Stat 306:
Finding Relationships in Data.
Lecture 18
Sections 5 – Case Studies

Two datasets :

- Suggested Price of used General Motors (GM) cars in 2005

- Price of Homes for sale in Eugene Oregon in 2005

Price of used General Motors (GM) cars in 2005

All cars in this data set were less than one year old when priced and considered to be in excellent condition.

Outcome Variable:

Price:
- Suggested retail price of the used 2005 GM car in excellent condition, according to the "Kelly Blue Book".

Possible Explanatory Variables:

- Mileage: number of miles the car has been driven
- Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet
- Trim (of car): specific type of car model such as SE Sedan 4D, Quad Coupe 2D
- Type: body type such as sedan, coupe, etc.
- Cylinder: number of cylinders in the engine
- Liter: a more specific measure of engine size
- Doors: number of doors
- Cruise: indicator variable representing whether the car has cruise control (1 = cruise)
- Sound: indicator variable representing whether the car has upgraded speakers (1 = upgraded)
- Leather: indicator variable representing whether the car has leather seats (1 = leather)

Price of used General Motors (GM) cars in 2005

Simple linear regression:

$$\text{price}_i = \beta_0 + \beta_1 \text{Mileage}_i + \varepsilon_i \text{, for } i = 1, \dots, n$$

Price of used General Motors (GM) cars in 2005

Simple linear regression:

$$price_i = \beta_0 + \beta_1 Mileage_i + \varepsilon_i \text{ , for } i = 1, \ldots, n$$

```
> summary(lm(Price~ Mileage, data= cars))

Call:
lm(formula = Price ~ Mileage, data = cars)

Residuals:
   Min      1Q Median     3Q    Max
-13905  -7254  -3520   5188  46091

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.476e+04  9.044e+02  27.383  < 2e-16 ***
Mileage     -1.725e-01  4.215e-02  -4.093 4.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared:  0.02046, Adjusted R-squared:  0.01924
F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
```

Price of used General Motors (GM) cars in 2005

Simple linear regression:

$$\text{price}_i = \beta_0 + \beta_1 \text{Mileage}_i + \varepsilon_i \text{, for } i = 1, \dots, n$$

1) In general, what happens to price when there is one more mile on the car?

2) Does the fact that b_1 is small (-0.17) mean mileage is not very important?

 - "How does the price change if two cars are identical except one has 100,000 more miles?"

3) Does mileage help predict price?

Price of used General Motors (GM) cars in 2005

Simple linear regression:

$$\text{price}_i = \beta_0 + \beta_1 \text{Mileage}_i + \varepsilon_i \text{ , for } i = 1, ..., n$$

```
> summary(lm(Price~ Mileage, data= cars))

Call:
lm(formula = Price ~ Mileage, data = cars)

Residuals:
   Min      1Q Median     3Q    Max
-13905   -7254   -3520   5188  46091

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.476e+04  9.044e+02   27.383  < 2e-16 ***
Mileage      -1.725e-01  4.215e-02   -4.093 4.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared:  0.02046, Adjusted R-squared:  0.01924
F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
```

Price of used General Motors (GM) cars in 2005

Simple linear regression:

$$\mathbf{price_i} = \beta_0 + \beta_1 \mathbf{Mileage_i} + \varepsilon_i \quad , \text{for } i = 1, \ldots, n$$

```
> summary(lm(Price~ Mileage, data= cars))

Call:
lm(formula = Price ~ Mileage, data = cars)

Residuals:
   Min      1Q Median     3Q    Max
-13905  -7254  -3520   5188  46091

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.476e+04  9.044e+02  27.383  < 2e-16 ***
Mileage     -1.725e-01  4.215e-02  -4.093 4.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9789 on 802 degrees of freedom
Multiple R-squared:  0.02046,  Adjusted R-squared:  0.01924
F-statistic: 16.75 on 1 and 802 DF,  p-value: 4.685e-05
```
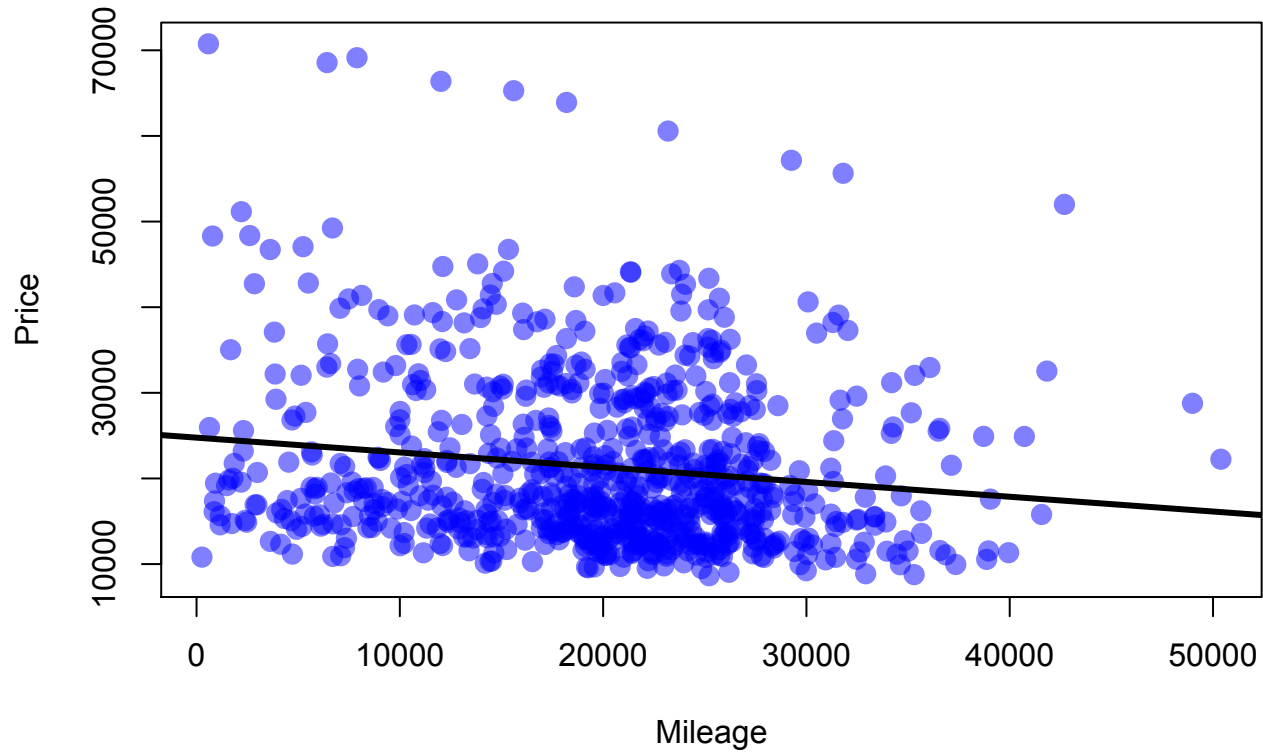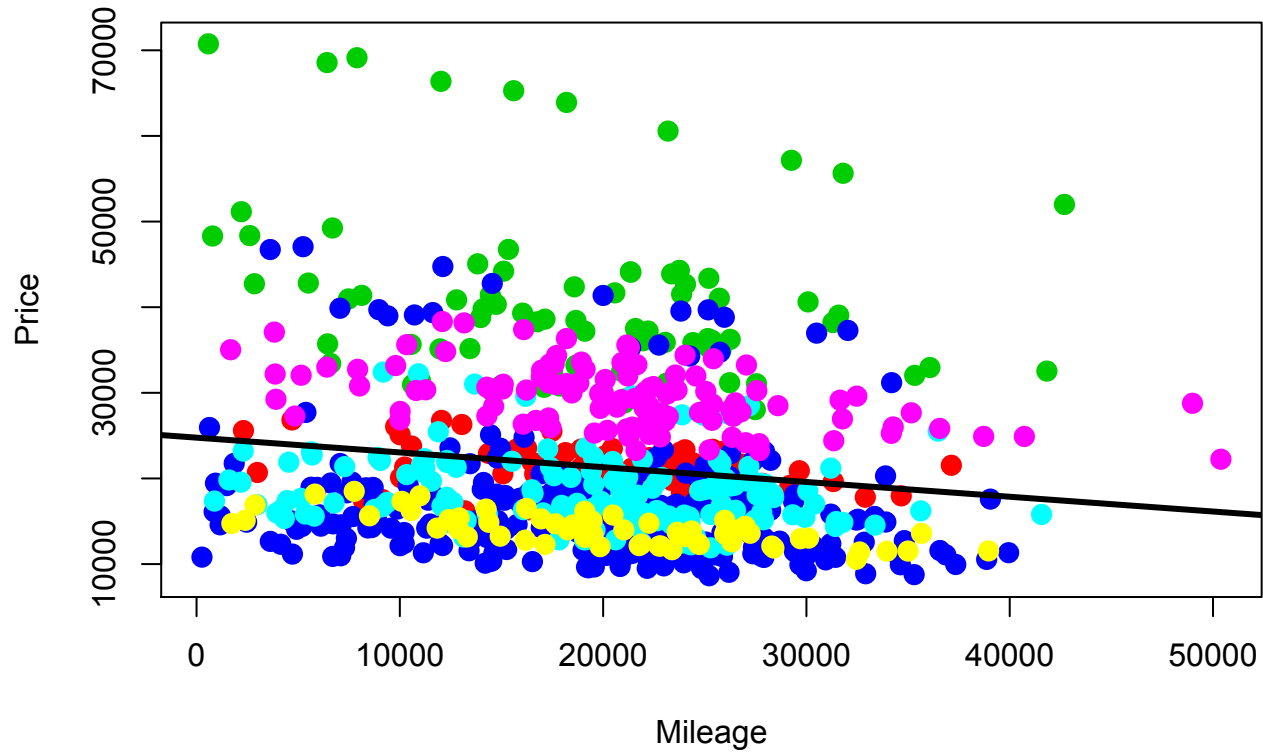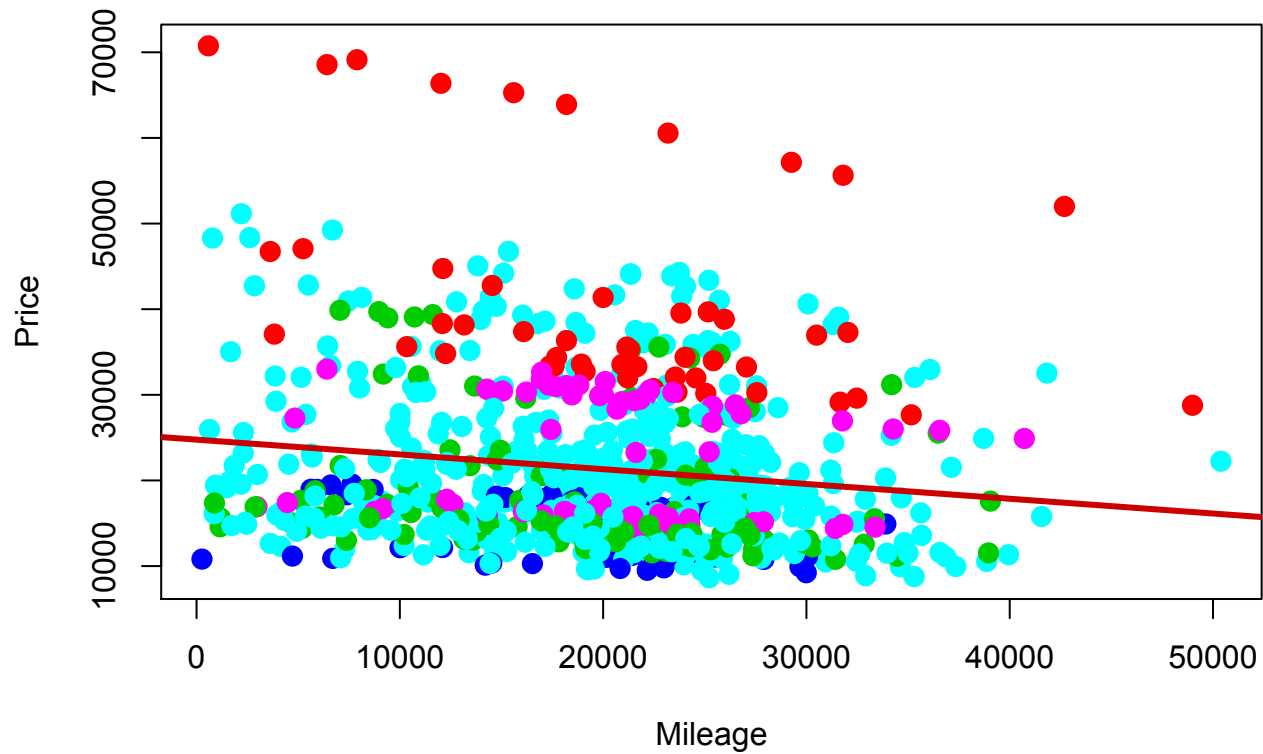
The R2 is only 2 %

Price of used General Motors (GM) cars in 2005

Price of used General Motors (GM) cars in 2005

Price of used General Motors (GM) cars in 2005

Price of used General Motors (GM) cars in 2005

```
> library(ggplot2)
> qplot(Mileage,Price, col= Make, data = cars)
```

Price of used General Motors (GM) cars in 2005

```
> # Create a model with Make and Mileage
> model1 <- lm( Price ~ Make + Mileage, data = cars)
> summary(model1)

Call:
lm(formula = Price ~ Make + Mileage, data = cars)

Residuals:
    Min      1Q   Median      3Q      Max
-11755.2  -3274.0   -701.8   1517.1  28174.1

Coefficients:
                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    2.431e+04   8.182e+02   29.705   < 2e-16  ***
MakeCadillac   1.986e+04   9.093e+02   21.844   < 2e-16  ***
MakeChevrolet -4.520e+03   7.185e+02   -6.290  5.22e-10  ***
MakePontiac   -2.592e+03   7.959e+02   -3.257   0.00117  **
MakeSAAB       8.771e+03   8.381e+02   10.465   < 2e-16  ***
MakeSaturn    -6.852e+03   9.813e+02   -6.983  6.10e-12  ***
Mileage       -1.709e-01   2.481e-02   -6.888  1.15e-11  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5746 on 797 degrees of freedom
Multiple R-squared:  0.6647,  Adjusted R-squared:  0.6621
F-statistic: 263.3 on 6 and 797 DF,  p-value: < 2.2e-16
```

Price of used General Motors (GM) cars in 2005

```
> # Create a model with Make and Mileage
> model1 <- lm( Price ~ Make + Mileage, data = cars)
> summary(model1)

Call:
lm(formula = Price ~ Make + Mileage, data = cars)

Residuals:
     Min       1Q   Median       3Q      Max
-11755.2  -3274.0   -701.8   1517.1  28174.1

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.431e+04  8.182e+02  29.705  < 2e-16 ***
MakeCadillac    1.986e+04  9.093e+02  21.844  < 2e-16 ***
MakeChevrolet  -4.520e+03  7.185e+02  -6.290 5.22e-10 ***
MakePontiac    -2.592e+03  7.959e+02  -3.257  0.00117 **
MakeSAAB        8.771e+03  8.381e+02  10.465  < 2e-16 ***
MakeSaturn     -6.852e+03  9.813e+02  -6.983 6.10e-12 ***
Mileage        -1.709e-01  2.481e-02  -6.888 1.15e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5746 on 797 degrees of freedom
Multiple R-squared:  0.6647,  Adjusted R-squared:  0.6621
F-statistic: 263.3 on 6 and 797 DF,  p-value: < 2.2e-16
```
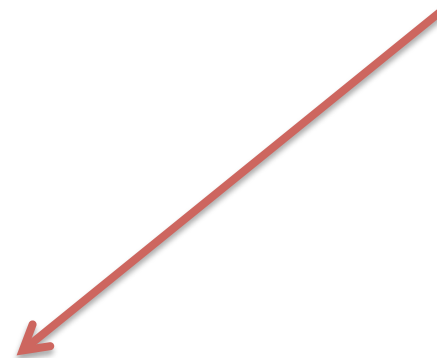
What is the reference category?

Price of used General Motors (GM) cars in 2005

```
> # Create a model with Make and Mileage
> model1 <- lm( Price ~ Make + Mileage, data = cars)
> summary(model1)

Call:
lm(formula = Price ~ Make + Mileage, data = cars)

Residuals:
     Min        1Q    Median        3Q       Max
-11755.2   -3274.0    -701.8    1517.1   28174.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.431e+04  8.182e+02  29.705  < 2e-16 ***
MakeCadillac   1.986e+04  9.093e+02  21.844  < 2e-16 ***
MakeChevrolet -4.520e+03  7.185e+02  -6.290 5.22e-10 ***
MakePontiac   -2.592e+03  7.959e+02  -3.257  0.00117 **
MakeSAAB       8.771e+03  8.381e+02  10.465  < 2e-16 ***
MakeSaturn    -6.852e+03  9.813e+02  -6.983 6.10e-12 ***
Mileage       -1.709e-01  2.481e-02  -6.888 1.15e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5746 on 797 degrees of freedom
Multiple R-squared:  0.6647,  Adjusted R-squared:  0.6621
F-statistic: 263.3 on 6 and 797 DF,  p-value: < 2.2e-16
```
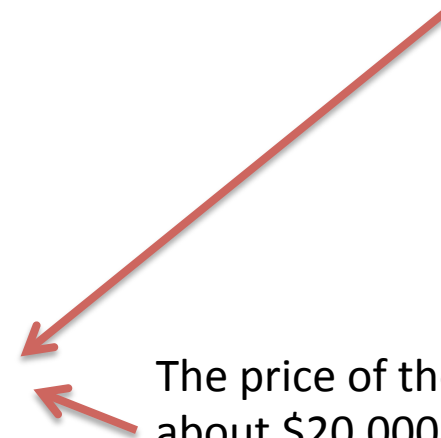
What is the reference category?

Buick

The price of the average Cadillac is about $20,000 more than the average Buick!

Price of used General Motors (GM) cars in 2005

```
> # Create a model with Make and Mileage
> model1 <- lm( Price ~ Make + Mileage, data = cars)
> summary(model1)

Call:
lm(formula = Price ~ Make + Mileage, data = cars)

Residuals:
     Min       1Q    Median       3Q       Max
-11755.2   -3274.0   -701.8    1517.1   28174.1

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.431e+04  8.182e+02  29.705  < 2e-16 ***
MakeCadillac    1.986e+04  9.093e+02  21.844  < 2e-16 ***
MakeChevrolet  -4.520e+03  7.185e+02  -6.290 5.22e-10 ***
MakePontiac    -2.592e+03  7.959e+02  -3.257  0.00117 **
MakeSAAB        8.771e+03  8.381e+02  10.465  < 2e-16 ***
MakeSaturn     -6.852e+03  9.813e+02  -6.983 6.10e-12 ***
Mileage        -1.709e-01  2.481e-02  -6.888 1.15e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5746 on 797 degrees of freedom
Multiple R-squared:  0.6647,   Adjusted R-squared:  0.6621
F-statistic: 263.3 on 6 and 797 DF,  p-value: < 2.2e-16
```
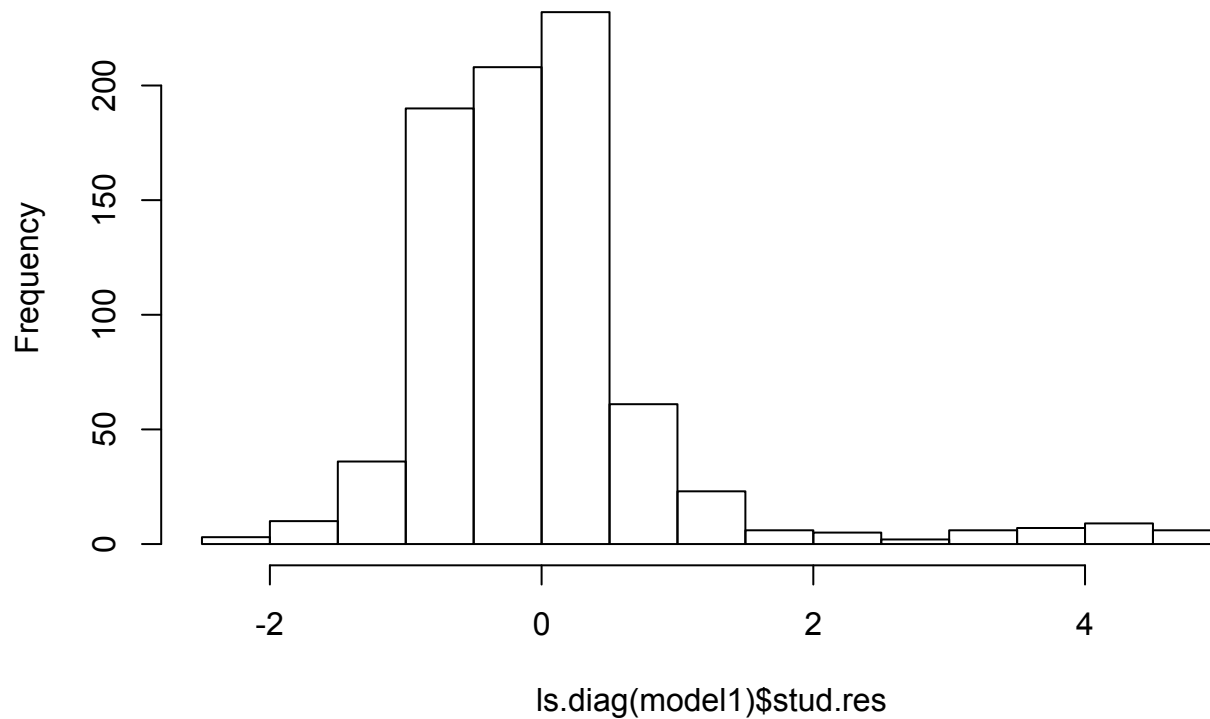
What is the reference category?

Buick

The price of the average Cadillac is about $20,000 more than the average Buick!

The R2 is now 66.5% !!

Price of used General Motors (GM) cars in 2005

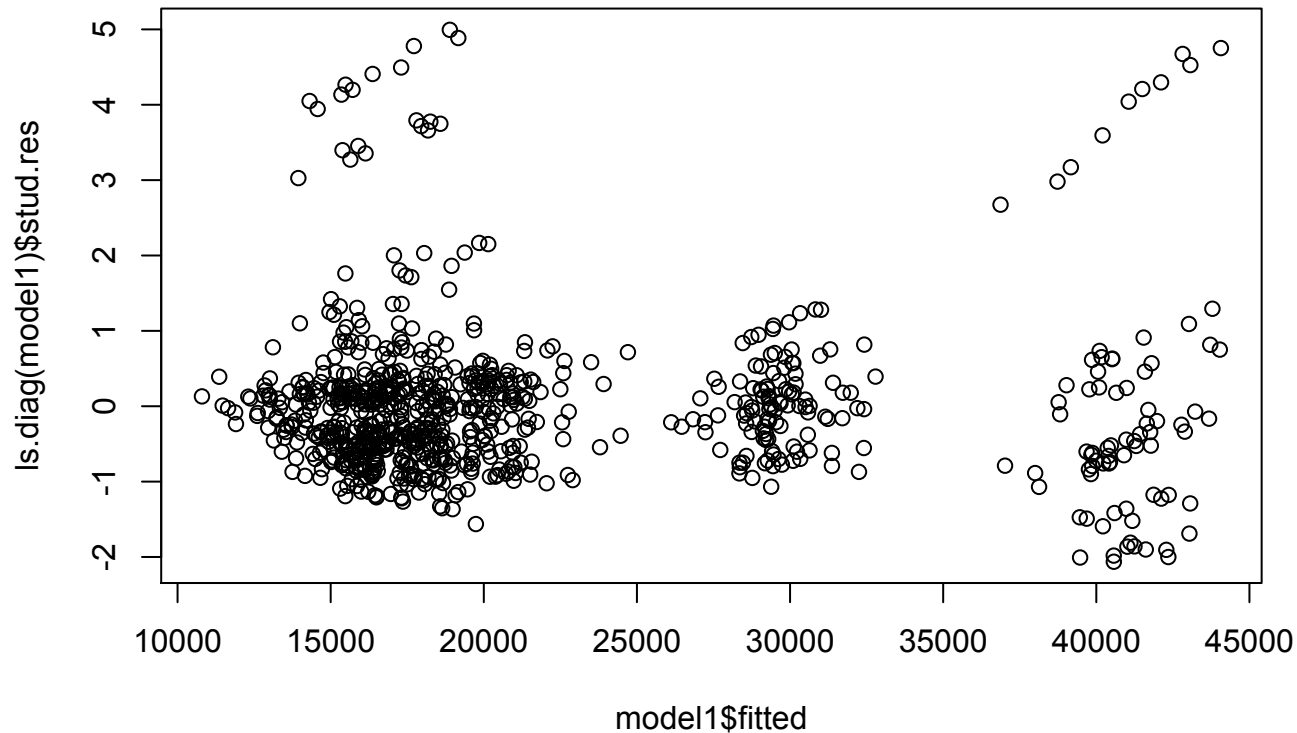```
> hist(ls.diag(model1)$stud.res)
> |
```



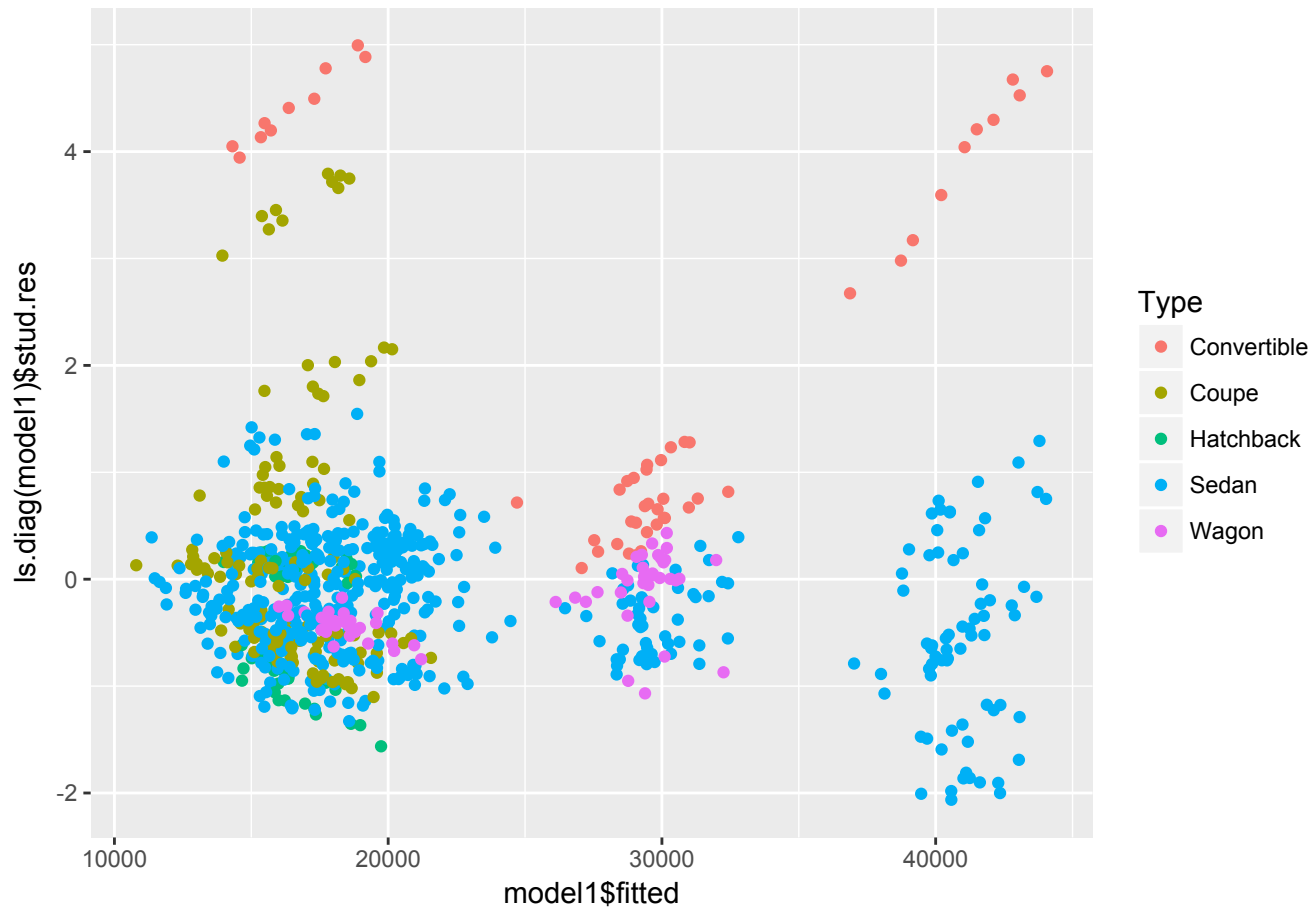**Histogram of ls.diag(model1)$stud.res**

# Price of used General Motors (GM) cars in 2005

```
> plot(model1$fitted,ls.diag(model1)$stud.res)
```

Price of used General Motors (GM) cars in 2005

```
> qplot(model1$fitted,ls.diag(model1)$stud.res, col=Type, data=cars)
>
```

Price of used General Motors (GM) cars in 2005

```
> summary(lm(Price~ Mileage + Make + Type, data= cars))

Call:
lm(formula = Price ~ Mileage + Make + Type, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-9889.1 -3190.3  -332.7  2720.5 20265.7

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.926e+04  1.006e+03  39.047  < 2e-16 ***
Mileage        -1.791e-01  2.010e-02  -8.909  < 2e-16 ***
MakeCadillac    1.800e+04  7.423e+02  24.251  < 2e-16 ***
MakeChevy      -5.217e+03  6.188e+02  -8.430  < 2e-16 ***
MakePontiac    -3.221e+03  6.682e+02  -4.820 1.72e-06 ***
MakeSaab        4.534e+03  7.475e+02   6.065 2.04e-09 ***
MakeSaturn     -7.494e+03  8.104e+02  -9.248  < 2e-16 ***
Typecoupe      -1.286e+04  8.576e+02 -15.002  < 2e-16 ***
Typehatchback  -1.646e+04  9.863e+02 -16.684  < 2e-16 ***
Typesedan      -1.479e+04  7.492e+02 -19.740  < 2e-16 ***
Typewagon      -1.362e+04  9.063e+02 -15.024  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4653 on 793 degrees of freedom
Multiple R-squared:  0.7812,   Adjusted R-squared:  0.7785
F-statistic: 283.2 on 10 and 793 DF,  p-value: < 2.2e-16
```
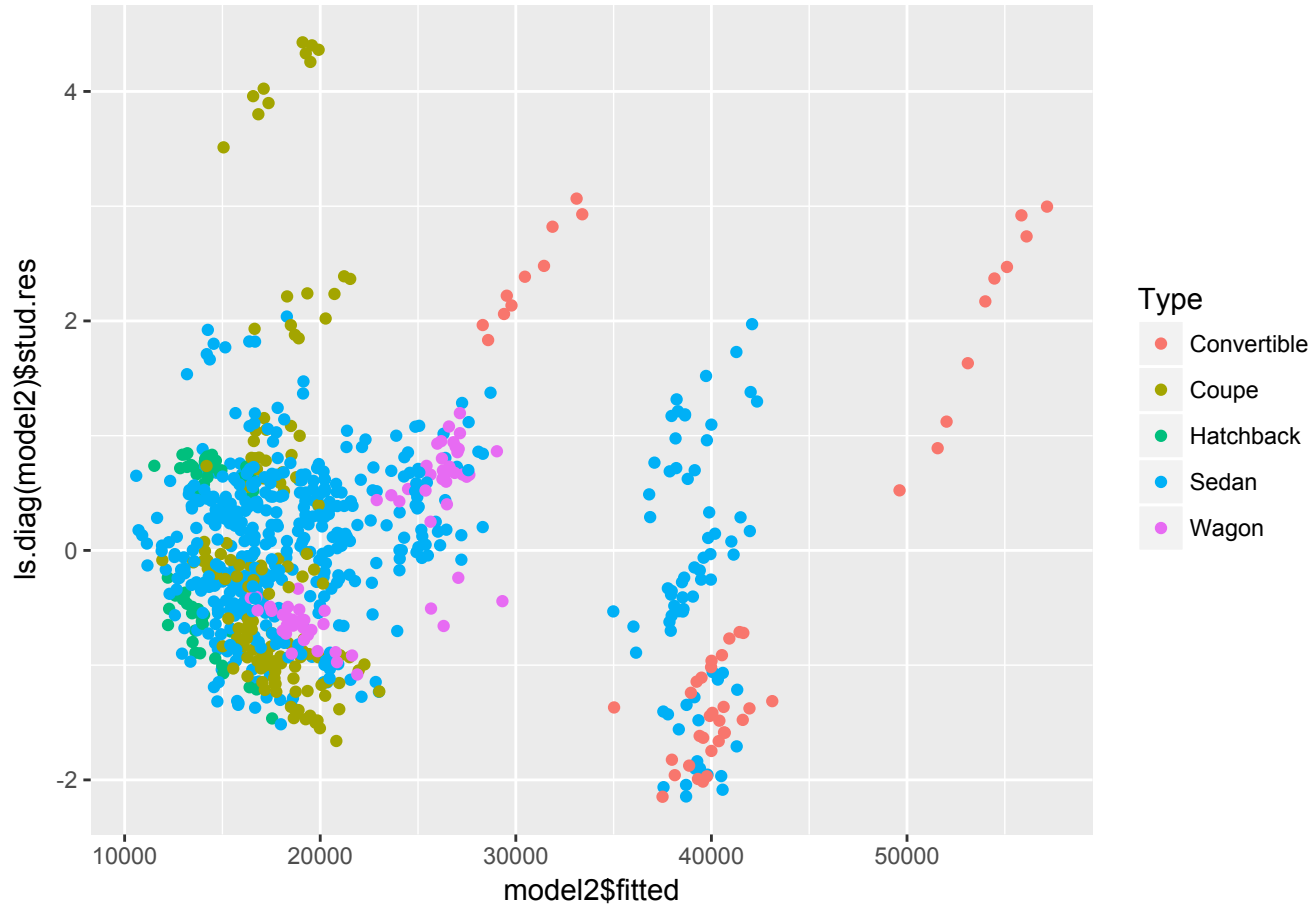
Price of used General Motors (GM) cars in 2005

Price of used General Motors (GM) cars in 2005

```
> model3 <- lm( Price ~ Mileage + Type + Make + Cylinder + Liter + Sound, data = cars)
> summary(model3)

Call:
lm(formula = Price ~ Mileage + Type + Make + Cylinder + Liter +
    Sound, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-8236.5 -1378.4    -6.2  1163.5 14890.5

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.295e+04  9.899e+02  23.181  < 2e-16 ***
Mileage        -1.846e-01  1.087e-02 -16.982  < 2e-16 ***
TypeCoupe      -1.199e+04  4.716e+02 -25.420  < 2e-16 ***
TypeHatchback  -1.235e+04  5.476e+02 -22.552  < 2e-16 ***
TypeSedan      -1.234e+04  4.099e+02 -30.093  < 2e-16 ***
TypeWagon      -8.266e+03  5.122e+02 -16.137  < 2e-16 ***
MakeCadillac    1.601e+04  4.636e+02  34.524  < 2e-16 ***
MakeChevrolet  -1.703e+03  3.498e+02  -4.868 1.36e-06 ***
MakePontiac    -1.866e+03  3.629e+02  -5.142 3.43e-07 ***
MakeSAAB        1.064e+04  4.457e+02  23.867  < 2e-16 ***
MakeSaturn     -1.261e+03  4.698e+02  -2.685   0.0074 **
Cylinder       -1.262e+03  3.127e+02  -4.037 5.95e-05 ***
Liter           5.825e+03  3.497e+02  16.658  < 2e-16 ***
Soundyes        3.238e+02  2.011e+02   1.610   0.1078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2516 on 790 degrees of freedom
Multiple R-squared:  0.9363,  Adjusted R-squared:  0.9352
F-statistic:   893 on 13 and 790 DF,  p-value: < 2.2e-16
```
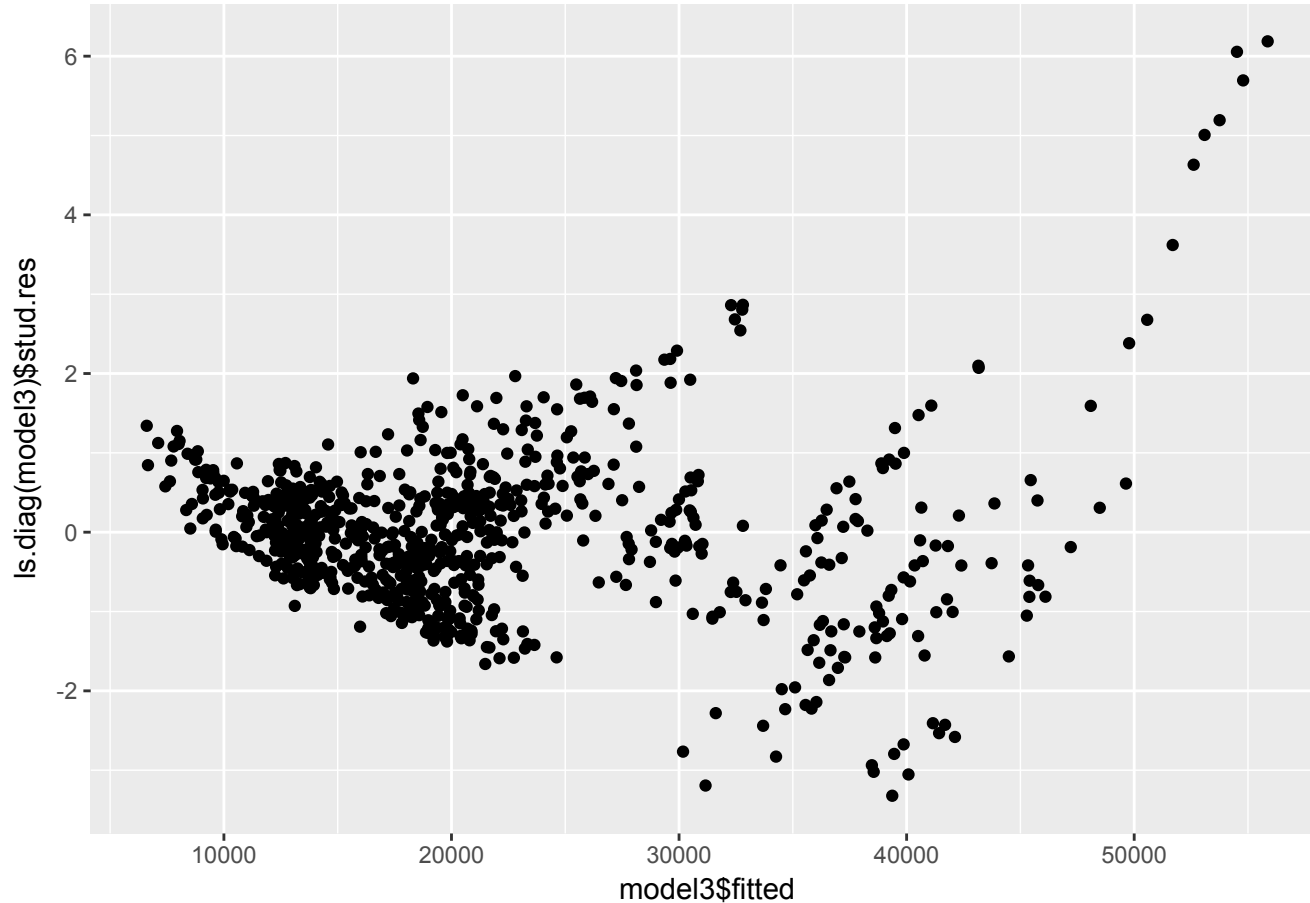
At this point, perhaps we should use a model selection algorithm to find other good variables…

Price of used General Motors (GM) cars in 2005

Studentized residuals don't look very homoscedastic!

Price of used General Motors (GM) cars in 2005

```
> # Create a model with log(Price) and Mileage + Type + Make + Cylinder + Liter + Sound
> model4 <- lm( log(Price) ~ Mileage + Type + Make + Cylinder + Liter + Sound, data = cars)
> summary(model4)

Call:
lm(formula = log(Price) ~ Mileage + Type + Make + Cylinder +
    Liter + Sound, data = cars)

Residuals:
     Min       1Q    Median       3Q       Max
-0.34362  -0.05734  0.00450  0.06054  0.28134

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     9.662e+00  3.564e-02 271.118  < 2e-16 ***
Mileage        -8.217e-06  3.914e-07 -20.996  < 2e-16 ***
TypeCoupe      -3.209e-01  1.698e-02 -18.902  < 2e-16 ***
TypeHatchback  -3.601e-01  1.972e-02 -18.266  < 2e-16 ***
TypeSedan      -3.270e-01  1.476e-02 -22.157  < 2e-16 ***
TypeWagon      -1.688e-01  1.844e-02  -9.152  < 2e-16 ***
MakeCadillac    4.675e-01  1.669e-02  28.006  < 2e-16 ***
MakeChevrolet  -1.306e-01  1.259e-02 -10.369  < 2e-16 ***
MakePontiac    -9.566e-02  1.306e-02  -7.322 6.02e-13 ***
MakeSAAB        5.508e-01  1.605e-02  34.325  < 2e-16 ***
MakeSaturn     -1.067e-01  1.691e-02  -6.307 4.72e-10 ***
Cylinder       -2.891e-02  1.126e-02  -2.568   0.0104 *
Liter           2.545e-01  1.259e-02  20.215  < 2e-16 ***
Soundyes        1.177e-02  7.240e-03   1.626   0.1044
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09056 on 790 degrees of freedom
Multiple R-squared:  0.952,   Adjusted R-squared:  0.9512
F-statistic:  1206 on 13 and 790 DF,  p-value: < 2.2e-16
```
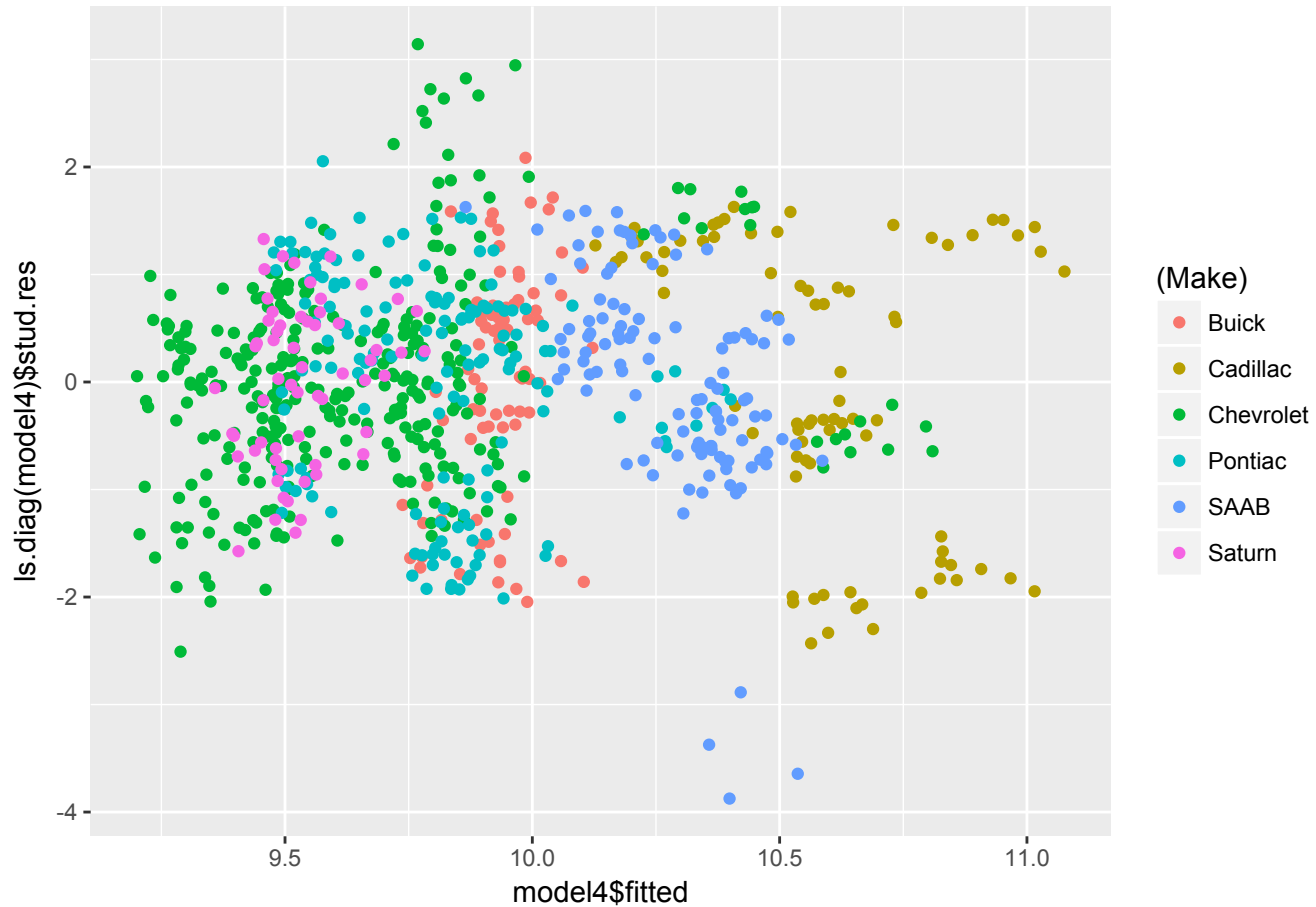
Price of used General Motors (GM) cars in 2005

Residuals vs. fitted plot looks much better now:

Residuals vs. other variables:
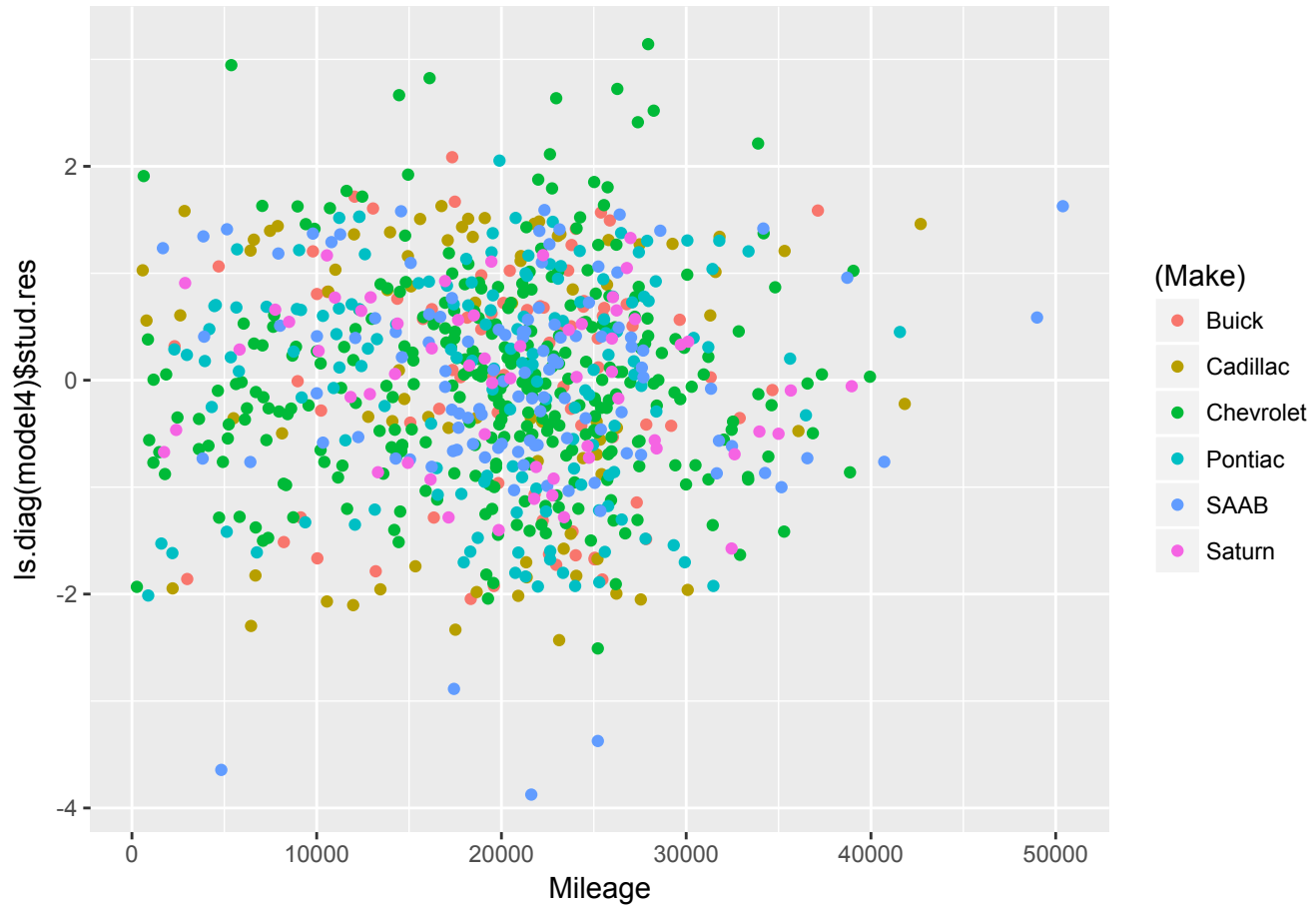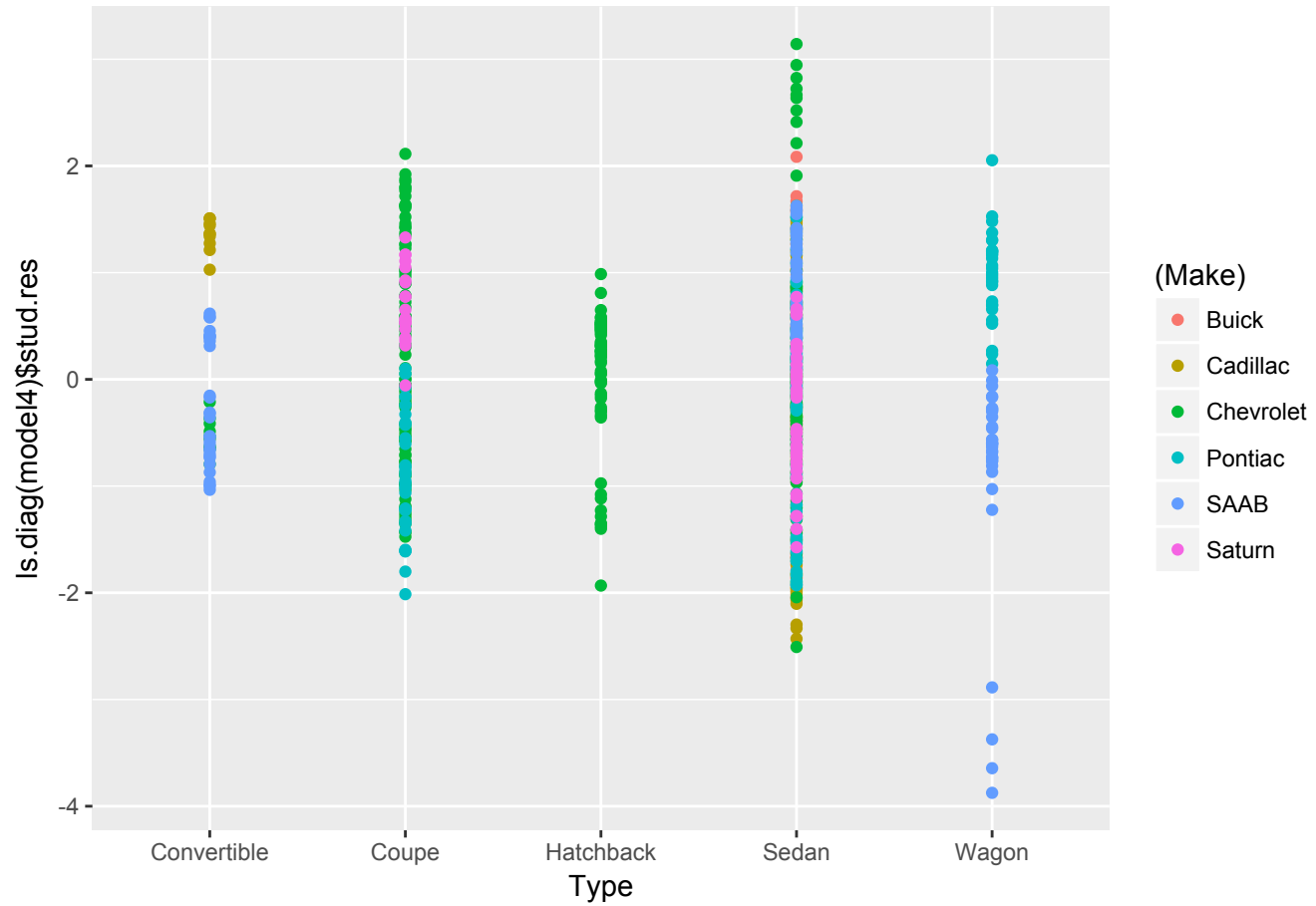
Residuals vs. other variables:

Residuals vs. other variables:
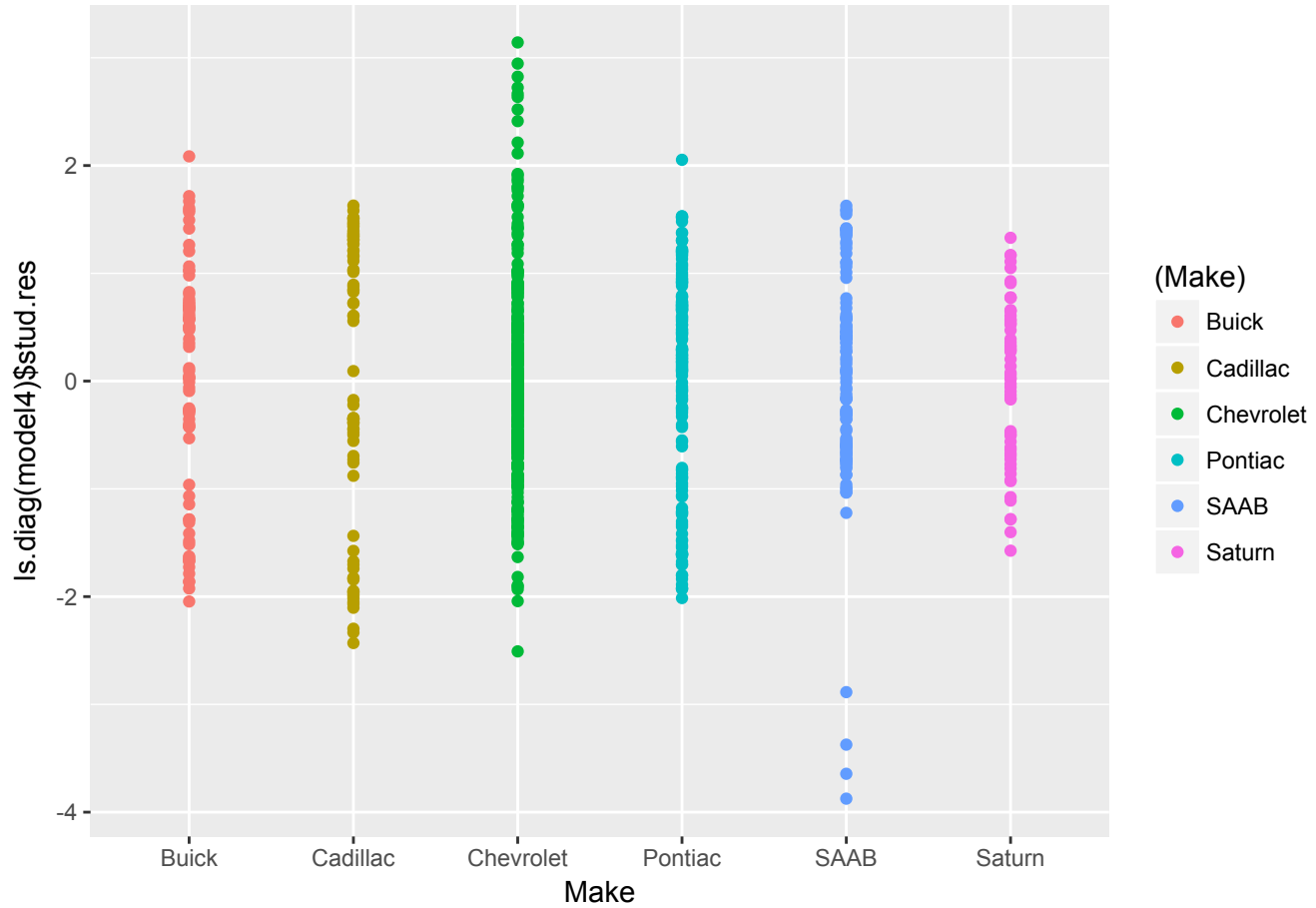
Residuals vs. other variables:
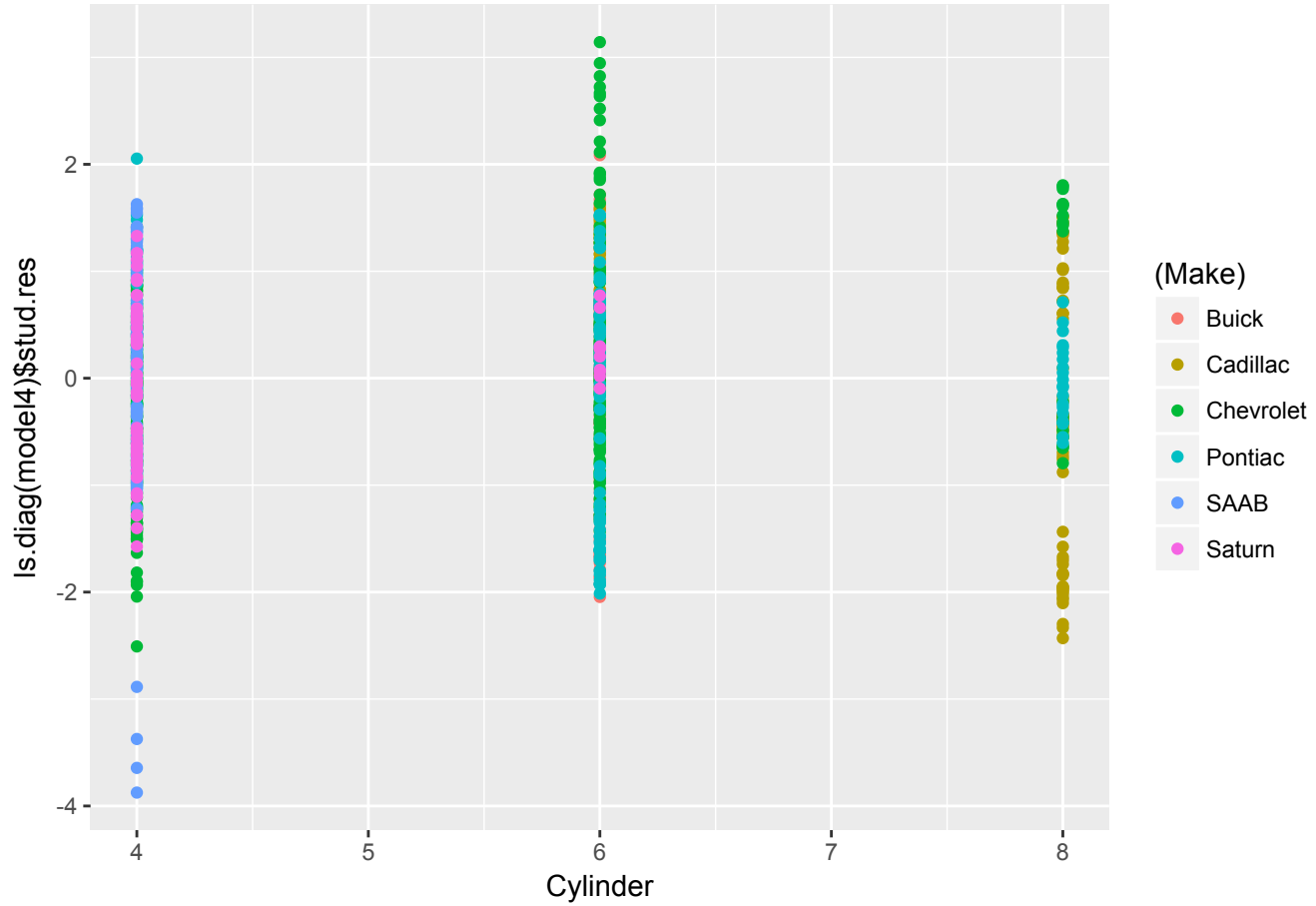
Residuals vs. other variables:
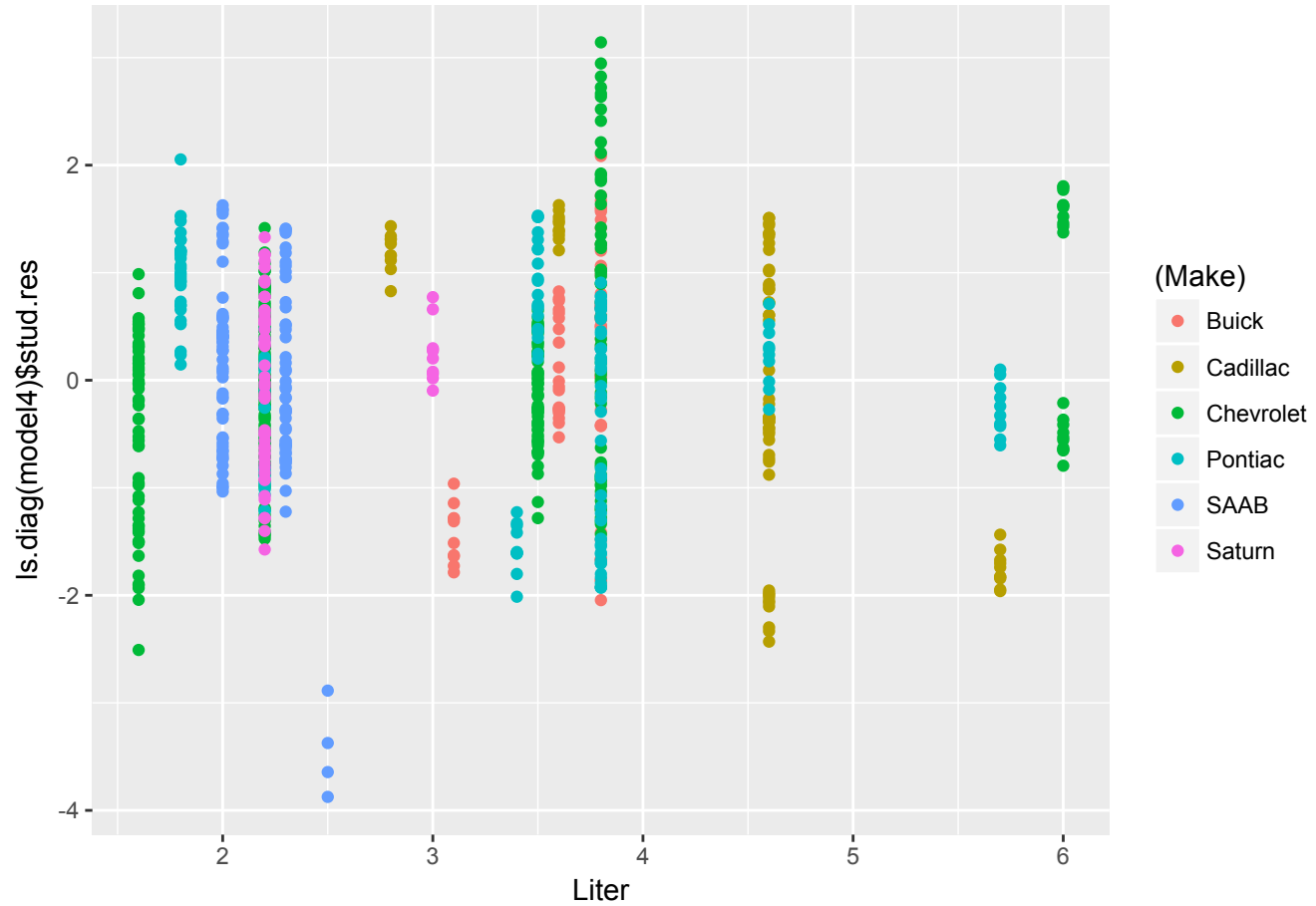
Residuals vs. other variables:
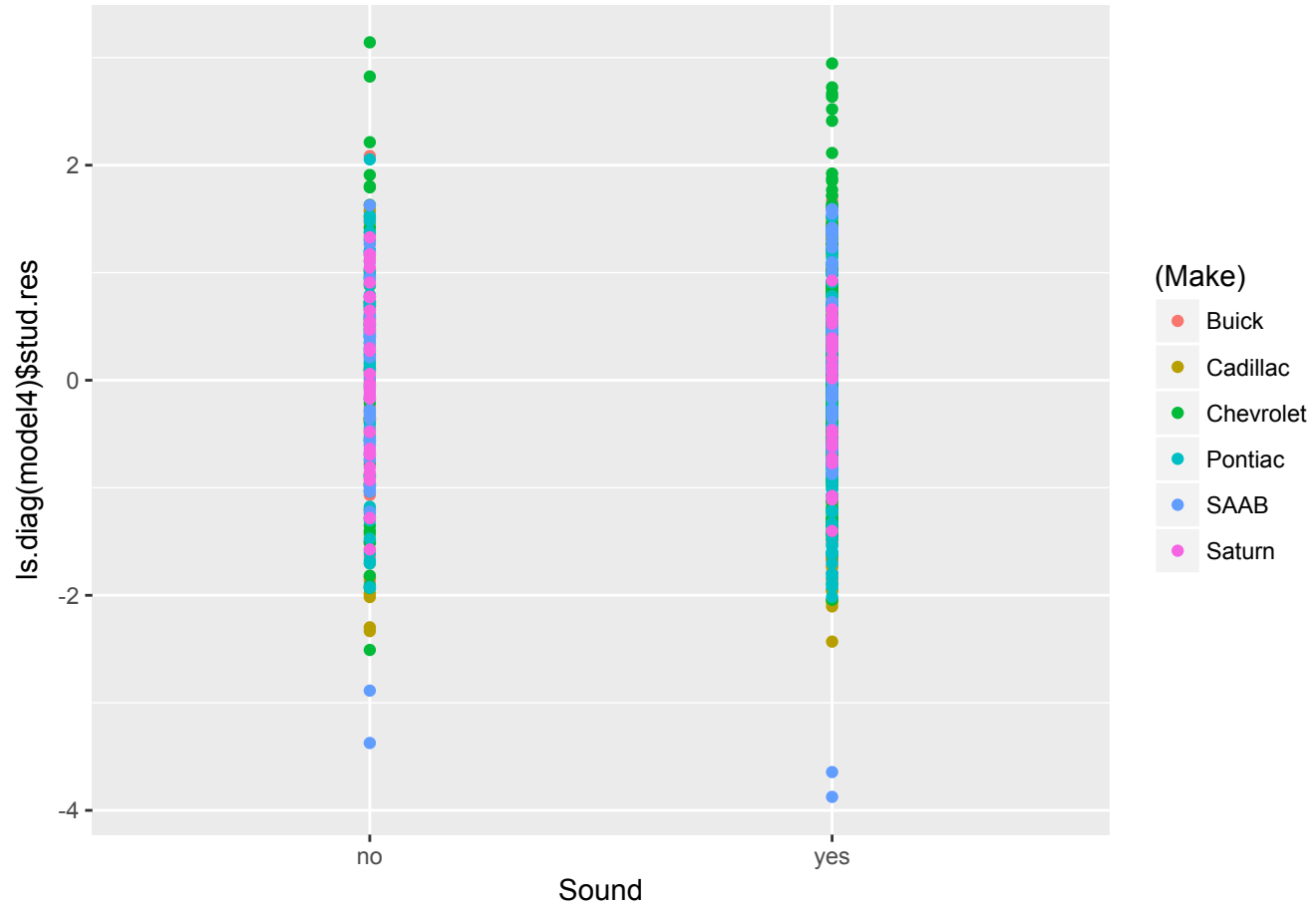
Price of used General Motors (GM) cars in 2005

Issues with highly leverage-points or influential points?
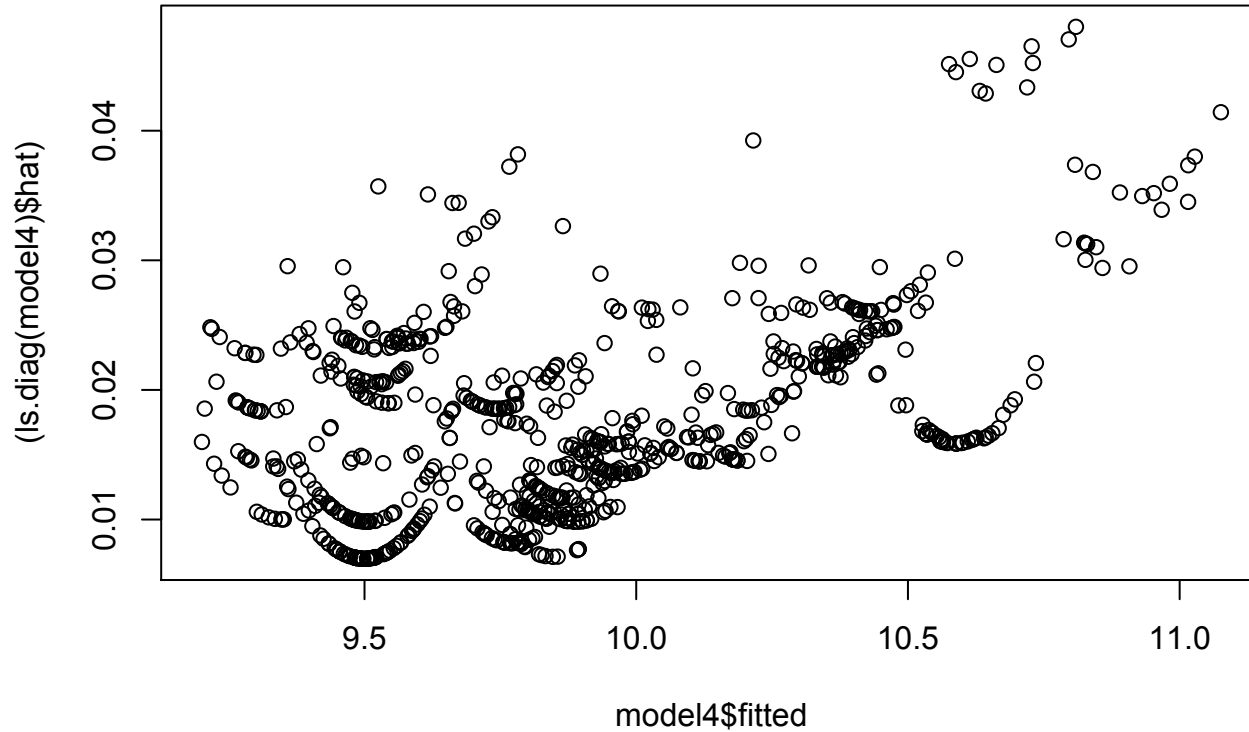
Price of used General Motors (GM) cars in 2005

Issues with highly leverage-points or influential points?
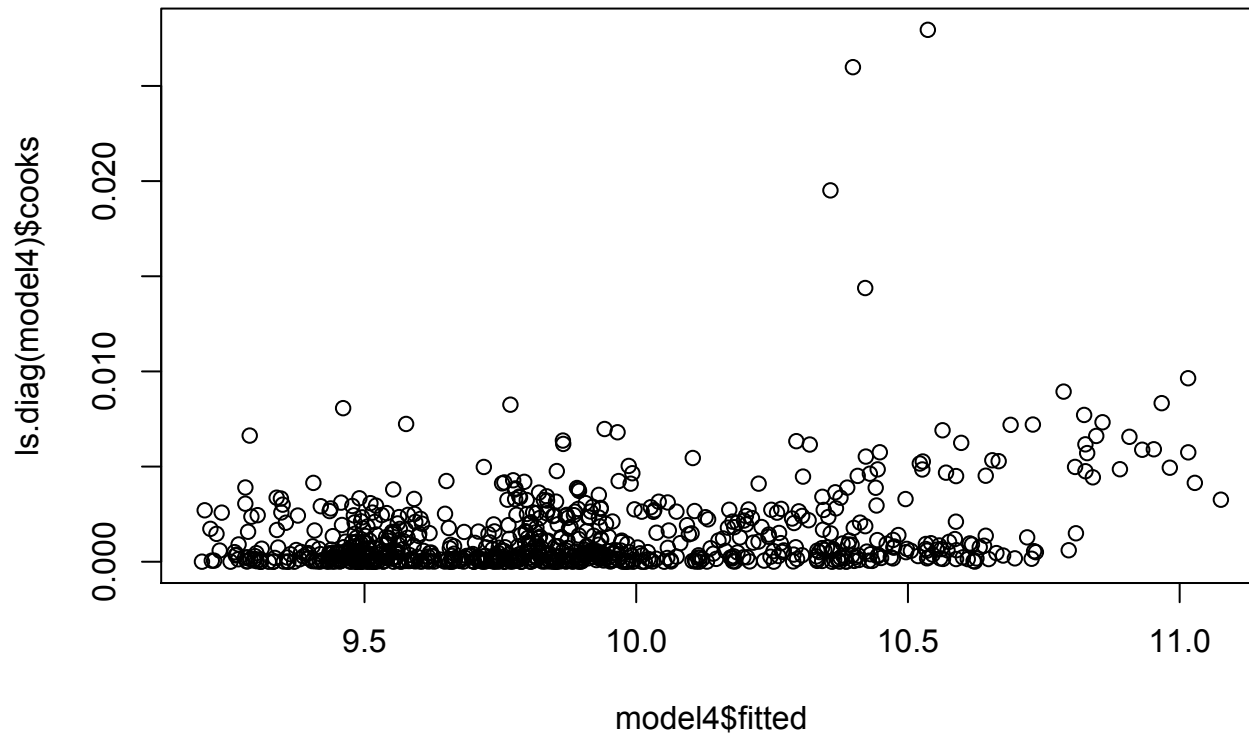

```
> cars[(ls.diag(model4)$cooks)>0.01,]
      Price Mileage Make      Model                   Trim    Type  Cylinder Liter Doors Cruise Sound Leather
741 27280.98    4836 SAAB 9-2X AWD Linear Wagon 4D Wagon         4   2.5     4    yes   yes      no
742 25959.12   17431 SAAB 9-2X AWD Linear Wagon 4D Wagon         4   2.5     4    yes    no     yes
743 23274.48   21616 SAAB 9-2X AWD Linear Wagon 4D Wagon         4   2.5     4    yes   yes      no
744 23329.21   25218 SAAB 9-2X AWD Linear Wagon 4D Wagon         4   2.5     4    yes    no     yes
>
```

Price of used General Motors (GM) cars in 2005

Interpretation

```
> round(summary(model4)$coefficients,3)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.662      0.036  271.12    0.000
I(Mileage/1000)    -0.008      0.000  -21.00    0.000
TypeCoupe          -0.321      0.017  -18.90    0.000
TypeHatchback      -0.360      0.020  -18.27    0.000
TypeSedan          -0.327      0.015  -22.16    0.000
TypeWagon          -0.169      0.018   -9.15    0.000
MakeCadillac        0.467      0.017   28.01    0.000
MakeChevrolet      -0.131      0.013  -10.37    0.000
MakePontiac        -0.096      0.013   -7.32    0.000
MakeSAAB            0.551      0.016   34.33    0.000
MakeSaturn         -0.107      0.017   -6.31    0.000
Cylinder           -0.029      0.011   -2.57    0.010
Liter               0.255      0.013   20.21    0.000
Soundyes            0.012      0.007    1.63    0.104
```

This is a log-linear model, therefore:

The effect of a p-unit increase in X_j is to multiply the mean of y by $\exp(p\beta_j)$.

Price of used General Motors (GM) cars in 2005

Interpretation

```
> round(summary(model4)$coefficients,3)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.662      0.036  271.12    0.000
I(Mileage/1000)    -0.008      0.000  -21.00    0.000
TypeCoupe          -0.321      0.017  -18.90    0.000
TypeHatchback      -0.360      0.020  -18.27    0.000
TypeSedan          -0.327      0.015  -22.16    0.000
TypeWagon          -0.169      0.018   -9.15    0.000
MakeCadillac        0.467      0.017   28.01    0.000
MakeChevrolet      -0.131      0.013  -10.37    0.000
MakePontiac        -0.096      0.013   -7.32    0.000
MakeSAAB            0.551      0.016   34.33    0.000
MakeSaturn         -0.107      0.017   -6.31    0.000
Cylinder           -0.029      0.011   -2.57    0.010
Liter               0.255      0.013   20.21    0.000
Soundyes            0.012      0.007    1.63    0.104
```

```
                                   0.899            0.
> cbind(exp(coef(model4)))
                                 [,1]
(Intercept)                 15708.516
I(Mileage/1000)                 0.992
TypeCoupe                       0.725
TypeHatchback                   0.698
TypeSedan                       0.721
TypeWagon                       0.845
MakeCadillac                    1.596
MakeChevrolet                   0.878
MakePontiac                     0.909
MakeSAAB                        1.735
MakeSaturn                      0.899
Cylinder                        0.972
Liter                           1.290
Soundyes                        1.012
>
```

This is a log-linear model, therefore:

For an added 10,000 miles the price of the car drops by about 8%.

```
> exp(10*(-0.00822))
[1] 0.921
```

Effect of multicollinearity?

Price of used General Motors (GM) cars in 2005

Interpretation

```
> round(summary(model4)$coefficients,3)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.662      0.036  271.12    0.000
I(Mileage/1000)   -0.008      0.000  -21.00    0.000
TypeCoupe         -0.321      0.017  -18.90    0.000
TypeHatchback     -0.360      0.020  -18.27    0.000
TypeSedan         -0.327      0.015  -22.16    0.000
TypeWagon         -0.169      0.018   -9.15    0.000
MakeCadillac       0.467      0.017   28.01    0.000
MakeChevrolet     -0.131      0.013  -10.37    0.000
MakePontiac       -0.096      0.013   -7.32    0.000
MakeSAAB           0.551      0.016   34.33    0.000
MakeSaturn        -0.107      0.017   -6.31    0.000
Cylinder          -0.029      0.011   -2.57    0.010
Liter              0.255      0.013   20.21    0.000
Soundyes           0.012      0.007    1.63    0.104
```

```
                           0.899              0.
> cbind(exp(coef(model4)))
                             [,1]
(Intercept)            15708.516
I(Mileage/1000)            0.992
TypeCoupe                  0.725
TypeHatchback              0.698
TypeSedan                  0.721
TypeWagon                  0.845
MakeCadillac               1.596
MakeChevrolet              0.878
MakePontiac                0.909
MakeSAAB                   1.735
MakeSaturn                 0.899
Cylinder                   0.972
Liter                      1.290
Soundyes                   1.012
> |
```

```
> library(DAAG)
> vif(model4)
     I(Mileage/1000)         TypeCoupe    TypeHatchback         TypeSedan        TypeWagon
                1.01              4.06             2.63              5.08             2.44
        MakeCadillac     MakeChevrolet      MakePontiac          MakeSAAB       MakeSaturn
                2.45              3.73             2.54              3.07             1.94
            Cylinder             Liter         Soundyes
               23.89             18.97             1.12
```

Price of used General Motors (GM) cars in 2005

Interpretation

```
> round(summary(model4)$coefficients,3)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.662      0.036  271.12    0.000
I(Mileage/1000)   -0.008      0.000  -21.00    0.000
TypeCoupe         -0.321      0.017  -18.90    0.000
TypeHatchback     -0.360      0.020  -18.27    0.000
TypeSedan         -0.327      0.015  -22.16    0.000
TypeWagon         -0.169      0.018   -9.15    0.000
MakeCadillac       0.467      0.017   28.01    0.000
MakeChevrolet     -0.131      0.013  -10.37    0.000
MakePontiac       -0.096      0.013   -7.32    0.000
MakeSAAB           0.551      0.016   34.33    0.000
MakeSaturn        -0.107      0.017   -6.31    0.000
Cylinder          -0.029      0.011   -2.57    0.010
Liter              0.255      0.013   20.21    0.000
Soundyes           0.012      0.007    1.63    0.104
```

```
                                  0.899           0.
> cbind(exp(coef(model4)))
                          [,1]
(Intercept)         15708.516
I(Mileage/1000)         0.992
TypeCoupe               0.725
TypeHatchback           0.698
TypeSedan               0.721
TypeWagon               0.845
MakeCadillac            1.596
MakeChevrolet           0.878
MakePontiac             0.909
MakeSAAB                1.735
MakeSaturn              0.899
Cylinder                0.972
Liter                   1.290
Soundyes                1.012
>
```

```
> library(DAAG)
> vif(model4)
I(Mileage/1000)        TypeCoupe     TypeHatchback        TypeSedan        TypeWagon
           1.01             4.06              2.63             5.08             2.44
   MakeCadillac    MakeChevrolet       MakePontiac          MakeSAAB        MakeSaturn
           2.45             3.73              2.54             3.07             1.94
       Cylinder            Liter          Soundyes
          23.89            18.97              1.12
>
```

Price of Homes for sale in Eugene, Oregon in 2005

Price of Homes for sale in Eugene, Oregon in 2005

Outcome Variable:

> **Price:**
> - Listed sale price.

Possible Explanatory Variables:

- *Size* = floor size (thousands of square feet)
- *Lot* = lot size category (from 1 to 11—explained below)
- *Bath* = number of bathrooms (with half-bathrooms counting as 0.1—explained below)
- *Bed* = number of bedrooms (between 2 and 6)
- *Age* = age (standardized: (year built - 1970)=10—explained below)
- *Garage* = garage size (0, 1, 2, or 3 cars)
- *Active* = indicator for "active listing" (reference: pending or sold)
- *Edison* = indicator for Edison Elementary (reference: Edgewood Elementary)
- *Harris* = indicator for Harris Elementary (reference: Edgewood Elementary)
- *Adams* = indicator for Adams Elementary (reference: Edgewood Elementary)
- *Crest* = indicator for Crest Elementary (reference: Edgewood Elementary)
- *Parker* = indicator for Parker Elementary (reference: Edgewood Elementary)

Price of Homes for sale in Eugene, Oregon in 2005

```
> model1 <- lm(Price~.,data=HDATA)
> summary(model1)

Call:
lm(formula = Price ~ ., data = HDATA)

Residuals:
   Min     1Q Median     3Q    Max
-87.26 -27.22  -0.22  30.13  95.58
```

What is the reference category?

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           95.61      58.39    1.64   0.1065
Size                  69.78      30.93    2.26   0.0276 *
Lot                   10.75       3.67    2.93   0.0047 **
Bath                   4.62      11.75    0.39   0.6958
Bed                  -12.46       9.12   -1.37   0.1768
Age                    1.63       3.32    0.49   0.6264
Garage                10.05       9.26    1.09   0.2818
Active                30.87      12.90    2.39   0.0197 *
`Edison Elementary`   79.73      17.63    4.52 0.000028 ***
`Harris Elementary`   46.32      16.29    2.84   0.0060 **
`Adams Elementary`    -7.47      29.05   -0.26   0.7980
`Crest Elementary`    -2.66      22.93   -0.12   0.9081
`Parker Elementary`  -19.37      15.65   -1.24   0.2203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.1 on 63 degrees of freedom
Multiple R-squared:  0.53, Adjusted R-squared:  0.441
F-statistic: 5.93 on 12 and 63 DF,  p-value: 0.000000891
```

Price of Homes for sale in Eugene, Oregon in 2005

```
> model1 <- lm(Price~.,data=HDATA)
> summary(model1)

Call:
lm(formula = Price ~ ., data = HDATA)

Residuals:
   Min     1Q Median     3Q    Max
-87.26 -27.22  -0.22  30.13  95.58

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           95.61      58.39    1.64   0.1065
Size                  69.78      30.93    2.26   0.0276 *
Lot                   10.75       3.67    2.93   0.0047 **
Bath                   4.62      11.75    0.39   0.6958
Bed                  -12.46       9.12   -1.37   0.1768
Age                    1.63       3.32    0.49   0.6264
Garage                10.05       9.26    1.09   0.2818
Active                30.87      12.90    2.39   0.0197 *
`Edison Elementary`   79.73      17.63    4.52 0.000028 ***
`Harris Elementary`   46.32      16.29    2.84   0.0060 **
`Adams Elementary`    -7.47      29.05   -0.26   0.7980
`Crest Elementary`    -2.66      22.93   -0.12   0.9081
`Parker Elementary`  -19.37      15.65   -1.24   0.2203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.1 on 63 degrees of freedom
Multiple R-squared:  0.53, Adjusted R-squared:  0.441
F-statistic: 5.93 on 12 and 63 DF,  p-value: 0.000000891
```
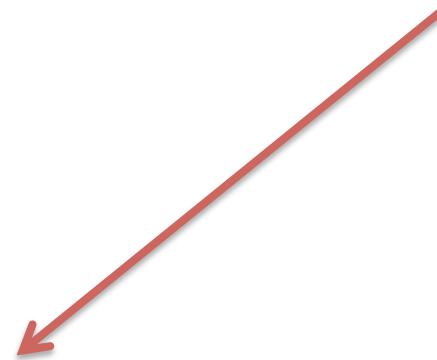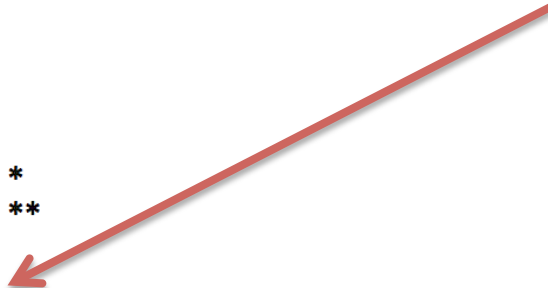
Why is the estimated coefficient for bed negative?

Price of Homes for sale in Eugene, Oregon in 2005
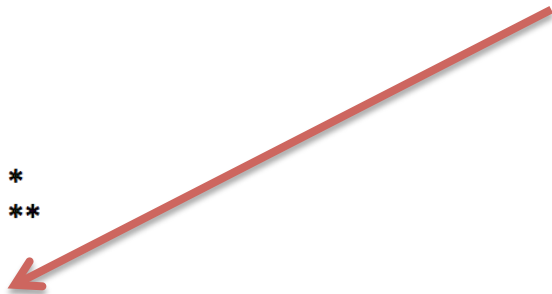
```
> model1 <- lm(Price~.,data=HDATA)
> summary(model1)

Call:
lm(formula = Price ~ ., data = HDATA)

Residuals:
   Min     1Q Median     3Q    Max
-87.26 -27.22  -0.22  30.13  95.58

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)             95.61      58.39    1.64   0.1065
Size                    69.78      30.93    2.26   0.0276 *
Lot                     10.75       3.67    2.93   0.0047 **
Bath                     4.62      11.75    0.39   0.6958
Bed                    -12.46       9.12   -1.37   0.1768
Age                      1.63       3.32    0.49   0.6264
Garage                  10.05       9.26    1.09   0.2818
Active                  30.87      12.90    2.39   0.0197 *
`Edison Elementary`     79.73      17.63    4.52 0.000028 ***
`Harris Elementary`     46.32      16.29    2.84   0.0060 **
`Adams Elementary`      -7.47      29.05   -0.26   0.7980
`Crest Elementary`
`Parker Elementary`     -1
---
Signif. codes:  0 '***'

Residual standard error:
Multiple R-squared:  0.5
F-statistic: 5.93 on 12
```

Why is the estimated coefficient for bed negative?

```
> library(DAAG)
> vif(model1)
                Size                      Lot                    Bath                      Bed
                1.59                     1.35                    1.65                     1.67
                 Age                   Garage                  Active      `Edison Elementary`
                2.25                     1.88                    1.37                     1.54
 `Harris Elementary`       `Adams Elementary`      `Crest Elementary`     `Parker Elementary`
                1.49                     1.20                    1.43                     1.45
```

Price of Homes for sale in Eugene, Oregon in 2005

```
> model1 <- lm(Price~.,data=HDATA)
> summary(model1)

Call:
lm(formula = Price ~ ., data = HDATA)

Residuals:
   Min     1Q Median     3Q    Max
-87.26 -27.22  -0.22  30.13  95.58

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           95.61      58.39    1.64   0.1065
Size                  69.78      30.93    2.26   0.0276 *
Lot                   10.75       3.67    2.93   0.0047 **
Bath                   4.62      11.75    0.39   0.6958
Bed                  -12.46       9.12   -1.37   0.1768
Age                    1.63       3.32    0.49   0.6264
Garage                10.05       9.26    1.09   0.2818
Active                30.87      12.90    2.39   0.0197 *
`Edison Elementary`   79.73      17.63    4.52 0.000028 ***
`Harris Elementary`   46.32      16.29    2.84   0.0060 **
`Adams Elementary`    -7.47      29.05   -0.26   0.7980
`Crest Elementary`    -2.66      22.93   -0.12   0.9081
`Parker Elementary`  -19.37      15.65   -1.24   0.2203
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 45.1 on 63 degrees of freedom
Multiple R-squared:  0.53, Adjusted R-squared:  0.441
F-statistic: 5.93 on 12 and 63 DF,  p-value: 0.000000891
```

Why is the estimated coefficient for bed negative?

Possible explanations:

Adding extra bathrooms to homes with just two or three bedrooms might be considered a waste of space...

But there might be a benefit for homes with four or five bedrooms to have more than one bathroom.

Price of Homes for sale in Eugene, Oregon in 2005

```
> model2 <- lm(Price~ Size+ Lot + Bath*Bed + Age + Garage + Active +EdisonElementary +
HarrisElementary + AdamsElementary + CrestElementary + ParkerElementary, data=HDATA)
> summary(model2)

Call:
lm(formula = Price ~ Size + Lot + Bath * Bed + Age + Garage +
    Active + EdisonElementary + HarrisElementary + AdamsElementary +
    CrestElementary + ParkerElementary, data = HDATA)

Residuals:
   Min     1Q Median     3Q    Max
-96.37 -25.81   0.18  33.87  90.79

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         336.42     129.71    2.59   0.0118 *
Size                 64.80      30.26    2.14   0.0362 *
Lot                   9.07       3.67    2.47   0.0162 *
Bath                -95.42      49.75   -1.92   0.0597 .
Bed                 -79.29      33.54   -2.36   0.0212 *
Age                   1.72       3.24    0.53   0.5968
Garage               10.64       9.03    1.18   0.2434
Active               27.60      12.68    2.18   0.0333 *
EdisonElementary     77.25      17.23    4.48 0.000032 ***
HarrisElementary     52.44      16.16    3.25   0.0019 **
AdamsElementary     -21.32      29.11   -0.73   0.4666
CrestElementary      -5.71      22.40   -0.25   0.7998
ParkerElementary     -6.90      16.41   -0.42   0.6757
Bath:Bed             29.29      14.18    2.07   0.0430 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44 on 62 degrees of freedom
Multiple R-squared:  0.561,   Adjusted R-squared:  0.469
F-statistic: 6.09 on 13 and 62 DF,  p-value: 0.000000382
```
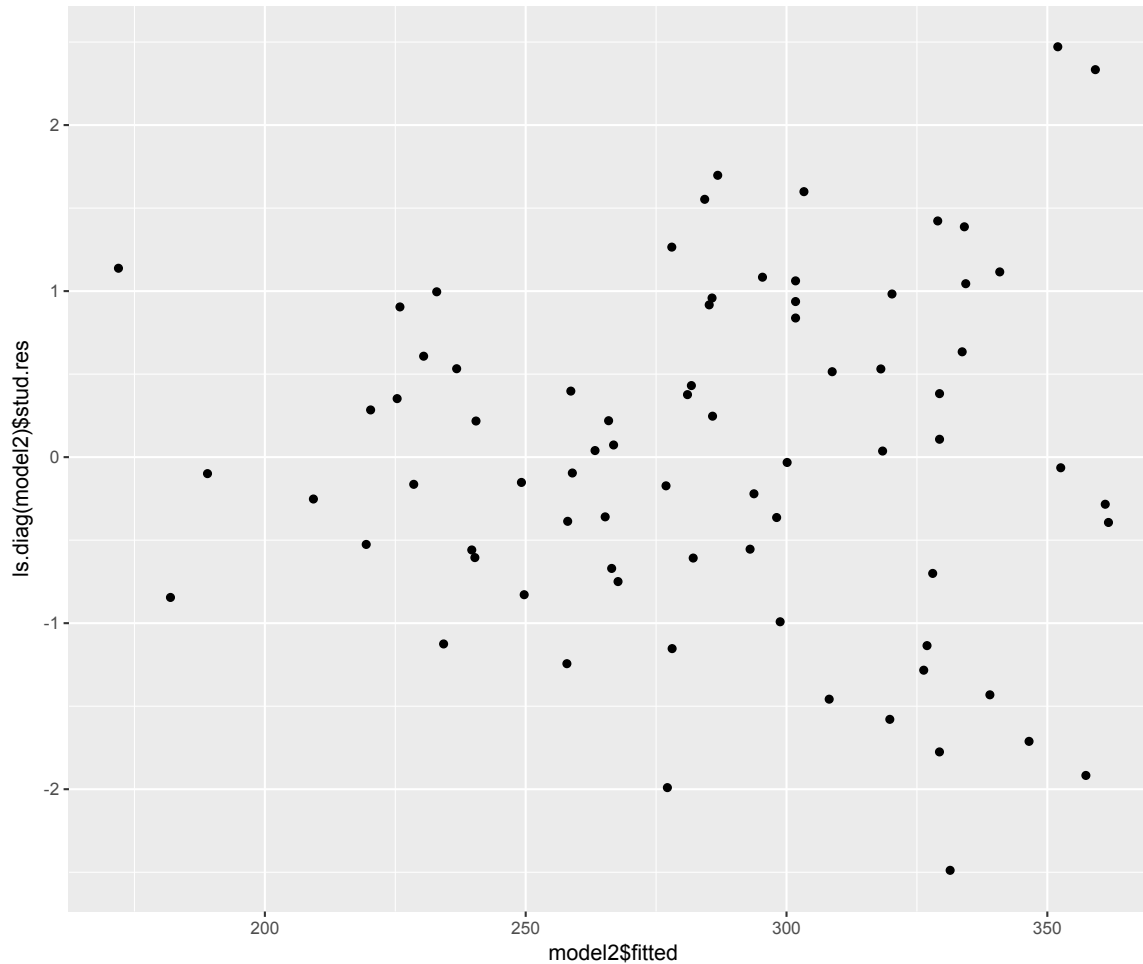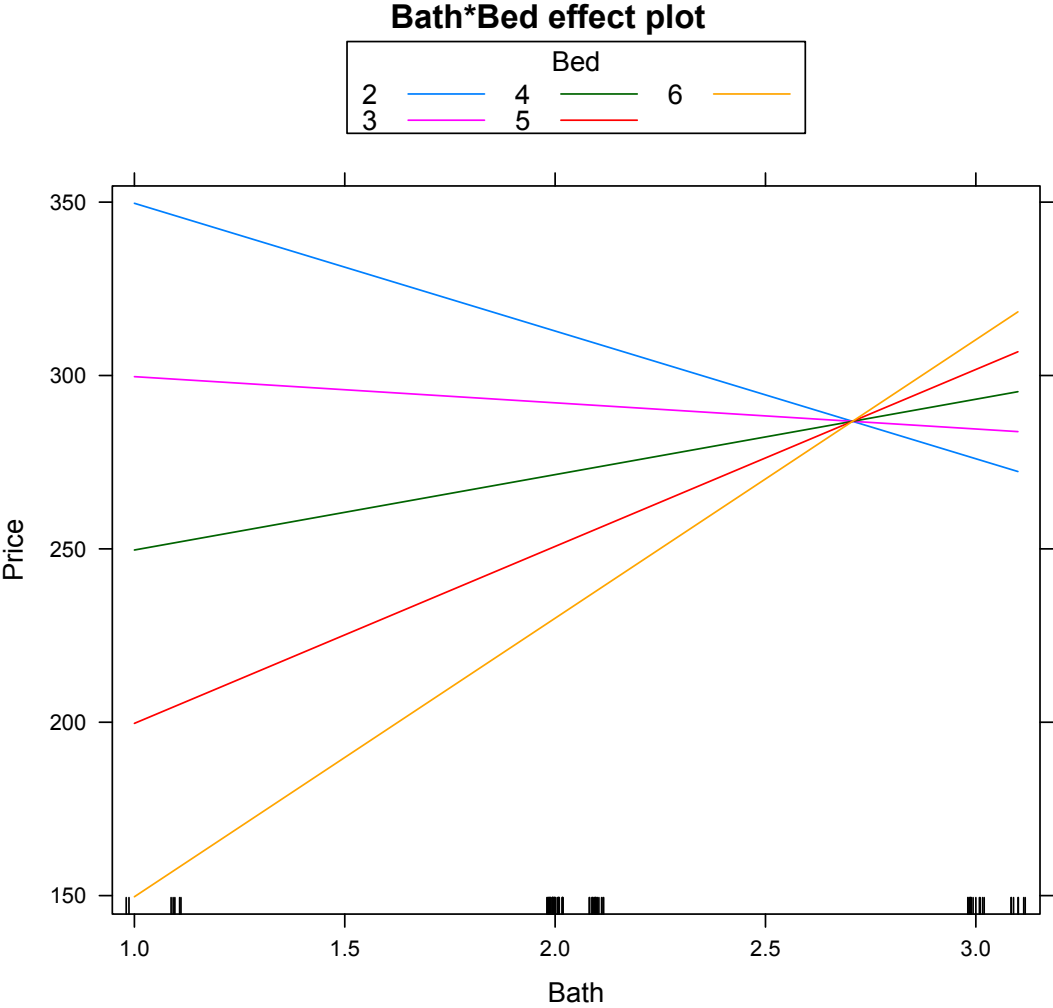
Price of Homes for sale in Eugene, Oregon in 2005

Price of Homes for sale in Eugene, Oregon in 2005

plot(effect(term="Bath:Bed",mod=model2,default.levels=20),multiline=TRUE)

Bath*Bed effect plot

Price of Homes for sale in Eugene, Oregon in 2005

```
> model3 <- lm(log(Price)~ Size+ Lot + Bath*Bed + Age + Garage + Active + EdisonElementary
+ HarrisElementary + AdamsElementary + CrestElementary + ParkerElementary, data=HDATA)
>
>
> summary(model3)

Call:
lm(formula = log(Price) ~ Size + Lot + Bath * Bed + Age + Garage +
    Active + EdisonElementary + HarrisElementary + AdamsElementary +
    CrestElementary + ParkerElementary, data = HDATA)

Residuals:
   Min      1Q Median     3Q    Max
-0.328 -0.104  0.010  0.121  0.243

Coefficients:
                  Estimate Std. Error t value            Pr(>|t|)
(Intercept)        5.98245    0.44323   13.50 < 0.0000000000000002 ***
Size               0.19946    0.10339    1.93             0.05829 .
Lot                0.02852    0.01253    2.28             0.02630 *
Bath              -0.37887    0.17001   -2.23             0.02948 *
Bed               -0.31840    0.11463   -2.78             0.00723 **
Age                0.00323    0.01108    0.29             0.77124
Garage             0.04571    0.03087    1.48             0.14374
Active             0.10776    0.04333    2.49             0.01558 *
EdisonElementary   0.26168    0.05887    4.45            0.000037 ***
HarrisElementary   0.19191    0.05522    3.48             0.00094 ***
AdamsElementary   -0.13009    0.09946   -1.31             0.19575
CrestElementary   -0.01052    0.07655   -0.14             0.89117
ParkerElementary  -0.00925    0.05606   -0.17             0.86944
Bath:Bed           0.11761    0.04844    2.43             0.01810 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.15 on 62 degrees of freedom
Multiple R-squared:  0.59, Adjusted R-squared:  0.504
F-statistic: 6.85 on 13 and 62 DF,  p-value: 0.0000000598
```
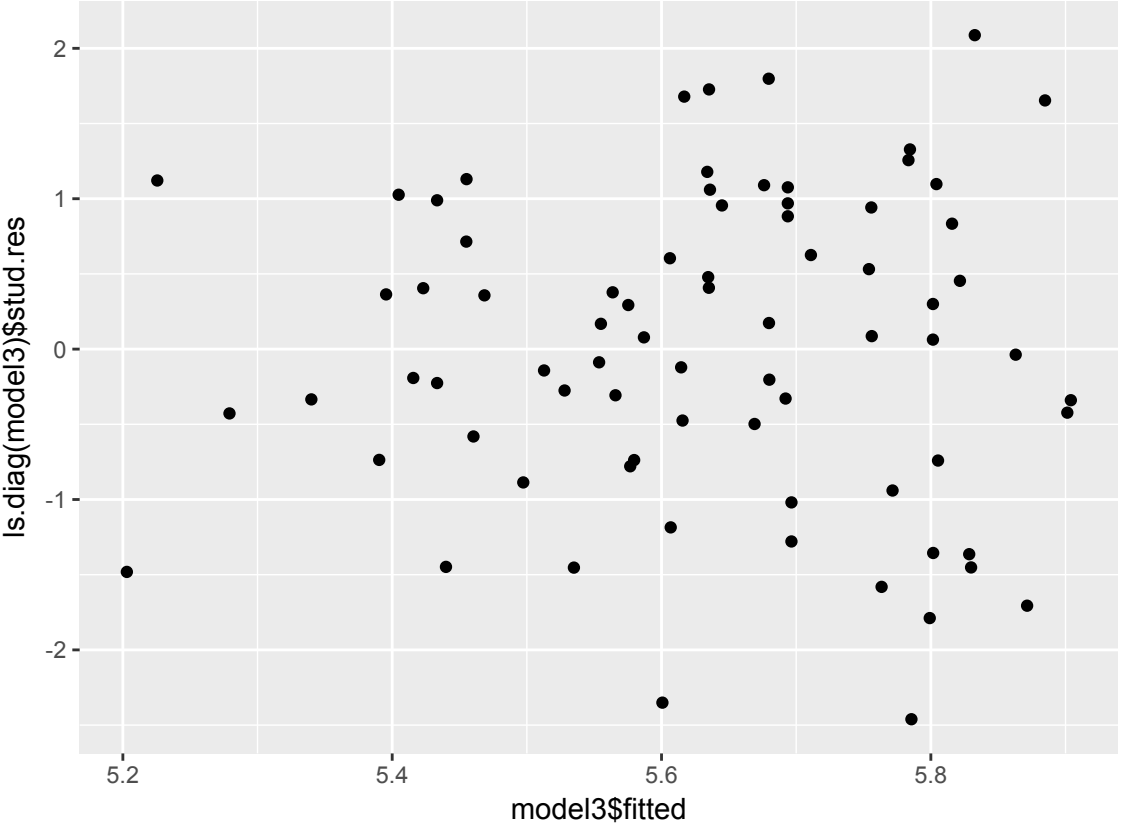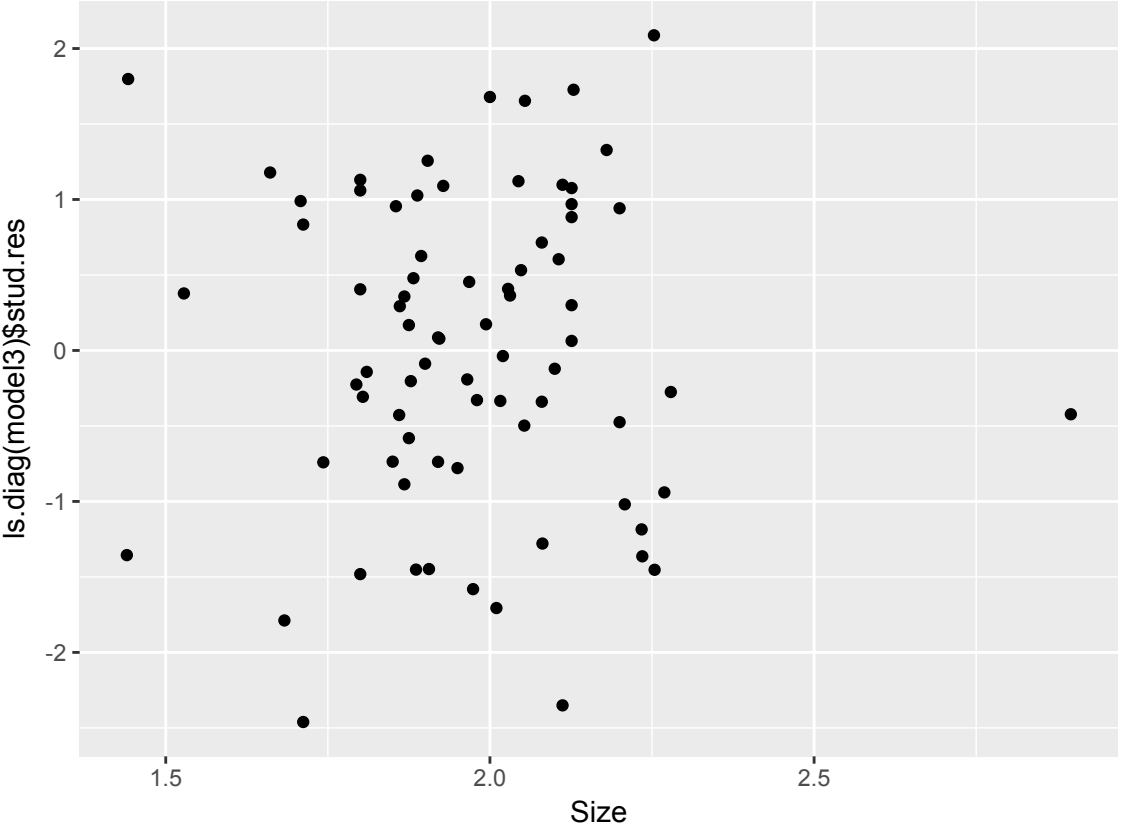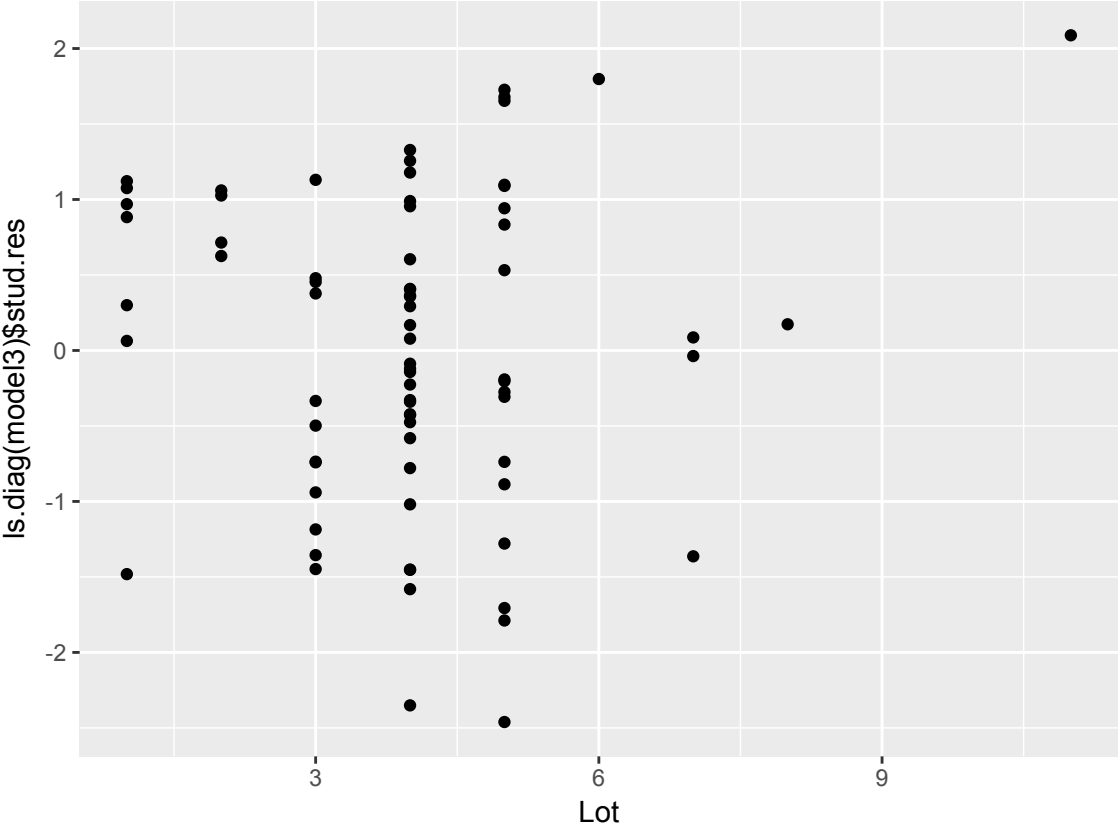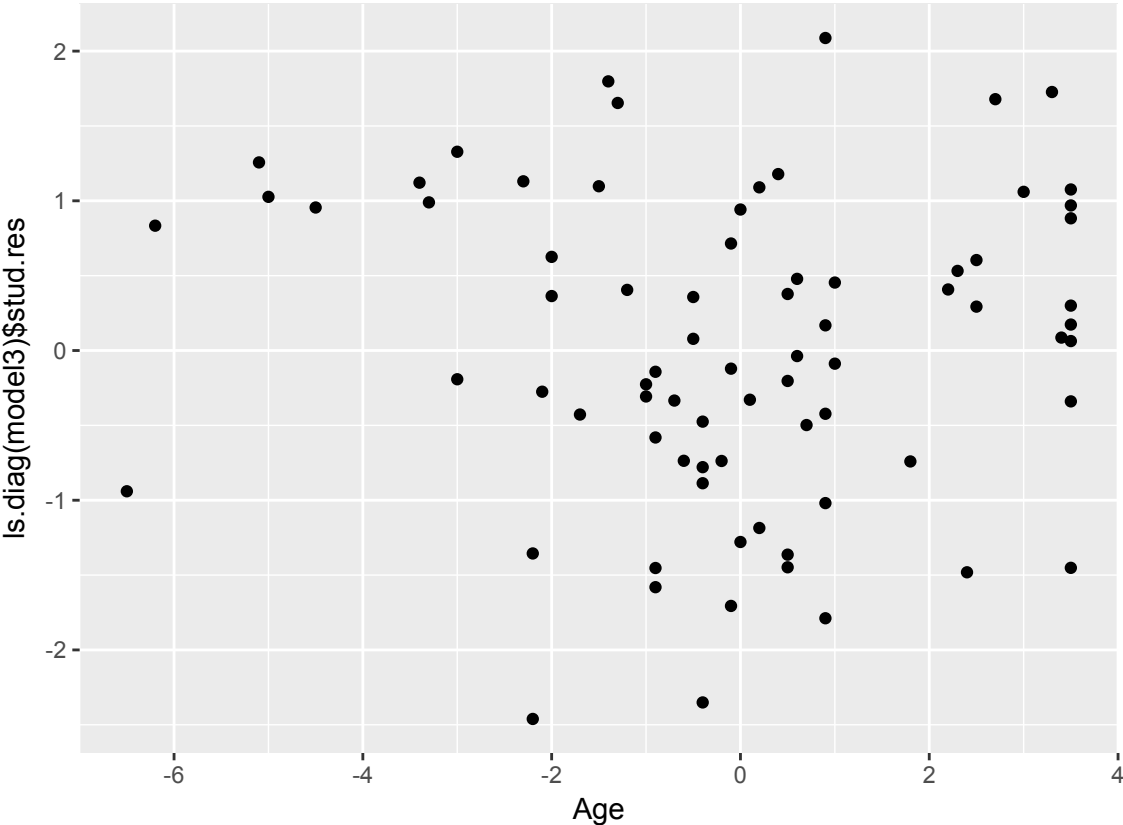
Price of Homes for sale in Eugene, Oregon in 2005

Price of Homes for sale in Eugene, Oregon in 2005

Price of Homes for sale in Eugene, Oregon in 2005

Price of Homes for sale in Eugene, Oregon in 2005

Price of Homes for sale in Eugene, Oregon in 2005

```
> summary(model4)

Call:
lm(formula = log(Price) ~ Size + Lot + Bath * Bed + Age + I(Age^2) +
    Garage + Active + EdisonElementary + HarrisElementary + AdamsElementary +
    CrestElementary + ParkerElementary, data = HDATA)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30591 -0.09073  0.00117  0.10335  0.30184

Coefficients:
                   Estimate Std. Error t value            Pr(>|t|)
(Intercept)         5.98715    0.42258   14.17 <0.0000000000000002 ***
Size                0.17674    0.09894    1.79              0.0790 .
Lot                 0.03340    0.01208    2.76              0.0075 **
Bath               -0.39137    0.16215   -2.41              0.0188 *
Bed                -0.31159    0.10931   -2.85              0.0059 **
Age                 0.01097    0.01095    1.00              0.3201
I(Age^2)            0.00687    0.00256    2.68              0.0093 **
Garage              0.04585    0.02943    1.56              0.1245
Active              0.11815    0.04149    2.85              0.0060 **
EdisonElementary    0.19369    0.06157    3.15              0.0026 **
HarrisElementary    0.14589    0.05536    2.64              0.0106 *
AdamsElementary    -0.15860    0.09542   -1.66              0.1016
CrestElementary    -0.02249    0.07312   -0.31              0.7594
ParkerElementary   -0.03578    0.05436   -0.66              0.5129
Bath:Bed            0.11900    0.04619    2.58              0.0124 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.143 on 61 degrees of freedom
Multiple R-squared:  0.633,   Adjusted R-squared:  0.549
F-statistic: 7.52 on 14 and 61 DF,  p-value: 0.00000000806
```
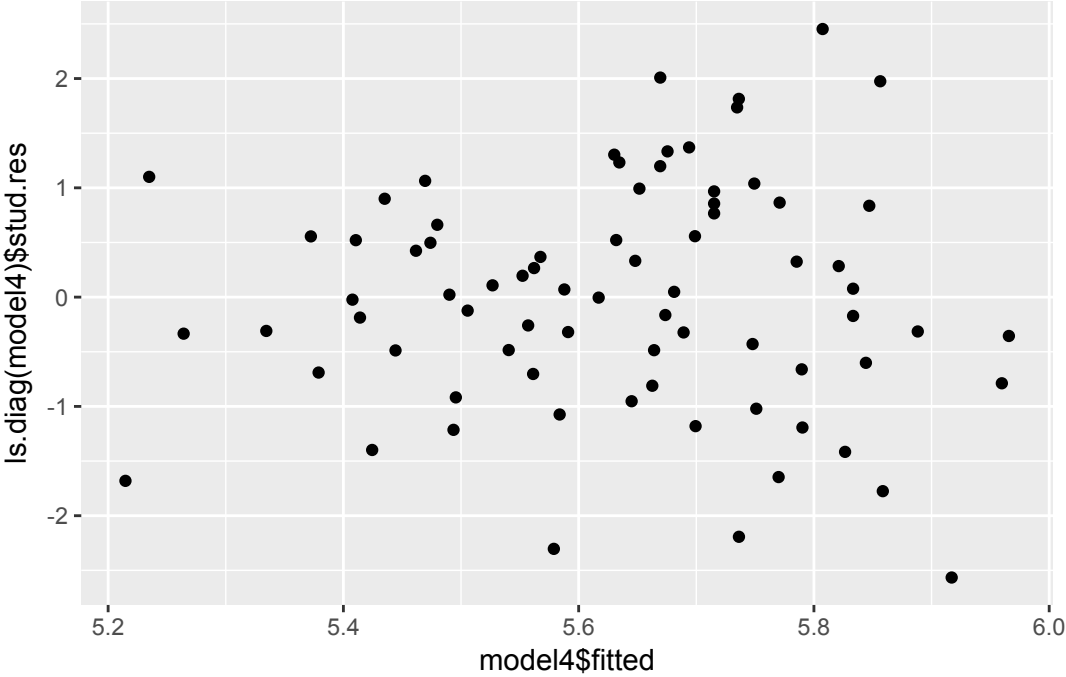
Price of Homes for sale in Eugene, Oregon in 2005

Studentized residuals vs. fitted values:

Price of Homes for sale in Eugene, Oregon in 2005

Normal Q-Q Plot