Choosing a Model

What criteria do we use to select one model from many?

- Scatter plots - do they indicate a linear relationship? When we consider two variables in the model, we may want to hold one explanatory variable constant and examine the relationship between the second explanatory variable and the response variable.

- Residual plots - do they indicate a lack of fit?

  Sketch below examples of nonlinear relationship, need for an additional variable in the regression, heteroscadasticity.

- Scientific backing for the model

- Best k variable fit.

  In the Kaiser example, the Surgeon General Report listed 6 variables. If we were looking for "the best" 2-variable model, how many regressions would we need to examine? what about the best 3-variable model?

- Reduction in explained variation

Fit	Residual sum of squares/n
birth weight on constant	337 oz^2
birth weight on constant, height	324 oz^2 = 337 - 13
birth weight on constant, weight	329 oz^2 = 337 - 8
birth weight on constant, smoking	
birth weight on constant, weight,height	322 oz^2 = 337 - 15
birth weight on constant, height, smoking	

Transformations

In the simple linear model we have,

$$\begin{aligned} E(Y|x) &= a + bx \\ Y &= a + bx + E \end{aligned}$$

But when we start to consider transformations, such as $\log(E(Y|x)$, what happens to the errors?

$$\begin{aligned} Y &= exp^{a+bx}U \\ \log(Y) &= \\ OR & \\ Y &= expa + bx + V \\ \log(Y) &= \end{aligned}$$

If U has mean 1 then $\log(U)$ may or may not have mean 0.

Consider transforations for each of the following:

- $$E(Y) = a/(b + cx)$$
- $$E(Y) = a * b^x$$
- $$E(Y) = x/(c + dx)$$
- $$E(Y) = 1/(1 + e^{bx})$$