**Topic 28**

**Least Squares Regression**

Least squares is a method for fitting a line to a set of data. The line represents the dynamics of any given system. Ultimately what one wants to do is predict an outcome, e.g. compute the theoretical values of a response variable given theoretical values of an explanatory variable.

So far we have only looked at data. Data represented the relationship. This time we are constructing a theoretical relationship based on data. In one situation we had 1784 patients receiving a medicine and 34 developed cancer. What if we have 1793 patients. Do we have to do the experiment all over again?

Of course not. We build an equation that gives us a theoretical response value for a proposed experimental value. We have already seen this and, in fact, it is part of our daily life. When we ask a sample of the population if they eat goose for Christmas, we expect to extrapolate this proportional value to the general population, or another similar sample. If 78% of group A like Dizzie dishwashing liquid, then we expect that approximately 78% of group B will.

In least squares regression, we formalize a similar relationship in terms of the equation for a line.

**Linear equation**

$\hat{y}$ = a + bx

 $\hat{y}$ Response Variable

 a y-intercept. Location where line crosses the y axis

 b slope

 x explanatory or predictor variable

 this can be written y = 2x -3, or y = -3 + 2x



**Minimize** $\sum\_{}^{}y\_{i} - \hat{y}$

The values of a and b that minimize these values become the values for the best line.

**Residual** is difference between an observed value, yi and an estimated $\hat{y}$, **ei = yi -** $\hat{y}$



**Slope Coefficient** (b or β) is the predicted change in the response variable, y, for one unit change in the explanatory variable, x.

**Intercept Coefficient**  the predicted value of the response variable, y, when x=0. This coefficient can be sometimes ignored.

**Extrapolation –** predicting the response variablevalues beyond the domain or range of the data.

**Influential Observation –** observation with values that substantially impact the regression equation.

**R2** , the **Ceofficient of Determination,** estimates the fraction of the variance in *the response variable*  that is explained by *the explanatory variable*  in a linear regression

**Fitted value** is the predicted value for a given explanatory variable, x.

Note that the x’s are never predicted. They keep their values. At the beginning of a regression analysis,

We have a set of x’s and y’s. At the end, we have, in addition to x’s and y’s, a $\hat{y}$ for every x, two new parameters, a and b, the intercept and slope respectively, and a residual, e, for ever y value.

**Residual Plot** (there are two, or more!) **e** vs. **x**, or **e** vs $\hat{y}$. The plots should appear RANDOM!!!! R~0.

Here is a pretty good plot



Here we are dealing with residuals from a bimodal distribution



Clearly, these data points are not randomly distributed



Here we have heteroscedasticity



**Transformation** – remapping of data.

y=x+b, z = y + 2 --> z = x+b+2 = x + (b+2)

y= 2x, z= 2 y, 🡪 z = 2(2x)=4x



