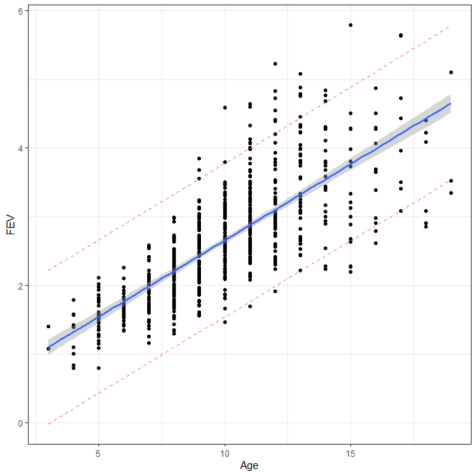
Stat 414 - Day 8
Unequal variances

Last Time:

- Interactions model the "effect" of one variable (slope) changing depending on the value of the other variable
  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2$
  - Coefficient of $x_2 = \hat{\beta}_2 + \hat{\beta}_3 x_1$
  - Graphs are very (most) helpful for explaining interactions

Example 1: FEV cont.

Recall the FEV data and fit a model with just Age.



(a) Which "band" shows the prediction intervals and which shows the confidence intervals? How are you deciding/what is the difference?

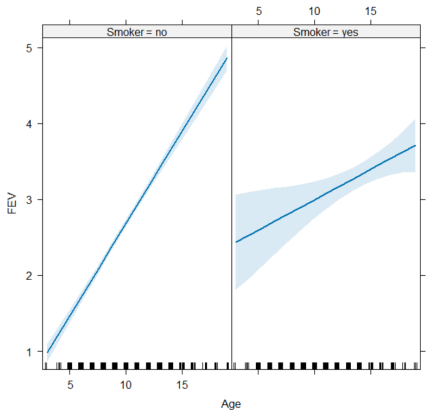
(b) Why do the bands look curved?

(c) Do the prediction intervals appear to be performing adequately?

Reconsider model 4 with the interaction

```
#Notice some short cuts I took here
model4 = lm(FEV ~ Smoker*scale(Age, center=TRUE),
data = FEVdata)
summary(model4)

#install.packages("effects")
library(effects)
plot(allEffects(model4))
```



(d) Are these confidence bands or prediction bands? Why do the widths differ so much between the smokers and non-smokers? (*Hint*: Two reasons at least)

So it's important that we estimate σ "correctly." In particular, the "basic linear model" assumes $Var(\epsilon) = \sigma^2$. If you don't believe you have sufficient heterogeneity, one approach is to make different distributional assumptions.

Example 2: Variance Stabilizing Transformation An electric utility is interested in developing a model relating peak-hour demand (in KW) to total energy usage during the month. This is an important planning problem because while most customers pay directly for energy usage (in kwh), the generation system must be large enough to meet the maximum demand imposed.
(a) What do you notice as we try the square root and then the "more severe" log transformations?

Example 3: Modeling Heterogeneity/Weighted Least Squares Smith et al. (2005) examined reproductive and somatic tissues in the squid *Loligo forbesi*. The data in Squid.txt include the dorsal mantel length (in mm) and testis weight from 768 male squid, over different months.
(a) How does Testis weight appear to change with DML? For which DML values do we have less variable measurements of Testis weight?

Fit a linear model predicting testis weight from dorsal mantel length.
(b) What do you conclude from the residual plots? (Also pay attention to the third one now.)

(c) Why might a transformation **not** be helpful here?

Another approach is to model the relationship between the variance and the explanatory variable. Because the variability in the residuals is increasing with DML values, we could try modelling the error variance as proportional to DML: $\sigma_i \times I \sim N(0, \sigma^2 \times DML_i)$.
We can think of this as a special case of weighted least squares which assumes $V(Y_i \dashv | x_i) = \sigma^2 w_i$ and if we know the w_i then we minimize $\Sigma w_i (Y_i - \beta_0 - \beta_1 X_i)^2$. For this dataset, we can take $w_i = 1/DML_i$ which will give more "weight" in the least squares estimation to squid with smaller DML values.
(g) How did the slope coefficient change? Is this what you predicted? Did R^2 and $\hat{\sigma}$ change?

To see whether this has sufficiently addressed the heterogeneity we saw in the residuals, we want to look again at the residual plots. However, with weighted least squares, we need to look at the *standardized* residuals rather than the non-standardized residuals.
(h) Have things improved? Be clear how you are deciding.

Notes:
- Can explore other "forms" (e.g., powers) of the variance covariates. (Watch for zero and negative values)
- The estimates of the coefficients will usually be nearly the same as the unweighted estimates, but the weights will impact the widths of prediction intervals.
- But there is a different problem with how we have done things here. We don't know the true σ_j values, we have only estimated them from the sample data. Instead we should use *generalized least squares* (Aiken, 1934) which is an iterative approach for simultaneously estimating the regression coefficients and estimating the variance terms….