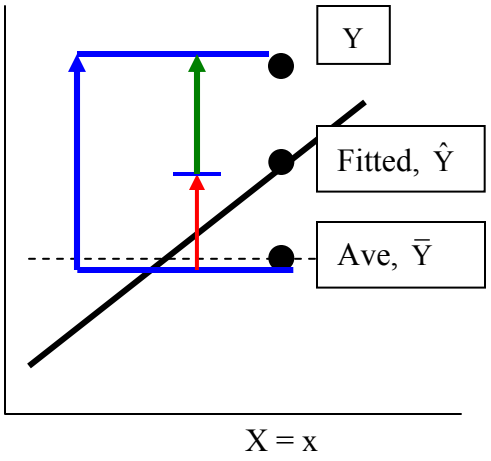
BE640 – Intermediate Biostatistics
Frequently Asked Questions
Topic 2 FAQ 1 – Regression and Correlation

1.
Here is a little schematic to give you a better sense of what the analysis of variance table is all about in the setting of simple linear regression.



In 540, we might have looked just at the Y's, without regard for the companion X's. The total variability in the Y data that we are looking to explain is the numerator of the familiar variance calculation. Here it is:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

This numerator is the "total variance, corrected" is the bottom row of the analysis of variance table. **Have a look again at the bottom of page 20 of your lecture notes.**

Consider one data point Y and have a look at the picture at left. If you look at the arrows, you can appreciate that

↑ = ↑ + ↑

$$(Y-\bar{Y}) = (\hat{Y}-\bar{Y}) + (Y-\hat{Y})$$

Thus, for this one individual Y, we have partitioned it's departure from $\bar{Y}$ into two portions. One is the departure of $\hat{Y}$ from $\bar{Y}$. The other is the departure of Y from $\hat{Y}$. Some extension of this to the world of squared differences (afterall, this is how we think about variability) and some algebra yields an amazing partitioning of the total variability

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

The terms on the right hand side are the "**regression**" and "**error**" rows of the table **at the bottom of page 20 of your lecture notes. Look!**

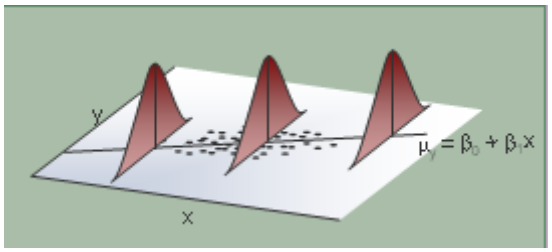
2.
The following is a picture of one of the assumptions required for hypothesis testing and confidence interval estimation in simple linear regression.

Assumption:
At each value of X (we'll call this "x"), Y is distributed Normal ($\beta_0+\beta_1 x$, $\sigma^2_{Y|X}$)

The picture below allows us to appreciate several things:

- **We're dealing with not just one normal distribution, but several.**
- **In particular, there is a separate normal distribution for the distribution of our outcome Y, for each value of our predictor variable X.**
- **Each of these normal distributions has its own mean parameter μ_x . Notice how we keep track of the separate means by attaching the little subscript x.**
- **The multiple, separate, μ_x are of a particular form – they all lie on a line!**

$$\mu_x=\beta_0+\beta_1 x$$



source: http://statistics.byu.edu/faculty/wal/public/511/PowerPoint/511_Chap8.ppt