MLR Examples

SDS 291

2/24/2020

Rail Trail Multiple Regression Example

We're still using data from a sample of 104 homes in Northampton, MA to see whether being close to the bike trail enhances the value of the home. Specifically, we're looking at the association between square feet (a house's size) and distance from the rail trail with the house's estimated value in 2014. The variables we're using are:

- `Price2014`: Zillow price estimate from 2014 (in thousands of dollars)
- `Distance`: Distance (in miles) to the nearest entry point to the rail trail network
- `SquareFeet`: Square footage of interior finished space (in thousands of sf)

```
library(Stat2Data)
data("RailsTrails")
m1<-lm(Price2014 ~ SquareFeet + Distance , data = RailsTrails)
summary(m1)
```

```
##
## Call:
## lm(formula = Price2014 ~ SquareFeet + Distance, data = RailsTrails)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -152.15  -30.27   -4.14   25.75  337.93
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   78.985     25.607   3.085  0.00263 **
## SquareFeet   147.920     12.765  11.588  < 2e-16 ***
## Distance     -15.788      7.586  -2.081  0.03994 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.55 on 101 degrees of freedom
## Multiple R-squared:  0.6574, Adjusted R-squared:  0.6506
## F-statistic: 96.89 on 2 and 101 DF,  p-value: < 2.2e-16
```

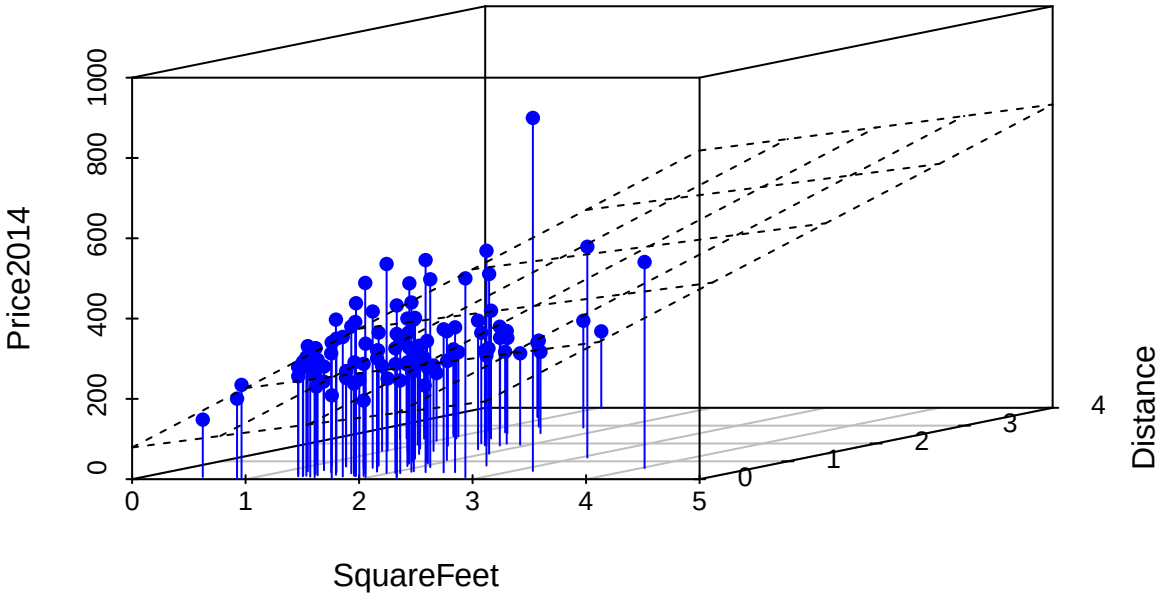
$\widehat{Price2014} = \hat{\beta}_0 + \hat{\beta}_1 \cdot SquareFeet + \hat{\beta}_2 \cdot Distance$

1. What price would this model predict for a 1000 square foot house that is *1 mile* from the rail trail? (Be cautious with the units)

$\widehat{Price2014} = 78.985 + 147.920 \cdot 1 - 15.788 \cdot 1 = 211.117$

2. What price would this model predict for a 1000 square foot house that is *2 miles* from the rail trail? (Be cautious with the units)

$\widehat{Price2014} = 78.985 + 147.920 \cdot 1 - 15.788 \cdot 2 = 195.329$



	1 mile	2 miles	Difference
1000 ft^2	211.117	195.329	-15.788

For every 1 additional mile away from the Rail Trail, the price of a house in 2014 would decrease by $15,788, on average, adjusted for house size in square feet.

Adjusting for Distance Group

Rather than distance in miles, what if we thought a more useful measure would be whether the house was closer (<1 mile from an entrance to the rail trail) or further away ($\geq$ 1 mile from a rail trail entrance)?

- `Price2014`: Zillow price estimate from 2014 (in thousands of dollars)
- `DistGroup`:
  - `Closer`: <1 mile to the nearest entry point to the rail trail network
  - `Farther Away`: $\geq$ 1 mile to the nearest entry point to the rail trail network
- `SquareFeet`: Square footage of interior finished space (in thousands of sf)

`R` treats `DistGroup` as a `factor` variable. It can also be treated as a numeric variable, where one category has the value of 0 and the other category has the value of 1. In other words, you can think that the numerical equivalent is: "Closer" = 0 and "Farther Away" = 1.

```
m2<-lm(Price2014 ~ SquareFeet+DistGroup , data = RailsTrails)
summary(m2)
```

```
##
## Call:
## lm(formula = Price2014 ~ SquareFeet + DistGroup, data = RailsTrails)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -136.55  -30.14   -2.14   22.17  321.40
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              80.10      23.13   3.463 0.000785 ***
## SquareFeet              150.50      11.83  12.724  < 2e-16 ***
## DistGroupFarther Away   -36.97      13.51  -2.736 0.007356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.59 on 101 degrees of freedom
## Multiple R-squared:  0.6673, Adjusted R-squared:  0.6607
## F-statistic: 101.3 on 2 and 101 DF,  p-value: < 2.2e-16
```

$$\widehat{Price2014} = \hat{\beta}_0 + \hat{\beta}_1 \cdot SquareFeet + \hat{\beta}_2 \cdot Far$$

1. What price would this model predict for a 1000 square foot house that is *Closer* from the rail trail?

$\widehat{Price2014} = 80.10 + 150.50 \cdot 1 - 36.97 \cdot 0 = 230.6$

2. What price would this model predict for a 1000 square foot house that is *Farther Away* from the rail trail?

$\widehat{Price2014} = 80.10 + 150.50 \cdot 1 - 36.97 \cdot 1 = 193.63$

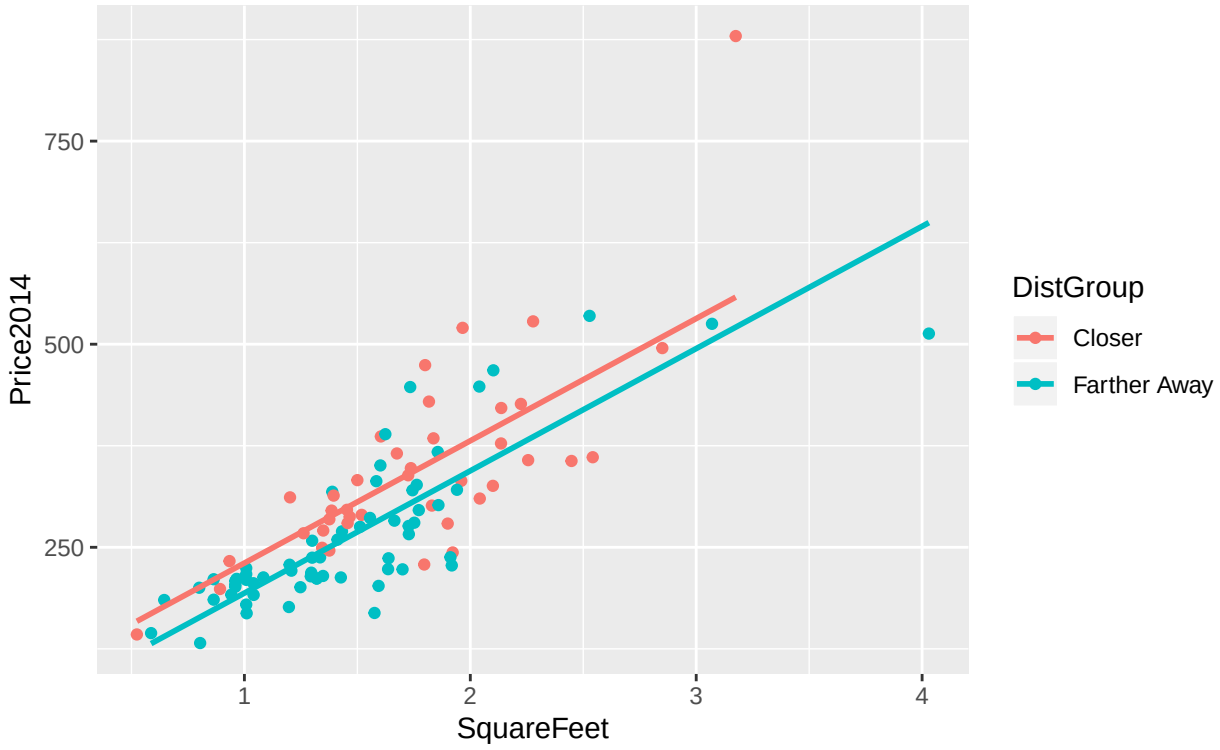
3. What price would this model predict for a 2000 square foot house that is *Closer* from the rail trail?

$\widehat{Price2014} = 80.10 + 150.50 \cdot 2 - 36.97 \cdot 0 = 381.1$

4. What price would this model predict for a 2000 square foot house that is *Farther Away* from the rail trail?

$\widehat{Price2014} = 80.10 + 150.50 \cdot 2 - 36.97 \cdot 1 = 344.13$

```
library(moderndive)
qplot(y=Price2014, x=SquareFeet, color=DistGroup, data=RailsTrails)+geom_parallel_slopes(se=FALSE)
```



ft^2	Close	Far	Difference
1000 ft^2	230.6	193.63	-36.97
2000 ft^2	381.1	344.13	-36.97
Difference	150.5	150.5	

For every additional 1000 ft^2 of house size, our model suggests the estimated 2014 price will increase, on average, by $150,500, adjusted for whether the house is closer or farther from the rail trail.

On average, a house that is farther away from the rail trail will be $-36,970 less than a house closer to the rail trail, adjusted for house size in square feet.

The slope of the line is defined by the house size (ft^2); the distance (far vs. close) affects the intercept. So the lines have parallel slopes, and the difference between two identically sized houses where one is close and one that is far will always be the same amount regardless of how big or small the house is.

Bedrooms

Rather than square feet, let's consider the number of bedrooms the house has, in addition to its distance from the rail trail.

- `Price2014`: Zillow price estimate from 2014 (in thousands of dollars)

- `BedGroup`: Categorical Variable of house type by group of bedrooms:
  - 1-2 bedrooms (reference),
  - 3 bedrooms,
  - 4+ bedrooms
- `Distance`: Distance (in miles) to the nearest entry point to the rail trail network

You can think about `BedGroup` similarly to `DistGroup` and consider the 3 bedroom group output in the model below akin to an indicator variable with the values of 0 or 1: 0 if the house doesn't have 3 bedrooms and 1 if it does have 3 bedrooms. Same for 4+ bedrooms.

```
m3 <- lm(Price2014 ~ Distance+BedGroup, data = RailsTrails)
summary(m3)
```

```
##
## Call:
## lm(formula = Price2014 ~ Distance + BedGroup, data = RailsTrails)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -195.28  -48.15  -13.19   26.02  509.02
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       283.56      26.67  10.633  < 2e-16 ***
## Distance          -42.30      10.18  -4.154 6.89e-05 ***
## BedGroup3 beds     39.88      26.63   1.498 0.137364
## BedGroup4+ beds   106.13      28.49   3.725 0.000323 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.12 on 100 degrees of freedom
## Multiple R-squared:  0.3153, Adjusted R-squared:  0.2948
## F-statistic: 15.35 on 3 and 100 DF,  p-value: 2.746e-08
```

$\widehat{Price2014} = \hat{\beta}_0 + \hat{\beta}_1 \cdot Distance + \hat{\beta}_2 \cdot 3Beds + \hat{\beta}_3 \cdot 4{+}Beds$

1 mile Distance

5. What price would this model predict for a house 1 mile away that has 1-2 Bedrooms?

$\widehat{Price2014} = 283.56 - 42.30 \cdot 1 + 39.88 \cdot 0 + 106.13 \cdot 0 = 241.26$

6. What price would this model predict for a house 1 mile away that has 3 Bedrooms?

$\widehat{Price2014} = 283.56 - 42.30 \cdot 1 + 39.88 \cdot 1 + 106.13 \cdot 0 = 281.14$

7. What price would this model predict for a house 1 mile away that has 4+ Bedrooms?

$\widehat{Price2014} = 283.56 - 42.30 \cdot 1 + 39.88 \cdot 0 + 106.13 \cdot 1 = 347.39$

2 mile Distance

8. What price would this model predict for a house 2 miles away that has 1-2 Bedrooms?

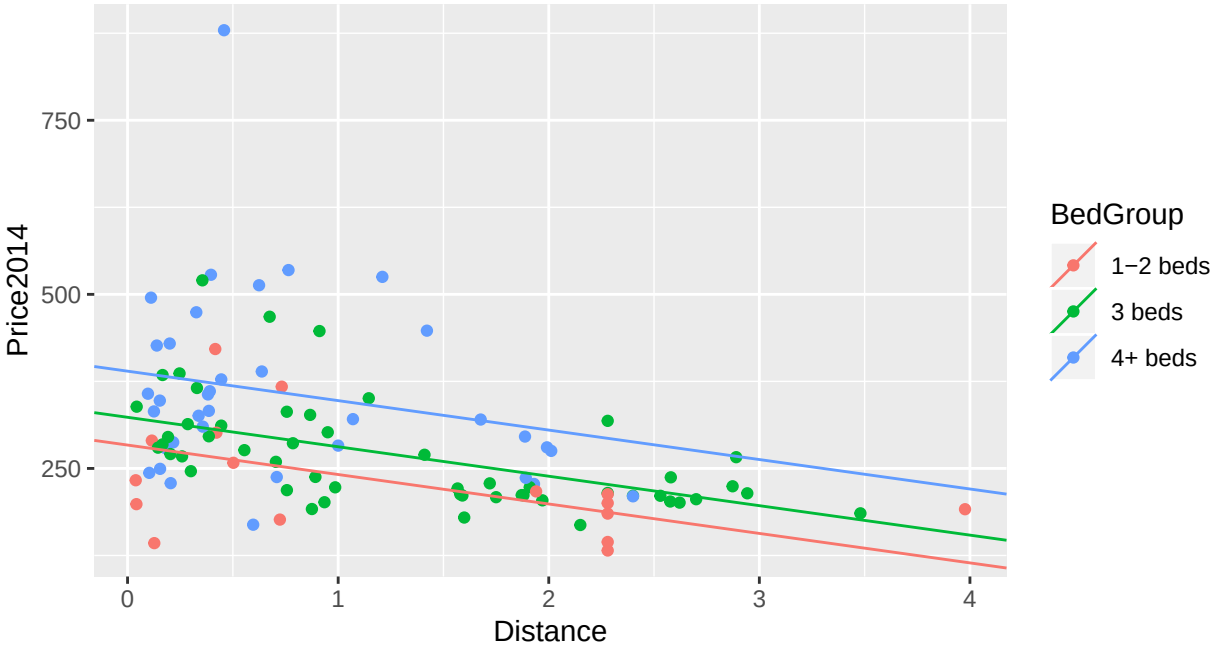
$\widehat{Price2014} = 283.56 - 42.30 \cdot 2 + 39.88 \cdot 0 + 106.13 \cdot 0 = 198.96$

9. What price would this model predict for a house 2 miles away that has 3 Bedrooms?

$\widehat{Price2014} = 283.56 - 42.30 \cdot 2 + 39.88 \cdot 1 + 106.13 \cdot 0 = 238.84$

10. What price would this model predict for a house 2 miles away that has 4+ Bedrooms?

$\widehat{Price2014} = 283.56 - 42.30 \cdot 2 + 39.88 \cdot 0 + 106.13 \cdot 1 = 305.09$



Distance	1-2 Beds	3 beds	Difference
1 mile	241.26	281.14	39.88
2 miles	198.96	238.84	39.88
Difference	-42.3	-42.3	

Distance	1-2 Beds	4+ beds	Difference
1 mile	241.26	347.39	106.13
2 miles	198.96	305.09	106.13
Difference	-42.3	-42.3	

Similar to the example above with a binary explanatory variable, the bed group categorical variable also affects the intercepts, which makes parallel lines for each of the three bedroom groups with the same slope based on distance to the rail trail.

Back to Distance Group

What if we thought that the relationship between Square Footage of a house and its price would *vary* by whether it's closer or further from the rail trail. A big house may not matter as much if it's really far from the rail trail, and a smaller house may be more valuable if it's closer to the rail trail than if it were further away.

You can also think about bigger houses in terms of an addition to a house – will adding another $500ft^2$ increase the value of your price the same amount regardless of where the house is located, or do you think that adding $500ft^2$ to a house close to the rail trail will matter more/less for the total house's price than if the house were further from the rail trail. Imagine you were a realtor and were asked whether a homeowner should build an addition to their house – how would you respond?

```
m4 <- lm(Price2014 ~ SquareFeet*DistGroup , data = RailsTrails)
summary(m4)
```

```
##
## Call:
## lm(formula = Price2014 ~ SquareFeet * DistGroup, data = RailsTrails)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -130.287  -32.792    0.084   23.018  282.596
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         32.16      36.11   0.891   0.3752
## SquareFeet                         177.82      19.75   9.003 1.51e-14 ***
## DistGroupFarther Away               32.46      42.57   0.763   0.4475
## SquareFeet:DistGroupFarther Away   -42.15      24.53  -1.718   0.0889 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.97 on 100 degrees of freedom
## Multiple R-squared:  0.6769, Adjusted R-squared:  0.6672
## F-statistic: 69.82 on 3 and 100 DF,  p-value: < 2.2e-16
```

$$\widehat{Price2014} = \hat{\beta}_0 + \hat{\beta}_1 \cdot SquareFeet + \hat{\beta}_2 \cdot FartherAway + \hat{\beta}_3 \cdot (SquareFeet \times FartherAway)$$

11. What price would this model predict for a 1000 square foot house that is *Closer* from the rail trail?

$\widehat{Price2014} = 32.16 + 177.82 \cdot 1 + 32.46 \cdot 0 - 42.15 \cdot 0 = 209.98$

12. What price would this model predict for a 1000 square foot house that is *Farther Away* from the rail trail?

$\widehat{Price2014} = 32.16 + 177.82 \cdot 1 + 32.46 \cdot 1 - 42.15 \cdot 1 = 200.29$

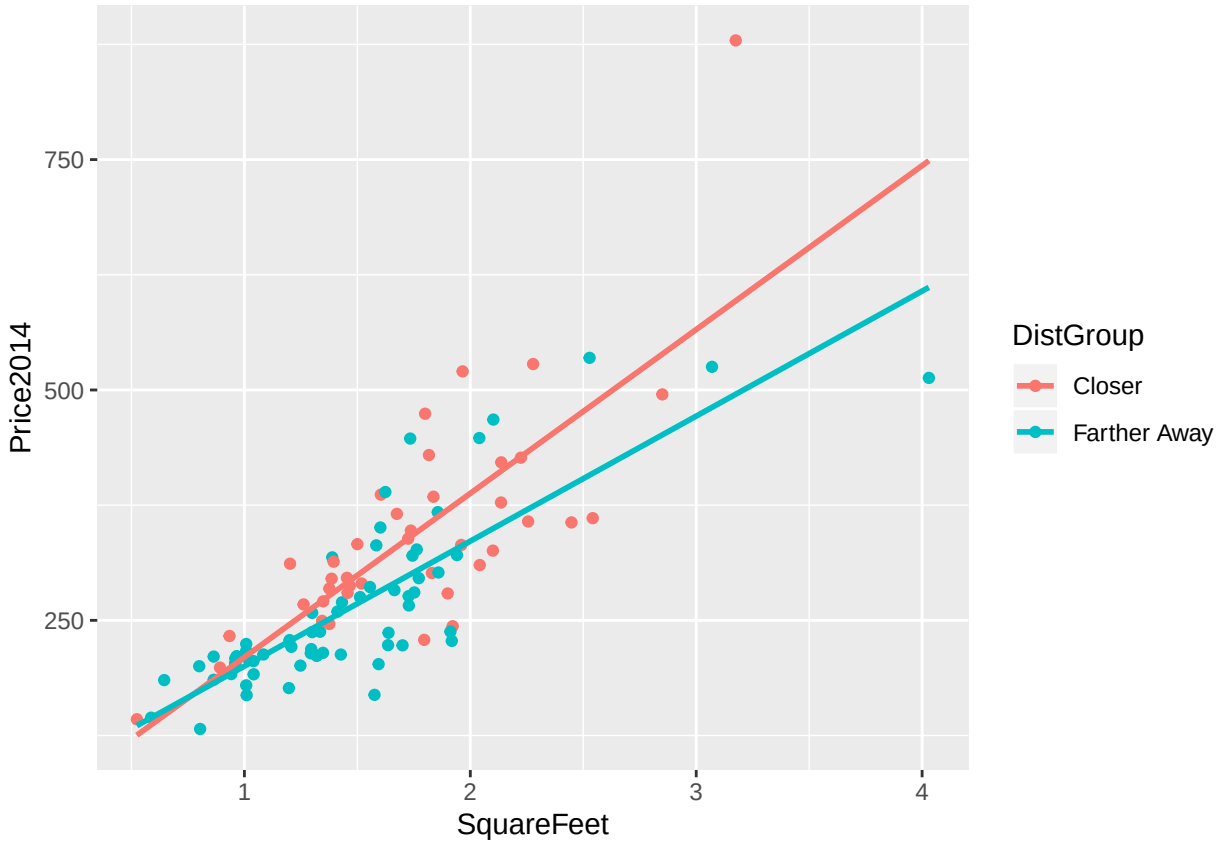
13. What price would this model predict for a 2000 square foot house that is *Closer* from the rail trail?

$\widehat{Price2014} = 32.16 + 177.82 \cdot 2 + 32.46 \cdot 0 - 42.15 \cdot 0 = 387.8$

14. What price would this model predict for a 2000 square foot house that is *Farther Away* from the rail trail?

$\widehat{Price2014} = 32.16 + 177.82 \cdot 1 + 32.46 \cdot 1 - 42.15 \cdot 2 = 335.96$

```
qplot(y=Price2014, x=SquareFeet, data=RailsTrails, color=DistGroup) +
  geom_smooth(method=lm,se=FALSE,fullrange = TRUE)
```



ft^2	Close	Far	Difference
1000 ft^2	209.98	200.29	-9.69
2000 ft^2	387.8	335.96	-51.84
Difference	177.82	135.67	

The model suggests that the relationship between SquareFeet and Price varies by whether the house is close or far from the rail trail. (This difference is not statistically significant, but let's not worry about that for now; let's just think about the coefficients and what they are telling us separate from their hypothesis tests.)

The $\hat{\beta}_2 Far$ variable affects the intercept. A house that has 0 square feet and is far from the rail trail will be worth \$32,458.25 more than a house of the same size that is closer.

For every additional 1000 ft^2 of house size, our model suggests the estimated 2014 price will increase, on average, though at different rates depending on whether the house is close or far from the rail trail. Houses far from the rail trail will increase more slowly in price as they have more square footage than houses closer to the rail trail do.

For every additional 1000 ft^2 in size, the price of a house that is *close* to the rail trail will, on average, increase by \$177,818.2.

For every additional 1000 ft^2 in size, the price of a house that is *far* to the rail trail will, on average, increase by $135,669.1.

The *marginal difference* in the house price for each 1000 square feet between houses that are far versus those that are closer is, on average, $-42,149.1