Predicting with regression, multiple covariates

Jeffrey Leek

May 18, 2016

Example: predicting wages

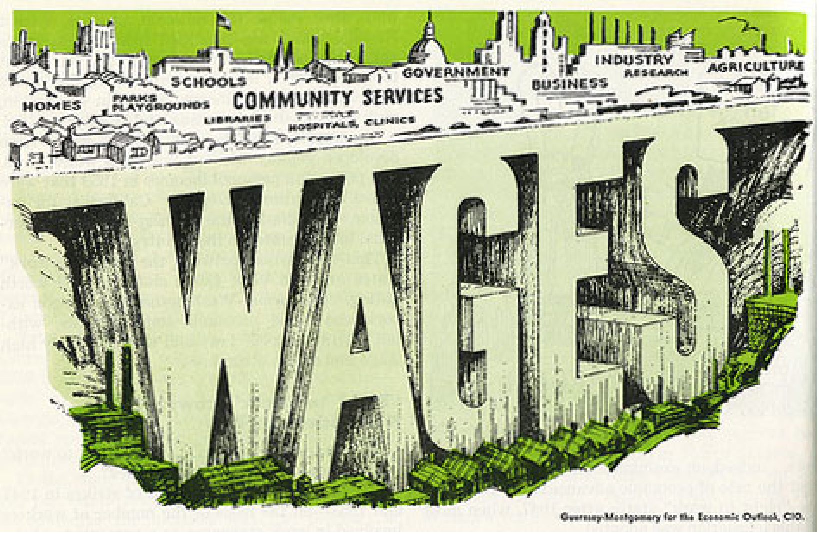


Image Credit

`http://www.cahs-media.org/the-high-cost-of-low-wages`

Data from: ISLR package from the book: Introduction to statistical

Example: Wage data

```
library(ISLR); library(ggplot2); library(caret);
```

```
## Loading required package: lattice
```

```
data(Wage); Wage <- subset(Wage,select=-c(logwage))
summary(Wage)
```

```
##       year           age                  sex
##  Min.   :2003   Min.   :18.00   1. Male  :3000   1. Neve
##  1st Qu.:2004   1st Qu.:33.75   2. Female:    0   2. Marr
##  Median :2006   Median :42.00                     3. Wido
##  Mean   :2006   Mean   :42.41                     4. Divo
##  3rd Qu.:2008   3rd Qu.:51.00                     5. Sepa
##  Max.   :2009   Max.   :80.00
##
##        race                   education
##  1. White:2480   1. < HS Grad         :268   2. Middle Atla
##  2. Black: 293   2. HS Grad           :971   1. New Englan
```

Get training/test sets

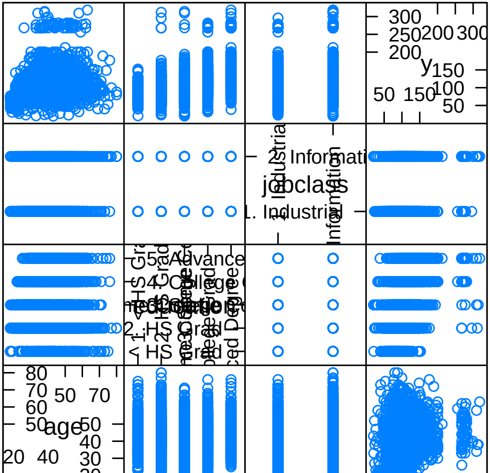
```r
inTrain <- createDataPartition(y=Wage$wage,
                              p=0.7, list=FALSE)
training <- Wage[inTrain,]; testing <- Wage[-inTrain,]
dim(training); dim(testing)
```

```
## [1] 2102   11
```

```
## [1] 898  11
```

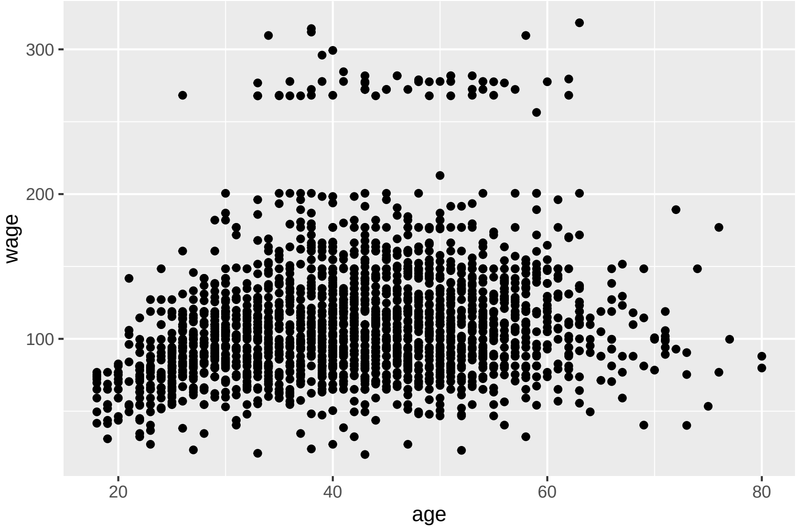
Feature plot

```
featurePlot(x=training[,c("age","education","jobclass")],
            y = training$wage,
            plot="pairs")
```



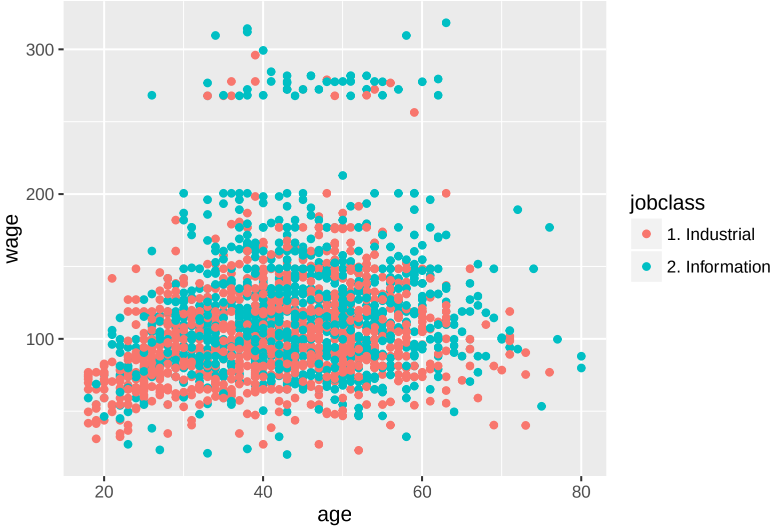
Plot age versus wage

```r
qplot(age,wage,data=training)
```



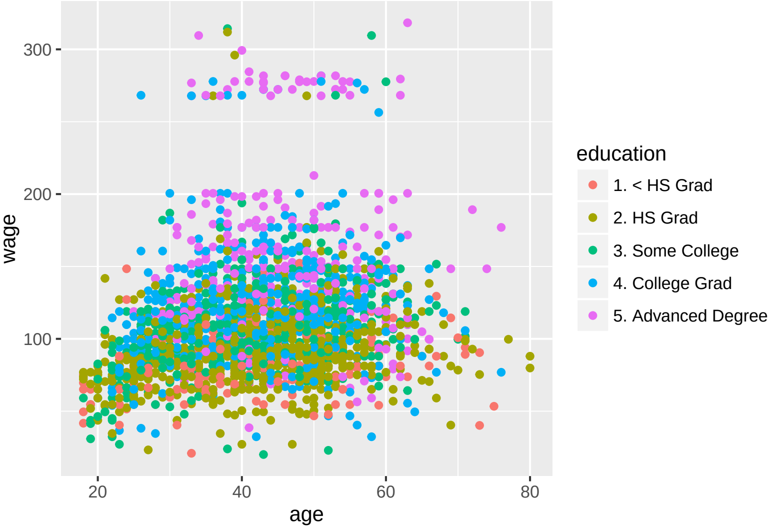
Plot age versus wage colour by jobclass

```r
qplot(age,wage,colour=jobclass,data=training)
```



Plot age versus wage colour by education

```
qplot(age,wage,colour=education,data=training)
```



Fit a linear model

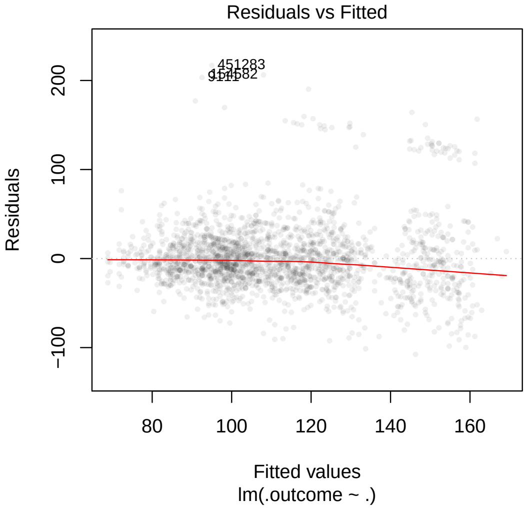
$$ED_i = b_0 + b_1 age + b_2 I(Jobclass_i = "Information") + \sum_{k=1}^{4} \gamma_k I(education_i =$$

```r
modFit<- train(wage ~ age + jobclass + education,
               method = "lm",data=training)
finMod <- modFit$finalModel
print(modFit)
```

```
## Linear Regression
##
## 2102 samples
##   10 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 2102, 2102, 2102, 2102, 2102, 2
## Resampling results:
```

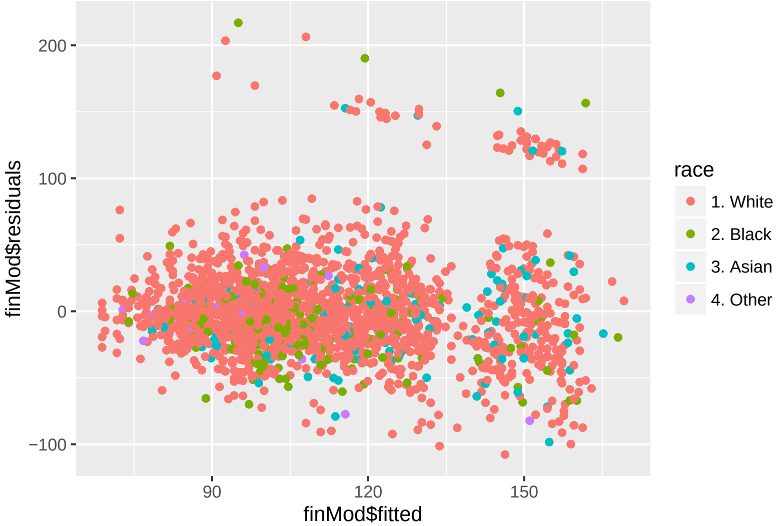
Diagnostics

```r
plot(finMod,1,pch=19,cex=0.5,col="#00000010")
```



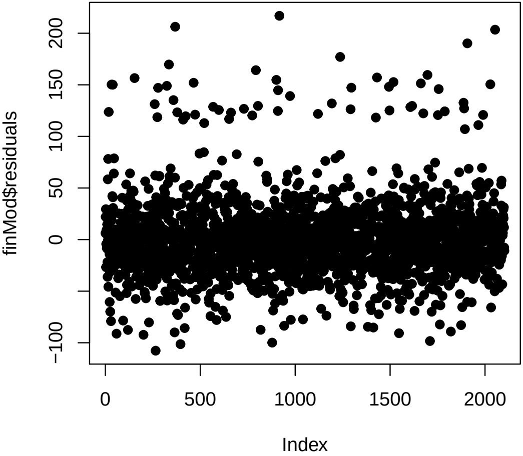
Color by variables not used in the model

```
qplot(finMod$fitted,finMod$residuals,colour=race,data=trai
```



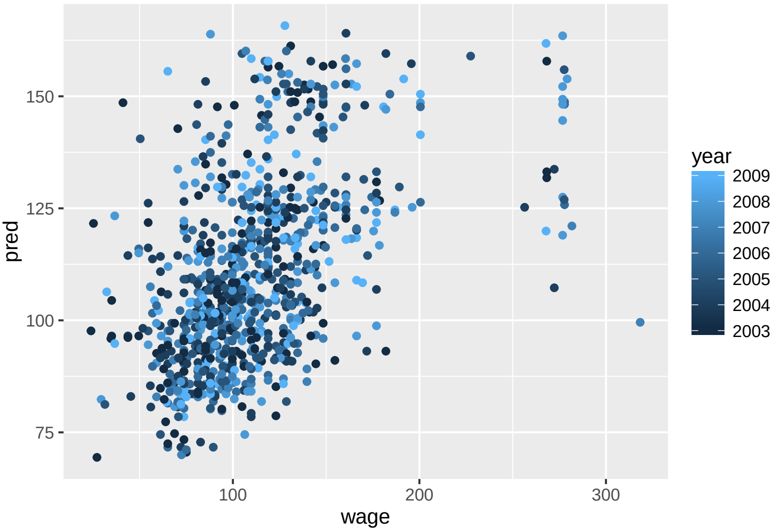
Plot by index

```
plot(finMod$residuals,pch=19)
```



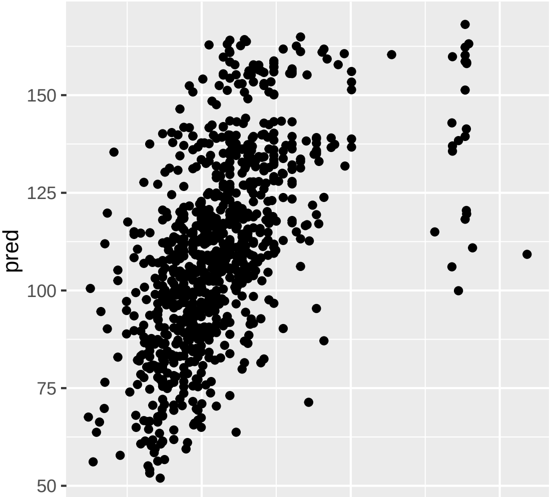
Predicted versus truth in test set

```
pred <- predict(modFit, testing)
qplot(wage,pred,colour=year,data=testing)
```



If you want to use all covariates

```
modFitAll<- train(wage ~ .,data=training,method="lm")
pred <- predict(modFitAll, testing)
qplot(wage,pred,data=testing)
```



Notes and further reading

- Often useful in combination with other models
- Elements of statistical learning
- Modern applied statistics with S
- Introduction to statistical learning