Week 5: On Regression Function Specification

Devesh Tiwari and Jeffrey Yau

September 26, 2016

Agenda

1. Instructor's announcement and any questions from students before we begin (10 - 15 minutes)
2. Mid-course evaluation (10 minutes)
3. Data scaling and its impact on the coefficients (10 - 20 minutes)
4. Feature Engineering (50 - 55)

1. Instructor's announcement and any questions from students before we begin (10 minutes)

- Lab 3 - first deliverable due 9/30
- Any questions before we begin?

2. Mid-course evaluation (10 minutes)

- Please kindy give us feedback on the course evaluation.

3. Data scaling and its impact on the coefficients (10 - 15 minutes)

** Breakout room: 10 minutes   Classwise discussion: 5 - 10 minutes**

- Consider the following simple data set and linear regression model, which tries quantify the relationship between actual weight, *weight*, and reported weight, *repwt*. This data set comes with the *car* library, and the name of the dataset is called *Davis*.

```
# Load the library
library(car)

# Display the structure of the data frame
str(Davis)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

```
nrow(Davis)
```

```
## [1] 200
```

```r
ncol(Davis)
```

```
## [1] 5
```

```r
# List a few rows of the data frame
head(Davis)
```

```
##   sex weight height repwt repht
## 1   M     77    182    77   180
## 2   F     58    161    51   159
## 3   F     53    161    54   158
## 4   M     68    177    70   175
## 5   F     59    157    59   155
## 6   M     76    170    76   165
```

```r
summary(Davis)
```

```
##  sex         weight          height          repwt            repht      
##  F:112   Min.   : 39.0   Min.   : 57.0   Min.   : 41.00   Min.   :148.0  
##  M: 88   1st Qu.: 55.0   1st Qu.:164.0   1st Qu.: 55.00   1st Qu.:160.5  
##          Median : 63.0   Median :169.5   Median : 63.00   Median :168.0  
##          Mean   : 65.8   Mean   :170.0   Mean   : 65.62   Mean   :168.5  
##          3rd Qu.: 74.0   3rd Qu.:177.2   3rd Qu.: 73.50   3rd Qu.:175.0  
##          Max.   :166.0   Max.   :197.0   Max.   :124.00   Max.   :200.0  
##                                          NA's   :17       NA's   :17
```

```r
class(Davis$sex)
```

```
## [1] "factor"
```

```r
levels(Davis$sex)
```

```
## [1] "F" "M"
```

```r
# Frequency table of sex
table(Davis$sex)
```

```
## 
##   F   M 
## 112  88
```

```r
prop.table(table(Davis$sex))
```

```
## 
##    F    M 
## 0.56 0.44
```

```r
# Note that the variables are recorded using the metric system:
# height and reported height are in cm
# weight and reported weight are in kg

df<-Davis
str(df)
```

```
## 'data.frame':    200 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

```r
# remove missing value (just for the sake of doing this exercise)
df2 <- df[complete.cases(df),]
str(df2)
```

```
## 'data.frame':    181 obs. of  5 variables:
##  $ sex   : Factor w/ 2 levels "F","M": 2 1 1 2 1 2 2 2 2 2 ...
##  $ weight: int  77 58 53 68 59 76 76 69 71 65 ...
##  $ height: int  182 161 161 177 157 170 167 186 178 171 ...
##  $ repwt : int  77 51 54 70 59 76 77 73 71 64 ...
##  $ repht : int  180 159 158 175 155 165 165 180 175 170 ...
```

```r
# Now, consider the following regression
davis.mod <- lm(weight ~ repwt, data=df2)
summary(davis.mod)
```

```
##
## Call:
## lm(formula = weight ~ repwt, data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -7.023  -1.822  -0.779   0.589 108.663
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.46083    3.05765   1.786   0.0758 .
## repwt        0.92636    0.04556  20.333   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.456 on 179 degrees of freedom
## Multiple R-squared:  0.6979, Adjusted R-squared:  0.6962
## F-statistic: 413.4 on 1 and 179 DF,  p-value: < 2.2e-16
```

Note * 1 kg = 2.2 lbs * 1 in = 2.54 cm

1. Write down the estimated regression and interpret the coefficient associcated with the variable *repwt*.
2. Rerun the above regression model but change the measurement unit to pounds. How is the coefficient estimate change? Is the change consistent with your intuition and what you read from the book? Please explain. How would it affect statistical inference, if at all?

4. Feature Engineering

We will practice the following funcational form transformation in this series of exericse: * 4.1 Funcational form transformation - log, log2, low order polynomial * 4.2 Allowing for different intercepts and different slopes for different subgroups in the sample

4.1 Funcational form transformation - log, log2, low order polynomial ** Breakout room: 10 minutes   Classwise discussion: 10 minutes **

Consider the following data set named *Prestige*, coming with the *car* library.

First, conduct a quick EDA, focusing on the varibles *income* and *prestige*

```
library(car)
str(Prestige)
```

```
## 'data.frame':    102 obs. of  6 variables:
##  $ education: num  13.1 12.3 12.8 11.4 14.6 ...
##  $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
##  $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
##  $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
##  $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

Now, consider the following variant of the models:

```
prestige.mod1a <- lm(prestige ~ education + income + women, data = Prestige)
summary(prestige.mod1a)
```

```
##
## Call:
## lm(formula = prestige ~ education + income + women, data = Prestige)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.8246  -5.3332  -0.1364   5.1587  17.5045
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.7943342  3.2390886  -2.098   0.0385 *
## education    4.1866373  0.3887013  10.771  < 2e-16 ***
## income       0.0013136  0.0002778   4.729 7.58e-06 ***
## women       -0.0089052  0.0304071  -0.293   0.7702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.846 on 98 degrees of freedom
## Multiple R-squared:  0.7982, Adjusted R-squared:  0.792
## F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

```
prestige.mod1b <- lm(prestige ~ education + log(income) + women, data = Prestige)
summary(prestige.mod1b)
```

```
##
## Call:
## lm(formula = prestige ~ education + log(income) + women, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.364  -4.429  -0.101   4.316  19.179
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -110.9658    14.8429  -7.476 3.27e-11 ***
## education      3.7305     0.3544  10.527  < 2e-16 ***
## log(income)   13.4382     1.9138   7.022 2.90e-10 ***
## women          0.0469     0.0299   1.568     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.093 on 98 degrees of freedom
## Multiple R-squared:  0.8351, Adjusted R-squared:   0.83
## F-statistic: 165.4 on 3 and 98 DF,  p-value: < 2.2e-16
```

```r
prestige.mod1c <- lm(prestige ~ education + log2(income)  + women, data = Prestige)
summary(prestige.mod1c)
```

```
##
## Call:
## lm(formula = prestige ~ education + log2(income) + women, data = Prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.364  -4.429  -0.101   4.316  19.179
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -110.9658    14.8429  -7.476 3.27e-11 ***
## education       3.7305     0.3544  10.527  < 2e-16 ***
## log2(income)    9.3147     1.3265   7.022 2.90e-10 ***
## women           0.0469     0.0299   1.568     0.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.093 on 98 degrees of freedom
## Multiple R-squared:  0.8351, Adjusted R-squared:   0.83
## F-statistic: 165.4 on 3 and 98 DF,  p-value: < 2.2e-16
```

```r
prestige.mod1d <- lm(prestige ~ education + income + I(income^2)  + women, data = Prestige)
summary(prestige.mod1d)
```

```
##
## Call:
## lm(formula = prestige ~ education + income + I(income^2) + women,
##     data = Prestige)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7474  -4.5061  -0.4951   3.9701  20.2235
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.348e+01  3.370e+00  -4.001 0.000123 ***
## education    3.292e+00  4.148e-01   7.936 3.67e-12 ***
## income       4.403e-03  7.657e-04   5.750 1.03e-07 ***
## I(income^2) -1.097e-07  2.563e-08  -4.281 4.37e-05 ***
## women        7.119e-02  3.370e-02   2.113 0.037211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.233 on 97 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8233
## F-statistic: 118.6 on 4 and 97 DF,  p-value: < 2.2e-16
```

- Interpret the coefficient estimate assoicate with the various function of the *income* variable in the above estimated regression models
- In the *prestige.mod1d* model, what is the effect of *income* on prestige? Ideally, plot the effect.

** 4.2 Allowing for different intercepts and different slopes for different subgroups in the sample   Breakout room: 15 minutes   Classwise discussion: 10 minutes **

Consider the following model:

```
prestige.mod2a <- lm(prestige ~ education + log(income) + type, data = Prestige)
```

Based on this regression model, what is the interpretation of the coefficient estimates associated with the type variables?

Lastly, - respecify in the model *prestige.mod2a* and allow for different slopes for different *type* (i.e. bc, wc, prof). - write down the respecified model - what is the interpretation of the coefficient estimates associated with the type variables?