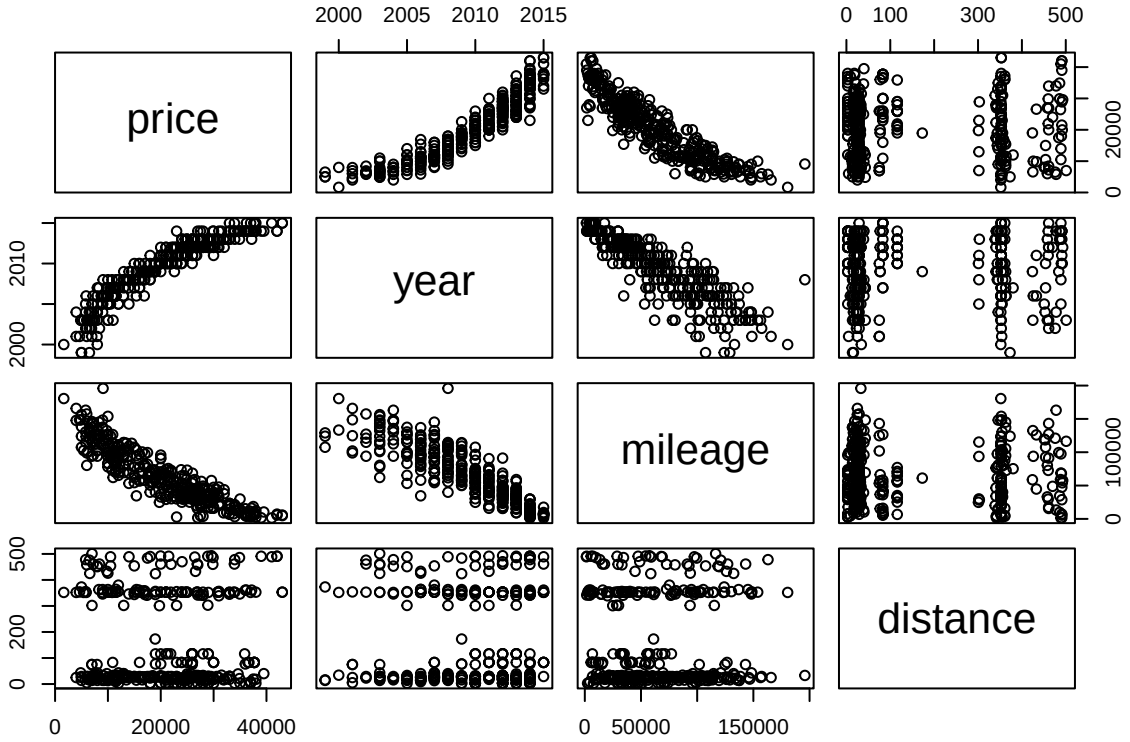
Homework 2

Alex Ojemann

2023-09-22

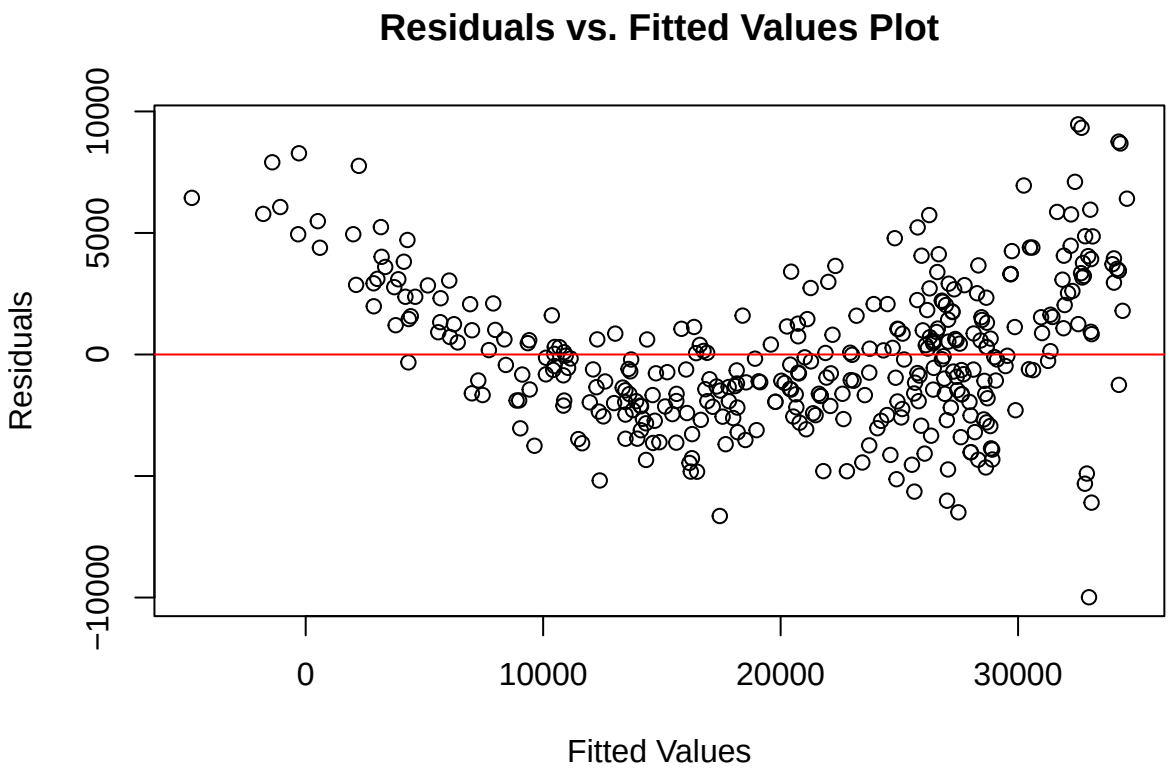
Problem 1

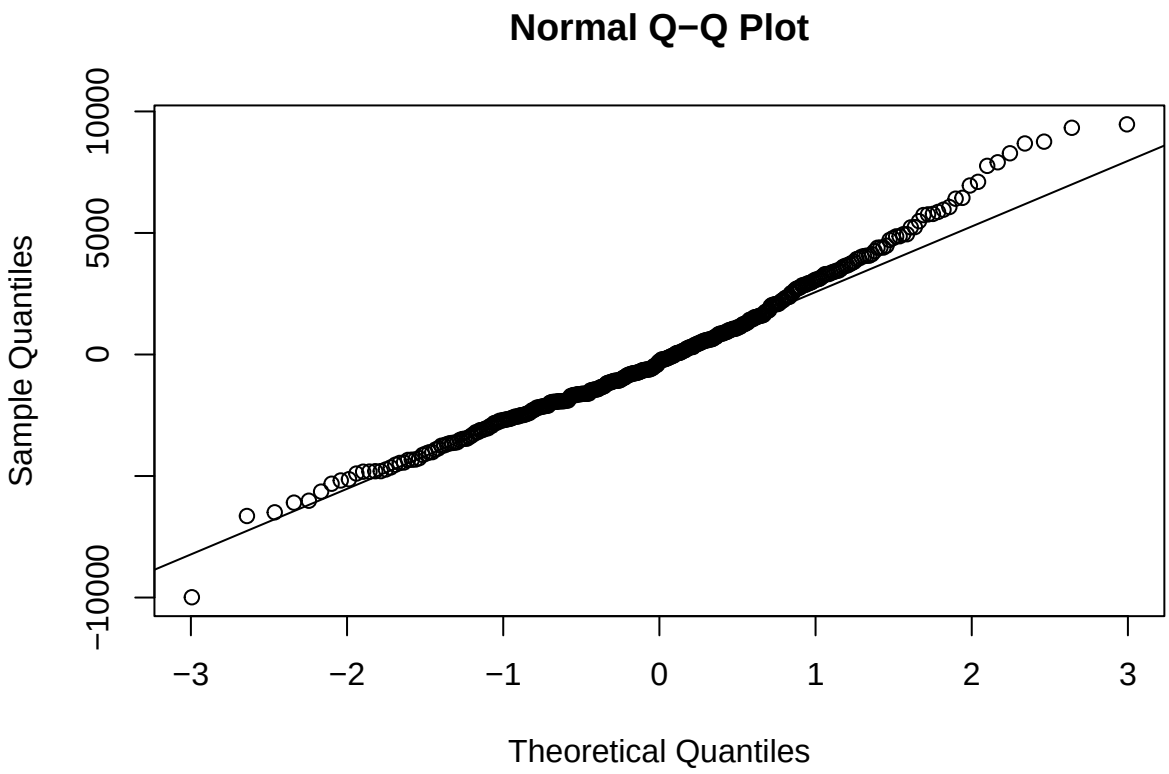
a. Asking price increases with year nonlinearly, possibly quadratically or exponentially. It also decreases with distance nonlinearly, possibly quadratically or exponentially. It has no visually apparent relationship with distance.



b. There is no evidence that distance from Boulder/Denver makes a difference because given a constant year and mileage, the probability of a value as far or further from 0 as the estimated coefficient for distance if there truly is no relationship between distnace and price is 0.696, which is greater than the 0.05 level of significance.

c. There is a very clear trend in the residuals vs fitted values plot. The plot also shows increasing variance as the fitted values get larger. Thus, this model is wrong. The normal QQ plot shows slightly heavier tails but isn't cause for additional concern.


Residuals vs. Fitted Values Plot
Residuals
10000
5000
0
−10000
0
10000
20000
30000
Fitted Values



d. I tried a model using year and mileage because those were the two useful predictors in part b and distance had no visually apparent relationship with price. I then tried a model with the interaction between those two. I then tried one with the degree two polynomial of year and one with the degree two polynomial of mileage because they appeared like they may have quadratic relationships to price in part a. I then tried one with both the degree two polynomial of year and the degree two polynomial of mileage and one with the interaction between the degree two polynomial of year and the degree two polynomial of mileage to see if they predict price effectively together as degree two polynomials. I then tried two more models the same as the previous two but with distance added to see if it could be a significant predictor when using degree two polynomials of year and mileage instead of linear factors as in part b.

e. The model with the lowest AIC and BIC was "price ~ poly(year, degree = 2) + poly(mileage, degree = 2) + distance." Distance from the front range does affect the price when controlling for mileage and year when both are polynomials of degree 2. It had an estimated coefficient of -1.905 and a p value of 0.00971, which is lower than the 0.05 level of significance.

f. To get the car that's most undervalued according to the most successful model from part e, I would find which one had the minimum residual. This is car number 157, which has a price of $22999, was made in 2014, has 2796 miles and is 355 miles from the front range. The model predicted the price to be $12026.76 higher.

g. When engine and color are added to the model, there are dozens of additional dummy variables created, most of which are not statistically significant when all are included. One way to limit the number of dummy variables would be to use forward or backward stepwise feature selection. Another way to reduce the dimensionality of the entire data set would be to use principal component analysis.

Appendix (R code)

a.

```
original_df = read.csv("AudiA4.csv")
View(original_df)
```

```
df <- original_df[, c("price","year","mileage","distance")]
pairs(df)
```

b.

```
model = lm(price ~ .,df)
summary(model)
```

c.

```
plot(fitted(model), residuals(model),
     main = "Residuals vs. Fitted Values Plot",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```

```
residuals = residuals(model)
qqnorm(residuals)
qqline(residuals)
```

d.

```
model2 = lm(price ~ poly(year, degree = 2),df)
summary(model2)
model3 = lm(price ~ poly(year, degree = 2)*poly(mileage,degree = 2),df)
summary(model3)
model4 = lm(price ~ poly(year, degree = 2) + poly(mileage,degree = 2),df)
summary(model4)
model5 = lm(price~year+mileage,df)
summary(model5)
model6 = lm(price ~ poly(mileage, degree = 2),df)
summary(model6)
model7 = lm(price ~ year*mileage,df)
summary(model7)
model8 = lm(price ~ poly(year, degree = 2)*poly(mileage,degree = 2)+distance,df)
summary(model8)
model9 = lm(price ~ poly(year, degree = 2) + poly(mileage,degree = 2)+distance,df)
summary(model9)
```

e.

```
print(AIC(model))
print(BIC(model))
print(AIC(model2))
print(BIC(model2))
print(AIC(model3))
print(BIC(model3))
print(AIC(model4))
print(BIC(model4))
print(AIC(model5))
print(BIC(model5))
print(AIC(model6))
print(BIC(model6))
print(AIC(model7))
print(BIC(model7))
print(AIC(model8))
print(BIC(model8))
print(AIC(model9))
print(BIC(model9))
```

f.

```
index = which.min(residuals(model9))
print(df[index,])
print(min(residuals(model9)))
```

g.

```
df2 <- original_df[, c("price","year","mileage","distance","engine","color")]
model_categorical = lm(price ~ .,df2)
summary(model_categorical)
```