Unit 7: Multiple linear regression

3. Transformations & case study

Sta 104 - Summer 2018, Term 1

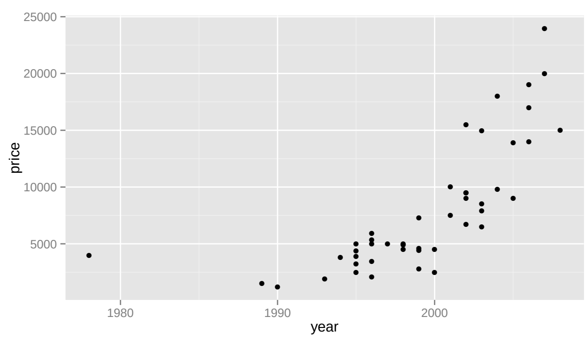
Duke University, Department of Statistical Science

Prof. White

Slides posted at *https://www2.stat.duke.edu/courses/Summer18/sta104.001-1/*

- Lab 8 is due today 12:45 pm.
- Problem Set 7 is due Monday 11:55 pm (This assignment is extraordinarily short)
- Performance Assessment 7 is due Monday 11:55 pm.
- Lab 9 is due Monday 11:55 pm
- Projects are due Monday 11:55 pm
- Project Presentations are Monday in lab
- Final Exam is Wednesday June 27 (2 - 5 pm)

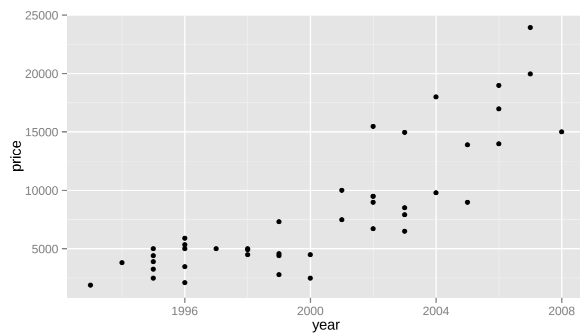
The scatterplot below shows the relationship between year and price of a random sample of 43 pickup trucks. Describe the relationship between these two variables.



From: *http://faculty.chicagobooth.edu/robert.gramacy/teaching.html*

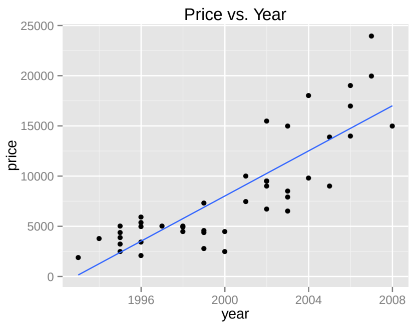
Let's remove trucks older than 20 years, and only focus on trucks made in 1992 or later.

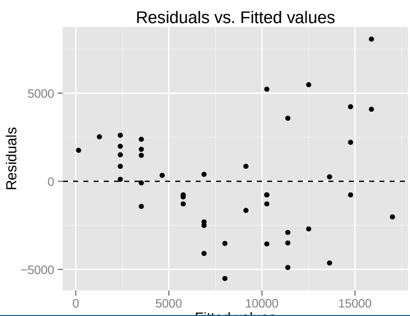
Now what can you say about the relationship?



Truck prices - linear model?

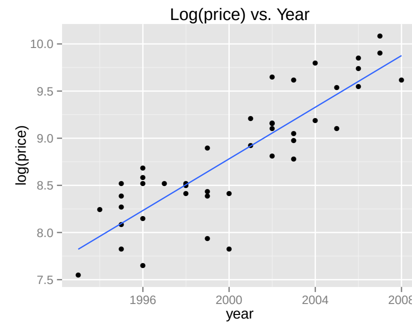
Model: $\widehat{price} = b_0 + b_1\ year$

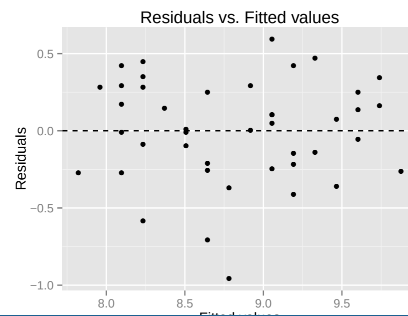




Truck prices - log transform of the response variable

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





Interpreting models with log transformation

	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: $\widehat{log(price)} = -265.073 + 0.137\ year$

- 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.
- which is not very useful...

Working with logs

- Subtraction and logs: $log(a) - log(b) = log(\frac{a}{b})$
- Natural logarithm: $e^{log(x)} = x$
- We can these identities to “undo” the log transformation

Interpreting models with log transformation (cont.)

The slope coefficient for the log transformed model is 0.137, meaning the log price difference between cars that are one year apart is predicted to be 0.14 log dollars.

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

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*.

Recap: dealing with non-constant variance

- Non-constant variance is one of the most common model violations, however 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^{b_1}.
- Another useful transformation is the square root: $\sqrt{y}$, especially useful when the response variable is counts.
- These transformations may also be useful when the relationship is non-linear, but in those cases a polynomial regression may also be needed.

Data from the ACS

1. income: Yearly income (wages and salaries)
2. employment: Employment status, not in labor force, unemployed, or employed
3. hrs_work: Weekly hours worked
4. race: Race, White, Black, Asian, or other
5. age: Age
6. gender: gender, male or female
7. citizens: Whether respondent is a US citizen or not
8. time_to_work: Travel time to work
9. lang: Language spoken at home, English or other
10. married: Whether respondent is married or not
11. edu: Education level, hs or lower, college, or grad
12. disability: Whether respondent is disabled or not
13. birth_qrtr: Quarter in which respondent is born, jan thru mar, apr thru jun, jul thru sep, or oct thru dec

Load and subset data

```
acs_emp <- acs %>%
  filter(employment == "employed", income > 0)
```

Aside: categorical (factor) variables in R

```
acs_emp %>%
  select(employment) %>%
  table()
```

```
not in labor force        unemployed          employed
                 0                 0               787
```

```
acs_emp <- droplevels(acs_emp) # overwrite acs_emp

acs_emp %>%
  select(employment) %>%
  table()
```

```
employed
     787
```


Full model

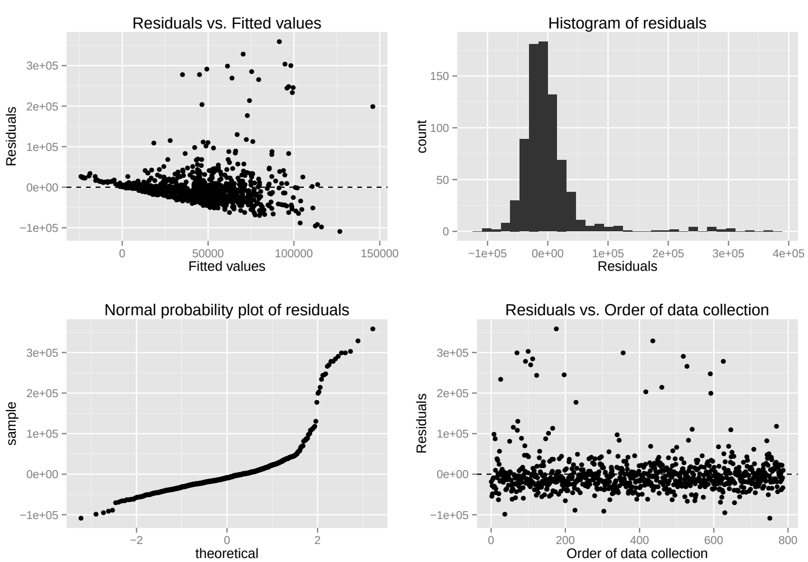
Suppose we only want to consider the following explanatory variables: hrs_work, race, age, gender, citizen.

```
m_full = lm(income ~ hrs_work + race + age + gender
    + citizen, data = acs_emp)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17215.60	11399.81	-1.51	0.13
hrs_work	1251.31	153.14	8.17	0.00
raceblack	-13202.39	6373.05	-2.07	0.04
raceasian	32699.34	8903.66	3.67	0.00
raceother	-12032.88	7556.78	-1.59	0.11
age	760.99	129.71	5.87	0.00
genderfemale	-17246.91	3887.17	-4.44	0.00
citizenyes	-9537.20	8360.85	-1.14	0.25

Diagnostics

What do you think?

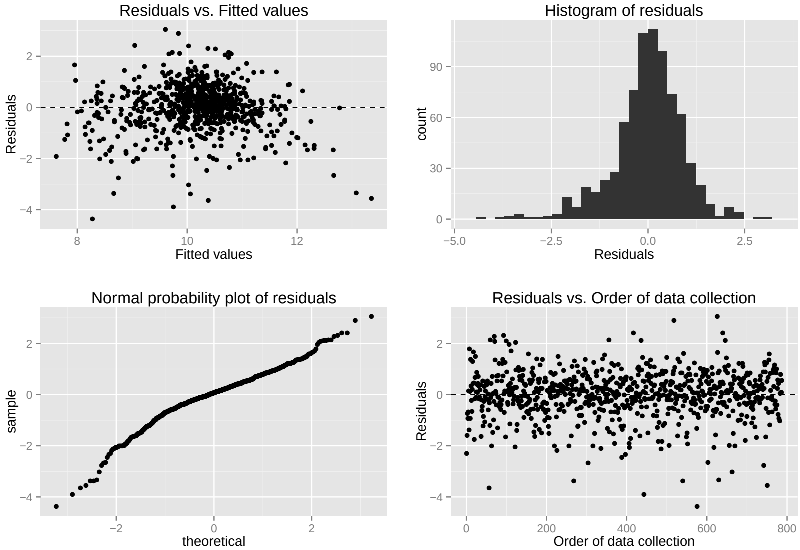


Diagnostics -- code

```
# residuals vs. fitted
qplot(data = m_full, y = .resid, x = .fitted, geom = "point") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals") +
  ggtitle("Residuals vs. Fitted values")

# histogram of residuals
qplot(data = m_full, x = .resid, geom = "histogram") +
  xlab("Residuals") +
  ggtitle("Histogram of residuals")

# normal prob plot of residuals
qplot(data = m_full, sample = .resid, stat = "qq") +
  ggtitle("Normal probability plot of residuals")

# order of residuals
qplot(data = m_full, y = .resid) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  ylab("Residuals") +
  xlab("Order of data collection") +
  ggtitle("Residuals vs. Order of data collection")
```

Log transformation

$$m\_full\_log = lm(log(income) \sim hrs\_work + race + age + gender + citizen, data = acs\_emp)$$



Application exercise: 7.4 Interpreting models with a transformed response

See course website for more details