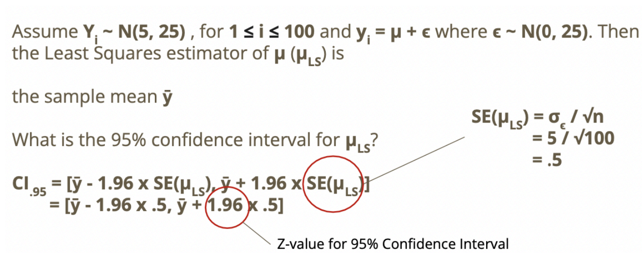
CAS CS 506
Lec 18

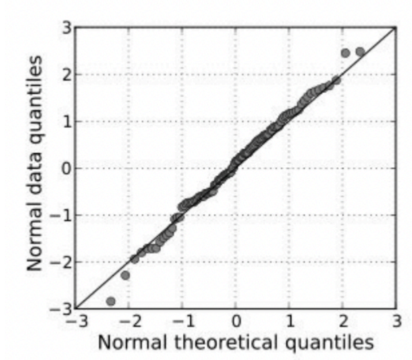
Linear Model Evaluation (cont.)

1. Confidence Intervals
    a. An interval that describes the uncertainty around an estimate (here this could be B hat)
    b. Goal: For a given confidence level, construct an interval around an estimate such that, if the estimation process were repeated indefinitely, the interval would contain the true value (that the estimate is estimating) 90% of the time
2. Z-Values
    a. These are the number of standard deviations from the mean of a N(0,1) distribution required in order to contain a specific % of values were you to sample a large number of times.
    b. To find the .95 z-value (the value z such that 95% of the observations lie within z standard deviations of the mean you need to solve

$$\int_{-z}^{z} \frac{1}{2\pi} e^{-\frac{1}{2}x^2} dx = .95$$

    c. The .95 z-value is 1.96
    d. This means 95% of observations from a N(mean, standard deviation) lie within 1.96 standard deviations of the mean(mean * 1.960 * standard deviation)
    e. If we get a sample from a N(mean, standard deviation) of size n, how would we create a confidence interval around the estimated mean?
        i. A:
    f. How do we confidence interval?
        i. 

3. QQ plot
    a. Quantiles are the values for which a particular % of values are contained below it
    b. For example the 50% quantile of a N(0,1) distribution is 0 since 50% of samples would be contained below 0 were you to sample a large number of times
    c. For all quantiles q, if sample.q == known_distribution.q then they have the same distribution

d. 

4. Constant Variance
    a. One of our assumptions was that our noise had constant variance. How can we verify this?
    b. We can plot residuals (noise estimates) for each fitted value y-hat$_i$
5. Extending Our Linear Model
    a. Changing the assumptions we made can drastically change the problem we are solving. A few ways to extend the linear model:
        i. Non-constant variance → used in WLS (weighted least squares)
        ii. Distribution of error is not normal → used in GLM (generalized linear models)