
```

```
##
## Call:
## lm(formula = RFFT ~ poly(Age, 3, raw = T), data = prevend)
##
## Coefficients:
##            (Intercept)  poly(Age, 3, raw = T)1  poly(Age, 3, raw = T)2
##              23.6263306               5.0430139              -0.1145177
## poly(Age, 3, raw = T)3
##               0.0006827
```

```r
# predict
x = min(prevend$Age):max(prevend$Age)
yhat = predict(lm.age.cubic, new=data.frame(Age=x))
plot(RFFT~Age, prevend)
lines(yhat~x,col="red",lwd=3)
```




```r
# extra sum of squares test
anova(lm.age, lm.age.cubic)
```

```
## Analysis of Variance Table
##
## Model 1: RFFT ~ Age
## Model 2: RFFT ~ poly(Age, 3, raw = T)
##   Res.Df     RSS Df Sum of Sq     F   Pr(>F)
## 1   4093 2027392
## 2   4091 2021958  2    5433.8 5.497 0.004129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

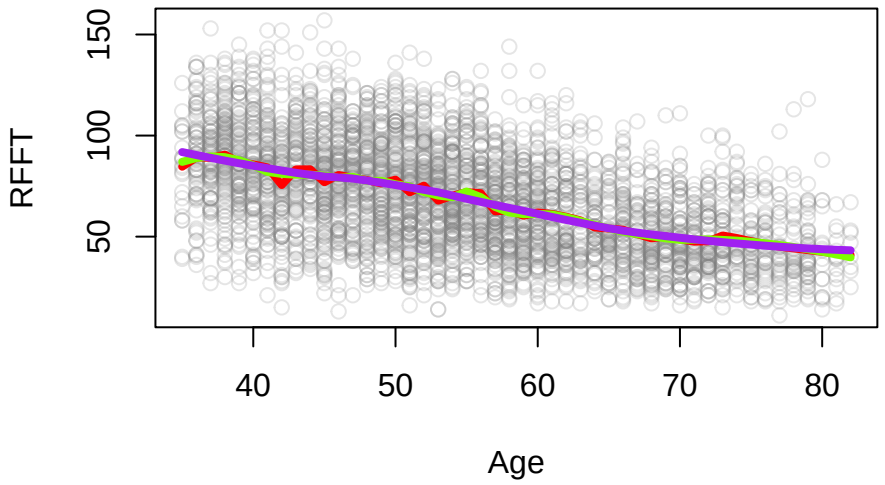
c) What are the implications of using a cubic model here? Why does it make sense mathematically based on the resulting plot?

d) Fit a loess model to predict RFFT score from Age. It is up to you to choose a well-suited value of `span` (include a visual to support your choice).

```
lo1.age <- loess(RFFT~Age, prevend, span=0.1)
lo2.age <- loess(RFFT~Age, prevend, span=0.2)
lo3.age <- loess(RFFT~Age, prevend, span=0.5)

x=min(prevend$Age):max(prevend$Age)
yhat1 = predict(lo1.age, new=data.frame(Age=x))
yhat2 = predict(lo2.age, new=data.frame(Age=x))
yhat3 = predict(lo3.age, new=data.frame(Age=x))

plot(RFFT~Age, prevend, col=rgb(0.5, 0.5, 0.5, 0.2))
lines(yhat1~x,col="red",lwd=4)
lines(yhat2~x,col="chartreuse",lwd=4)
lines(yhat3~x,col="purple",lwd=4)
```



Question 4: Interactions

This problem investigates the relationship between RFFT score (`RFFT`), age (`Age`), and diabetes (`DM`).

a) Fit a linear model that regresses RFFT score on age and diabetes status.

```
lm.age.dm <- lm(RFFT~Age+DM, prevend)
summary(lm.age.dm)
```

```
##
## Call:
## lm(formula = RFFT ~ Age + DM, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.436 -15.634  -0.827  14.733  78.647
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 130.90436    1.68904   77.50  < 2e-16 ***
## Age          -1.13019    0.03055  -36.99  < 2e-16 ***
## DM           -8.26679    1.46563   -5.64 1.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.17 on 4092 degrees of freedom
## Multiple R-squared:  0.2752, Adjusted R-squared:  0.2748
## F-statistic: 776.7 on 2 and 4092 DF,  p-value: < 2.2e-16
```

i. According to the model, how does the average RFFT score for a 60-year-old compare to that of a 50-year-old, if both have diabetes?

Lower by 11.13 points

ii. According to the model, how does the average RFFT score for a 60-year-old compare to that of a 50-year-old, if both do not have diabetes?

Same as above, because we did not include an interaction term in the model so the change is the same in RFFT score for both groups.

b) Fit a linear model for RFFT score from age, diabetes status, and the interaction term between age and diabetes status.

```
lm.age.and.dm <- lm(RFFT~Age*as.factor(DM), prevend)
summary(lm.age.and.dm)
```

```
##
## Call:
## lm(formula = RFFT ~ Age * as.factor(DM), data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.776 -15.571  -1.033  14.627  78.759
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        132.42948    1.72258  76.879  < 2e-16 ***
## Age                 -1.15842    0.03119 -37.143  < 2e-16 ***
## as.factor(DM)1     -48.51672    9.49994  -5.107 3.42e-07 ***
## Age:as.factor(DM)1   0.63364    0.14777   4.288 1.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.13 on 4091 degrees of freedom
## Multiple R-squared:  0.2784, Adjusted R-squared:  0.2779
## F-statistic: 526.1 on 3 and 4091 DF,  p-value: < 2.2e-16
```

i. Write the overall estimated model equation.

$$\widehat{RFFT} = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 DM + \hat{\beta}_3 AgeDM$$

ii. Simplify the model equation for diabetics. Simplify the model equation for non-diabetics.

iii. How does fitting an interaction term change the model? Specifically, how do the interpretations from parts a) i. and ii. change when the model has an interaction term?

c) Fit a model to predict RFFT score from age, educational attainment, and the interaction between the two. Formally test whether the interaction term(s) provide a statistically significant improvement in prediction accuray as measured by R^2 (you will need to fit a second model). Create a plot for the interaction model and summarize the model results.

```r
edu.interact = lm(RFFT~Age*as.factor(Education), prevend)
summary(edu.interact)
```

```
##
## Call:
## lm(formula = RFFT ~ Age * as.factor(Education), data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.910 -14.249  -1.393  13.641  89.329
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  95.04751    6.15117  15.452  < 2e-16 ***
## Age                          -0.75856    0.09658  -7.854 5.10e-15 ***
## as.factor(Education)1         9.97123    6.93510   1.438 0.150570
## as.factor(Education)2        26.46364    6.83174   3.874 0.000109 ***
## as.factor(Education)3        40.42818    6.74464   5.994 2.22e-09 ***
## Age:as.factor(Education)1    -0.05756    0.11062  -0.520 0.602852
## Age:as.factor(Education)2    -0.20813    0.11146  -1.867 0.061932 .
## Age:as.factor(Education)3    -0.27674    0.11024  -2.510 0.012096 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.68 on 4087 degrees of freedom
## Multiple R-squared:  0.3701, Adjusted R-squared:  0.369
## F-statistic:   343 on 7 and 4087 DF,  p-value: < 2.2e-16
```

```r
edu.age = lm(RFFT~Age+as.factor(Education), prevend)
summary(edu.age)
```

```
##
## Call:
## lm(formula = RFFT ~ Age + as.factor(Education), data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.459 -14.101  -1.178  13.407  86.280
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              105.34425    2.13934  49.241  < 2e-16 ***
## Age                       -0.92257    0.02981 -30.950  < 2e-16 ***
## as.factor(Education)1      5.88488    1.20236   4.894 1.02e-06 ***
## as.factor(Education)2     13.86215    1.24977  11.092  < 2e-16 ***
## as.factor(Education)3     24.38332    1.22821  19.853  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.71 on 4090 degrees of freedom
## Multiple R-squared:  0.3682, Adjusted R-squared:  0.3676
## F-statistic:   596 on 4 and 4090 DF,  p-value: < 2.2e-16
```

```
anova(edu.age, edu.interact)
```

```
## Analysis of Variance Table
##
## Model 1: RFFT ~ Age + as.factor(Education)
## Model 2: RFFT ~ Age * as.factor(Education)
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1   4090 1753410
## 2   4087 1748317  3    5092.9 3.9685 0.007771 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

d) Visually assess the linearity assumption for the two models you used in the test in the previous part. How do they compare?