Unit 5: The General Linear Model

3. Transformations

(Chapter 1.6 and 6.4)

4/19/2021

Recap from last time

1. For many real-world problems, a lot of variables contribute a little bit
2. Stepwise approaches try to correct for this, but there is no “one true way”
3. We can check assumptions for multiple regression using plots

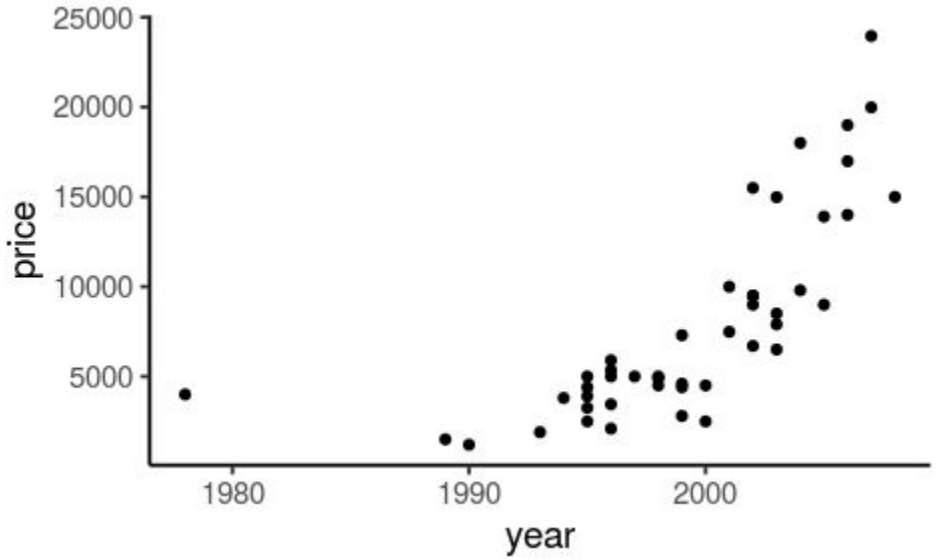
Key ideas

1. When your diagnostics show problems with constant variance, transformations can help to normalize data
2. The log transformation is a common solution to right-skewed data
3. Transformed models are often hard to interpret in linear units, reversing the transformation can help

Truck prices

The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks.

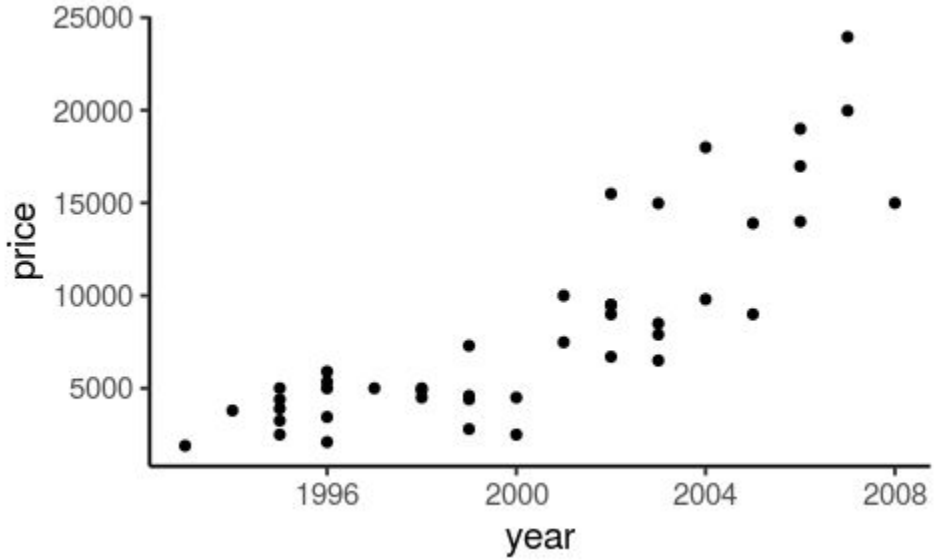
What is the relationship between these two variables?



Remove unusual observations

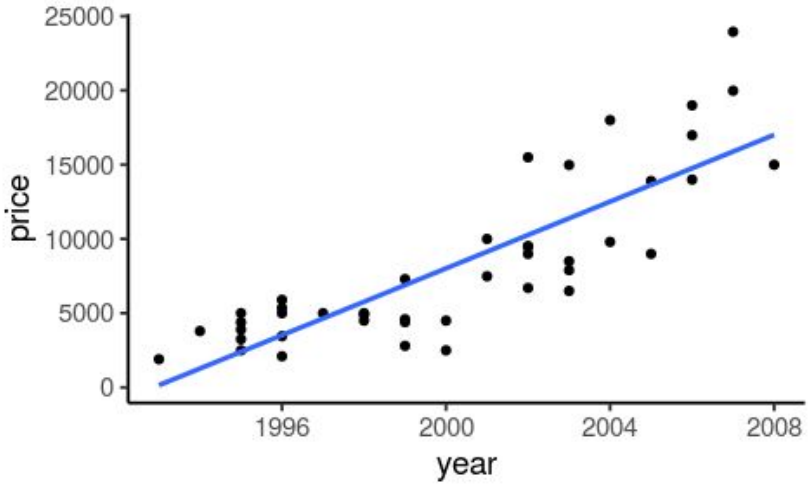
Let's the older trucks and only focus on trucks made in 1992 or later.

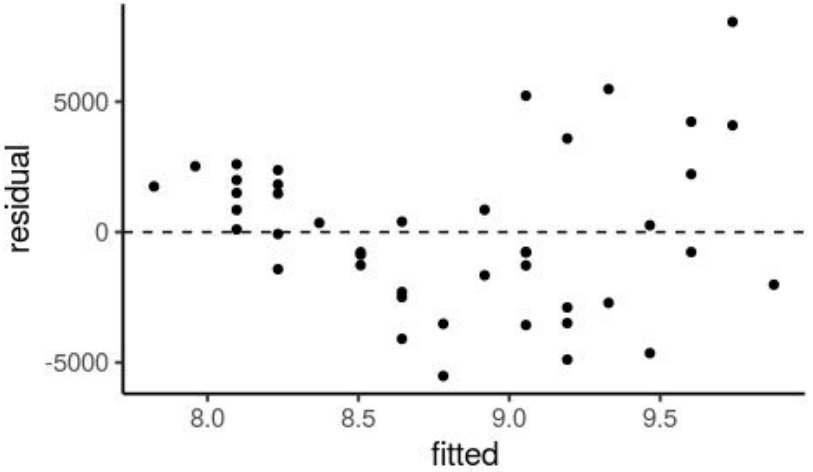
Now what can you say about the relationship?



A linear model for truck prices

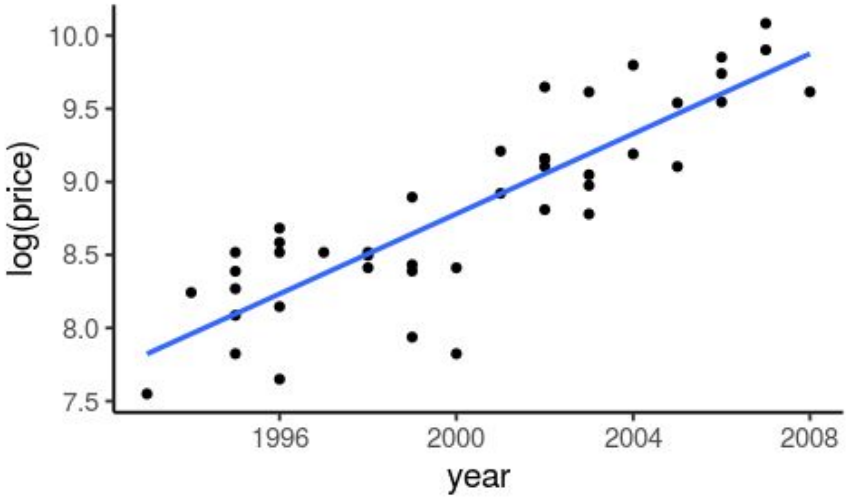
Model: $\widehat{price} = b_0 + b_1 year$

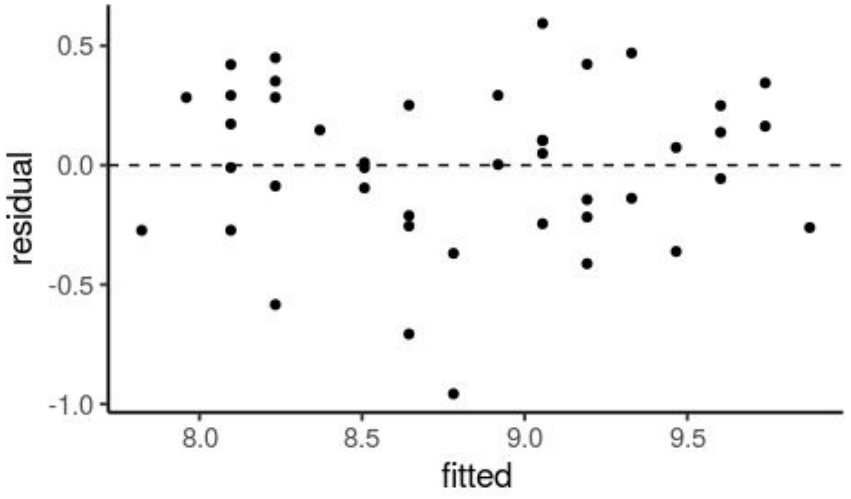




A Log-transform of price

Model: $\widehat{log(price)} = b_0 + b_1 year$





Interpreting models with log transforms

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-265.073	25.042	-10.585	0.000
year	0.137	0.013	10.937	0.000

Model: $\widehat{log(price)} = -265.073 + .137 \cdot year$

For each additional year the car is newer (for each year decrease in car's age) we would expect the log price of the car to increase on average by 0.137 log dollars.

Why is that not very useful?

Working with logs

- Subtraction and logs: $log(a) - log(b) = log\left(\frac{a}{b}\right)$
- Natural log: $e^{log(x)} = x$

We can these identities to “undo” the log transformation

Re-interpreting models with log transformations

The slope for the log transformed model is 0.137. So the log price difference between cars one year apart is predicted to be 0.14 log dollars.

$$\begin{aligned}
\log(\text{price at year } x+1) - \log(\text{price at year } x) &= 0.137 \\
log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right) &= 0.137 \\
e^{log\left(\frac{\text{price at year } x+1}{\text{price at year } x}\right)} &= e^{0.137} \\
\frac{\text{price at year } x+1}{\text{price at year } x} &= 1.15
\end{aligned}$$

For each additional year the car is newer (for each year decrease in car's age) we would expect the price of the car to increase on average by a **factor of 1.15**.

Dealing with non-constant variance

- Non-constant variance is one of the most common model violations, but it is usually fixable by transforming the response (y) variable
- The most common variance stabilizing transform is the log transformation: log(y), especially useful when the response variable is (extremely) right skewed.
- When using a log transformation on the response variable the interpretation of the slope changes:
  - For each unit increase in x, y is expected on average to decrease/increase by a factor of e^{b1}
- Other transformations can sometimes fix other common issues:
  - exp(y) can help when the dependent variable is left skewed
  - sqrt(x) is often used with count data that are non-normal

Key ideas

1. When your diagnostics show problems with constant variance, transformations can help to normalize data
2. The log transformation is a common solution to right-skewed data
3. Transformed models are often hard to interpret in linear units, reversing the transformation can help