STA 371G TA Review session

Han, Jared & Sai

Houses Dataset

Open the houses dataset at

```
houses_extended =
read.csv('https://raw.githubusercontent.com/brianlukoff/sta371g/master/data/HousesExtended.csv')
```

Multiple linear regression- Interpreting coefficients and p-values

In multiple regression, the coefficient of a predictor assesses the effect on Y of a unit increase in that particular predictor when all other predictors remain constant.

Run the dataset: Houses_Extended = read.csv('https://raw.githubusercontent.com/brianlukoff/sta371g/master/data/HousesExtended.csv')

Here's an extended houses dataset, where Sales price of houses is predicted using appraisal value of the house (thousands of dollars), Square footage (hundreds of sq.ft.) and the number of bedrooms.

Statistical significance from the summary() statement

- Statistical significance of a predictor variable relates to the p-values of the individual predictor variables.
    - Null hypothesis H0: The coefficient of the predictor variable is **not** significantly different from zero, i.e. the predictor variable has no additional significance in predicting Y given the other predictors.
    - Alternative hypothesis H1: The coefficient **is** significantly different than 0, that is, predictor variable has significance in predicting Y in addition to the other predictors

Practical significance from the summary() statement

Practical significance of a predictor variable refers to

- R^2 which shows the practical significance of the whole model
- The value of the coefficient of the predictor variable (which indicates the amount by which the Y variable changes when the X variable changes by 1 unit, and all other variables held constant)

Tasks

1. Run the full model (without interactions) and determine which predictors are statistically significant and update the reduced model.
2. R-Squared value has marginally decreased, but the reduced model is still a better model. Why?
3. How do you interpret the intercept of the reduced model?
4. How do you interpret the slopes of the predictor variables in the reduced model?

Confidence and Prediction intervals

We consider two types of **confidence intervals** in regression:

1. CI for a predictor's coefficient, using confint()
2. CI for the population mean of a forecasted value, using predict.lm()

A **prediction interval** tells you about the distribution of reasonable values for a predicted data point, not the uncertainty in determining the population mean of the forecast (that would be #2 above).

Note, with predict.lm(), prediction intervals are generally wider than confidence intervals.

Tasks

1. Find the confidence intervals of the predictors’ coefficients in the reduced model at 95%.
2. Estimate the Sales Price for the house at 123 Lotus Avenue, which is appraised at $150,000 and is 2500 sq.ft. and give a 90% prediction interval for the same house.
3. Consider that the house at 123 Lotus Avenue undergoes expansion, and 500 sq.ft. are being added to the house. Predict the **change** in Sales Price due to this expansion.

State Dataset

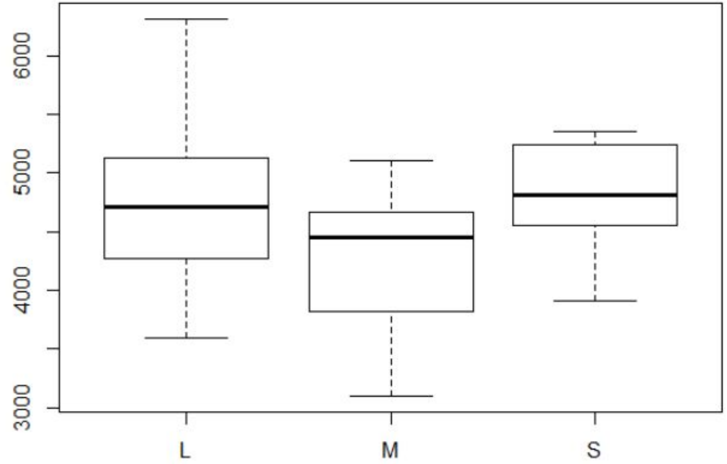
Open the State dataset

```
state <-
read.csv("https://raw.githubusercontent.com/brianlukoff/sta371g/master/data/state.csv")
```

Dummy variables

Take the value 0 or 1 to indicate the absence or presence of some categorical effects

Do boxplot of Income vs size



Dummy variables

Predict the Income using size.

Model = lm(Income~Size, data = state)

SizeL is the reference term

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4766.00     202.51  23.535   <2e-16 ***
SizeM        -508.27     225.73  -2.252   0.0291 *  
SizeS          29.22     278.32   0.105   0.9168    
```

How to interpret the coefficient of SizeS

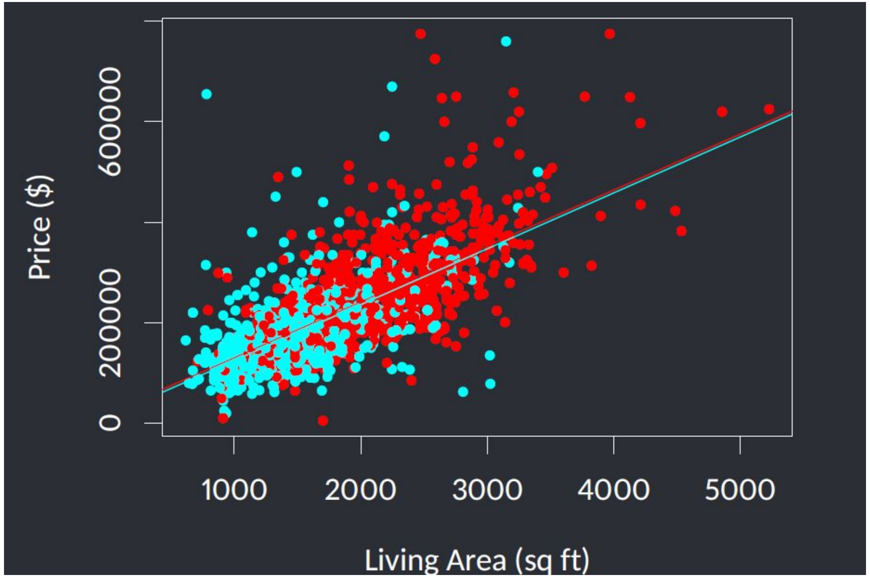
Interaction

$$\text{Price} = \beta_0 + \beta_1 \cdot \text{Living.Area} + \beta_2 \cdot \text{FireplaceYes} + \beta_3 \cdot \text{Living.Area} \cdot \text{FireplaceYes} + \epsilon_i.$$

The interaction term is the product of two variables.

The coefficient is called the interaction effect. The coefficient of individual variable is called main effect.

For a dummy variable Fireplace, the interaction term means we fit two lines in the same regression model.



Task

Use previous state dataset

Build a linear regression model for predicting Income using Illiteracy, Murder and their interaction term.

Build a linear regression model for predicting Income using Size (categorical! R will make dummy variables for you), Population and their interaction term.

What’s the effect of a unit increase of population for L size stage

Coefficients:

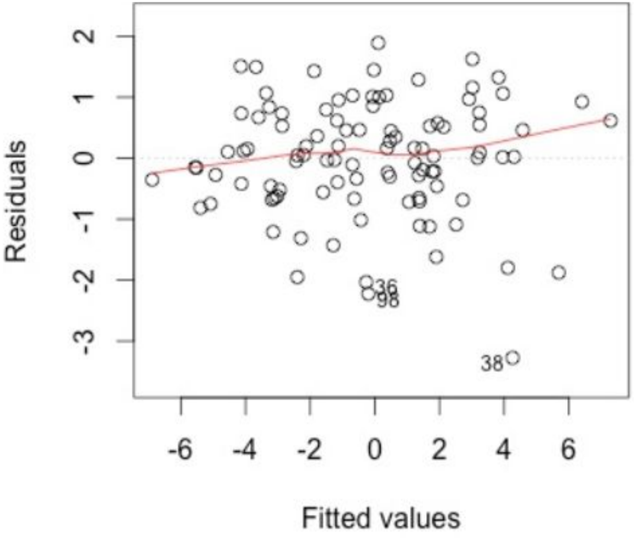
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.776e+03	2.398e+02	19.917	< 2e-16	***
SizeM	-7.625e+02	2.809e+02	-2.715	0.00944	**
SizeS	-2.783e+02	3.630e+02	-0.767	0.44741	
Population	-1.919e-03	2.740e-02	-0.070	0.94446	
SizeM:Population	5.668e-02	3.695e-02	1.534	0.13218	
SizeS:Population	1.134e-01	8.041e-02	1.410	0.16545	

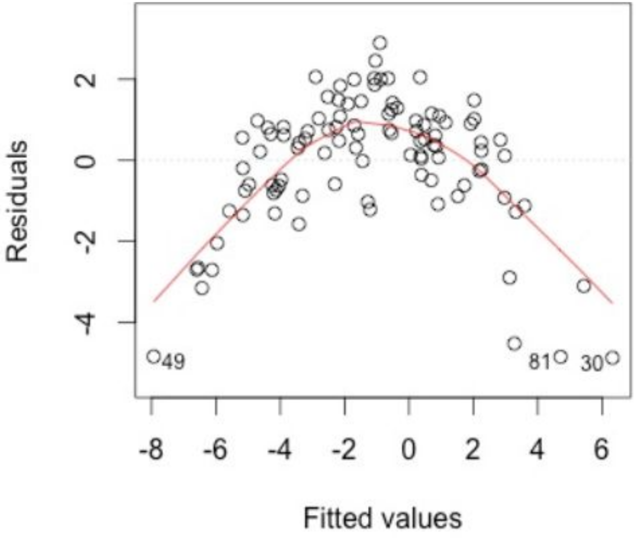
The assumptions for linear regression models

- Linearity
  - Linear relationships between Y and X's
- Independence
  - Data points are sampled independently (we don't have a great way to test this! Only by context of the dataset)
- Normality
  - Normality of residuals
- Equal Variance (homoscedasticity)
  - Variance of the residuals doesn't change with X/fitted values

The assumptions for linear regression models

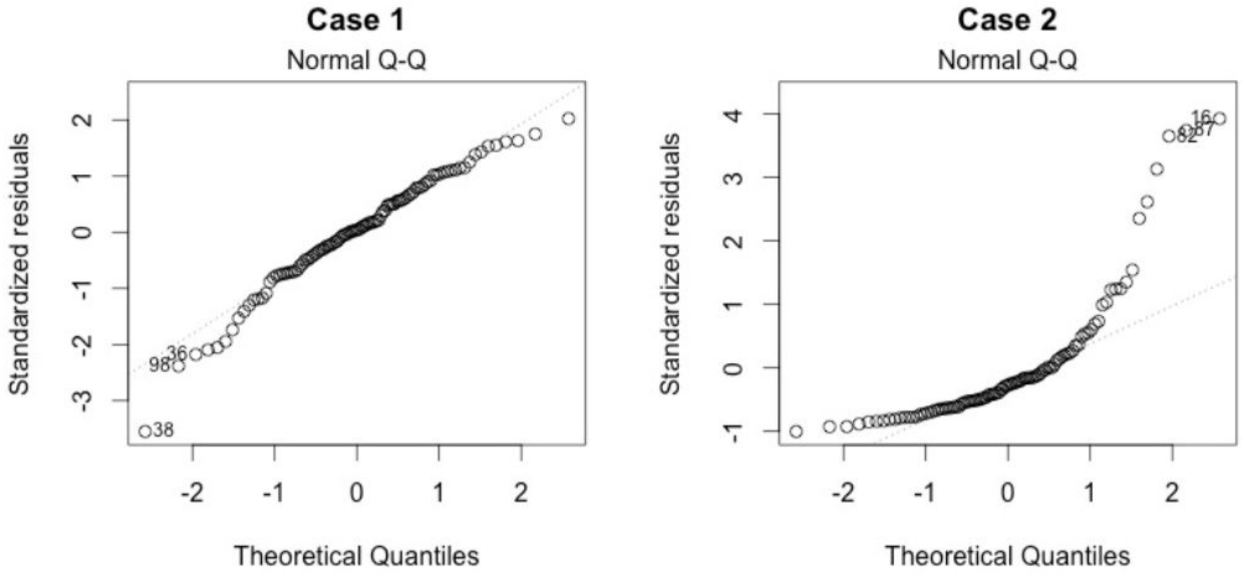
Linearity: There is no distinctive pattern in Case 1, but there is a non-linear relationship in Case 2.





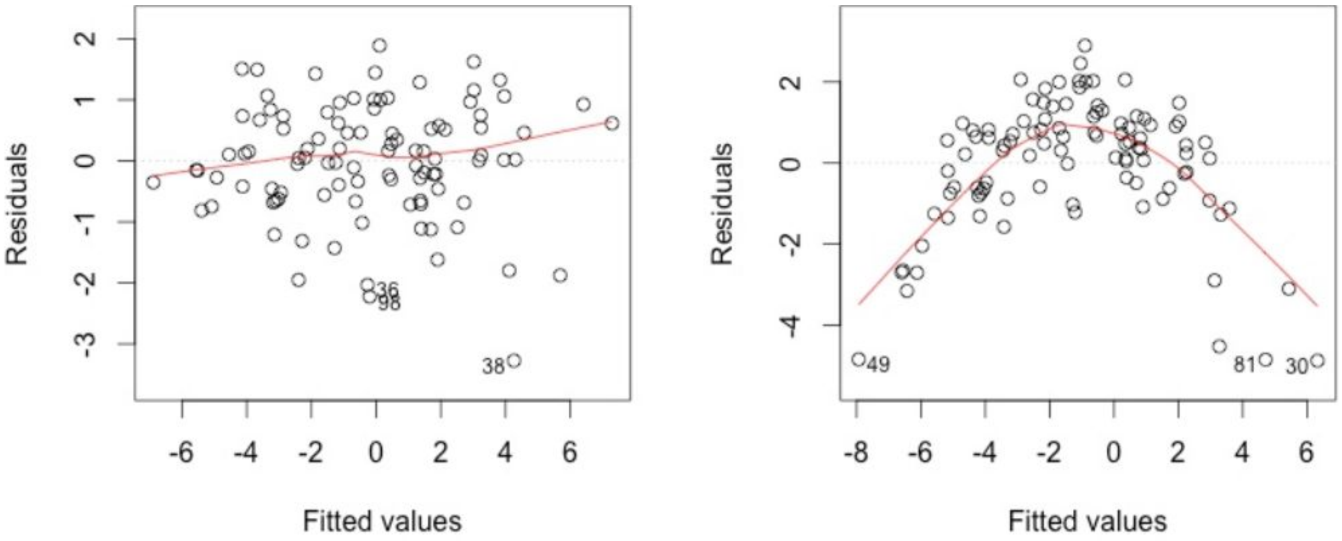
The assumptions for linear regression models

Normality: Case 1 shows the data is normally distributed. Case 2 has problems



The assumptions for linear regression models

Homoscedasticity: This plot shows if residuals are spread equally along the ranges of predictors



Cars Dataset

Open the cars dataset at

```
cars =
read.csv('https://faculty.mccombs.utexas.edu/carlos.carvalho/teaching/Cars.csv')
```

The assumptions for linear regression models

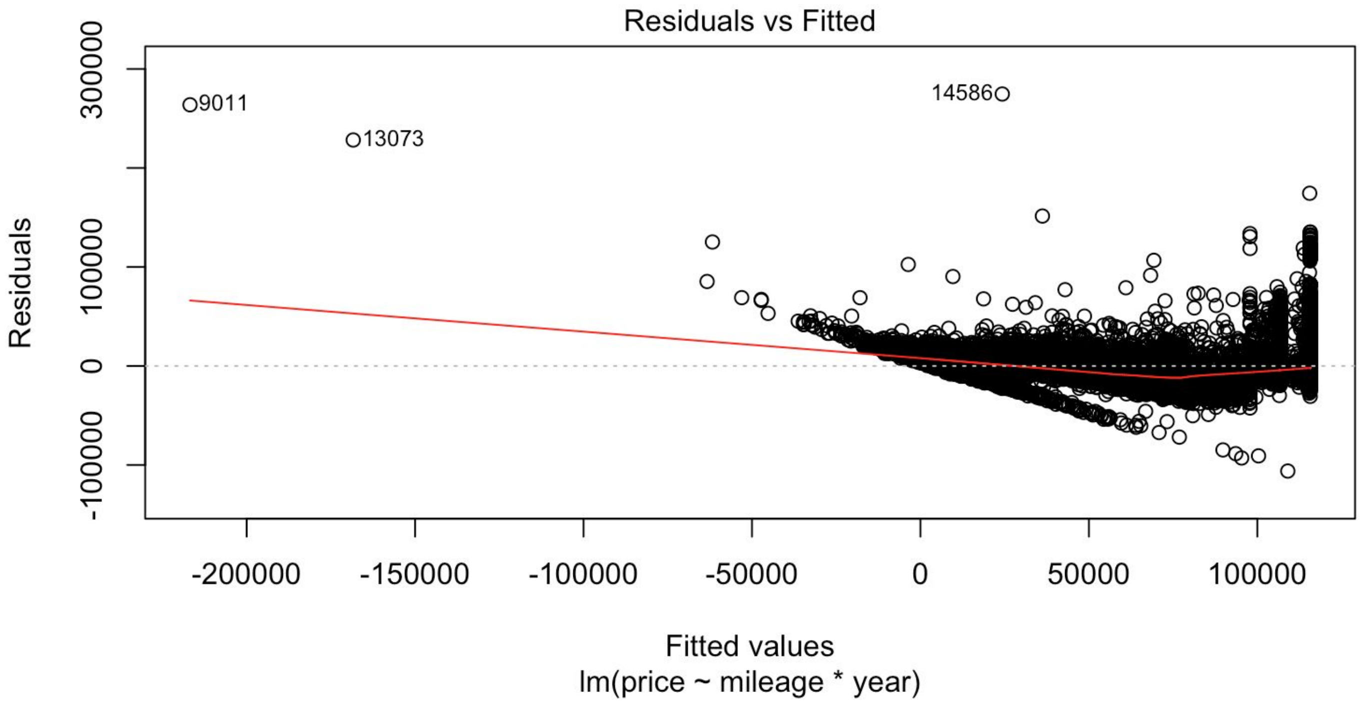
Brief intermission to R code, as Jared runs through lots of example plots of violations (or not!) of assumptions.

Tasks

```
cars =
read.csv('https://faculty.mccombs.utexas.edu/carlos.carvalho/teaching/Cars.csv')
```

Using the cars dataset, model price with the interaction of mileage and year.

Is there anything wrong in this residual plot?

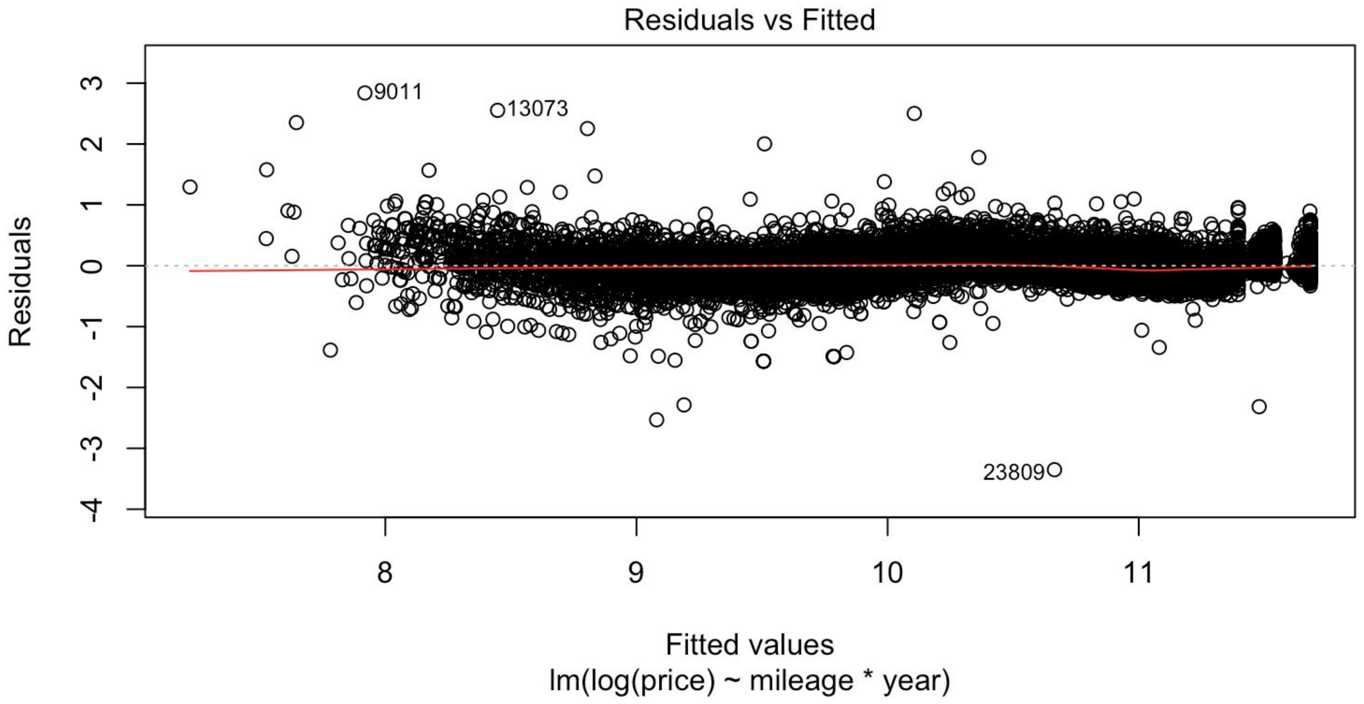


From Lecture 17

Transformation strategy

- If the model has two or three of the equal variance, normality and linearity issues, try transforming Y.
- Transforming the response often fixes nonlinearity in addition to fixing normality and equal variance issues.
- After transforming the response, if the nonlinearity is not fixed, try transforming the predictor(s) as well.
- There is no rule for which transformations will work in all cases; trial and error may be required.
- Remember, the interpretations of the coefficients will change after you transform one or more variables!

Does taking log(y) help?



Interpreting transformed models

Let’s say our model is

$$\log(\text{Income}) = 10 + 0.03*\text{GPA}$$

What is your change in Income for increasing your GPA from 3.0 to 4.0?

Hint: it’s not 0.03

Interpreting transformed models

log(Income) = 10 + 0.03*GPA

Is equivalent to:

Income = $e^{10+0.03GPA} = e^{10}\ e^{0.03GPA}$

So:

$e^{10}\ e^{0.03(4)} - e^{10}\ e^{0.03(3)} = e^{10}\ (e^{0.03(4)} - e^{0.03(3)})$

When in doubt:
exp(predict.lm(model,list(GPA=4))) - exp(predict.lm(model,list(GPA=3)))

***Be sure to exponentiate, meaning **exp()** (even I forget sometimes ;)

Interpreting transformed models

NOW, Let's say our model is

$$\log(\text{Income}) = 10 + 0.8*\log(\text{GPA})$$

What is your change in Income for increasing your GPA from 3.0 to 3.3?

Hint: it's not 0.8

Interpreting transformed models

$\log(\text{Income}) = 10 + 0.8*\log(\text{GPA})$

Is equivalent to:

$\text{Income} = e^{10+0.8 \log(\text{GPA})} = e^{10}\, e^{0.8 \log(\text{GPA})} = e^{10}\, \text{GPA}^{0.8}$

So:

$e^{10}\, (3.3)^{0.8} - e^{10}\, (3.0)^{0.8}$

When in doubt:
exp(predict.lm(model,list(GPA=3.3))) - exp(predict.lm(model,list(GPA=3)))

***assuming you put log(GPA) in the regression model, and didn't create a new variable called GPA_ln or similar

Model Selection

Why is there anything but R^2?

- Because worthless predictors never decrease R^2
- In other words, models are never penalized for having too many predictors
- Adjusted R^2, AIC and BIC are ways to fix this general problem, but each is a little different!
- Large values of R^2 and Adjusted R^2 are good
- Small values of AIC and BIC are good: -600 is better than -500.

Any Questions?