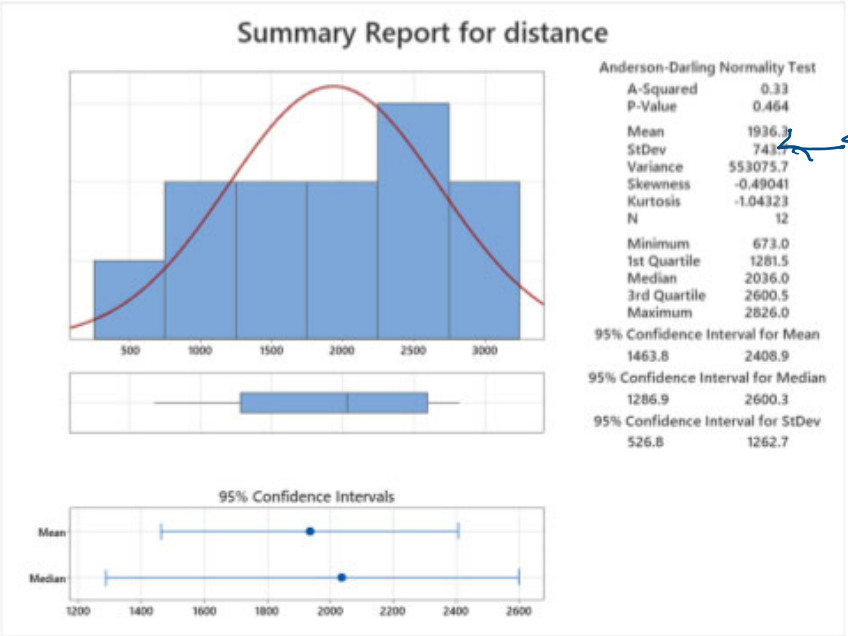
**Stat 414 - Day 1**
**Review of Simple Linear Regression**

**Recall:**

- Single level simple linear regression model: $E(Y_i) = \beta_0 + \beta_1 X_i + \epsilon_i$ where $\epsilon_i$ is assumed to be normally distributed with mean 0 and variance σ.
  - We can think of the model as connecting the expected values of the populations of responses at each "fixed" *x* value. We assume these populations are normally distributed and all have the same variability $\sigma^2$.
- The "fitted" model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and the residuals $e_i = y_i - \hat{y}_i$ represent the leftover unexplained variation in the responses.
- With least squares regression, we use $MSE = \frac{SSE}{n-2} = \frac{\Sigma(e_i^2)}{n-2}$

**Example:** Airfares from San Luis Obispo to a "random" sample of 12 major U.S. cities as found March 31, 2017 on Travelocity.com for travel on May 8-May 12, 2017 are found in airfare.txt.




← s=typical prediction error

(a) What *one number* would you use to estimate the airfare of an individual flight? How accurate do you think your estimate would be?

mean? median? ← use 2S?

(b) By what criteria is the value in (a) the best value?

$\min \Sigma|y_i - k|$ ← median

$\min \Sigma(y_i - k)^2$ ← mean

(c) Is the distribution of the airfare variable normally distributed? Is this a problem? What are some steps we can take if we think this is a problem?

not always necessary (in regression, it's the errors not y that need to be normal) or transform

(d) If I assume the prices follow a normal distribution in the population, what *one number* would you use to estimate the airfare of an individual flight?

mean (same as median)

Open the Two Quantitative Variables applet (this link has the airfare data or type airfare.txt in the data window and press Use Data).

(e) Is the association linear? Is this a problem? What are some things we can do if we think this (linearity) is a problem?

enough or transform

(e) Check **Show Movable Line** and use the green squares to adjust the line until you find the bit fit line. How are you deciding?

$SAE = \sum |y_i - \hat{y}_i|$ $SSE = \sum (y_i - \hat{y}_i)^2$

(f) What is the *percentage reduction in the unexplained variation* when you include the distance variable in the model?

$1 - \frac{SE(residual)^2}{SD(y)^2} \times \frac{10}{11} = 1 - \frac{SS\,Error}{SS\,Total}$

(g) Check the **Regression SE** box. What does this value tell you?

typical prediction error
SD of residuals

(h) Report and interpret, in context, the intercept. Is this interpretation meaningful? How could we improve the interpretation?

predicted price if distance = 0
like a fixed set up cost?

(i) Report and interpret, in context, the slope coefficient. Is this interpretation meaningful? How could we improve the interpretation?

predicted increase in price for each additional mile (could use per 100 miles)

**To Do:**

- Skim Ch. 2
- Submit Quiz 1 and Complete Computer Problem 1
- Install R and RStudio