King County Housing Data

Multiple Linear Regression

```
house = read.csv("~/Dropbox/Priya-PhD- Documents/Courses/Data Analysis and Visualization-Spring 2019/Da
colnames(house)
```

```
##  [1] "id"            "date"          "price"         "bedrooms"
##  [5] "bathrooms"     "sqft_living"   "sqft_lot"      "floors"
##  [9] "waterfront"    "view"          "condition"     "grade"
## [13] "sqft_above"    "sqft_basement" "yr_built"      "yr_renovated"
## [17] "zipcode"       "lat"           "long"          "sqft_living15"
## [21] "sqft_lot15"
```

```
head(house[,c("price","sqft_living","sqft_lot","bathrooms","bedrooms","grade")])
```

```
##     price sqft_living sqft_lot bathrooms bedrooms grade
## 1  221900        1180     5650      1.00        3     7
## 2  538000        2570     7242      2.25        3     7
## 3  180000         770    10000      1.00        2     6
## 4  604000        1960     5000      3.00        4     7
## 5  510000        1680     8080      2.00        3     8
## 6 1225000        5420   101930      4.50        4    11
```

```
house = house[complete.cases(house),]
```

Predict the sale price from other variables.

```
house_lm = lm(price ~ sqft_living + sqft_lot + bathrooms + bedrooms + grade, data = house, na.action = 
house_lm
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_lot + bathrooms + bedrooms +
##     grade, data = house, na.action = na.omit)
##
## Coefficients:
## (Intercept)  sqft_living     sqft_lot    bathrooms     bedrooms
##  -4.716e+05    2.313e+02   -3.254e-01   -2.797e+04   -4.074e+04
##       grade
##   9.559e+04
```

Interpretation:

Adding an extra finished square foot to a house increases the estimated value by roughly $231, adding 1000 finished square feet implies the value will increase by $231,300

Assessing the model.

```r
summary(house_lm)
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_lot + bathrooms + bedrooms +
##     grade, data = house, na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1037833  -135336   -22451    97778  4618420
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.716e+05  1.490e+04 -31.652  < 2e-16 ***
## sqft_living  2.313e+02  3.622e+00  63.872  < 2e-16 ***
## sqft_lot    -3.254e-01  4.154e-02  -7.835 4.92e-15 ***
## bathrooms   -2.797e+04  3.479e+03  -8.041 9.37e-16 ***
## bedrooms    -4.074e+04  2.295e+03 -17.754  < 2e-16 ***
## grade        9.559e+04  2.313e+03  41.320  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 247800 on 21607 degrees of freedom
## Multiple R-squared:  0.5446, Adjusted R-squared:  0.5445
## F-statistic:  5169 on 5 and 21607 DF,  p-value: < 2.2e-16
```

With the housing data, older sales are less reliable than more recent sales. We can compute a weight as the number of years since 2005

Weighted Linear Regression

```r
house$weight = abs(house$yr_built - 2005)
class(house$yr_built)
```

```
## [1] "integer"
```

```r
summary(house$weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     9.0    30.0    35.2    54.0   105.0
```

```r
house_wt = lm(price ~ sqft_living + sqft_lot + bathrooms + bedrooms + grade, data = house, weight = wei
round(cbind(house_lm = house_lm$coefficients, house_wt = house_wt$coefficients), digits = 3)
```

```
##                house_lm    house_wt
## (Intercept) -471575.692 -540483.510
## sqft_living     231.350     246.066
## sqft_lot         -0.325      -0.414
## bathrooms    -27973.439  -17010.094
## bedrooms     -40744.142  -41282.021
## grade         95586.697  108078.686
```

Model Selection and Stepwise Regression

```
library(MASS)

house_full = lm(price ~ sqft_living + sqft_lot + bathrooms + bedrooms + grade + yr_renovated + yr_built
step_lm = stepAIC(house_full, direction = "both")
```

```
## Start:  AIC=608982.4
## price ~ sqft_living + sqft_lot + bathrooms + bedrooms + grade + 
##     yr_renovated + yr_built + sqft_basement
## 
##                 Df  Sum of Sq        RSS    AIC
## - yr_renovated   1 1.3296e+11 3.5747e+16 608080
## <none>                        3.5747e+16 608082
## - sqft_lot       1 3.7678e+13 3.5785e+16 608103
## - bathrooms      1 8.3463e+13 3.5831e+16 608131
## - sqft_basement  1 1.0331e+14 3.5850e+16 608143
## - bedrooms       1 7.5846e+14 3.6506e+16 608534
## - sqft_living    1 5.4354e+15 4.1183e+16 611140
## - grade          1 5.9915e+15 4.1739e+16 611429
## - yr_built       1 7.5362e+15 4.3283e+16 612215
## 
## Step:  AIC=608980.5
## price ~ sqft_living + sqft_lot + bathrooms + bedrooms + grade + 
##     yr_built + sqft_basement
## 
##                 Df  Sum of Sq        RSS    AIC
## <none>                        3.5747e+16 608080
## + yr_renovated   1 1.3296e+11 3.5747e+16 608082
## - sqft_lot       1 3.7595e+13 3.5785e+16 608101
## - bathrooms      1 8.6728e+13 3.5834e+16 608131
## - sqft_basement  1 1.0390e+14 3.5851e+16 608141
## - bedrooms       1 7.6291e+14 3.6510e+16 608535
## - sqft_living    1 5.4406e+15 4.1188e+16 611140
## - grade          1 5.9969e+15 4.1744e+16 611430
## - yr_built       1 8.0569e+15 4.3804e+16 612471
```

```
step_lm$coefficients
```

```
##   (Intercept)   sqft_living      sqft_lot     bathrooms      bedrooms 
##  7.389614e+06  2.296850e+02 -1.892725e-01  2.384940e+04 -3.923810e+04 
##         grade      yr_built sqft_basement 
##  1.329728e+05 -4.185122e+03 -3.420837e+01
```

```
update(step_lm,.~.,-sqft_living  -sqft_basement -bathrooms)
```

```
## 
## Call:
## lm(formula = price ~ sqft_living + sqft_lot + bathrooms + bedrooms + 
##     grade + yr_built + sqft_basement, data = house, weights = weight, 
##     na.action = na.omit)
## 
## Coefficients:
##   (Intercept)    sqft_living       sqft_lot      bathrooms       bedrooms  
##     7.390e+06      2.297e+02     -1.893e-01      2.385e+04     -3.924e+04
##         grade       yr_built  sqft_basement
##     1.330e+05     -4.185e+03     -3.421e+01
```

Confounding Variables

```
library(magrittr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
zip_groups = house %>%
  mutate(resid = residuals(house_lm)) %>%
  group_by(zipcode) %>%
  summarize(med_resid = median(resid),
            cnt = n()) %>%
  arrange(med_resid) %>%
  mutate(cum_cnt = cumsum(cnt),
         zipgroup = ntile(cum_cnt,5))
house = house %>%
  left_join(select(zip_groups,zipcode,zipgroup),by = "zipcode")



lm(price ~sqft_living + sqft_lot + bathrooms + bedrooms + grade + zipgroup, data = house, na.action = na
```

```
##
## Call:
## lm(formula = price ~ sqft_living + sqft_lot + bathrooms + bedrooms +
##     grade + zipgroup, data = house, na.action = na.omit)
##
## Coefficients:
## (Intercept)  sqft_living     sqft_lot    bathrooms     bedrooms
##  -6.736e+05    2.366e+02    8.574e-02   -1.300e+04   -3.135e+04
##       grade     zipgroup
##   7.200e+04    1.011e+05
```

Interactions and Main Effects

```
lm(price ~ sqft_living *zipgroup + sqft_lot + bathrooms + bedrooms +
    grade  , data = house, na.action = na.omit)
```

```
##
## Call:
## lm(formula = price ~ sqft_living * zipgroup + sqft_lot + bathrooms +
##     bedrooms + grade, data = house, na.action = na.omit)
##
## Coefficients:
##          (Intercept)           sqft_living              zipgroup
##           -2.409e+05             1.038e+00            -4.390e+04
##             sqft_lot             bathrooms              bedrooms
##            2.251e-01            -1.348e+04            -3.466e+04
##                grade  sqft_living:zipgroup
##            8.027e+04             7.038e+01
```