Multiple Linear Regression
Transformations and Qualitative Models

Laura Barnes & Julianne Quinn

University of Virginia
Charlottesville, VA

Agenda

1. Review of Multiple Regression
2. Transformation of Response and Predictors
3. Qualitative Models

Review of Multiple Regression

- Multiple Linear regression :

$$y = f(X) + \epsilon = X\beta + \epsilon$$

- Regression is the solution to an optimization problem. We find a linear fit to the data that minimizes the sum of square errors.
- Model assumptions
- Metrics of models: R^2, Adjusted-R^2, AIC, BIC
- Model comparison: Partial F test, test sets, and cross-validation

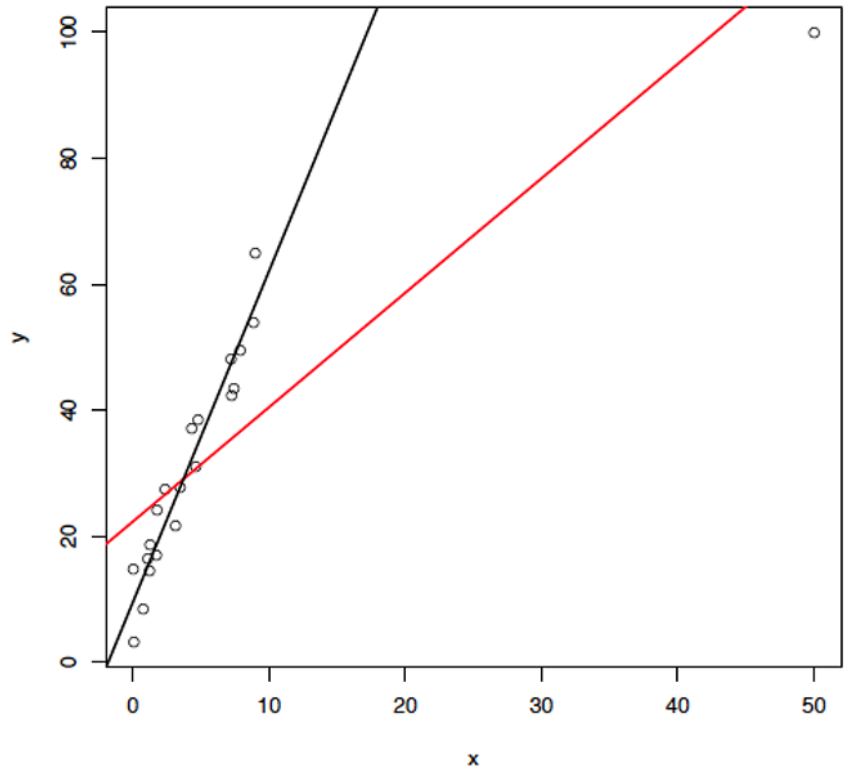
Model Diagnostics

- Graphical Diagnostics:
  - Residuals vs. fitted
  - Scale-Location plot of square root of absolute standardized residuals vs. fitted
  - QQ plot of standardized residuals
  - Residual-Leverage plot
- Analytical Diagnostics:
  - We can also diagnose problems in the regression by analyzing the results: coefficient values, performance tests.
  - Multicollinearity is apparent in the flipped signs for coefficients and for large changes in coefficient values with small changes to values in the observations.
  - Diagnosing Simpson's Paradox with coefficient values: coefficients show contradictory patterns; adding variables or replacing variables causes major changes to coefficient values.

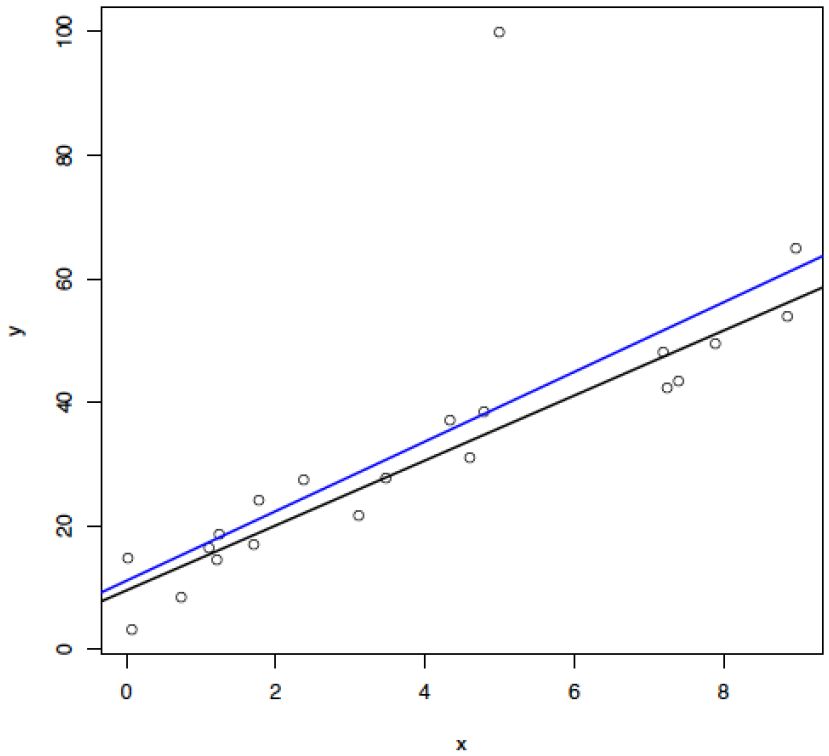
Influential Observations and Outliers

- Influential observations, outliers and high leverage points:
  - Influential points can change the values of parameter estimates and other results by their removal.
  - Outliers do not fit the model well.
  - Observations with high leverage are far away from the average of predictor values.
  - High leverage observations are potential influential points.

Example of Influential Point



Example of Outlier



Leverage and Cook's Distance

- Cook's distance attempts to identify influential observations using leverage.
- Cook's distance is calculated from the leverage and standardized residual of an observation:

$$C_i = r_i^2 \frac{h_i}{p(1-h_i)}$$

where r_i is the standardized residual, and p is the number of parameters including β_0.

- What does a large Cook's distance mean?

When the Tests Fail

- Transform the response: Box-Cox plot
- Transform the predictors;
- Add new variables (i.e., new models) for Simpson's paradox and lack of fit;
- Remove or combine variables for multicollinearity;
- Principal components regression or ridge regression for multicollinearity; and
- Use robust methods for regression.

Response Transformations: Box-Cox Plots

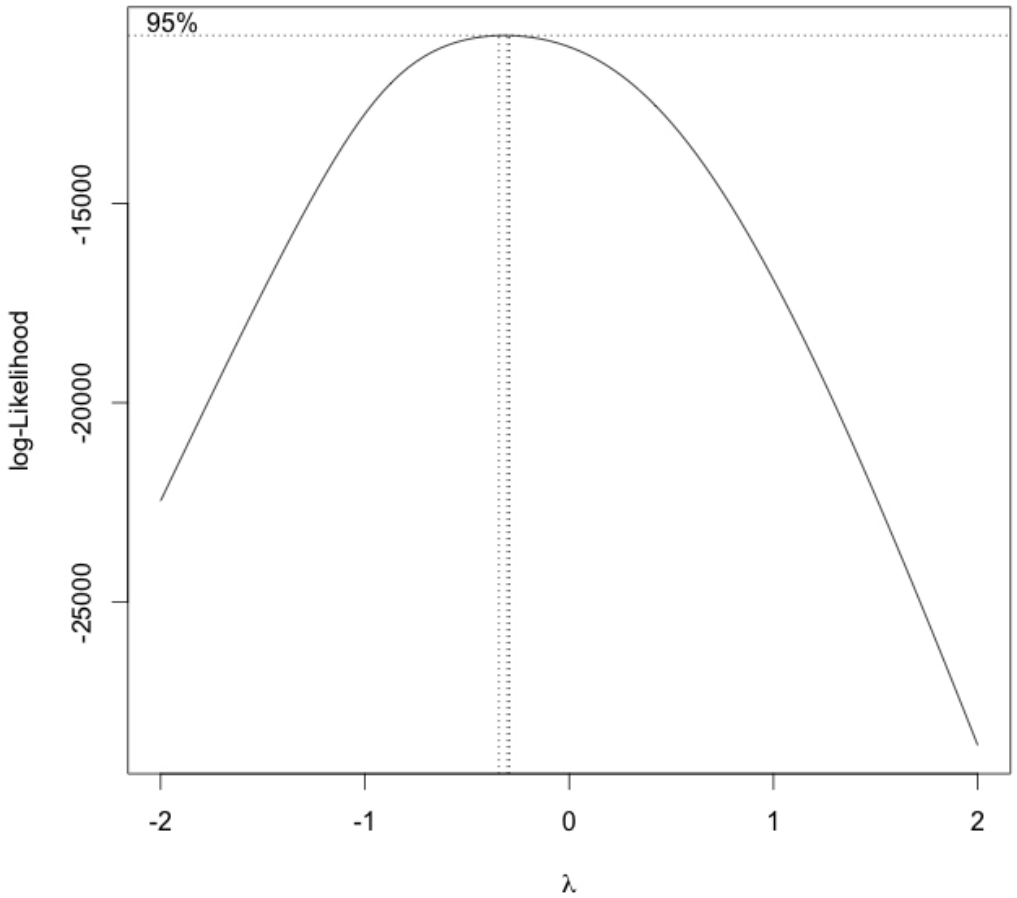
- When lack of Gaussian errors or lack of fit are detected consider transformation of the response. This can be helped with Box-Cox plots.
- Box-Cox plots allow us to consider power transformations of the form:

$$t_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases}$$

- The plot shows the 95% confidence interval for the λ. Look for the integers that this interval brackets. If it is around 1 what do we do? No transformation.
- If it is around zero what do we do?Log transformation.

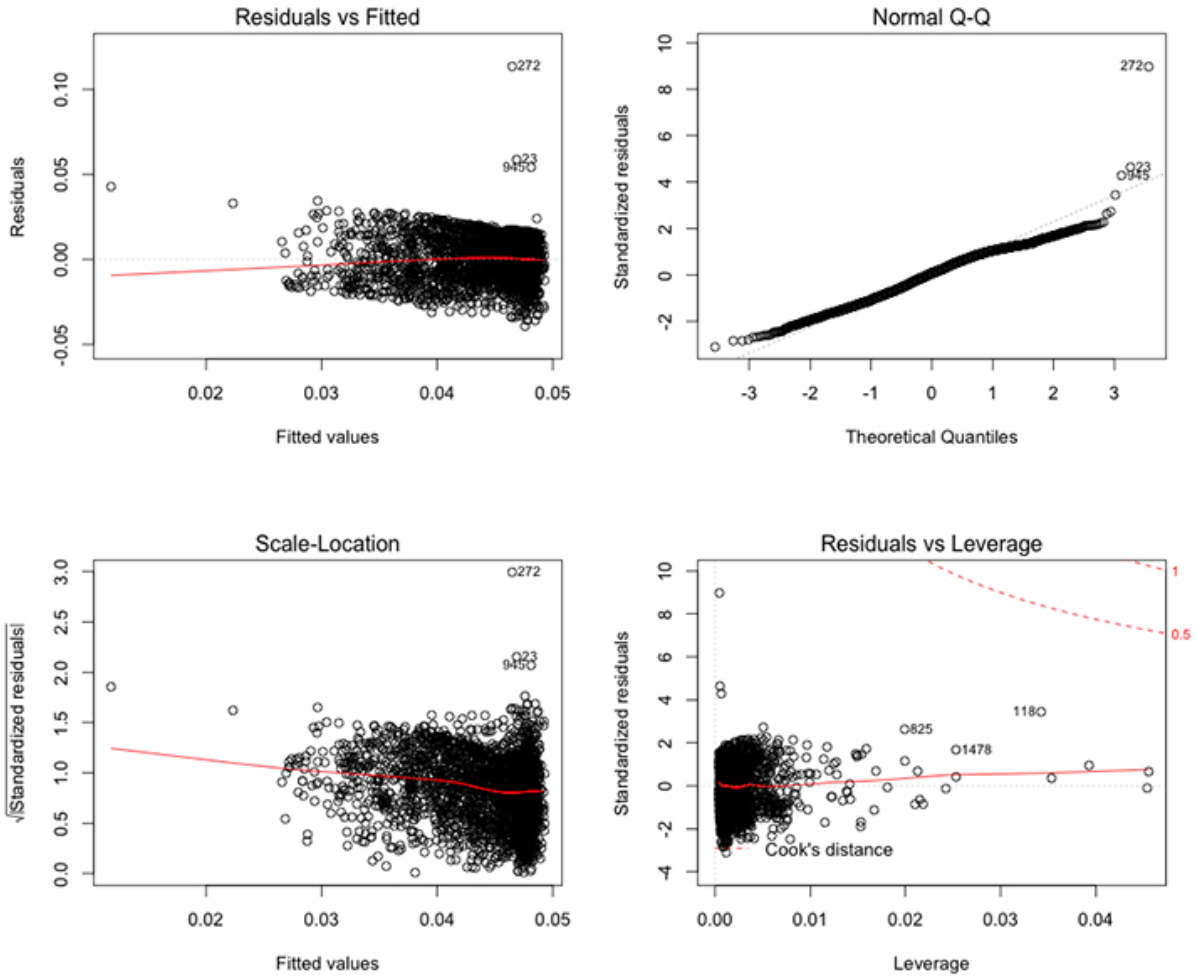
Example of Box-Cox Plot

Model: ACCDMG TEMP+TRNSPD+TONS+CARS+HEADEND1



Example of Diagnostic Plots: Log transformation

Model: log(ACCDMG) $\sim$ TEMP+TRNSPD+TONS+CARS+HEADEND1



Quantitative Regression Models

- We begin with linear models but the world for our analyses has many nonlinearities.
- Least square regression requires that the models are linear in parameters, but not in variables. So, we can still model nonlinear processes with least squares regression.
- We apply nonlinear transformations to the predictor variables as we have used them for the response variable.
- Transformations of the predictors provide for model fit and reduce model bias.

First Order Models

Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

- In first order models, all of the exponents are 1 on the x variables (predictors)

First Order Models

- Consider the prediction of task performance of a person as a function of sleep.
- Let,

$$Y = \text{Task Performance}$$
$$X = \text{Sleep}$$

- The first order model to show this relationship:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Draw the first order model. Do you think this is correct?

Second Order Models

Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2 + \cdots + \beta_k x_k + \beta_k x_k^2 + \epsilon$$

- Write a second order model.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Draw the second order model. Which of these models is more correct?

Review

- What kind of model is this one?

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- What kind of model is this one?

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

- Are either of these models nonlinear in parameters? No
- Nonlinear in variables? Yes

More than one predictor

- Now, add caffeine intake as a predictor to the task performance example:

$$Y = \text{Task Performance}$$
$$X_1 = \text{Sleep}$$
$$X_2 = \text{Caffeine Intake}$$

- A linear (main effects) model with the two predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

Interaction

- Suppose we think that the effect of sleep on task performance depends on caffeine intake.
- This implies that these variables interact in their relationship with the response.
- A model with main effects and interaction terms for the task performance problem:

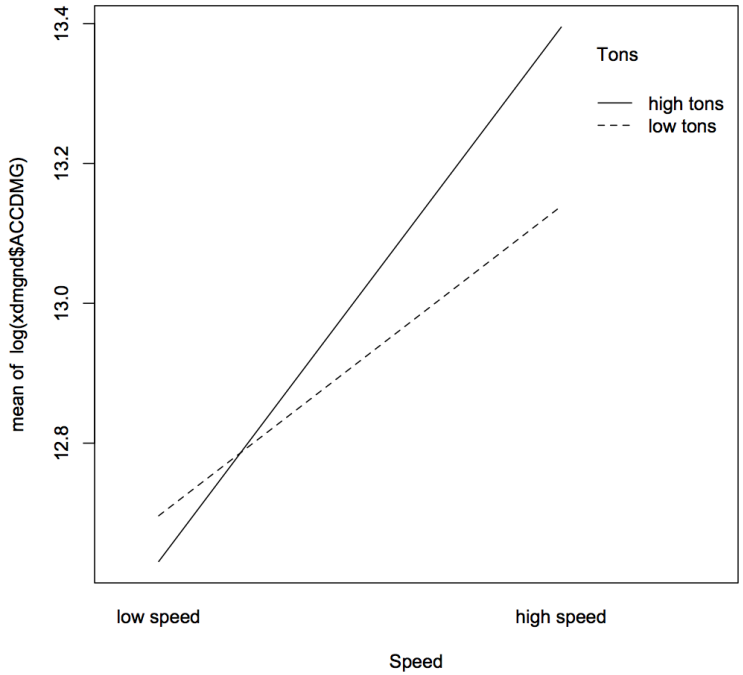
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- Is this model nonlinear in variables?Yes
- In parameters?No

Interaction Terms

- An interaction term means the effect of a variable on the response depends on the values of the other variables in the interaction.
- For the task performance example an interaction between sleep and caffeine intake means that the effect of sleep on task performance depends on caffeine intake. It also means that the effect of caffeine intake on task performance depends on sleep.
- Can you develop an example of a possible interaction for your project?

Interaction Plot



Complete Second Order Models

- Complete second order models include all the first and second order predictors (including interaction terms).
- Write a complete second order model of exam performance versus sleep.
- Write a complete second order model of exam performance versus sleep and caffeine intake.
- Suppose we add a third predictor variable, hours studied. How many terms (coefficients or parameters) do we have in a complete second order model with 3 variables? 9 coefficients and 10 parameters

Complexity

- In general, for a complete second order model with $p, p > 1$ variables, we have $2p + \binom{p}{2}$ coefficients.
- As we increase the number of variables the number of terms is growing as p^2.
- Choose second order variables with caution. Okham's Razor!
- We may want to consider transformations other than polynomials and interactions.

Qualitative Regression Models

- Qualitative predictors or categorical predictors require coding to use in regression models.
- The different values of the qualitative variable are called levels. Suppose we have a qualitative variable color with three values: red, white and blue. This variable has 3 levels.
- The only nonlinear relationship between qualitative variables and the response is with interactions. Why don't we use squared terms?

Treatment Contrasts

- A convenient method to represent categorical variables by numbers is with treatment contrasts or dummy variables.
- The default coding for qualitative variables in R is with treatment contrasts or dummy variables.
- If the qualitative variable has m levels then it can be encoded with $m - 1$ dummy variables. Each dummy variable represents a level of the qualitative variable.

Treatment Contrasts

- Each dummy variable takes on one of two values: 1 if the value of the qualitative variable equals the level represented by the dummy variable and 0 otherwise.
- One level for the qualitative variable is represented by all 0 values for the dummy variables. As we will see the choice of which level is represented this way is important.
- R will create the $m - 1$ dummy variables automatically using lexicographic selection.

Example Qualitative Regression Model

- Consider MPG of different cars. Suppose we have cars in three brands: Toyota, Ford and Chevrolet.
- Brand is a qualitative variable that may be related to MPG.
- Code the brand variable with 2 dummy variables as follows:

$$X_1 = \begin{cases} 1 & \text{if Ford} \\ 0 & \text{else} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if Toyota} \\ 0 & \text{else} \end{cases}$$

- What if we have another brand, Honda?

Tests of Understanding

- Which value is the base case? Is this the value that would be chosen by *R*?
- The main effects model with MPG (Y) as a function of brand: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- For your project, how do you code the variable *CAUSE*?
- *CAUSE* has 5 levels: E, H, M, S, T.
- Write a linear model with ACCDMG as a function of CAUSE (with five levels)

Example Qualitative Regression Model with 2 Variables

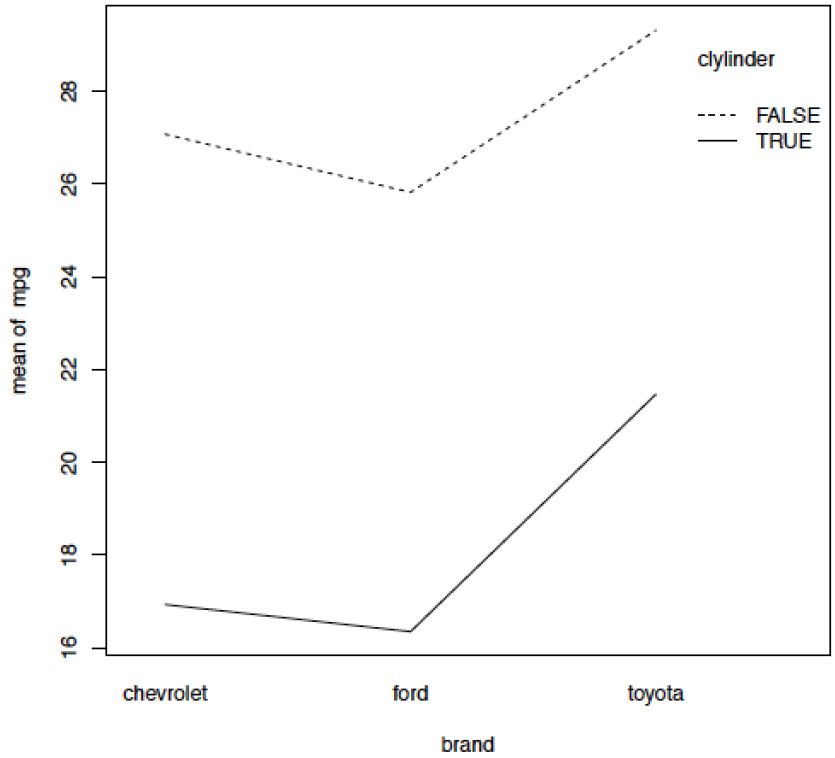
- We now add another categorical variable *Cylinder* to the MPG problem. Cylinder has two levels: more than 5 cylinders or not.
- For cylinder, define the dummy variable:

$$X_3 = \begin{cases} 1 & \text{if more than 5 cylinders} \\ 0 & \text{else} \end{cases}$$

Tests of Understanding

- Write a linear, main effects model with MPG (Y) as a function of brand and cylinder.
- Write a main effects plus interaction model with MPG (Y) as a function of brand and cylinder.
- Write a complete second order model with MPG (Y) as a function of brand and cylinder.

Interaction Plot of MPG



Regression Results of MPG

- Model: mpg $\sim$ brand
- Results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.4721	0.9042	22.641	< 2e-16 ***
ford	-0.7780	1.2275	-0.634	0.527
toyota	7.8999	1.4912	5.298	5.65e-07 ***

- Interpret the coefficient for *ford*
- Interpret the coefficient for *toyota*
- Based on this result, what's your conclusion?

Regression Results of MPG

- Model: mpg $\sim$ brand + cylinder
- Results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.6994	0.8171	32.676	<2e-16 ***
ford	-0.8172	0.8367	-0.977	0.3307
toyota	2.8202	1.1066	2.549	0.0121 *
greaterthan5	-9.5633	0.8240	-11.607	<2e-16 ***

- What is the base case?
- Interpret the above result, what's your conclusion?

Regression Results - Train Accidents

- Model: ACCDMG $\sim$ Cause (5 levels: E, H, M, S, T)
- Results:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	169976	24368	6.975	3.82e-12 ***
CauseH	-100517	27384	-3.671	0.000247 ***
CauseM	-44111	29130	-1.514	0.130073
CauseS	-130979	60260	-2.174	0.029825 *
CauseT	11273	28265	0.399	0.690050

- What is the base case?
- Interpret the above result, what's your conclusion?