Lecture 7: Gelman Hill Ch 4.1 - 4.3

For various reasons, data transformations may be necessary or result in better interpretations for regression models.

Linear Transformations

Linear transformations of predictors can be formulated as:

$$X^* = a + bX, \quad X \text{ is the predictor on the original scale}$$

X^* transformed predictor

Linear transformations of the predictors do not influence the fit or predictions of a regression model.

The changes are canceled out with XB (predicted value)

Recall the general interpretation of the regression coefficients is "the average difference in y when comparing units that differ by one unit, on predictor j, and are otherwise identical." However, consider two covariates:

1. Living space (sqft)
2. # of bedrooms

```
library(readr)
library(arm)
Seattle <- read_csv('http://math.montana.edu/ahoegh/teaching/stat408/datasets/SeattleHousing.csv')
```

```
summary(lm(price ~ bedrooms + sqft_living, data = Seattle))
```

```
##
## Call:
## lm(formula = price ~ bedrooms + sqft_living, data = Seattle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1585674  -215744   -14056   181162  2847989
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -110617.57   48122.61  -2.299   0.0218
## bedrooms     -75232.00   17593.19  -4.276 2.11e-05
## sqft_living     465.52      14.09  33.031  < 2e-16
##
## Residual standard error: 391900 on 866 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.619
## F-statistic:   706 on 2 and 866 DF,  p-value: < 2.2e-16
```

Furthermore, the interpretation of the intercept is still a little confusing. In this case, we are looking at a house with zero bedrooms and zero square feet of living space.

BUT THE PRICE IS NEGATIVE!

Standardization

One common option is to standardize the predictors using a z-scale.

$$X_z = \frac{X - \bar{X}}{\hat{\sigma}_x}$$

$X_z = 1$ house that is 1 sd larger than normal

0 : average sized house

-1

```r
library(dplyr)
Seattle <- Seattle %>% mutate(sqft_z = (sqft_living - mean(sqft_living))/sd(sqft_living),
                              bedrooms_z = (bedrooms - mean(bedrooms))/sd(bedrooms))
```

Note it is important to interpret the zero values for each covariate, so the mean living space is 2114 and the mean number of bedrooms is 3.2.

```r
lm_standard <- lm(price ~ bedrooms_z + sqft_z, data = Seattle)
summary(lm_standard)
```

```
##
## Call:
## lm(formula = price ~ bedrooms_z + sqft_z, data = Seattle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1585674  -215744   -14056   181162  2847989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   632592      13293  47.589  < 2e-16
## bedrooms_z    -69442      16239  -4.276 2.11e-05
## sqft_z        536401      16239  33.031  < 2e-16
##
## Residual standard error: 391900 on 866 degrees of freedom
## Multiple R-squared:  0.6199, Adjusted R-squared:  0.619
## F-statistic:   706 on 2 and 866 DF,  p-value: < 2.2e-16
```

Now the intepretation of the parameters is:

- (Intercept): The predicted price of a house with average square footage (2114 sqft) and the average # of bedrooms (3.2).

- (bedrooms_z): The average difference in price when comparing houses that differ by one sd for the number of bedrooms AND all other predictors (sqft) is constant. Note one sd for bedrooms is 0.9.

- (sqft_z): The average difference in price when comparing houses that differ by 1 sd for living space and # bedrooms is the same.

  1 sd for living space (1152).

Note when summarizing the coefficients for homework, exams, or projects, make sure to talk about the size of the difference and include confidence intervals in the discussion.

```
confint(lm_standard)
```

```
##                  2.5 %    97.5 %
## (Intercept)  606501.5 658681.45
## bedrooms_z  -101315.1 -37569.34
## sqft_z       504527.9 568273.61
```

The data can also be centered and/or standardized using different approaches.

* Use reasonable scales, ... divides sqft /1000.

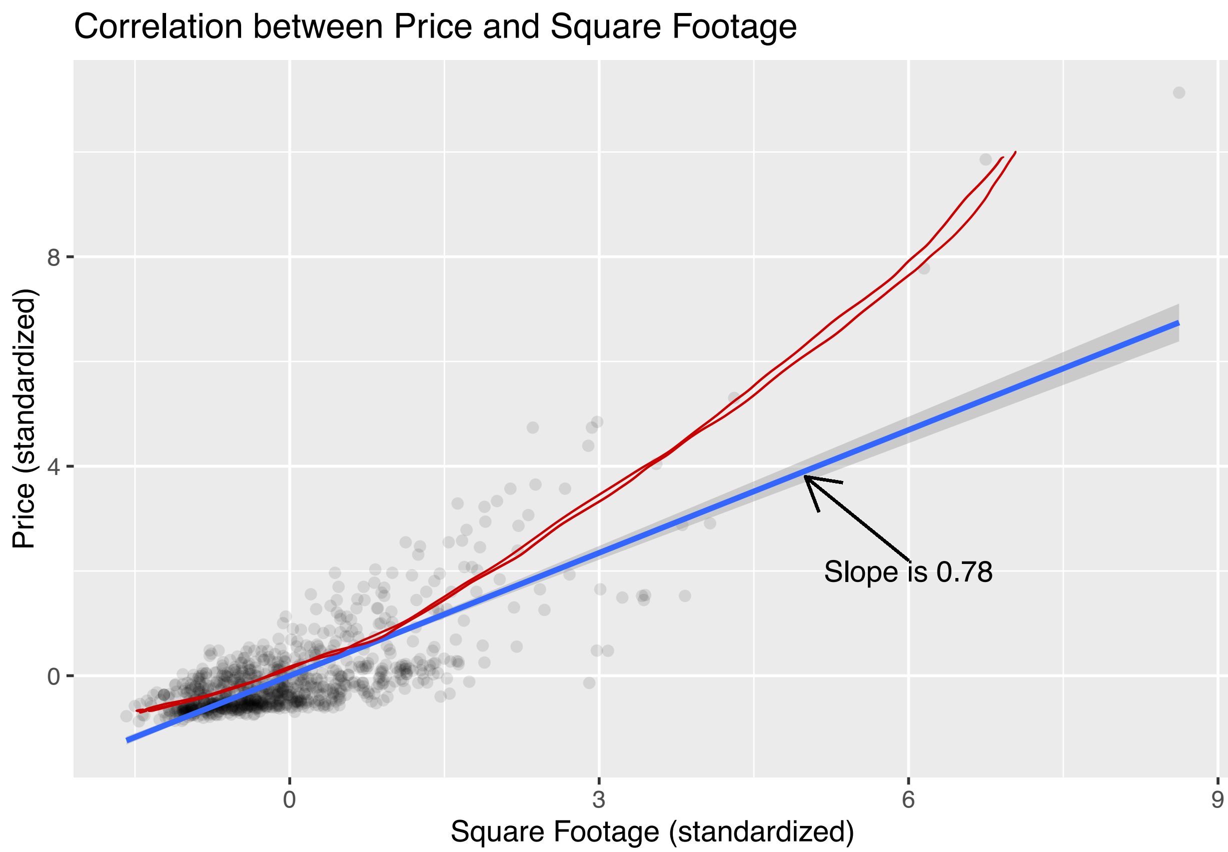
* By subtracting the mean of the predictor.

* Use a conventional centering point, say 2000 sqft 3 bedroom

Correlation

Consider a regression line $y = \beta_0 + \beta_1 x$, where both x and y are standardized.

```
Seattle <- Seattle %>% mutate(y = (price - mean(price))/sd(price),
                              x = (sqft_living - mean(sqft_living))/sd(sqft_living))
library(ggplot2)
Seattle %>% ggplot(aes(y = y, x=x)) + geom_point(alpha = .1) + geom_smooth(method = 'lm') +
  ggtitle('Correlation between Price and Square Footage') + ylab('Price (standardized)') + xlab('Square
```




```
display(lm(y~x, data = Seattle))
```

```
## lm(formula = y ~ x, data = Seattle)
##             coef.est coef.se
## (Intercept) 0.00     0.02
## x           0.78     0.02
## ---
## n = 869, k = 2
## residual sd = 0.62, R-Squared = 0.61
```

```
cor(Seattle$y,Seattle$x)
```

```
## [1] 0.7821967
```

Regression to the mean: