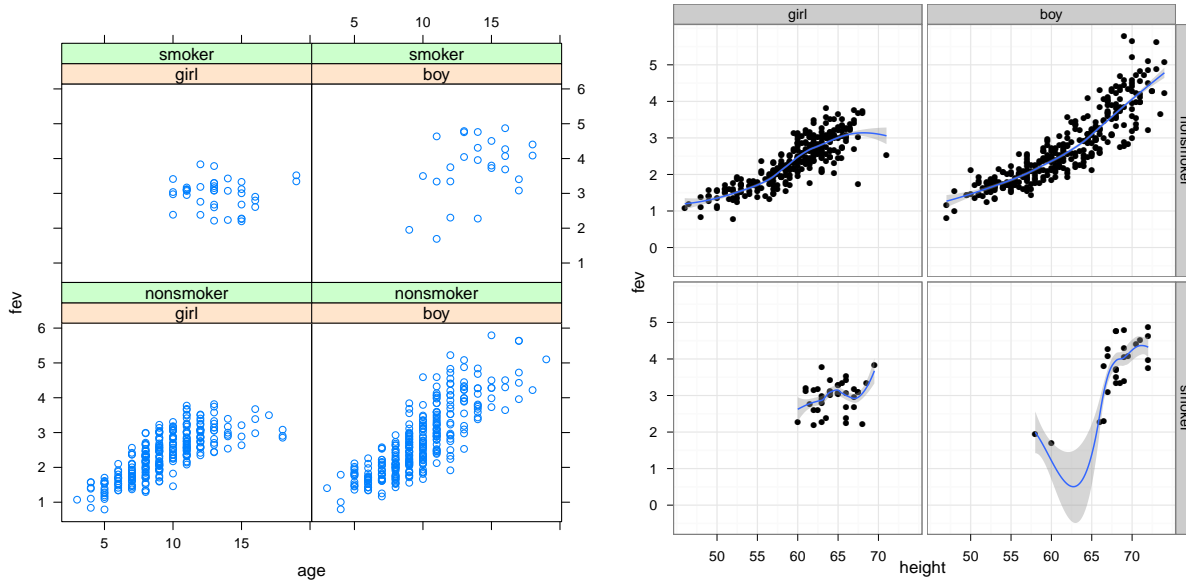


## Homework 4 STAT 505 Fall 2011

A report on a study of lung volume in children.

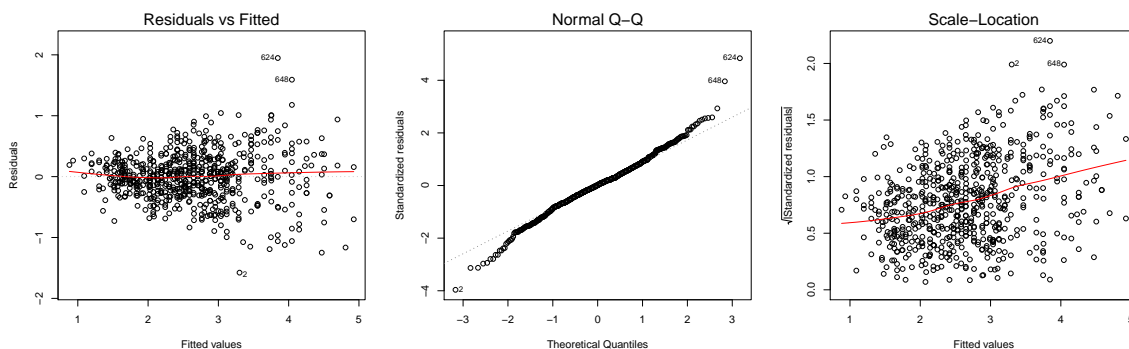
We are tasked with determining relationships between several possible predictors and lung volume as measured by FEV, forced expiration volume. The presenting question is “Is there an effect of smoking on forced expiration volume after adjusting for other variables and if so, is it related to gender?” I began with plots of the response versus predictors.



The plots above show me several things about the data:

- FEV increases with both age and height. Height is a more precise measure than age, and lung volume is expected to increase with physical size (height) so I view height as a more precise predictor than age and will favor its inclusion in models. The correlation between height and age is 0.792.
- The relationship between fev and height is not just linear; there is some curvature.
- I'm glad to see that no smokers are less than 9 years old (10 for girls). This will affect our analysis in that the smoker variable is partially confounded with age.
- Unsurprisingly, the larger heights and larger fev's are in the boys.

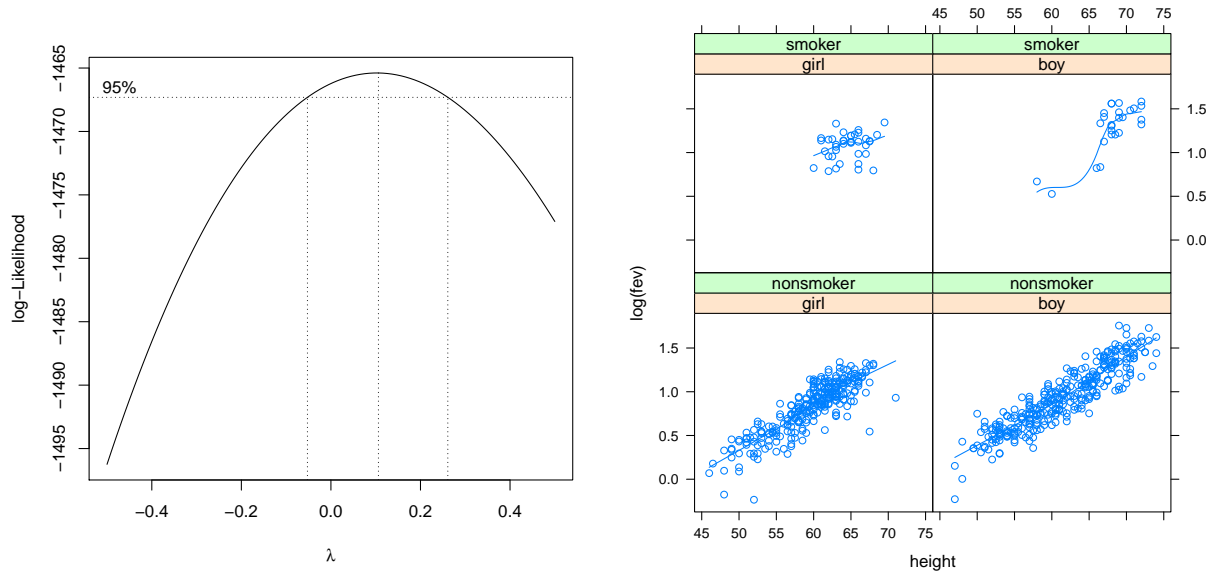
To look for residual problems, I fit a big model with sex, smoker, and quadratic in height and all 2 and 3-way interactions.



The diagnostics for this model show a distinct fan shape in the residuals and an increasing trend in the scale–location plot. We need to use either a transformation or weighted regression.

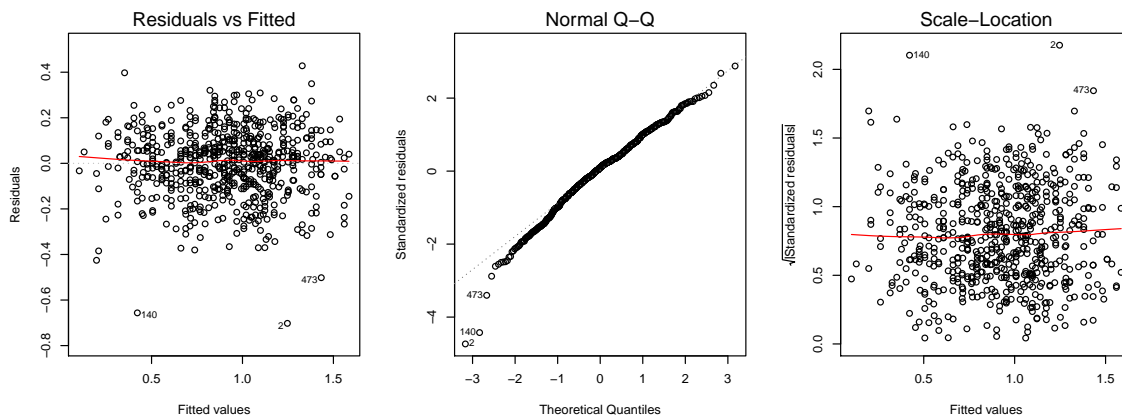
**Transformation:**

Using Venables and Ripley’s `boxcox` function, I see that the optimal power transformation is close to a log transform. I like that because physical lung volume is a product of height, width and breadth, and transforming to the log scale makes the relationship additive. We now have a new response variable, so I would redo the above plots.



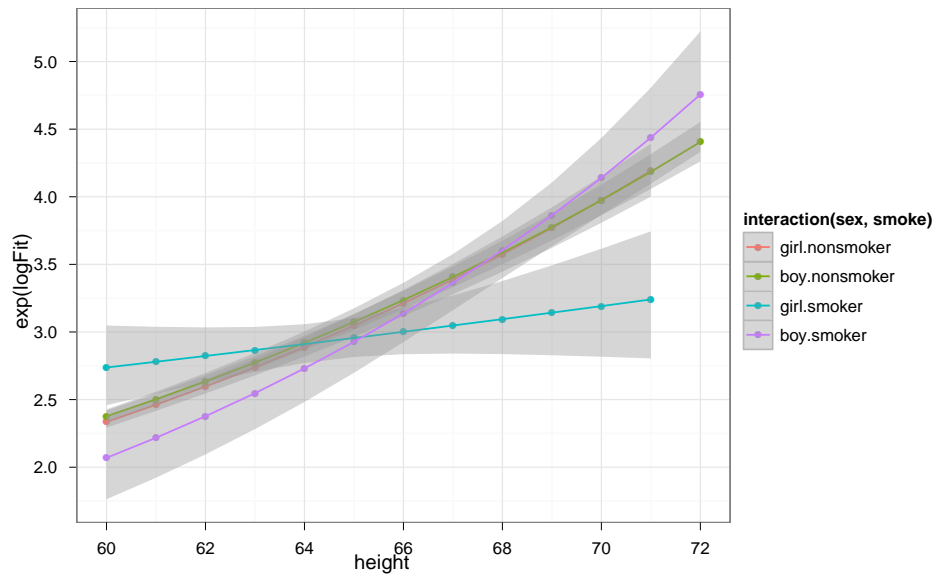
I no longer see a need for a quadratic term (and I did check the model and the t-stats agreed with me). I then refit the model, and with sex, smoke, and height main effects and all interactions. The 3-way interaction has a very small p-value, so I will leave all terms in the model. We could add in age, but I have an easier time with interpretation, so I am going to leave it out.

Diagnostic plots:



The transformation solves the problem with the diagnostic plots, and has a nice physical interpretation. The residuals are now left-skewed, but only slightly.

Let’s compare fitted values and look more deeply at the smoking effects. We need to back-transform the log-fits and look at a multiplicative model. The action in these data all occurs in the older kids, so we need predicted values for the older ages, or for heights over 60 inches. The smoke and sex factors subdivide the sample into four groups which we can compare.



Conclusions: At heights 64 to 66 inches, the model estimates all overlap. Smoking girls do seem to have lower fev than the other 3 groups for heights 66 to 69.5, after which we would be extrapolating. Smoking boys do not seem to differ much from nonsmokers, but there is a suggestion othat smoking might be associated with larger fev.

These data came from

Rosner, B. (1999), *Fundamentals of Biostatistics*, 5th ed., Pacific Grove, CA: Duxbury.

as reported by

Kahn, M. (2005), An exhalent problem for teaching statistics. *Journal of Statistics Education*, **13**:2

URL:<http://www.amstat.org/publications/jse/v13n2/datasets.kahn.html>

## R code

```
### code chunk number 1: fev-smoker1
#####
fev <- read.table("data/fev.dat",head=T)
fev$sex = factor(fev$sex, labels=c("girl","boy"))
fev$smoke <- factor(fev$smoke, labels = c("nonsmoker","smoker"))
require(lattice)
print(xyplot( fev ~ age|sex*smoke, data = fev))

### code chunk number 2: fev-smoker2
#####
require(ggplot2)
theme_update(theme_bw())
print(qplot( x=height,y=fev, data = fev, facets= smoke~sex) + geom_smooth())

### code chunk number 3: diagnostics
#####
fev.fullfit <- lm( fev ~ sex * smoke * (I(height-61) + I((height-61)^2)),
  data = fev)
par(mfrow=c(1,3))
plot(fev.fullfit,which=1:3)

### code chunk number 4: fev-smokerBoxCox
#####
MASS::boxcox( fev.fullfit,lambda=seq(-.5,.5,.1))

### code chunk number 5: fev-smoker3
#####
print(xyplot( log(fev) ~ height|sex*smoke, data = fev,
  type =c("p","smooth") ))

### code chunk number 6: fev-logfit
#####
summary(logfev.fullfit <- lm( log(fev) ~ sex * smoke * height,
  data = fev))$coef
par(mfrow=c(1,3))
plot( logfev.fullfit, which=1:3)

### code chunk number 7: newDF1
#####
newFEV <- data.frame(height = rep(60:72,4),
  smoke=gl(2,26,52,label=c("nonsmoker","smoker")),
  sex = gl(2,13,52,lab=c("girl","boy")))
newFEV <- subset(newFEV, sex == "boy" | height < 71.5)
newYs <- predict(logfev.fullfit, new = newFEV, se.fit=TRUE)
newFEV$logFit <- newYs$fit
newFEV$SE <- newYs$se.fit
newFEV$upr <- newYs$fit + 2 * newYs$se.fit
newFEV$lwr <- newYs$fit - 2 * newYs$se.fit

print( qplot(x=height, y=exp(logFit), colour = interaction(sex,smoke),
  type="l",data=newFEV) + theme_bw() + geom_smooth(aes(ymin = exp(lwr),
  ymax = exp(upr)), stat="identity"))
```