Project 2

Sol Rabine and Keqing Lu

9/12/2022

1. JWHT Chapter 2. Exercise 5.

What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

#Very flexible models have the best and most accurate fit to the data itself. They are preffered when there is a large number of observations and when the goal is to make predictions. Very flexible approaches tend to be less legible. A less flexible approach, like a linear approach may be preferred when giving a broad summary of the data, or when you are interested in the average of the data.

#2. Faraway Chapter 2. Exercise 2. The dataset uswages is drawn as a sample from the Current Population Survey in 1988. Fit a model with weekly wages as the response and years of education and experience as predictors in linear regression. Report and give a simple interpretation to the regression coefficient for years of education. Now fit the same model but with logged weekly wages. Give an interpretation to the regression coefficient for years of education. Which interpretation is more natural?

```
library(faraway)
data("uswages")
```

```
pred_wage<-lm(wage~educ+exper,uswages)
summary(pred_wage)
```

```
##
## Call:
## lm(formula = wage ~ educ + exper, data = uswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1018.2  -237.9   -50.9   149.9  7228.6
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
## educ          51.1753     3.3419  15.313  < 2e-16 ***
## exper          9.7748     0.7506  13.023  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 427.9 on 1997 degrees of freedom
## Multiple R-squared:  0.1351, Adjusted R-squared:  0.1343
## F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
coef(pred_wage)
```

```
## (Intercept)        educ       exper
## -242.799412   51.175268    9.774767
```

The regression coefficient for education is 51.1753. This means that each increase of one year of education is associated with a 51.1753 unit increase in wage when experience is fixed.

```
log_wage <- log(uswages$wage)
wlogs<- bind_cols(uswages, log_wage)
pred_log_wage<-lm(log_wage~educ+exper,wlogs)
summary(pred_log_wage)
```

```
##
## Call:
## lm(formula = log_wage ~ educ + exper, data = wlogs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7533 -0.3495  0.1068  0.4381  3.5699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.650319   0.078354   59.35   <2e-16 ***
## educ        0.090506   0.005167   17.52   <2e-16 ***
## exper       0.018079   0.001160   15.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6615 on 1997 degrees of freedom
## Multiple R-squared:  0.1749, Adjusted R-squared:  0.174
## F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

```
coef(pred_log_wage)
```

```
## (Intercept)        educ       exper 
##  4.65031905  0.09050628  0.01807855
```

increasing education by one changes the log odds by .09506