(F23) PSTAT 126: Project Step 2

Anthony Cu and William Mahnke

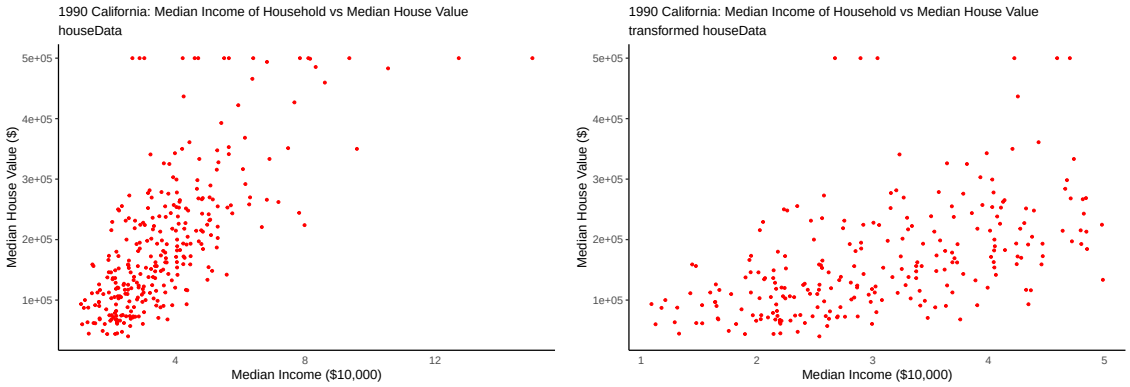
2023-11-12

Introduction

In our project, we explore the 1990 California Housing dataset, providing information on a specified district in the state. The survey data describes homes in a district of California in 1990 in order to represent the larger population of this district during the 1990's overall. Each independent observation corresponds to a different block within the district.

We notice there exists outliers in our dataset, which may result in our fitted model not satisfying the model assumptions. So, we transform our data by filtering only observations with a median income under $50,000.

```
houseData[houseData$median_income < 5, ]
```



We notice the transformed scatterplot has fewer points horizontally lying on the top.

Hypotheses

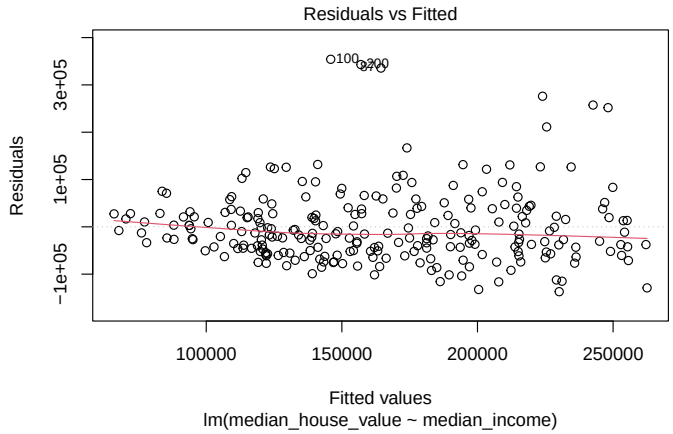
We intend to test whether our *response*, median house value, is linearly related with our *predictor*, median income (measured in $10,000). In other words, the null hypothesis is that median house value is linearly dependent on median income. The alternate hypothesis is that median house value is not linearly dependent on median income.

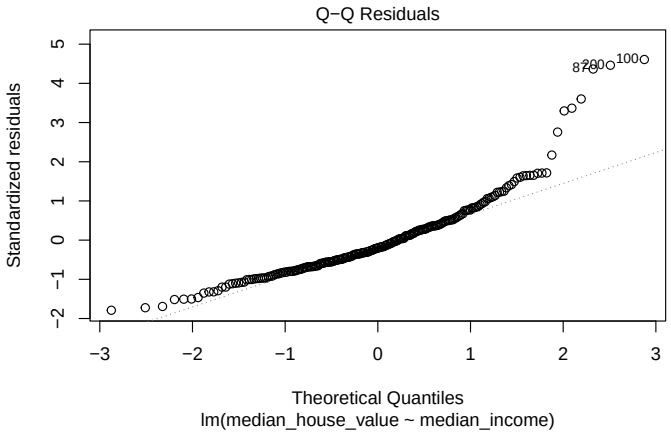
$$\begin{cases} H_0 : \beta_1 = 0 \\ H_a : \beta_1 \neq 0 \end{cases}$$

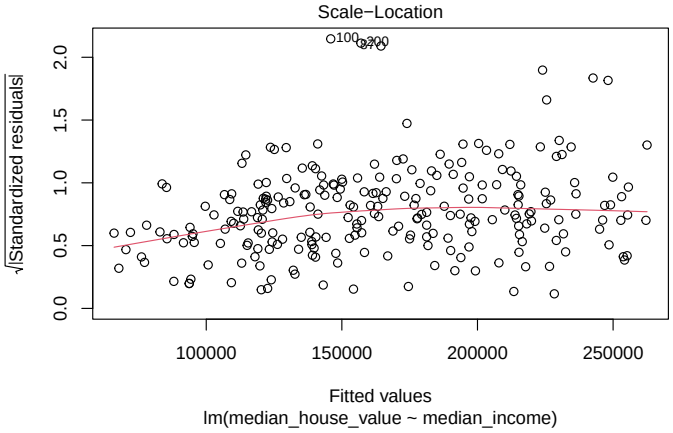
Assumptions for Linear Regression

We use plot() on our fitted model to test these model assumptions.

```
plot(lm(median_house_value ~ median_income,
        data = houseData[houseData$median_income < 5, ]), which = 1:3)
```







The **Residuals vs Fitted plot** shows that the red line closely resembles the dashed line, signifying that the linearity assumption holds, i.e. we can assume $\mathbb{E}[\varepsilon_i] = 0$ for every i.

The **Q-Q Residuals plot** showcases that because the observations fall on the percentile-matched line, they are normally distributed, satisfying the Normal assumption, i.e. we can assume $\varepsilon_i \sim N(0, \sigma^2\mathbb{I})$, for every i.

The **Scale-Location plot** shows a red line that is approximately horizontal, implying that the average magnitude of the standard residuals is not changing much. In addition, the spread around the red line does not vary with the fitted values. Thus means that the variance is constant, satisfying the homoscedasticity assumption, i.e. we can assume $\text{Var}(\varepsilon_i) = \sigma^2$ for every i.

Confidence Interval of β_1

We fit our linear model using the lm() function.

```
fit_houseData <- lm(median_house_value ~ median_income,
                    data = houseData[houseData$median_income < 5, ])
```

```r
fit_houseData %>% summary()
```

```
##
## Call:
## lm(formula = median_house_value ~ median_income, data = houseData[houseData$median_income <
##     5, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -137032  -50663  -15281   30866  354106
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11052      16273   0.679    0.498
## median_income    50435       5100   9.889   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77020 on 246 degrees of freedom
## Multiple R-squared:  0.2845, Adjusted R-squared:  0.2816
## F-statistic:  97.8 on 1 and 246 DF,  p-value: < 2.2e-16
```

We create a confidence interval for β_1.

```r
confint(fit_houseData, 'median_income', level = 0.95)
```

```
##                   2.5 %   97.5 %
## median_income 40390.21 60480.26
```

With 95% confidence, a \$10,000 increase in median income will increase the average median house value between \$40,390 and \$60,480.

Our analysis shows $\beta_1 = 5.0435237 \times 10^4$, which lies within the confidence interval of β_1 calculated above, [40390.21, 60480.26]. We observe that our confidence interval for β_1 does not contain 0, leading us to conclude that there is evidence of a linear relationship between the predictor and response as $\beta_1 \neq 0$.

The p-value for this hypothesis test is $1.2412509 \times 10^{-19} < 0.05$. Thus, the null hypothesis is rejected so median income is a significant predictor of the median house value is linear when not considering the observations where the median income is greater than \$50,000.

Fit of Model

We observe that the $R^2 = 0.2844711$, with the adjusted $R^2 = 0.2815624$. This value indicates that 28.16% of variance in the median house value is explained by the model, so 71.84% of the variance in the median house value is explained by the noise. This model doesn't consider observations where the median income is \$50,000 or greater.

Confidence and Prediction Interval

Here we calculate and plot our confidence and prediction interval for the transformed data set.

We now create a data grid, a sequence of unique values from the median_income column. Using add_predictions(), we will append a column of model predictions.

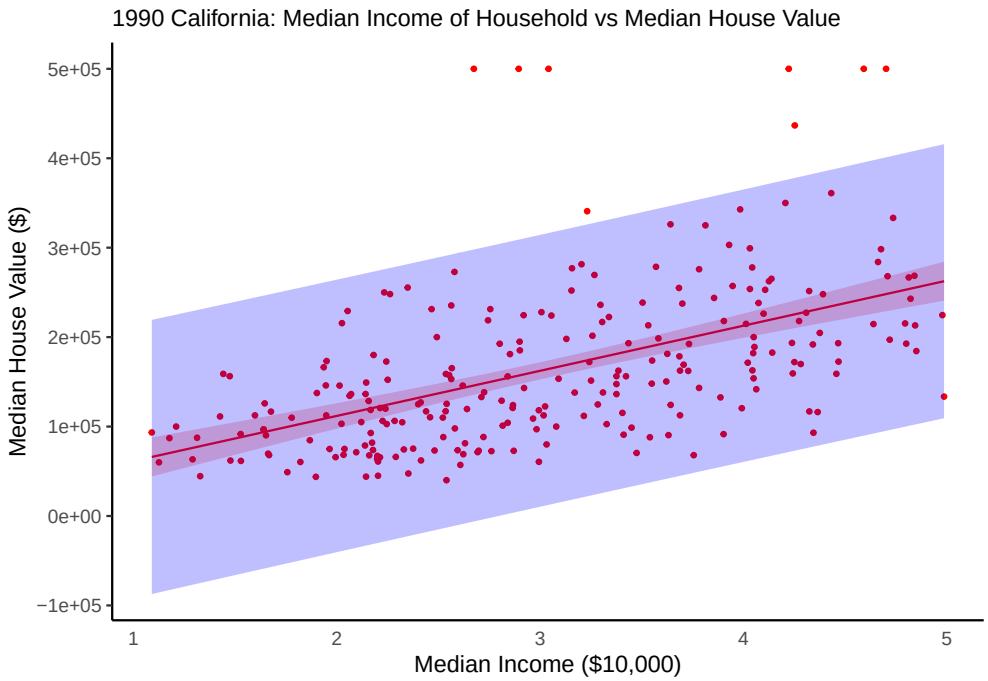
We create a confidence interval for the **mean response**, displaying the interval for the median income of $12,904.

```
##      fit      lwr      upr
## 76133.37 56062.57 96204.16
```

We proceed with creating a prediction interval for a **specific response**, displaying the interval for the median income of $12,904.

```
##       fit       lwr       upr
##  76133.37 -76892.40 229159.13
```

Finally, we visualize our 95% confidence and prediction interval.



With 95% confidence, the mean median house value when the median income is $12,904 is between $56,062.57 and $96,204.16. With 95% confidence, the predicted median house value when the median income is $12,904 is between -$76,892.40 and $229,159.10.

Conclusion

Our results suggest that there is a positive linear relationship between the median income and the median house value (not considering the observations where the median income is $50,000 or greater). We were interested in how the linear relationship between the variables appeared only after filtering out some of the observations to ensure the normality assumption for the noise was satisfied. The data was what we had expected, as the model explained a large percentage of the variance in the observations. Some other questions we would like to ask about the data include:

- Is there a linear relationship between other variables and the median house value?
- How does considering all of the median income observations change its relationship with the median house value?
- What variables are significant in predicting the median house value?