California Housing Price

Data from 1990

Predicting House Prices in 1990 California

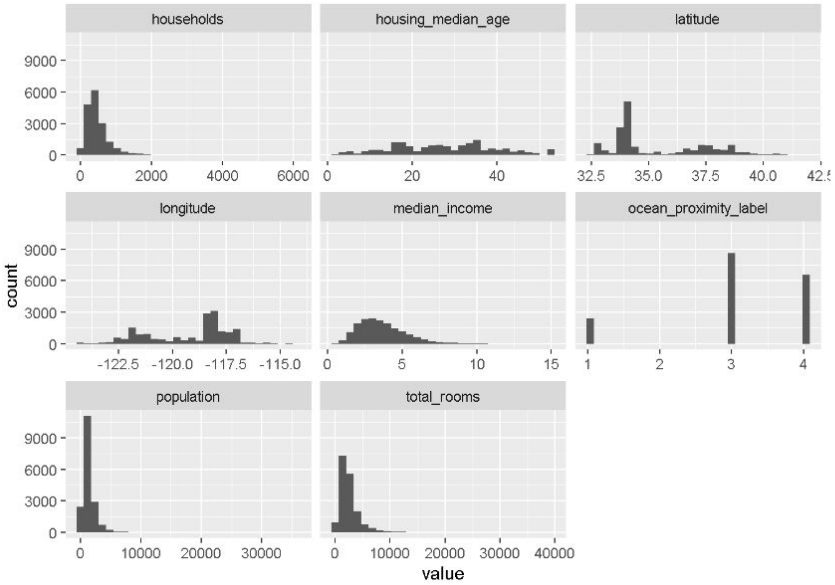
Our dataset's information was collected from the 1990 Census from every block group in California which holds over 19,000 block groups. In this dataset, a block group on average includes 1425 individuals living in a geographically compact area. Using linear regression methods we will ask two questions:

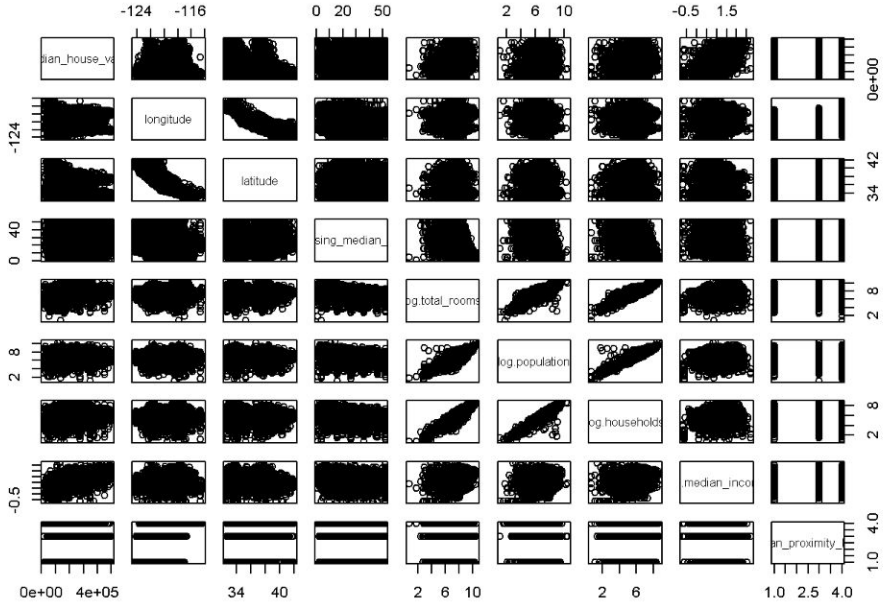
1. What combination of our features best predict median house value?
2. How precise are those predictions?

Exploring variables

- Response variable
  - Median house value – in dollars
- Predictor variables in the data set are (per block)
  - Households
  - Population
  - Median income
  - Latitude
  - Longitude
  - Total rooms
  - Median house age
  - Ocean proximity (1 = near ocean, 2 = near bay, 3 = <1 one hr from the ocean, 4 = inland)
- There are 19643 blocks in this dataset

Distributions and Correlations





The distributions of households, median income, population, and total rooms are skewed and need to be transformed

- There is a moderate correlation between median house value and median income.
- There is a strong inverse correlation between longitude and latitude.
- Total_rooms, population, and households are highly correlated with each other.

First-Order Model

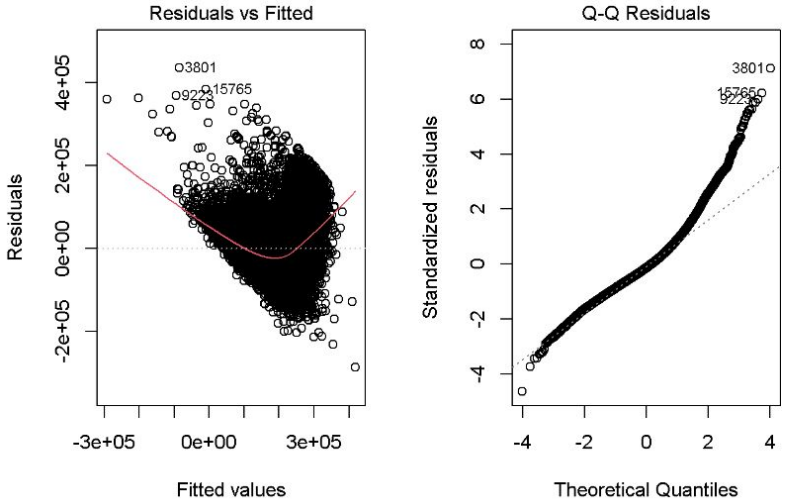
- Response variable is Median House Value
- All variables are statistically significant except log.total_rooms
- Ocean Proximity is the only categorical variable and has been one-hot encoded
- The adjusted R-squared of 0.576 suggests the model explains 57.6% of the variance in median house value. The residual standard error of $61,930 relative to a range of $14,999 to $499,100 is of moderate magnitude.

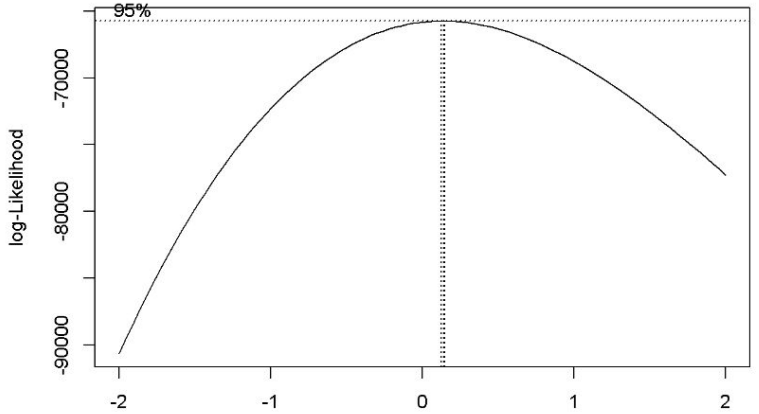
```
### First-order model with all predictors
fit1 = lm (median_house_value ~ longitude + latitude + housing_median_age + log.total_rooms + log.population +
            log.households + log.median_income + ocean_proximity_label, data=housing_data)
summary (fit1)
```

```
##
## Call:
## lm(formula = median_house_value ~ longitude + latitude + housing_median_age +
##     log.total_rooms + log.population + log.households + log.median_income +
##     ocean_proximity_label, data = housing_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -286803  -39961   -8486   30484  436232
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.823e+06  7.625e+04 -37.027   <2e-16 ***
## longitude             -3.540e+04  8.563e+02 -41.338   <2e-16 ***
## latitude              -3.600e+04  8.203e+02 -43.886   <2e-16 ***
## housing_median_age     5.913e+02  4.319e+01  13.692   <2e-16 ***
## log.total_rooms        3.912e+03  2.317e+03   1.688   0.0914 .
## log.population        -7.432e+04  1.792e+03 -41.469   <2e-16 ***
## log.households         7.126e+04  2.880e+03  24.748   <2e-16 ***
## log.median_income      1.159e+05  1.425e+03  81.298   <2e-16 ***
## ocean_proximity_label -8.014e+03  6.466e+02 -12.393   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61930 on 17546 degrees of freedom
## Multiple R-squared:  0.5765, Adjusted R-squared:  0.5763
## F-statistic:  2985 on 8 and 17546 DF,  p-value: < 2.2e-16
```

First-Order Model

1. The residual vs fitted plot violates the linearity assumption of constant residual variance. It is also curvy.
2. Residuals are not normally distributed with one tail highly being spread out
3. The box-cox suggest log transformation of the response variable (median house value)





First-Order Model Log House Value

- Response is Log Median House Value
- Every predictor variable except the log total room is statistically significant
- Adjusted R squared of 0.665 is moderately high and is significantly higher than model 1's value of 0.576
- Residual standard error of 0.308 is moderate considering the range of log transformed median house value

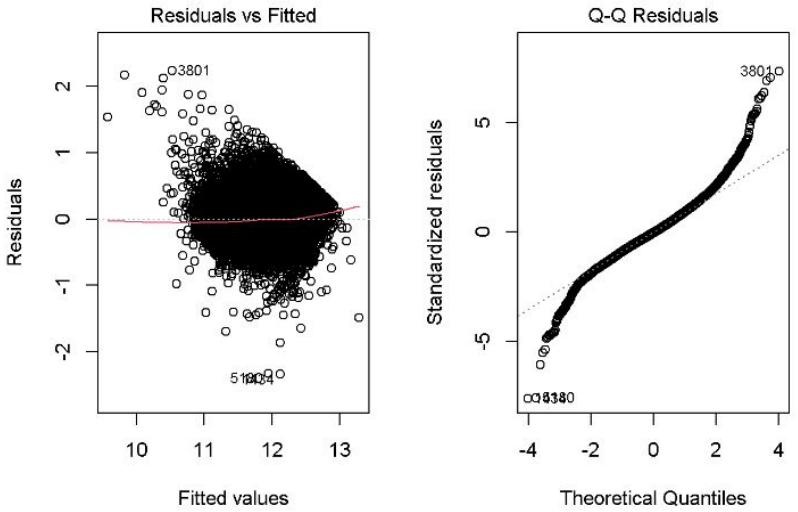
```
fit2 = lm (log(median_house_value) ~ longitude + latitude + housing_median_age + log.total_rooms + log.population
+
            log.households + log.median_income + as.factor (ocean_proximity_label),
             data=housing_data)
model2_summary = summary (fit2)

model2_summary
```

```
##
## Call:
## lm(formula = log(median_house_value) ~ longitude + latitude +
##     housing_median_age + log.total_rooms + log.population + log.households +
##     log.median_income + as.factor(ocean_proximity_label), data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34643 -0.18718 -0.01224  0.17853  2.23902
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.2638149  0.4091660  -0.645   0.5191
## longitude                         -0.1408634  0.0047015 -29.961  < 2e-16 ***
## latitude                          -0.1399435  0.0046568 -30.052  < 2e-16 ***
## housing_median_age                 0.0011800  0.0002158   5.468 4.60e-08 ***
## log.total_rooms                    0.0197217  0.0116344   1.695   0.0901 .
## log.population                    -0.3717623  0.0089145 -41.703  < 2e-16 ***
## log.households                     0.3647769  0.0143980  25.335  < 2e-16 ***
## log.median_income                  0.6294471  0.0072178  87.207  < 2e-16 ***
## as.factor(ocean_proximity_label)3  0.0488084  0.0073523   6.639 3.26e-11 ***
## as.factor(ocean_proximity_label)4 -0.2735894  0.0103153 -26.523  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.308 on 17545 degrees of freedom
## Multiple R-squared:  0.6654, Adjusted R-squared:  0.6653
## F-statistic:  3877 on 9 and 17545 DF,  p-value: < 2.2e-16
```

First-Order Model Log House Value

The residuals in the residuals vs fitted plot have more constant variance and decreased curvature. the Q-Q plots of model 1 and 2 are similar but model 2's left tail is more skewed.



Final Model

Determining the Final Model: The final model was determined through exploratory data analysis, model selection, and stepwise regression. This process involved adding or removing predictor variables to identify a statistically significant set of features,

Interaction Effects: Interaction effects were also used to understand the influence of multiple predictors on the median house value.

Residual SE and Adjusted R squared: Very similar to the first order model 2 but with slightly better RSE and moderately better R squared.

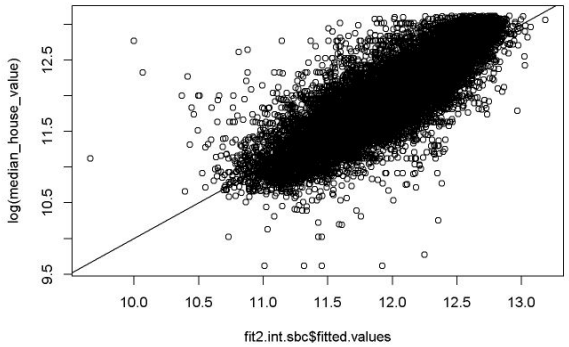
```
summary(fit2.int.sbc)
```

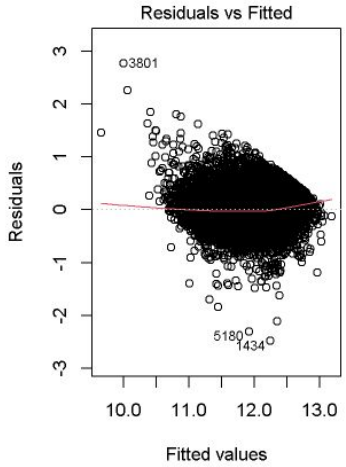
```
##
## Call:
## lm(formula = log(median_house_value) ~ latitiude.c + longitude.c +
##     housing_median_age.c + log.population.c + log.households.c +
##     log.median_income.c + as.factor(ocean_proximity) + latitiude.c:longitude.c +
##     latitiude.c:housing_median_age.c + latitiude.c:log.households.c +
##     latitiude.c:log.median_income.c + latitiude.c:as.factor(ocean_proximity) +
##     longitude.c:housing_median_age.c + longitude.c:log.median_income.c +
##     longitude.c:as.factor(ocean_proximity) + housing_median_age.c:log.population.c +
##     housing_median_age.c:log.households.c + log.population.c:log.households.c +
##     log.population.c:log.median_income.c + log.population.c:as.factor(ocean_proximity) +
##     log.households.c:as.factor(ocean_proximity) + log.median_income.c:as.factor(ocean_proximity),
##     data = housing_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48006 -0.17452 -0.00988  0.17105  2.76558
##
```

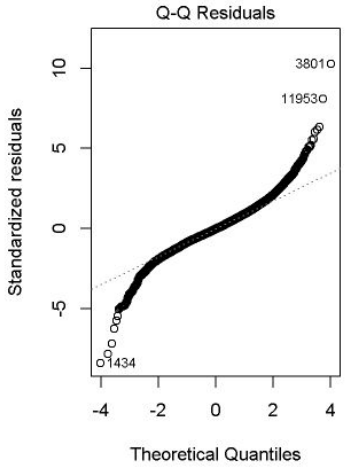
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2952 on 17526 degrees of freedom
## Multiple R-squared:  0.693,  Adjusted R-squared:  0.6925
## F-statistic:  1413 on 28 and 17526 DF,  p-value: < 2.2e-16
```

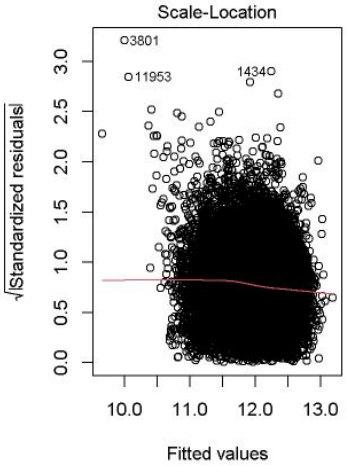
Final Model

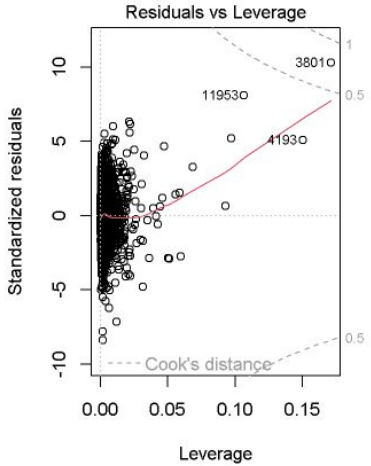
- No obvious patterns (good linearity)
- Constant variance
- Minimal curvature
- The distribution is not normal
- Due to the presence of influential observations we will consider removing data points that have cook's' distance values greater than 3 standard deviations above and below the mean cook's distance value.











Final Model Influential Points

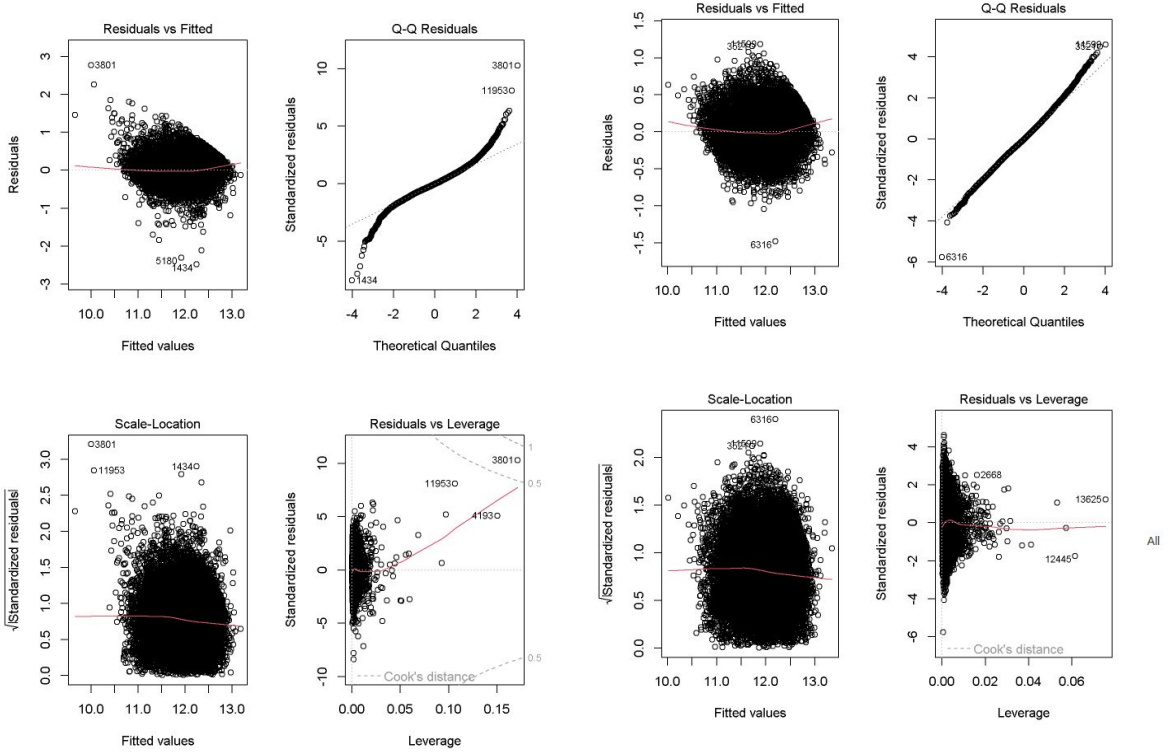
Pros of Removing Influential Points:

- Improved Normality n the Q-Q plot
- More symmetrical leverage plot suggests the model may be more stable

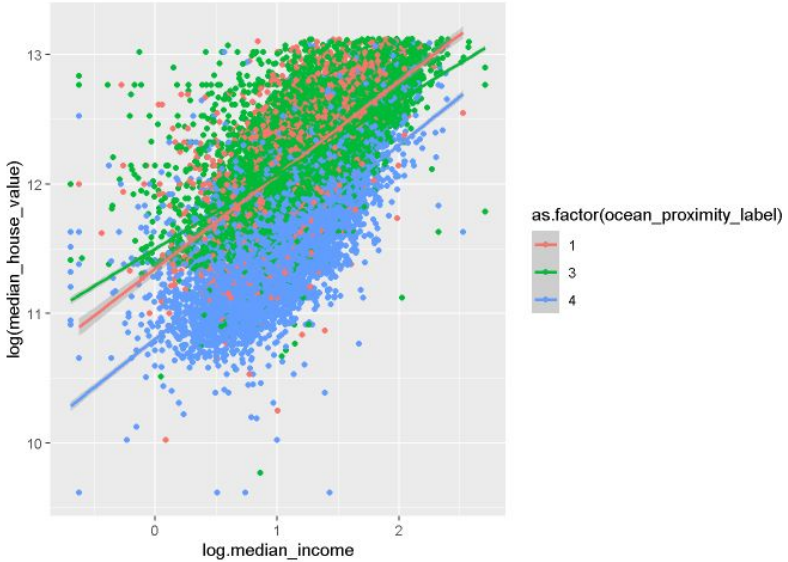
Cons:

- Similar Residual Plots: Since the scale-location and fitted vs. residuals plots are similar we can argue that the removal of points doesn't improve these aspects.
- Although we will not test these models on external datasets there is the possibility that the removal leads to overfitting
- Since the influential data points appear valid when observing them and come from a reliable source(1990 census), they may represent important variability in the data that could be important for accurate predictions.

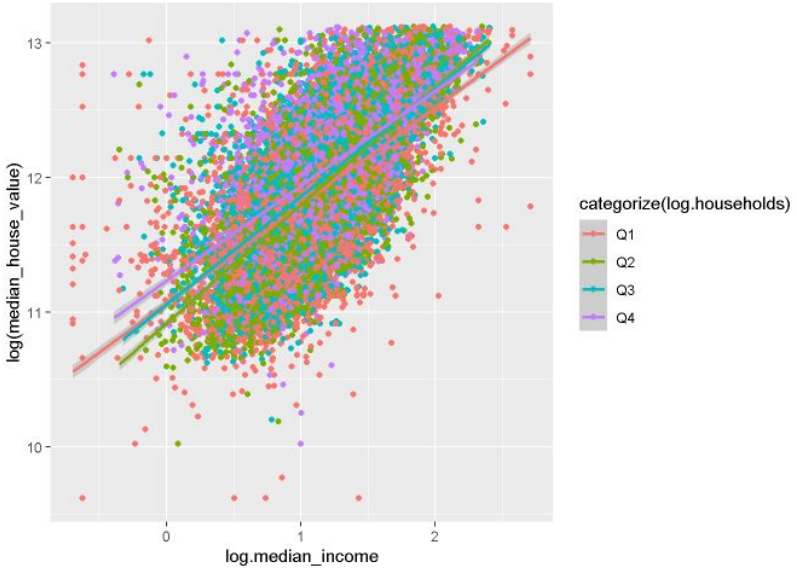
After considering these pros and Cons we have decided that there is no clear benefit to removing influential points



Final Model – Interaction Effects



Log median house value vs median income by ocean proximity: This plot shows that the relationship between log median house value and log median income is stronger for houses that are inland (4) that the houses in other areas.



Log median house value vs median income by number of households: This plot shows that relationship between log of median house value and log of median income is stronger with the number of household not having too much (Q4) or too less (Q1) per block.

Final Model- Example Predictions

Low-Priced Houses: For homes with a low median house value (e.g., around $110,000), the model predicts the price with a precision of approximately +/- $65,000 (95% prediction limits).

Mid-Priced Houses: For homes in the mid-price range (e.g., around $250,000), the precision is approximately +/- $150,000.

High-Priced Houses: For high-priced homes (e.g., above $300,000), the precision becomes wider, approximately +/- $200,000.

- The prediction intervals are wide and become wider as the median house value increases. There is more uncertainty in predicting higher values. The model is overall suboptimal at predicting median house value
- Prediction Accuracy: The overall success rate of 0.9505 indicates that in about 95% of the cases, the actual median house value falls within the predicted interval.
- Final model has a residual standard error of 0.295 and an Adjusted R squared of 69.25%

```
##       households median_income longitude latitude median_house_value
## 209          642        2.8438   -121.98    38.37              111500
## 662          219        1.9018   -119.98    38.94              109700
## 5687        4072        6.6288   -117.87    34.04              339700
## 497          428        3.8438   -122.41    37.66              250700
## 16808        851        3.9722   -122.31    37.55              354100
## 4558         335        4.0000   -118.28    33.80              165500
##       pred.median_house_value housing_median_age population pred.lower
## 209                110336.01                 21       2018   61854.41
## 662                 79730.32                 25        503   44672.59
## 5687               311507.95                  7      15037  174409.61
## 497                238973.09                 37       1255  133946.58
## 16808              272743.39                 27       1877  152848.78
## 4558               213075.68                 38       1207  119461.38
##       pred.upper
## 209     196817.6
## 662     142300.3
## 5687    556375.3
## 497     426350.1
## 16808   486683.4
## 4558    380049.5
```

```
housing_data$in.interval = ifelse (housing_data$pred.lower <= housing_data$median_house_value &
                                   housing_data$median_house_value <= housing_data$pred.upper,
                                   1, 0)
mean (housing_data$in.interval)
```

```
## [1] 0.9505554
```