Model Building

Isaac Scott

2022-04-04

Model Building: Modelling the Price of Cars from the "jaybob" Dataset.

Model 1

$$Price_i \sim \beta_0 \ + \beta_1 Age_i \ + \ \beta_2 Odometer_i$$

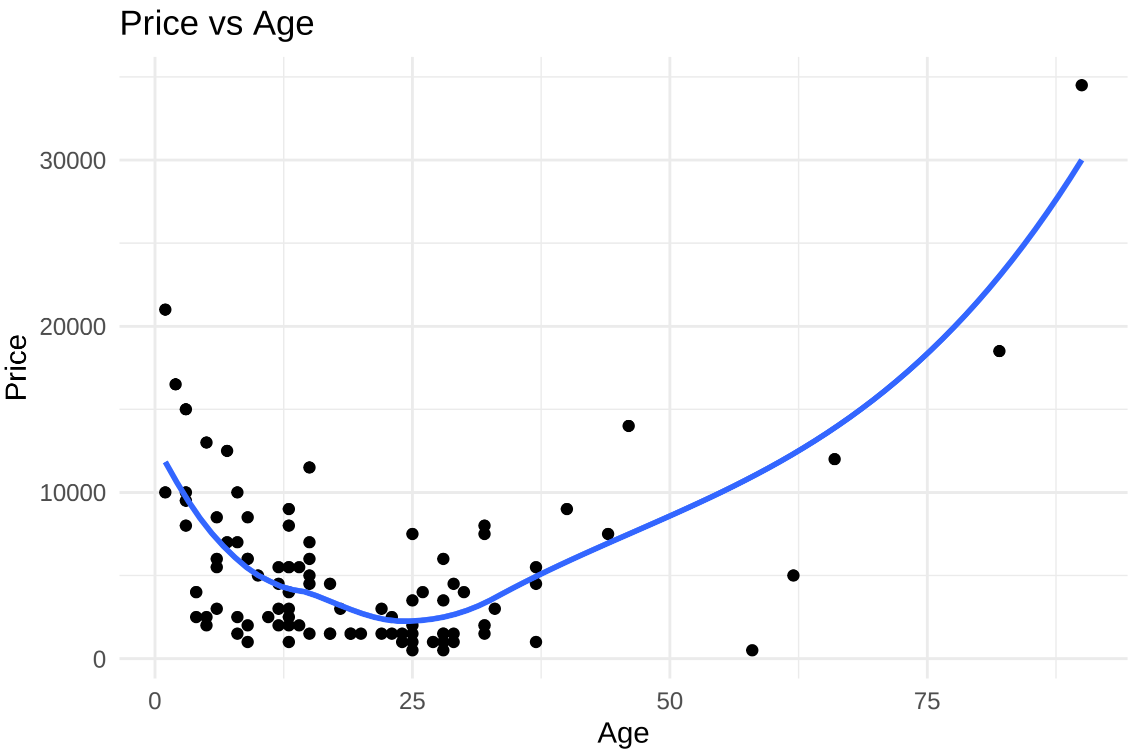
To start with, we model Price just with two variables: Age and Odometer.

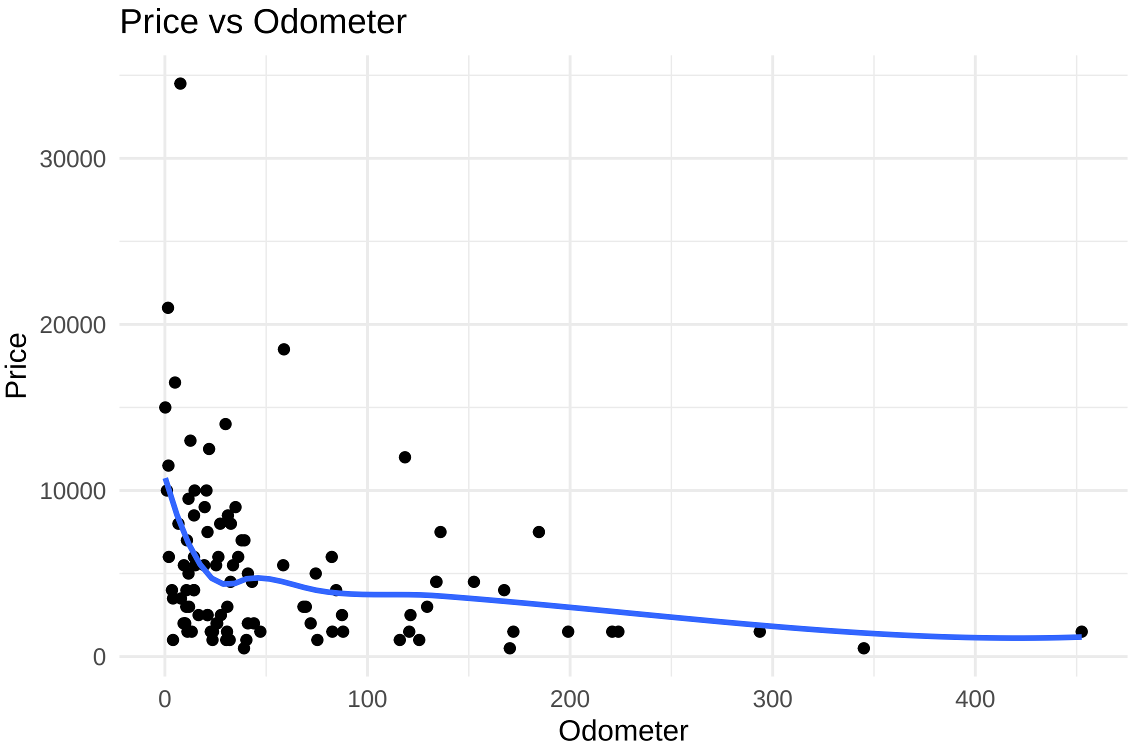
```
## (Intercept)         Age    Odometer 
##  4615.90038    98.92154   -23.02873
```

```
## 
## Call:
## lm(formula = Price ~ Age + Odometer, data = jaybob)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5996.3 -3372.8  -556.9  1529.2 21157.5 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4615.900    792.153   5.827 7.36e-08 ***
## Age           98.922     29.974   3.300 0.001352 ** 
## Odometer     -23.029      6.284  -3.665 0.000404 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4680 on 97 degrees of freedom
## Multiple R-squared:  0.1707, Adjusted R-squared:  0.1536 
## F-statistic:  9.98 on 2 and 97 DF,  p-value: 0.0001145
```

As we can see from the p-values, all the p-values seem significant, but the fit is bad, as evidenced by the Adjusted R^2 value of 0.1536.

This may be down to the relationship of the two variables to price. To find out if this is the case we must plot them.





We can see two things from these plots.

From the first plot, we can see that the relationship between the Age and Price appears to be a non-linear one. It resembles a parabolic curve around age where price decreases with age up to an age of 20 then price begins to increase again.

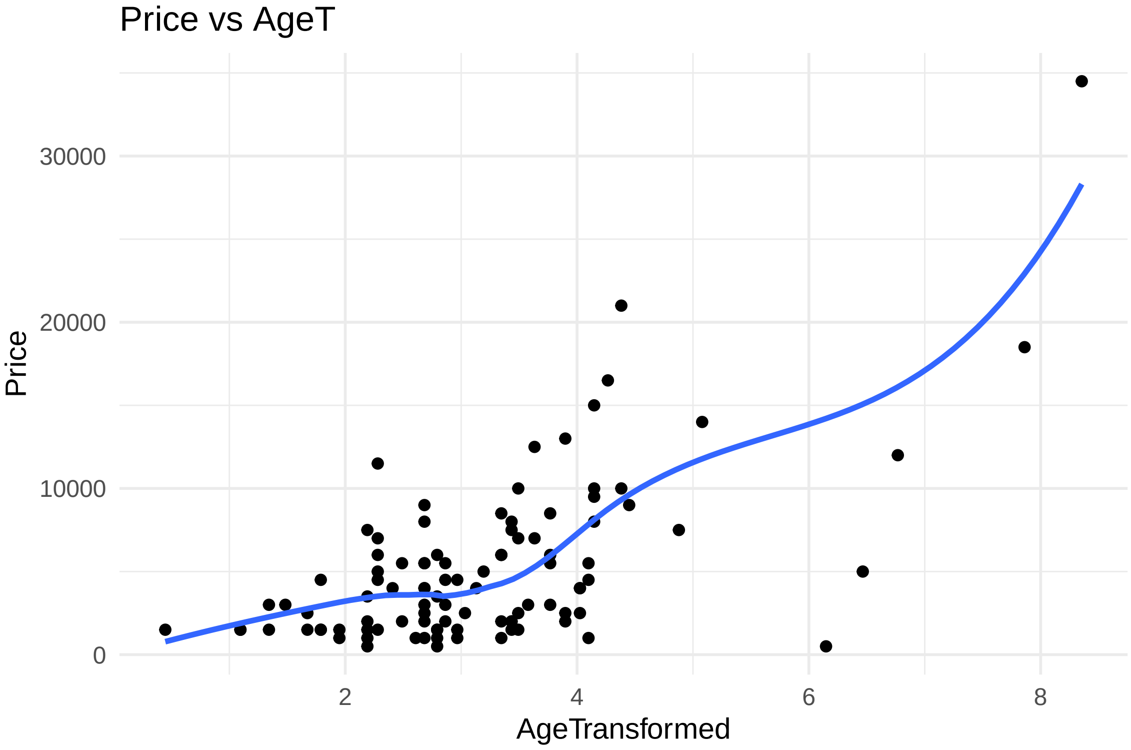
We can also see that the relationship between Odometer and Price is non-linear. This is evidenced by a sharp decrease in price through lower odometer values before a more gradual price decline at higher odometer values.

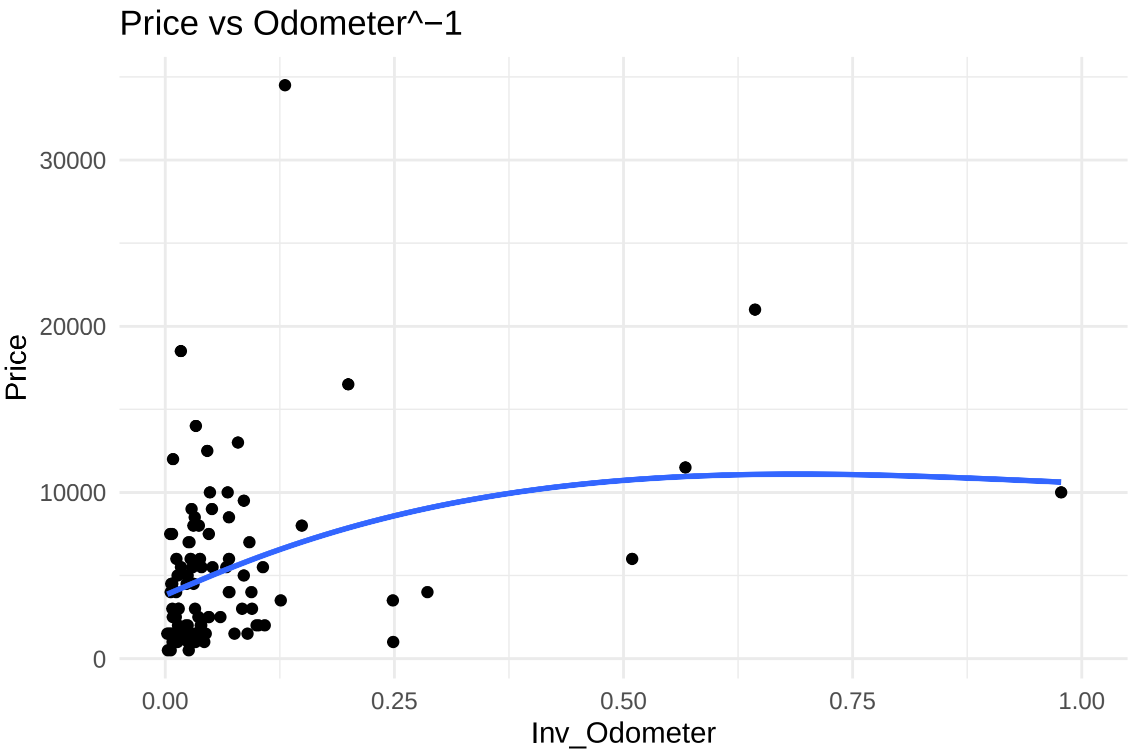
In order to improve the model, we must transform the variables within Model 1 to have a more linear relationship with Price.

Model 2

$$Price_i \sim \beta_0 \ + \beta_1 \sqrt{|(Age_i - A\bar{g}e)|} + \beta_2 (1/(Odometer_i))$$

In the second model, we include a transformed Age variable, as well as inverting the Odometer variable. The transformation on the Age variable first re-centers Age around 0 and takes the absolute value of this. We can then square root these new values, in an attempt to make the relationship more linear. Throughout the rest of the report, I will refer to this variable as AgeT for brevity.





We can see from the plots of both AgeT and Odometer^{-1} vs Price that the relationships between the two explanatory variables are now much more linear than before.

There are still some concerns with the relationship between the Odometer^{-1} variable and the response variable, but we proceed with assessing the model fit.

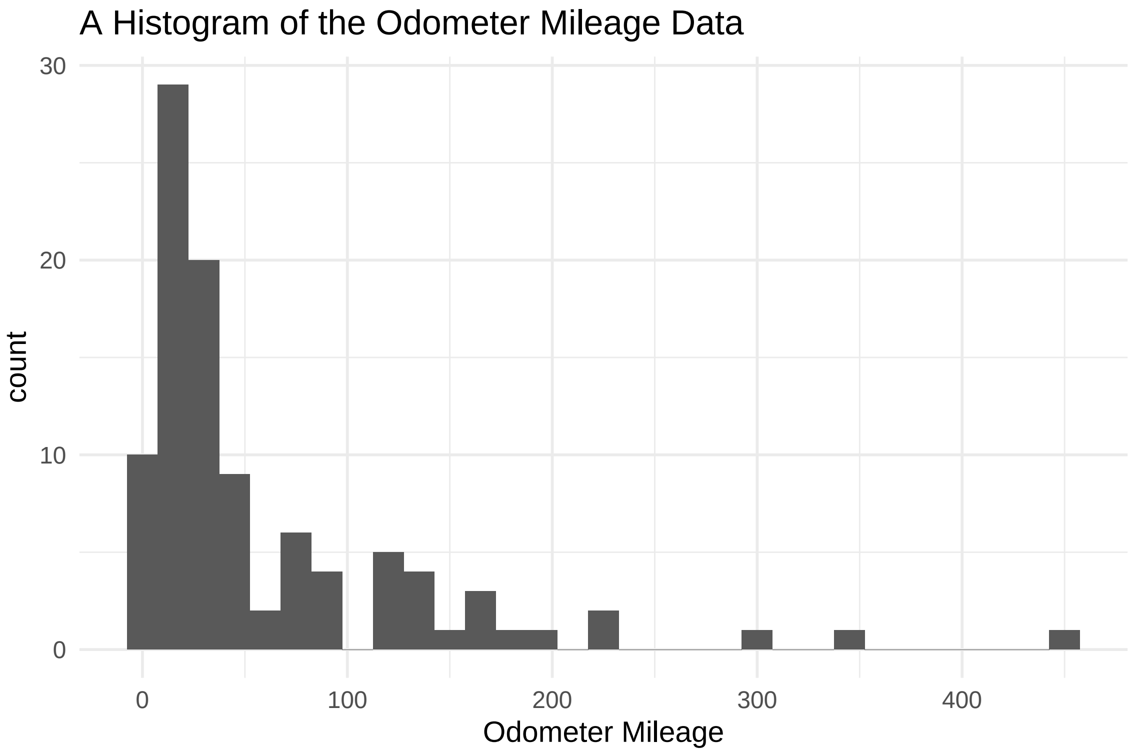
```
##
## Call:
## lm(formula = Price ~ AgeTransformed + Inv_Odometer, data = jaybob)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11616.3  -2608.6   -162.9   1966.7  16803.2
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2718.1     1039.9  -2.614  0.01038 *
## AgeTransformed   2410.8      304.5   7.916 4.04e-12 ***
## Inv_Odometer     2093.0      764.3   2.738  0.00735 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3851 on 97 degrees of freedom
## Multiple R-squared:  0.4384, Adjusted R-squared:  0.4268
## F-statistic: 37.85 on 2 and 97 DF,  p-value: 7.057e-13
```

As we can see, the p-values indicate that all the variables are significant again, but this time our Adjusted R^2 value is much higher: 0.4268, indicating a much better fit for Model 2 when compared to Model 1.

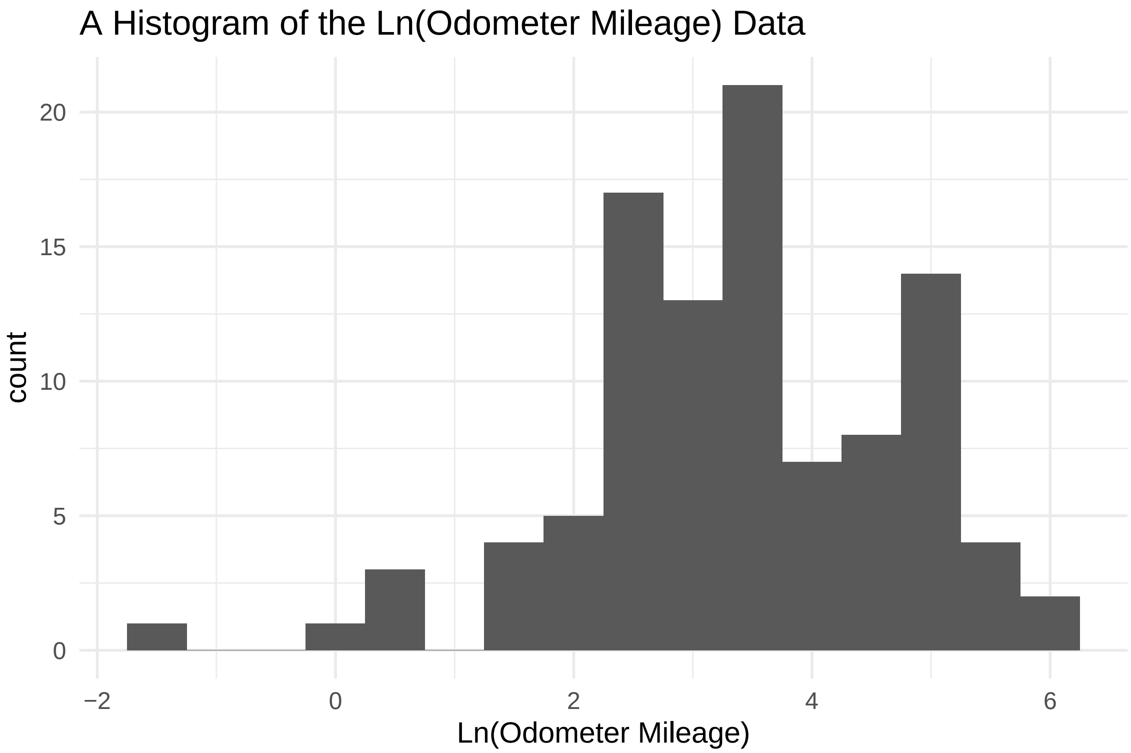
We may be able to improve on the model fit though by doing a different transformation to the Odometer variable: taking its logarithm.

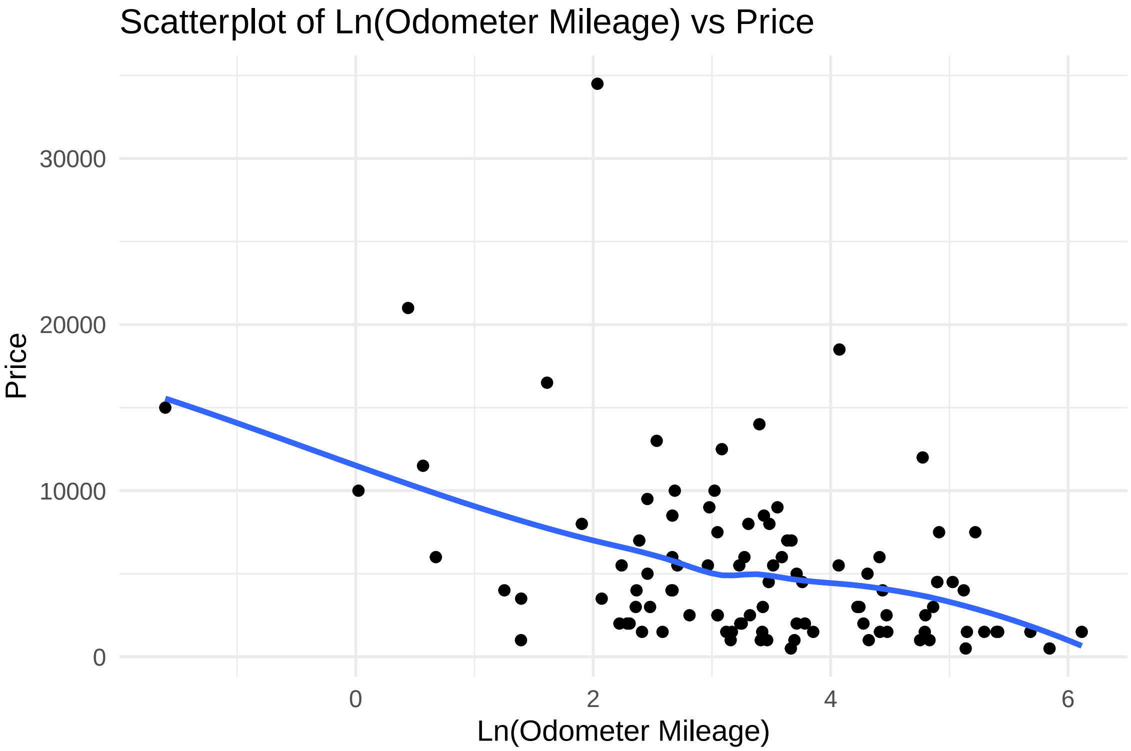
Model 3

$$Price_i \sim \beta_0 \ + \beta_1 AgeT_i + \beta_2 Ln(Odometer_i)$$



We can see from the above plot that the distribution of Odometer mileage is very heavily right-skewed, possibly compromising the assumption of linearity placed upon the variables within the model. In order to negate this skew, a good idea may be to transform the Odometer variable by taking the logarithm of it.





From the histogram of log(Odometer Mileage), we can see clearly that the data is now much more Normally distributed, and from the scatter plot we can see that the relationship between Ln(Odometer) and Price is much more linear in nature. Thus, we proceed with assessing the model fit.

```
##
## Call:
```

```
## lm(formula = Price ~ AgeTransformed + Log_Odometer, data = jaybob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9505.1 -2482.8    11.1  2100.6 16017.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1740.7     1527.9   1.139  0.25738
## AgeTransformed    2275.4      297.9   7.639 1.56e-11 ***
## Log_Odometer     -1114.4      286.0  -3.896  0.00018 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3717 on 97 degrees of freedom
## Multiple R-squared:  0.4768, Adjusted R-squared:  0.466
## F-statistic:  44.2 on 2 and 97 DF,  p-value: 2.264e-14
```

We can see again that this transformation improves the Adjusted R^2 value of the model to 0.466, showing that the log transformation improves the fit as expected. We now move to checking how the other variables, "Pink Slip" and "Sold?" affect the fit of the model.

Model 4

$$Price_i \sim \beta_0 \ + \beta_1 AgeT_i + \beta_2 Ln(Odometer_i) \ + \ \beta_3 (PinkSlip)_i$$

Within Model 4, we include the categorical Pink Slip variable, and see how this affects the fit of the model.

```
##
## Call:
## lm(formula = Price ~ AgeTransformed + Log_Odometer + `Pink slip`,
##     data = jaybob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8603.2 -2721.0  -150.4  2110.1 15705.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         512.6     1722.7   0.298 0.766689
## AgeTransformed     2288.9      296.1   7.731 1.05e-11 ***
## Log_Odometer      -1067.0      285.9  -3.732 0.000322 ***
## `Pink slip`TRUE   1331.0      882.9   1.507 0.134979
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3693 on 96 degrees of freedom
## Multiple R-squared:  0.4889, Adjusted R-squared:  0.4729
## F-statistic: 30.61 on 3 and 96 DF,  p-value: 5.669e-14
```

It would seem from the R^2 score of 0.4729 that the fit of the model is improved by the inclusion of the Pink Slip variable. However, we see from the variable's p-value of 0.134979 that it is insignificant, so we do not include it within our model, as it may be causing overfitting.

Next, we assess whether the "Sold?" variable has any meaningful effect on the fit of the model.

Model 5

$$Price_i \sim \beta_0 \ + \beta_1 AgeT_i + \beta_2 Ln(Odometer_i) \ + \ \beta_3 (Sold)_i$$

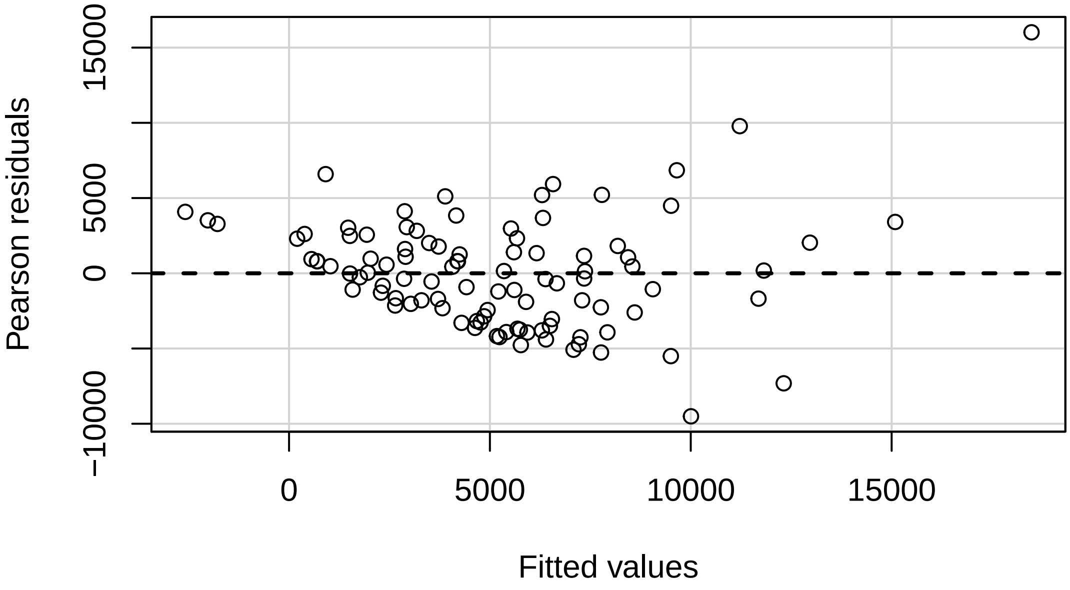
```
##
## Call:
## lm(formula = Price ~ AgeTransformed + Log_Odometer + `Sold?`,
##     data = jaybob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9144.8 -2389.5  -120.4  2146.4 16012.1
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2245.4     1667.6   1.346 0.181316
## AgeTransformed    2206.1      312.0   7.071 2.48e-10 ***
## Log_Odometer     -1075.3      291.2  -3.693 0.000369 ***
## `Sold?`TRUE       -640.1      837.9  -0.764 0.446767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3725 on 96 degrees of freedom
## Multiple R-squared:   0.48,  Adjusted R-squared:  0.4637
## F-statistic: 29.54 on 3 and 96 DF,  p-value: 1.291e-13
```

Again within this model, we see the "Sold?" variable is insignificant, with an accompanying p-value of 0.446767. This means we can discard this variable, as even though it has a relatively strong R^2 value of 0.4637, it may again lead to overfitting.

This means we select Model 3 as our most accurate model of Price from the "jaybob" data set. We then move to assessing the model assumptions are satisfied for this model.

Analysis of Residuals

To thoroughly assess the goodness of fit for any model we must analyse the residuals. The following plot shows the fitted values against the residuals. The residuals should be normally distributed around 0 and have consistant variance across the range of fitted values.



We see from this residual plot that Model 3 seems to have some issue with the distribution of residuals. There appears to be differences in variance of the resdiuals as the fitted values inicrease, hinting at heteroscedasticity. As well as this the residuals do not appear to be normally distributed around 0. This would indicate that assumptions for a well fitted model do not hold.

Confirming heteroscedasticity.

To further check the potential heteroscedasticity observed in the residuals plot, we can employ both the Breusch-Pagan and Goldfeldt-Quant tests.

First, we look at the Goldfeldt-Quant test, with the null hypothesis that the model has homoscedasticity.

```
##
##  Goldfeld-Quandt test
##
## data:  model3
## GQ = 0.49985, df1 = 37, df2 = 37, p-value = 0.981
## alternative hypothesis: variance increases from segment 1 to 2
```

From this test we obtain a p-value of 0.981 which is greater than the required 0.05 to reject H_0. Thus, from the Goldfeldt-Quant test, it would appear that we have homoscedasticity.

We proceed to applying the Breusch-Pagan test, also with a null hypothesis that homoscedasticity is present within our model.

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 27.045, df = 2, p-value = 1.34e-06
```

Here, we obtain a p-value of 0.000000136, which is much less than than 0.05, indicating that we reject H_0. That is, there is heteroscedasticity present within our model.

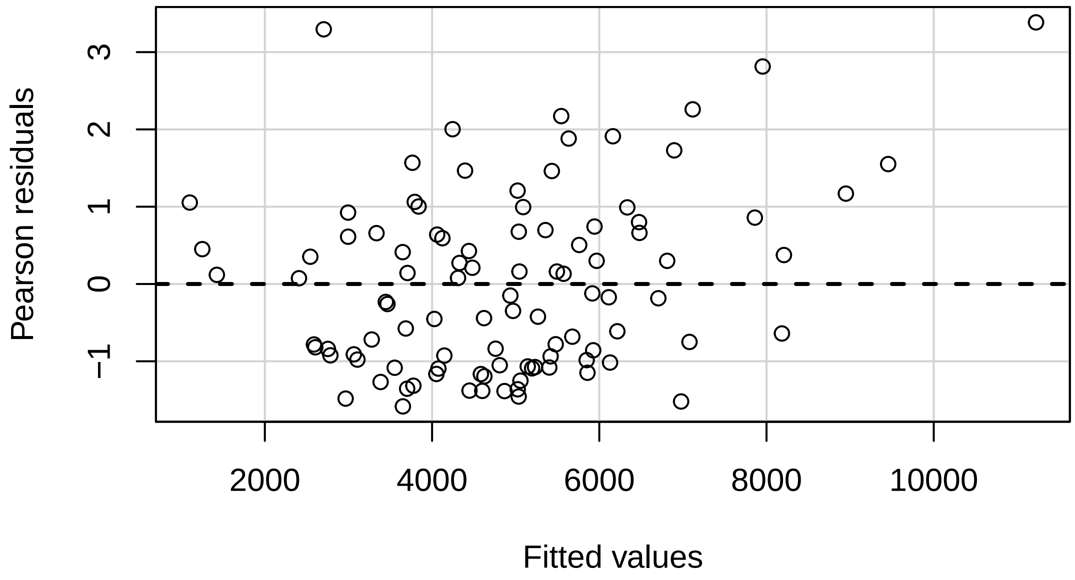
With the residuals plot and the Breusch-Pagan indicating heteroscedasticity it is worth addressing this. To rectify this heteroscedasticity, we can apply weighted regression to our model.

We apply our weighted regression, and then assess the model fit.

```
##
```

```
## Call:
## lm(formula = Price ~ AgeTransformed + Log_Odometer, data = jaybob,
##     weights = wt)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5836 -0.9799 -0.1610  0.6816  3.3846
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3544.3     1154.4   3.070  0.00277 **
## AgeTransformed   1067.6      204.8   5.212 1.05e-06 ***
## Log_Odometer     -609.2      204.0  -2.986  0.00358 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.156 on 97 degrees of freedom
## Multiple R-squared:  0.4006, Adjusted R-squared:  0.3883
## F-statistic: 32.42 on 2 and 97 DF,  p-value: 1.653e-11
```

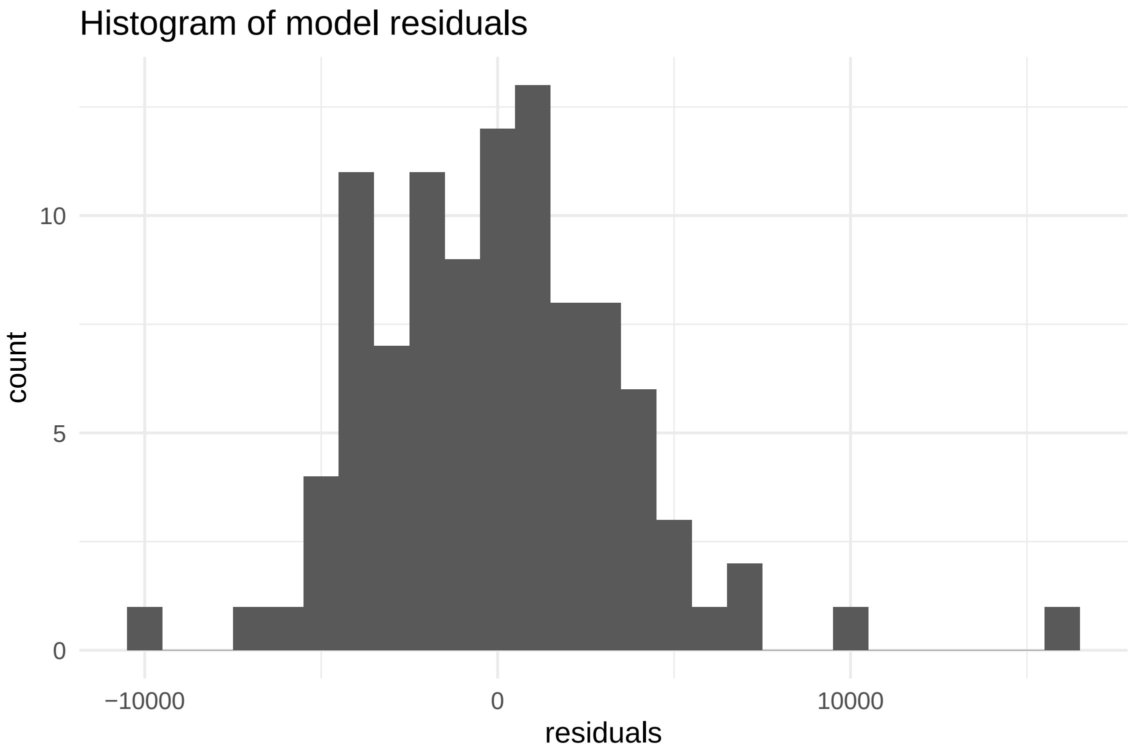
We see from the weighted least squares regression that our Adjusted R^2 value has decreased to 0.3883, as opposed to the previous 0.466, indicating that our use of a weighted model has not improved the fit. However, this is balanced by the fact that our model no longer exhibits heteroscedasticity.



Assessing the Normality of Error Terms

Assessing the Normality of the error terms in our model is of low priority as we have a large n for our data set. We test anyway though, for completion.

We do this first by simply by plotting the residuals of the model.



From our plot we can see that the residuals appear to be Normally Distributed, which is encouraging.

As a final check, we apply the Kolmogrov-Smirnov test, comparing the residuals to a Normal distribution and using a Null Hypothesis that the two samples are drawn from the same distribution.

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  rnorm(1000) and res3
## D = 0.5, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Applying this test we acquire a p-value of close to 0. This indicates that we reject H_0, and that our error terms are not Normally distributed.

To rectify this lack of Normality, we can perform a log transformation to the dependent variable Price in our model, to produce the model below:

$$Ln(Price_i) \sim \beta_0 \ + \beta_1 AgeT_i + \beta_2 Ln(Odometer_i)$$

We fit this model and observe the results.

```
##
## Call:
## lm(formula = Log_Price ~ AgeTransformed + Log_Odometer, data = jaybob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4203 -0.5361  0.1085  0.5564  1.4482
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.03964    0.29871  26.914  < 2e-16 ***
## AgeTransformed  0.28856    0.05824   4.955 3.07e-06 ***
## Log_Odometer   -0.22947    0.05592  -4.103 8.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7267 on 97 degrees of freedom
## Multiple R-squared:  0.3455, Adjusted R-squared:  0.332
## F-statistic:  25.6 on 2 and 97 DF,  p-value: 1.18e-09
```

Here, there is not a drastic change in R^2 value for the model, so we proceed to checking the Normality assumption again with the Kolmogrov-Smirnov test.

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  rnorm(1000) and res6
## D = 0.135, p-value = 0.07276
## alternative hypothesis: two-sided
```

Here we can see that the p-value is 0.06927. Thus, H_0 holds for our logarithmic model.

Model Interpretation.

```
##
## Call:
## lm(formula = Log_Price ~ AgeTransformed + Log_Odometer, data = jaybob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4203 -0.5361  0.1085  0.5564  1.4482
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.03964    0.29871  26.914  < 2e-16 ***
## AgeTransformed  0.28856    0.05824   4.955 3.07e-06 ***
## Log_Odometer   -0.22947    0.05592  -4.103 8.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7267 on 97 degrees of freedom
## Multiple R-squared:  0.3455, Adjusted R-squared:  0.332
## F-statistic:  25.6 on 2 and 97 DF,  p-value: 1.18e-09
```

From the coefficients of Model 6, we can draw:

For every year increase in AgeT, the value of Ln(Price) increases by 0.28856, meaning that the price increases by around $1334, holding the other variables constant.

As the Odometer mileage increases by 1%, we see that the price decreases by 0.22947%, holding the other variables constant.