Transformations

Patrick Breheny

January 17

Violations of linear regression assumptions

- When learning about linear regression, one is usually taught to examine diagnostic plots in order to check whether or not the model assumptions have been violated
- Suppose that these assumptions have been violated – what options are available to you?
- To a large extent, answering that question is one of the purposes of this course

Remedial measures

Generally speaking, you have three avenues down which you could proceed:

- Transforming variables
- Fitting a more flexible model
- Using a different method

Our motivating data set for today consists of daily measurements of air quality (in terms of ozone concentration) taken in New York during the summer of 1973

Ozone data

- While the ozone layer in the upper atmosphere is beneficial and protects us from ultraviolet light, in the lower atmosphere it is a pollutant that has been linked to a number of respiratory diseases as well as heart attacks and premature death
- The EPA's national air quality standard for ozone concentration is 75 parts per billion (ppb); in Europe, the standard is 60 ppb; and according to some studies, at-risk individuals may be adversely affected by ozone levels as low as 40 ppb
- Ozone concentrations, however, are not constant, and fluctuate quite a bit from day to day, depending on many factors

Ozone data set

The file `ozone.txt` contains the following variables:

- `Ozone`: Ozone concentration (in ppb)
- `Solar`: Solar radiation (in Langleys)
- `Wind`: Average wind speed (in miles/hour)
- `Temp`: Daily high temperature (in Fahrenheit)
- `Day`: Day of the year, with January $1 = 1$, February $1 = 32$, etc.

We will be considering ozone concentration to be the outcome variable and the rest as explanatory variables

Scatterplot matrices

- A useful way of visualizing multivariate relationships is with *scatterplot matrices*:
    - In R,

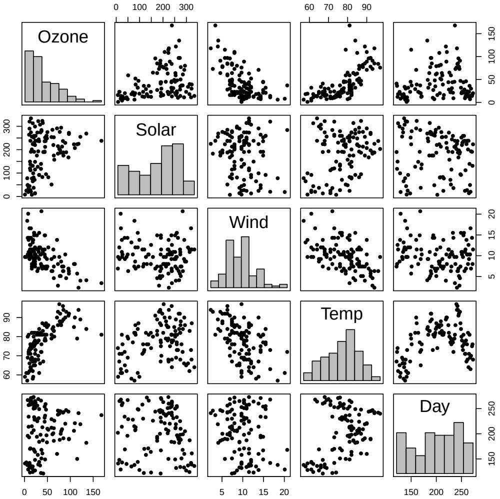
    ```
    pairs(ozone)
    ```

    - In SAS,

    ```
    PROC SGSCATTER;
        MATRIX Ozone Solar Wind Temp Day;
    RUN;
    ```

- Both SAS (using the `DIAGONAL` option) and R (using the `diag.panel` option) come with options allowing you to also plot things like histograms and kernel density estimates along the diagonal

Scatterplot matrix for the ozone data



The regression model

- Letting Y_i represent the ozone concentration on day i, x_{i1} represent solar radiation on day i, x_{i2} represent wind speed on day i, and so on, let's fit the linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

- In R,

```
fit <- lm(Ozone~Solar+Wind+Temp+Day, data=ozone)
```

- In SAS,

```
PROC REG DATA=ozone;
  MODEL Ozone = Solar Wind Temp Day;
RUN;
```

Prediction

- Let's consider two hypothetical days:
  - A: Solar=180, Wind=5, Temp=90, Day=274
  - B: Solar=180, Wind=15, Temp=70, Day=274
- We could estimate the mean ozone concentration of these two days in R using

```
Days <- data.frame(Solar=180, Wind=c(5,15),
                   Temp=c(90,70), Day=274)
predict(fit, Days, interval="confidence")
```

	Mean	Lower	Upper
A	74.9	66.3	83.5
B	5.2	-5.5	15.9

- In SAS, we can add these rows to the data set and set `Ozone` to missing (.), then use the `/ P` option

Modeling log ozone levels

- Things look reasonable for day A, but for day B, our model suggests that the average ozone concentrations may be negative, an unreasonable conclusion
- One way to fix this is to consider a transformation of the response: rather than model Y, perhaps we should model, say, $\log(Y)$:

$$\log(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

- In R, we can fit this model with

```
fit <- lm(log(Ozone)~Solar+Wind+Temp+Day, data=ozone)
```

SAS does not support this kind of on-the-fly transformation; we would have to create a variable called `logOzone` in a `Data` step prior to calling `PROC REG`

Confidence intervals for the log ozone model

- When we calculate confidence (or prediction) intervals, they will be on the log scale:

	Mean	Lower	Upper
A	4.3	4.1	4.5
B	2.6	2.4	2.9

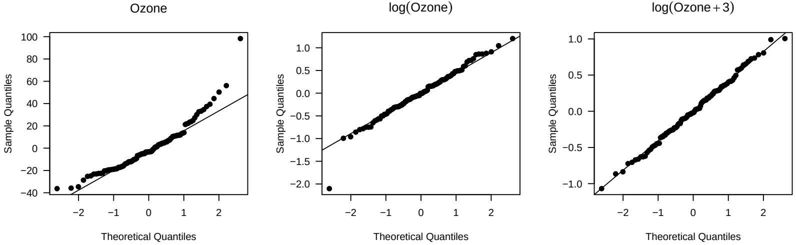
- Now, when we transform back onto the original scale (by exponentiating the above table), we avoid negative estimates:

	Mean	Lower	Upper
A	71.2	57.8	87.7
B	13.8	10.7	17.9

Normalizing transformations

- In addition to avoiding estimates that fall outside the logical range of the parameter space, such transformations often cause the response to appear more normally distributed
- This is a common strategy throughout statistics: when a variable (or parameter) does not follow a normal distribution, work with a transformation of it that does (such transformations are said to be *normalizing*)
- In addition, models that better reflect the data and its true distribution typically produce more powerful results as well

Effect on normality

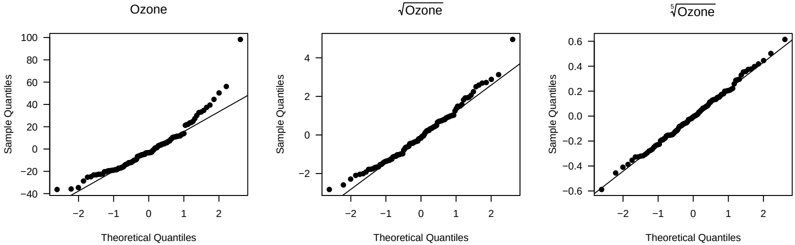


Ozone

log(Ozone)

log(Ozone + 3)

Power transformations



This approach is known as the *Box-Cox procedure*, after two statisticians who identified an automatic method for identifying the optimal normalizing exponent to which y should be raised

Effect of transformation on R^2, p-values

Effect on R^2:

	Transformation				
	None	log	log+3	$\sqrt{\ }$	$\sqrt[5]{\ }$
R^2	0.616	0.667	0.688	0.679	0.686
R^2_{adj}	0.602	0.654	0.676	0.667	0.675

Effect on p-values:

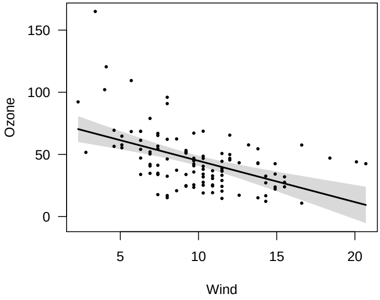
	Ozone	$\log(\text{Ozone} + 3)$	$\sqrt[5]{\text{Ozone}}$
Solar	.03	.0001	.0002
Wind	1×10^{-6}	2×10^{-5}	1×10^{-5}
Temp	1×10^{-9}	7×10^{-13}	7×10^{-13}
Day	.1	.2	.2

Transforming explanatory variables

- We may benefit from transforming the explanatory variables as well
- Here, there are no distributional considerations (linear regression makes no assumptions about the distribution of $\mathbf{X}$)
- However, the model does assume a linear relationship between x and $\mathrm{E}(Y|x)$, and this may not hold

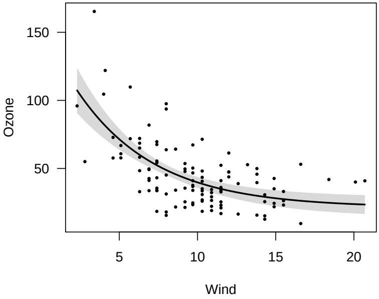
Ozone vs. wind

For example, consider the original linear model we discussed and let's look at what the model estimates about the relationship between wind and ozone:



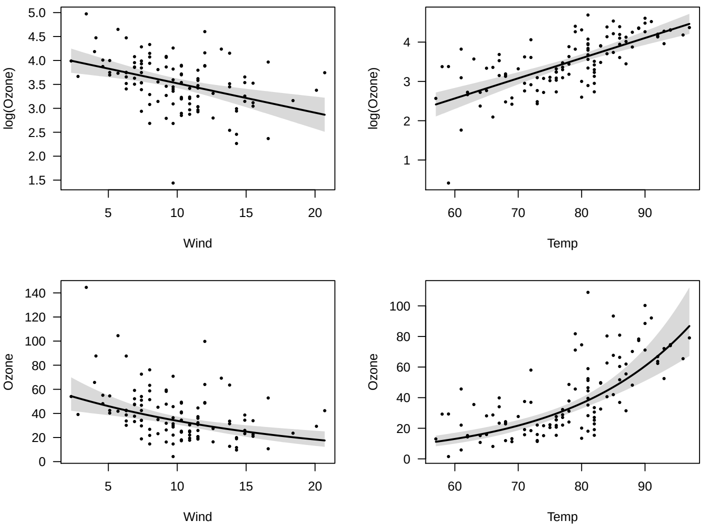
Ozone vs. wind

The previous plot suggests a vaguely exponential relationship between wind and ozone; if we replace Wind with $\exp(-\text{Wind}/5)$ in the model, we obtain:



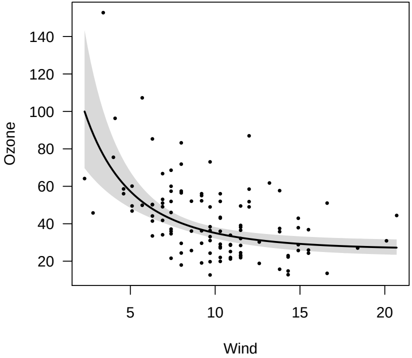
The R^2 of the model also improves, from 62% to 69%

Response transformation also introduces nonlinearity



Transforming both response and explanatory variables

There is no problem, of course, in transforming both response and explanatory variables:



Note that this model yields the highest R^2 so far, 71%

"Linear" regression?

- If we've got terms like Wind^2 and $\log(\text{Ozone})$ in the model, is our model still "linear"?
- Yes; our focus in statistics is always on the unknown parameters $\boldsymbol{\beta}$, and with respect to the parameters, the model is still linear
- Transforming observable quantities presents no barrier to model fitting, testing, estimation, or finding confidence intervals, and ordinary least squares linear regression can be applies to all these models in the exact same way (although you may need to transform back to original units at the end)

Assumptions and flexibility

- Now to the more complicated issue of fitting more flexible and complicated models to the data (but still sticking with linear regression models)
- A “more flexible” model is one that makes fewer assumptions, and may include:
    - Adding explanatory variables to the model
    - Adding interaction terms
    - Adding nonlinear functions of explanatory variables (*i.e.*, in addition to, as opposed to replacing, the linear terms)

The bias-variance tradeoff

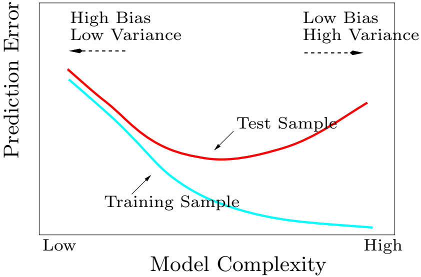
- When we make assumptions that are false, this introduces bias into the parameter estimates
- This makes a more flexible model attractive, but it comes at a price: the variability of the parameters we are trying to estimate goes up
- This illustrates what is perhaps the central concept in statistical modeling: the *bias-variance tradeoff*

Overfitting

- This is a fundamental idea in the notion of statistical *inference* – just because you can describe the sample well does not mean that this description can be generalized to the population
- As I'm sure you saw last semester, when you add more parameters to a model, RSS goes down and the R^2 goes up
- But this does not mean that more complicated models are more successful at predicting outcomes that lie *outside* our sample
- This phenomenon is referred to as *overfitting*; a model that describes the sample very well, but generalizes poorly, is said to be *overfit*

Bias-variance tradeoff – illustration

An illustration of this phenomenon, courtesy of *The Elements of Statistical Learning*, by Hastie, Tibshirani, and Friedman:



Here *training sample* refers to the data used to fit the model, and *test sample* on an external sample used to test the accuracy of the model

Summary

In summary, when the assumptions of linear regression are not met, you can:

- Try transforming the response and/or explanatory variables
- Add parameters to the model (at the risk of increasing variance)
- Use a method other than linear regression

This final point is the main focus of this class – to develop a more flexible class of models than linear regression, called *generalized linear models*