TREND ANALYSIS

ERT 474/574

Open-Source Hydro Data Analytics

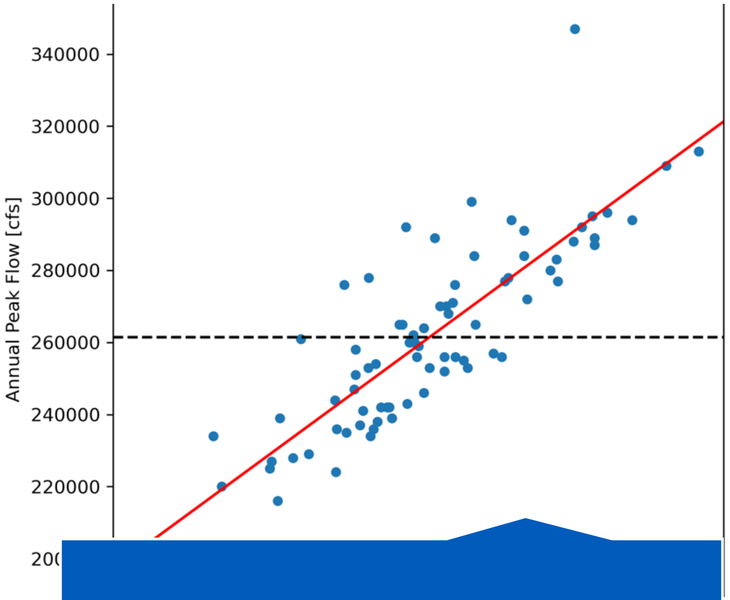
Sep 29th 2025

University at Buffalo The State University of New York

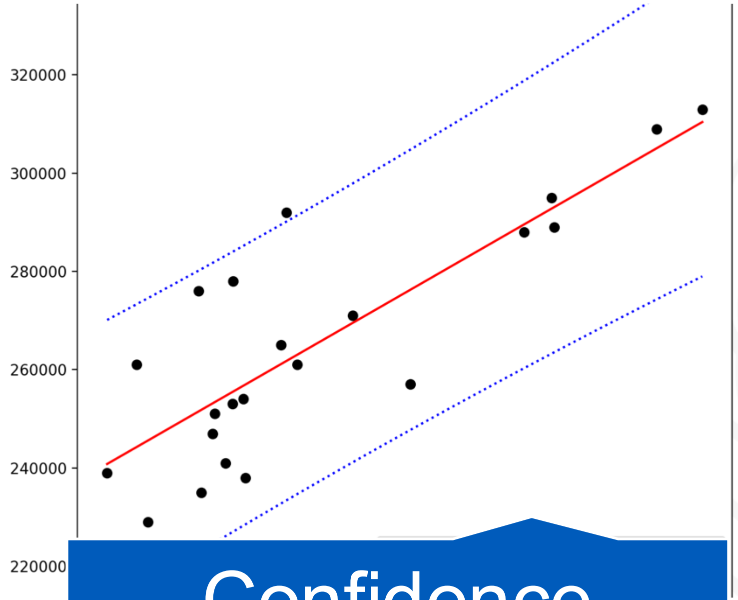
Announcements

- **Homework #3**
  - Due date: Wednesday (Oct 1)

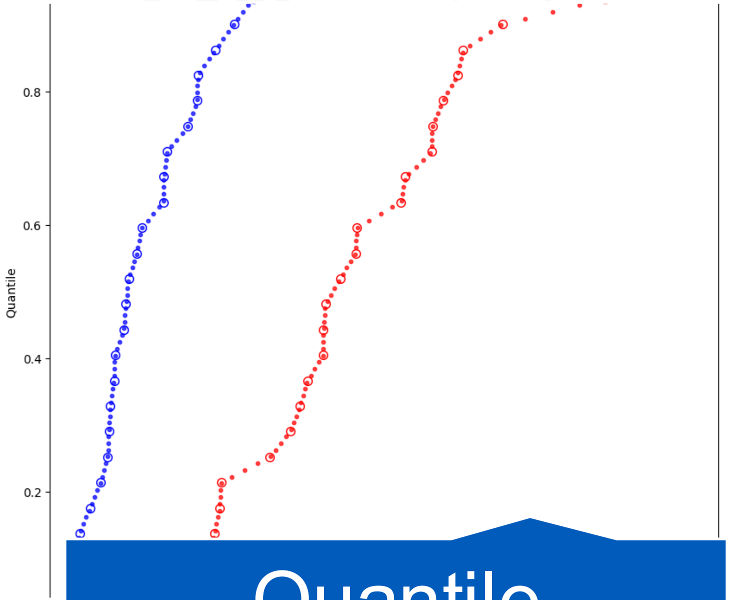
Trend analysis



Linear regression



Confidence intervals



Quantile regression

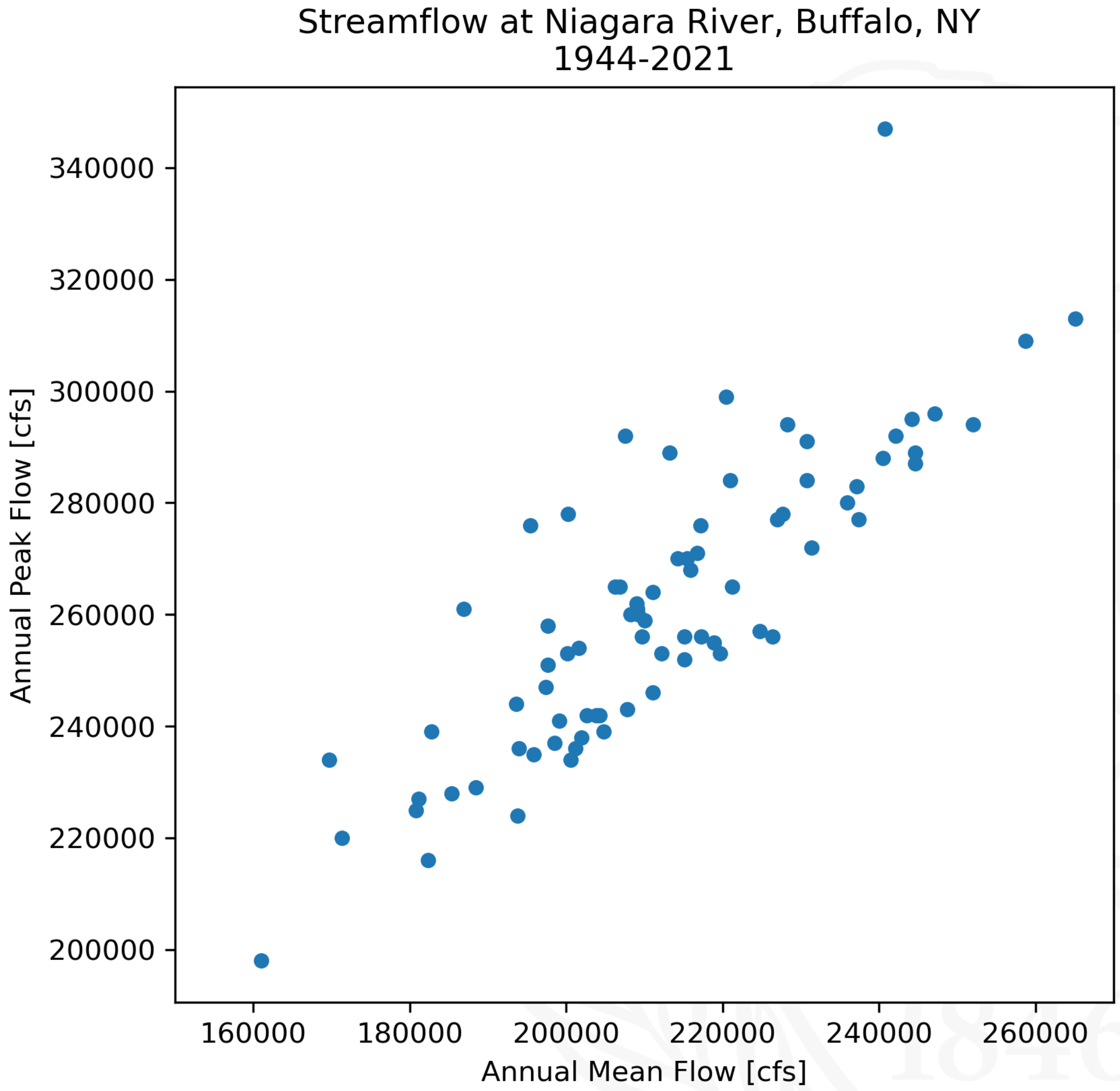
Least Squares Linear Regression

- In this approach we posit a linear relationship between an "independent" or "explanatory" variable x and some "dependent" variable y:

$$y = B_0 + B_1 x$$

- The first step in this process is to check whether a linear model approximation is reasonable. A good way to do this is to make a scatter plot of the available data

Can we use mean annual flow to predict peak flow?



Least Squares Linear Regression

Fitting of parameters

The parameters: B_0 and B_1

are selected so that the sum of the squared errors of the model are minimized for the available data. i.e. minimize:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Taking partial derivatives with respect to B_0 and B_1 and setting equal to zero yields:

$$nB_0 + \left(\sum_{i=1}^{n} x_i\right) B_1 = \left(\sum_{i=1}^{n} y_i\right)$$

$$\left(\sum_{i=1}^{n} x_i\right) B_0 + \left(\sum_{i=1}^{n} {x_i}^2\right) B_1 = \left(\sum_{i=1}^{n} x_i y_i\right)$$

Least Squares Linear Regression

Solving for B_0 and B_1 yields:

$$B_1 = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\left(\sum_{i=1}^{n} {x_i}^2\right) - (\sum_{i=1}^{n} x_i)^2}$$

$$B_0 = \frac{(\sum_{i=1}^{n} y_i) - B_1(\sum_{i=1}^{n} x_i)}{n} = \bar{y} + B_1 \bar{x}$$

$Let\ \hat{y}_i = B_0 + B_1 x_i$

Then the quantity $(y_i - \hat{y}_i)$ is called the "ith residual" .

Least Squares Linear Regression

SSE = Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

SST = Total Sum of Squares

How much variance is there about the mean.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Standard Error

$$\sigma^2 = s^2 = \frac{SSE}{(n-2)}$$

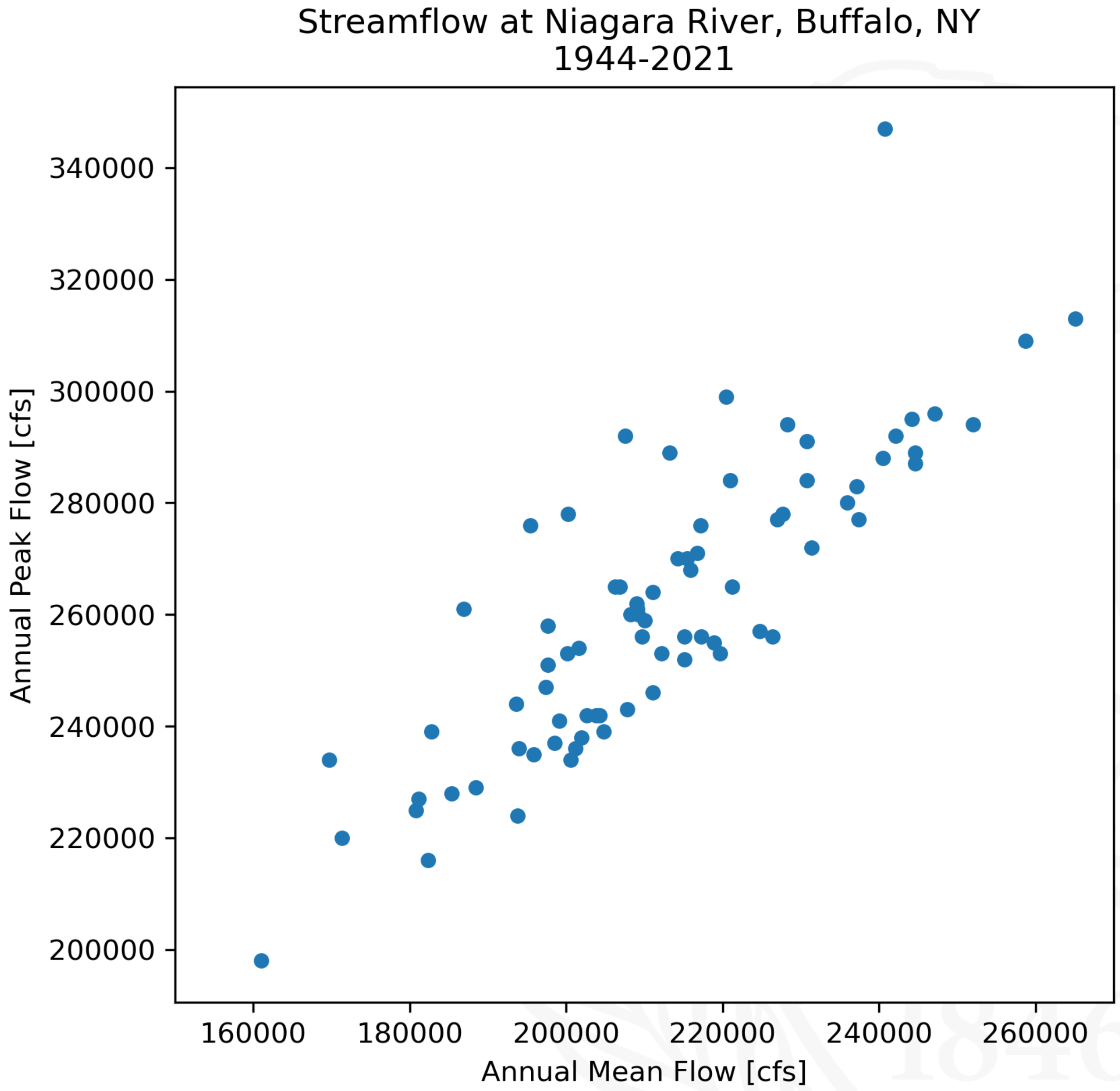
$$\sigma = \sqrt{\frac{SSE}{(n-2)}}$$

Correlation Coefficient

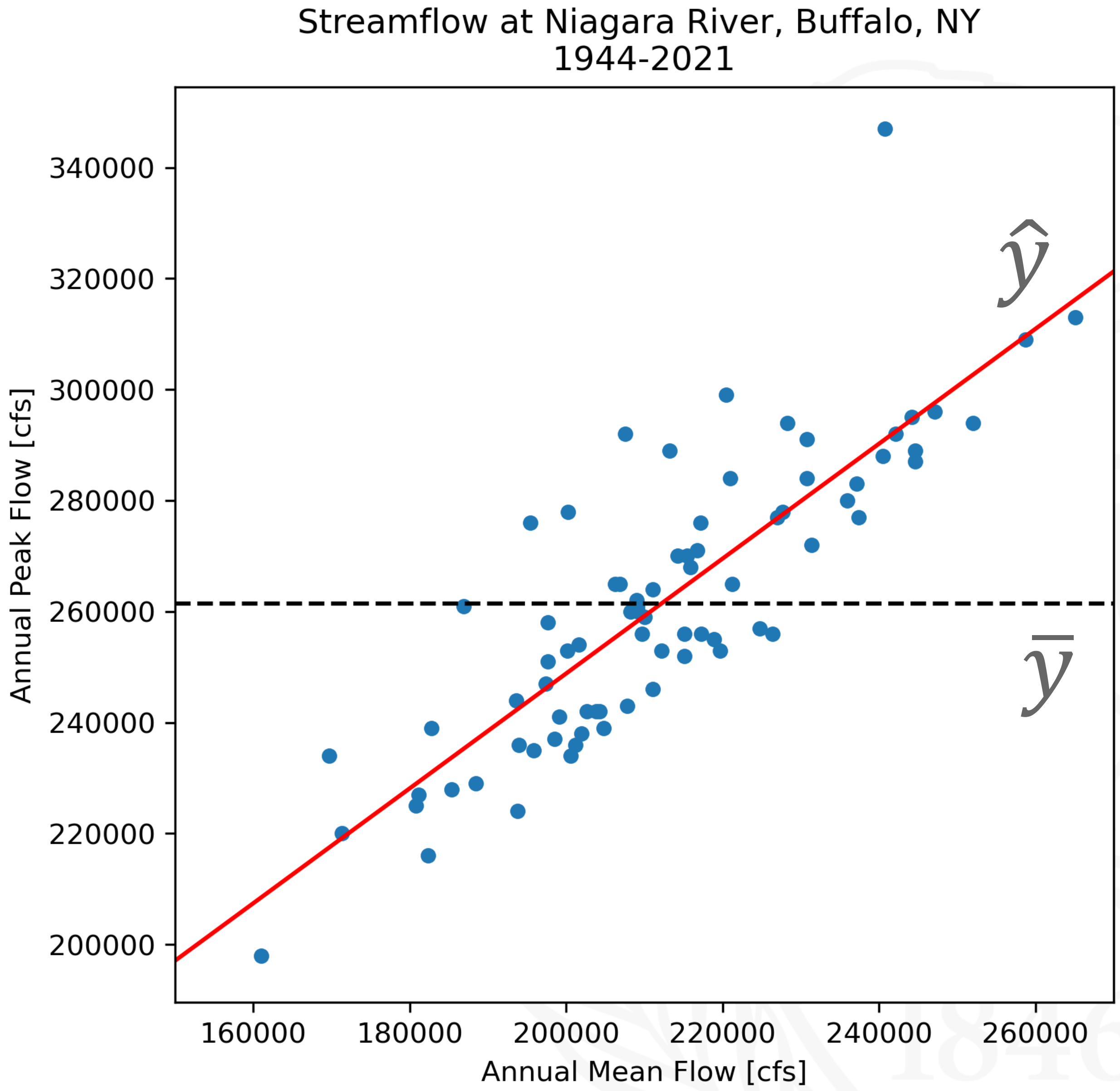
(Variance explained by the model)

$$R^2 = 1 - \frac{SSE}{SST}$$

Can we use mean annual flow to predict peak flow?

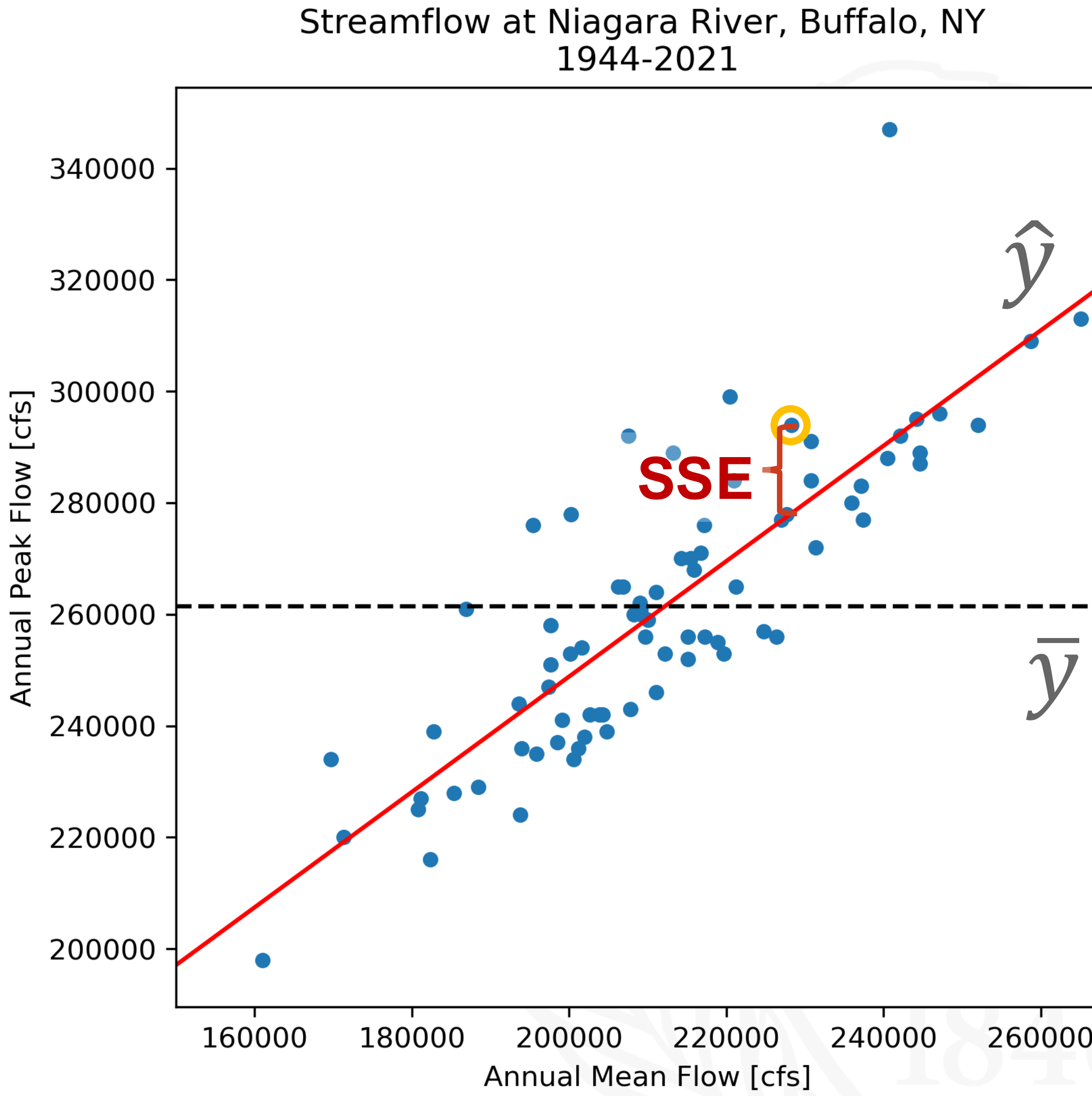


Can we use mean annual flow to predict peak flow?



Can we use mean annual flow to predict peak flow?

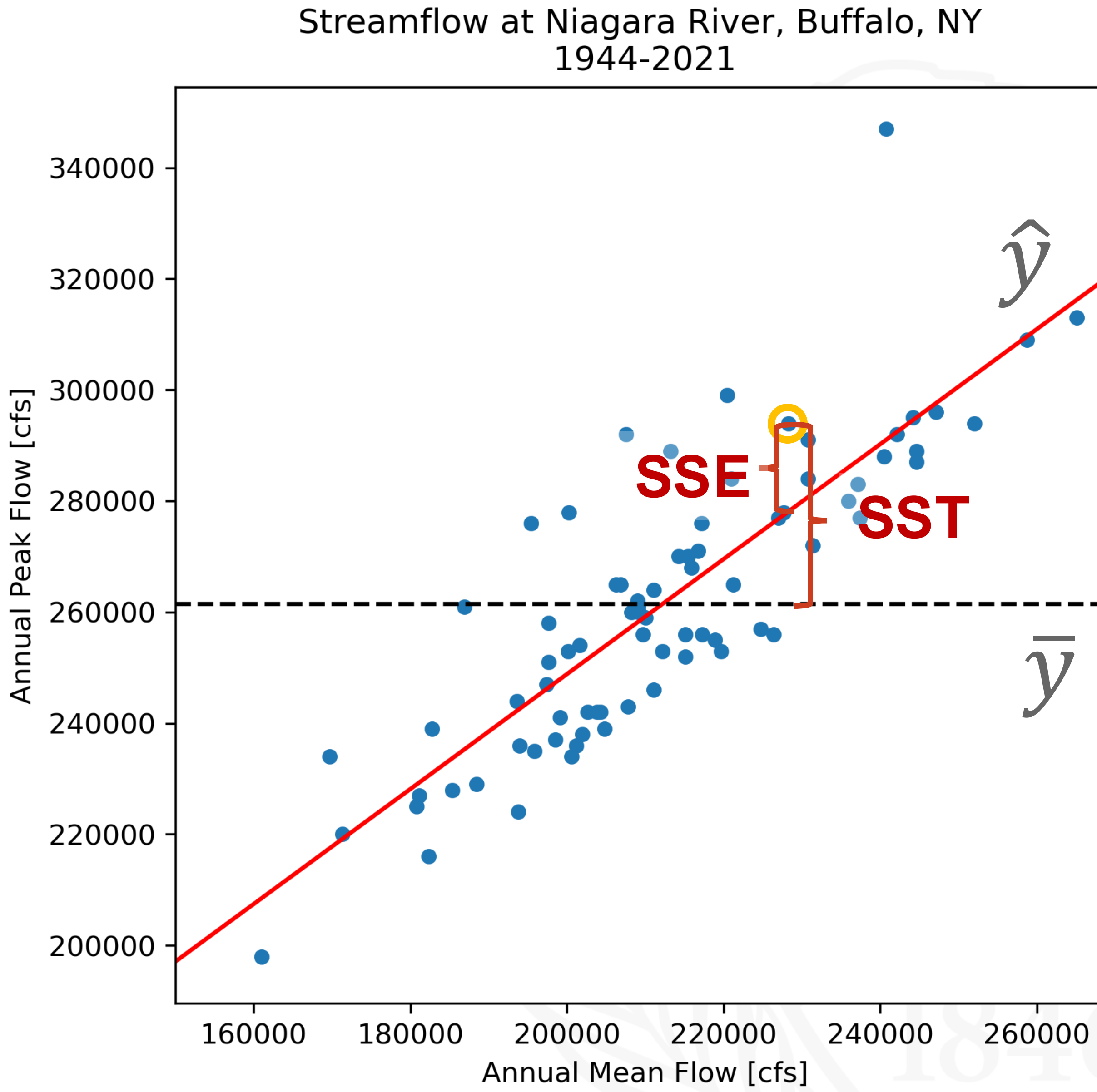
$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Can we use mean annual flow to predict peak flow?

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

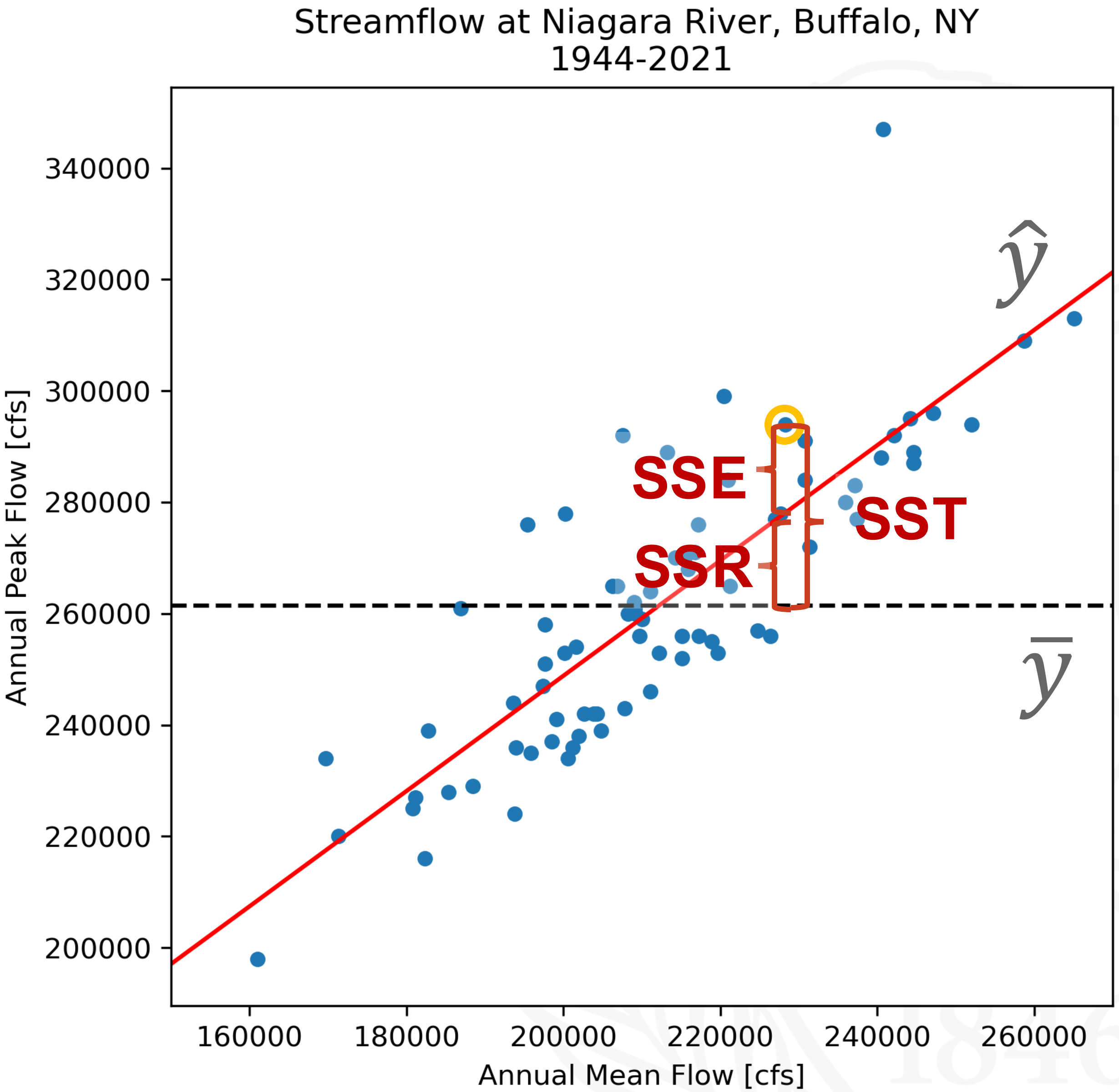


Can we use mean annual flow to predict peak flow?

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

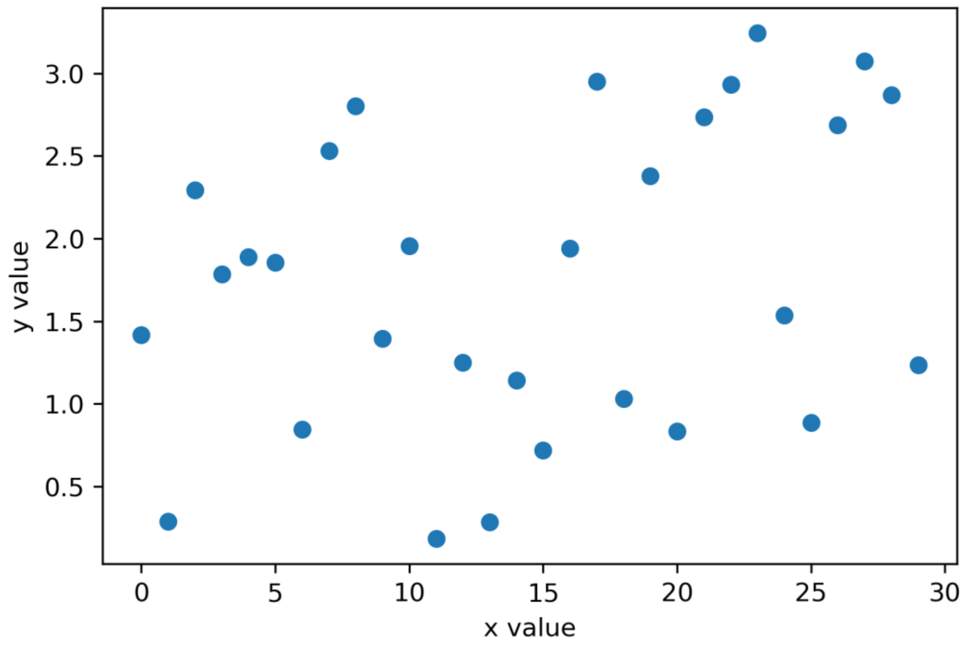
$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

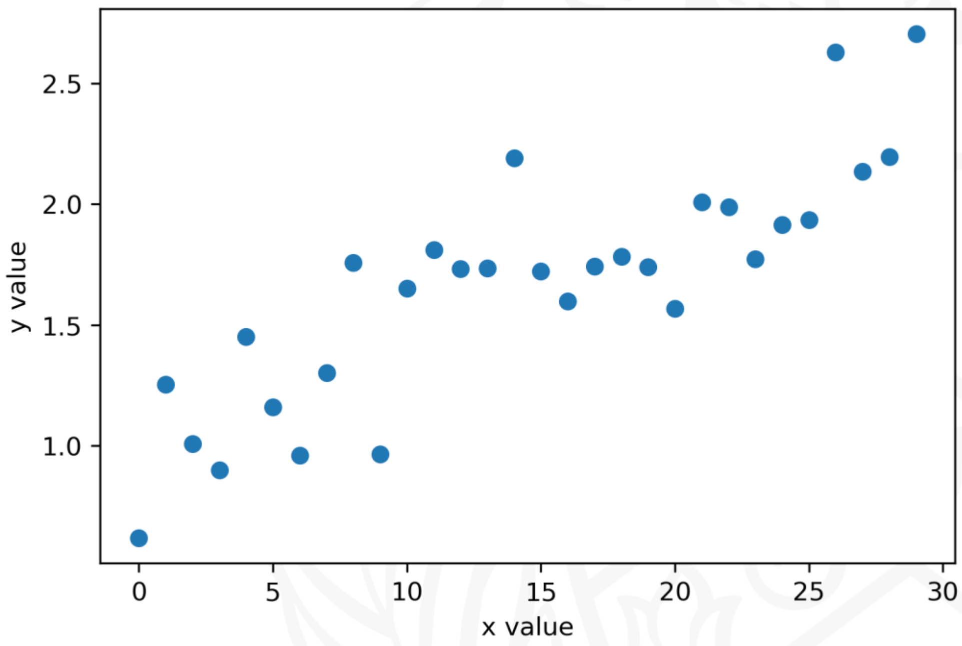
$$SSR = SST - SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$



Confidence Bounds on Regression Parameters

If you see scatter plots like below, how confident are you to claim that there is an increasing trend?





Confidence Bounds on Regression Parameters

- The variance of the regression parameter $\hat{B}_1$ is a function of the standard error *and* the “spread” of the x values.

$$S_{B_1}{}^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Confidence Bounds on Regression Parameters

- The variance of the regression parameter $\hat{B}_1$ is a function of the standard error *and* the “spread” of the x values.

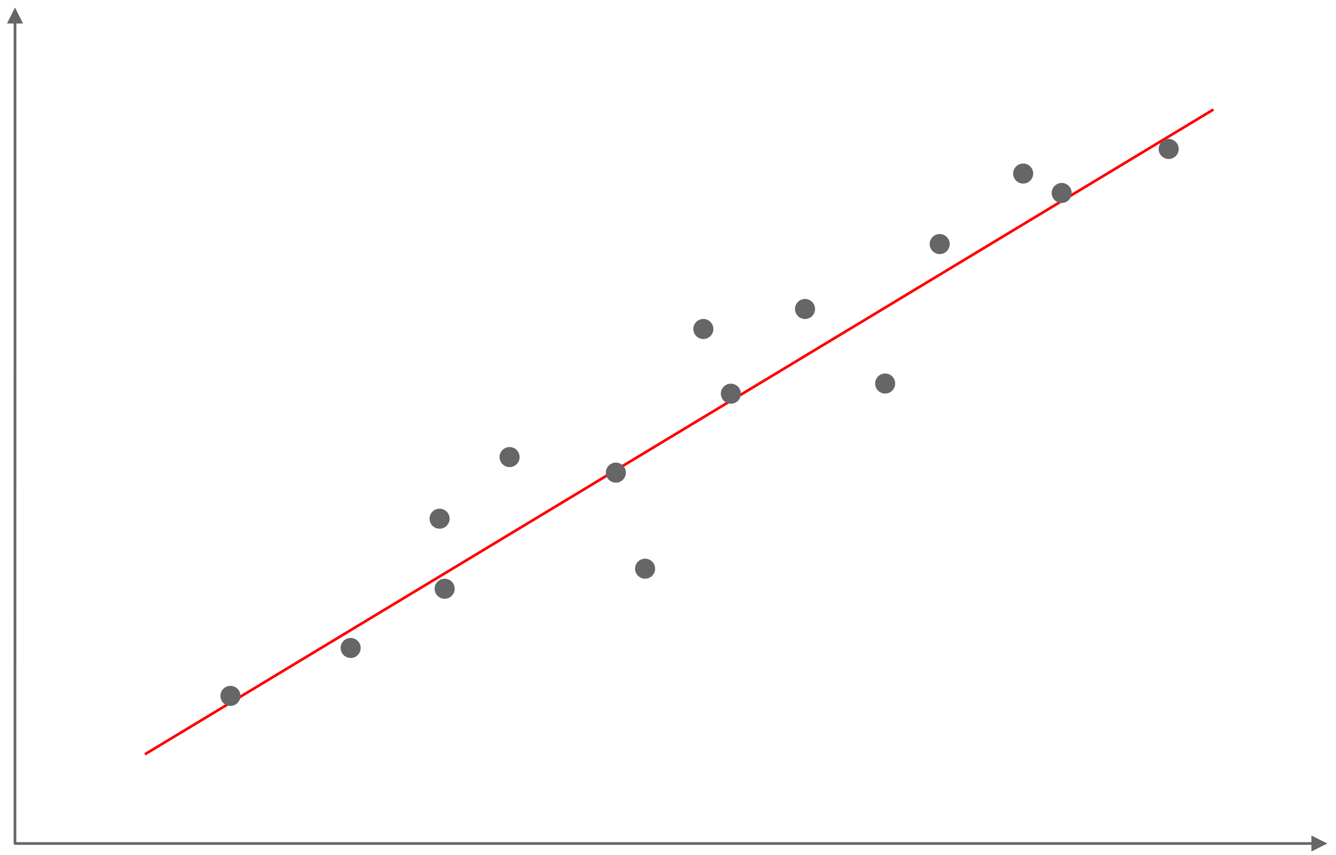
$$S_{B_1}{}^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- And $\frac{(\hat{B}_1 - B_1)}{S_{B_1}}$ is T distributed with n-2 degrees of freedom.

- So a confidence interval for B_1 is: $\hat{B}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot S_{B_1}$

What do the confidence bounds on the B1 Parameter look like?

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x$$

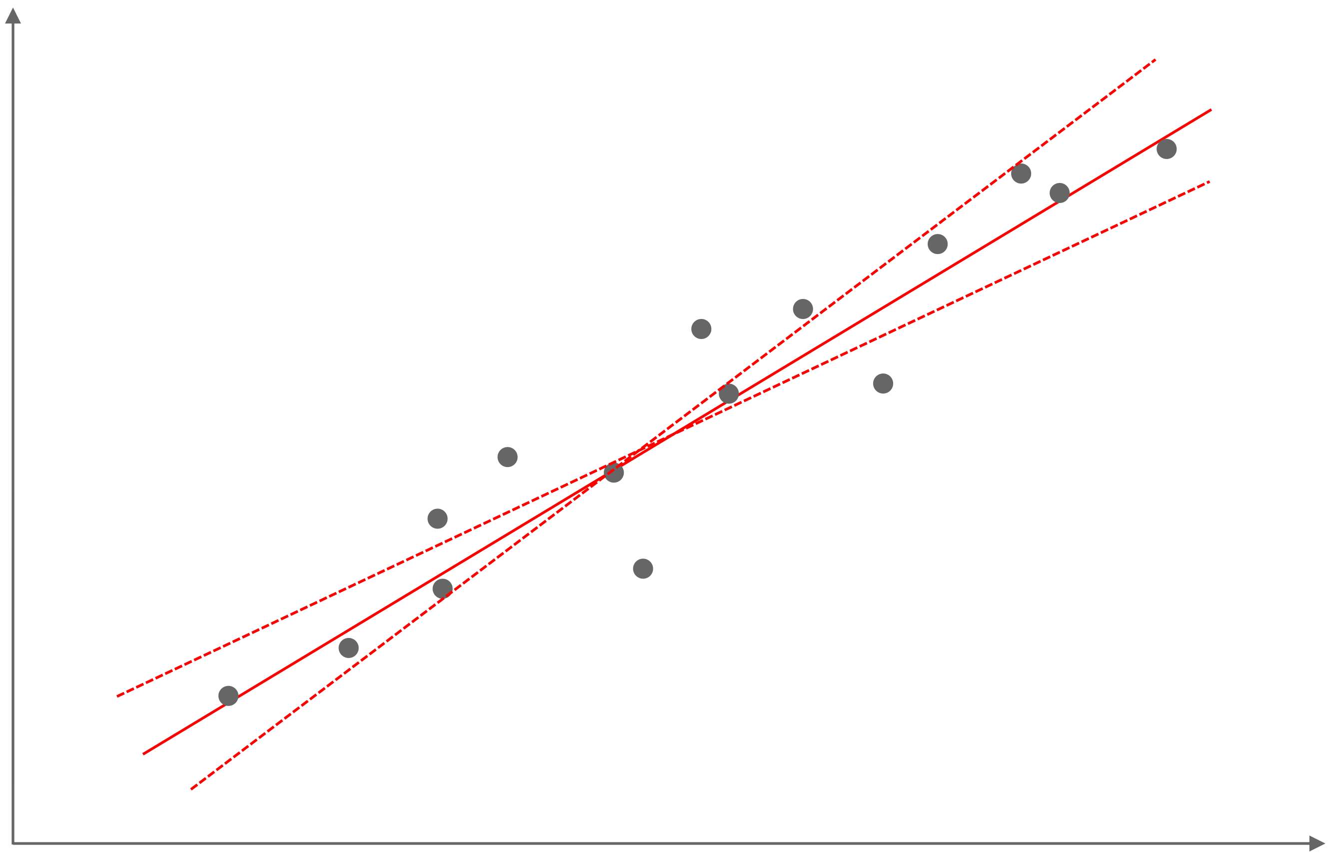


$$\hat{B}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot S_{B_1}$$

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x$$

What do the confidence bounds on the B1 Parameter look like?

And what's this point where the B1 slope values are pivoting?

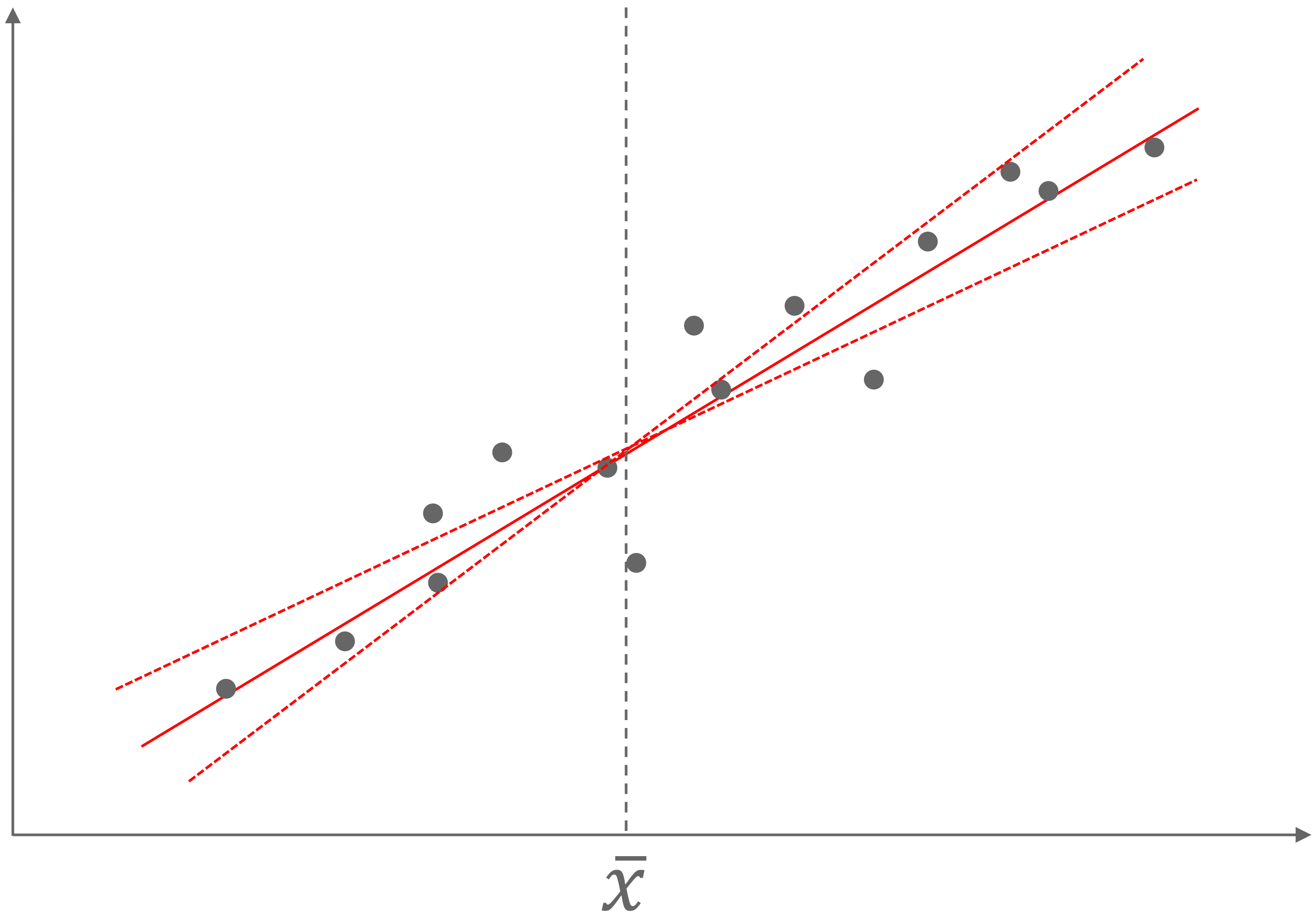


$$\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot S_{B_1}$$

What do the confidence bounds on the B1 Parameter look like?

And what's this point where the B1 slope values are pivoting?

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x$$



$$\hat{B}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot S_{B_1}$$

Hypothesis test for the estimator $\hat{B}_1$

Asking, "Does my regression line really have a slope of B_1?"

Null Hypothesis: $\hat{B}_1 = B_1$

α = Significance level (1– confidence level), number of degrees of freedom = (n-2)

Test statistic: $t = \frac{(\hat{B}_1 - B_1)}{s_{B_1}}$

Alternate Hypothesis:

- $\hat{B}_1 > B_1$
- $\hat{B}_1 < B_1$
- $\hat{B}_1 \neq B_1$

Rejection Region:

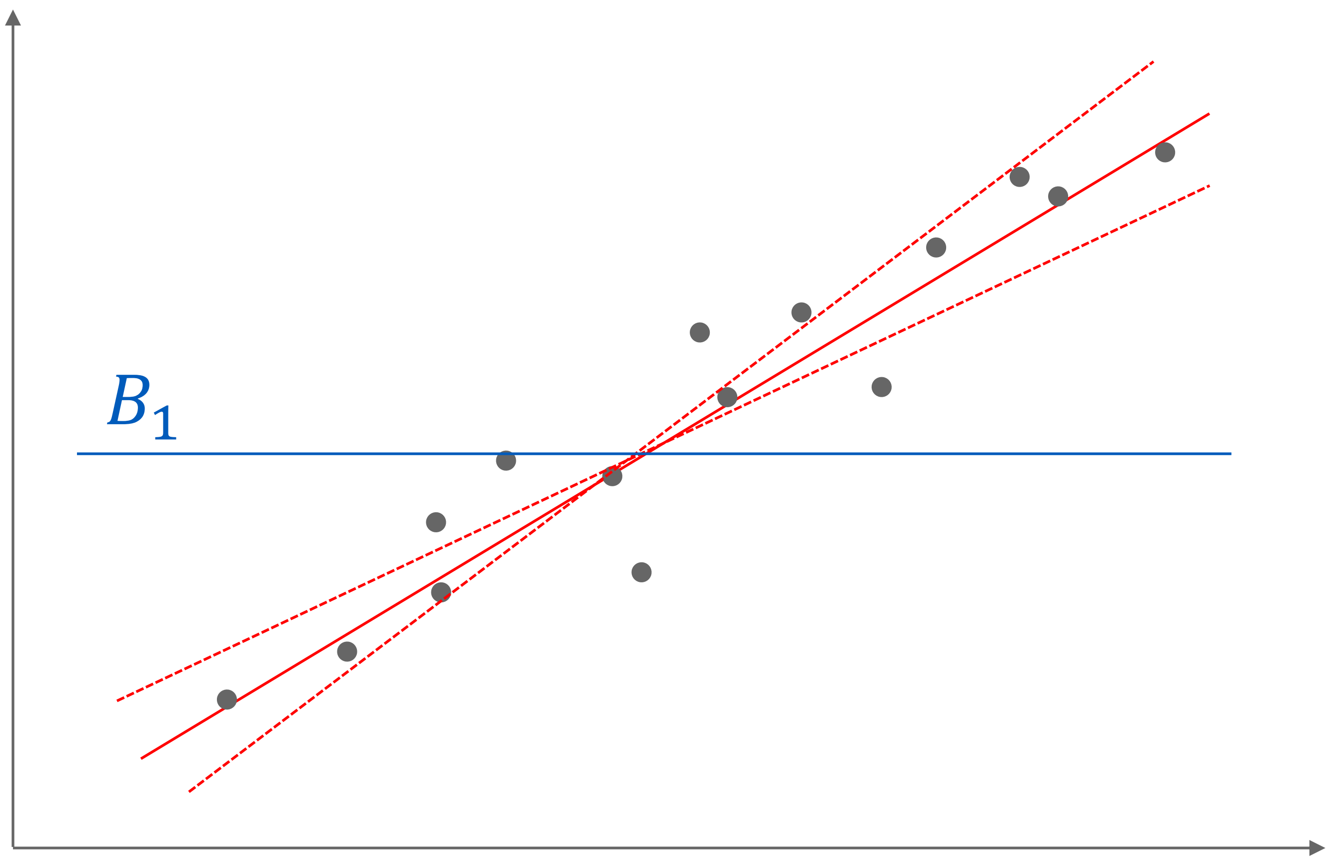
- $t \geq t_{\alpha,n-2}$
- $t \leq -t_{\alpha,n-2}$
- $t \leq -t_{\frac{\alpha}{2},n-2}$ or $t \geq t_{\frac{\alpha}{2},n-2}$

Example: Can we reject the null hypothesis?

H_0: $\hat{B}_1 = B_1 = 0$

H_A: $\hat{B}_1 \neq B_1$

$t \leq -t_{\frac{\alpha}{2},n-2}$ or $t \geq t_{\frac{\alpha}{2},n-2}$



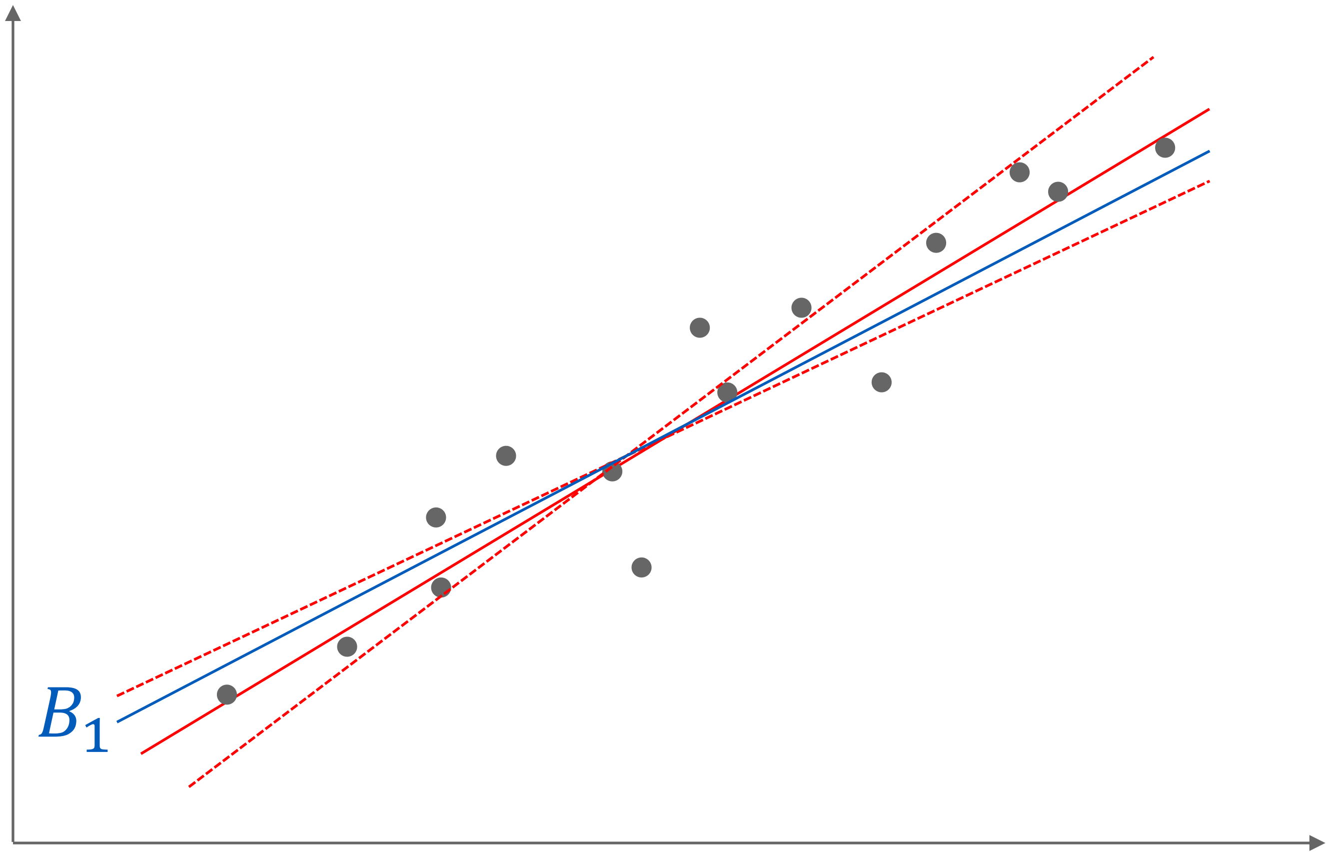
$$\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot S_{B_1}$$

Example:
Can we reject the null hypothesis?

H_0: $\hat{B}_1 = B_1$

H_A: $\hat{B}_1 \neq B_1$

$t \leq -t_{\frac{\alpha}{2},n-2}$ or $t \geq t_{\frac{\alpha}{2},n-2}$



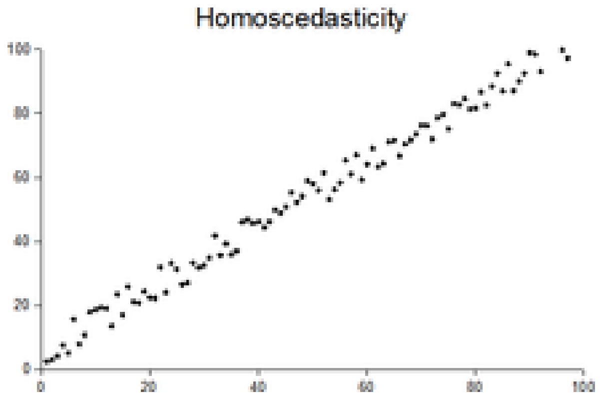
$$\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot S_{B_1}$$

Estimating the Trend Using a Least Squares Linear Model

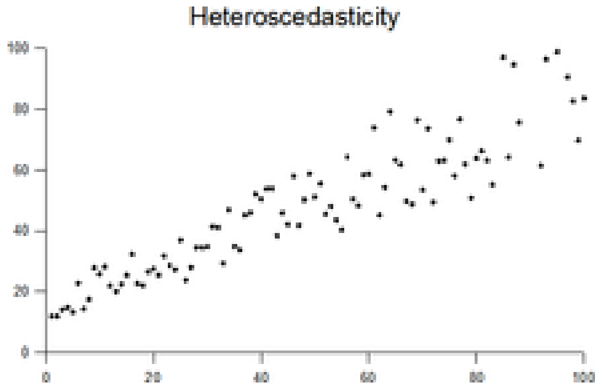
Some conditions that should be met for good results (see Helsel et al. for more)

- Data should not be strongly auto correlated
- There shouldn't be any dramatic expansion in the variance over x.
- A linear model should fit reasonably well (use a scatter plot to confirm)
- The residuals for the linear model should be approximately normally distributed and shouldn't have large trends in them (plot these to get a sense of whether there are problems).

Homoscedasticity: random variables in a sequence have the same finite variance.



Heteroscedasticity: subpopulations have different variance from others.



Graphs from wikipedia: https://en.wikipedia.org/wiki/Heteroscedasticity

Estimating the Trend Using a Least Squares Linear Model

Some conditions that should be met for good results (see Helsel et al. for more)

- Data should not be strongly auto correlated
- There shouldn't be any dramatic expansion in the variance over x.
- A linear model should fit reasonably well (use a scatter plot to confirm)
- The residuals for the linear model should be approximately normally distributed and shouldn't have large trends in them (plot these to get a sense of whether there are problems).

Procedures:

- Calculate B_1 (the trend) in the normal manner. (What are the units?)
- Use hypothesis tests on B_1 to see whether the trend is significantly different from 0 (i.e. no trend).
- Use the confidence interval around the estimate of B_1 to express the uncertainty in the trend.

Confidence Bounds for the Predicted Values of Y

Constructing an Interval for the Predicted Values of Y

For some value x^* we want to predict a corresponding y^* using our model

$$\hat{y}^* = \hat{B}_0 + \hat{B}_1 x^*$$

The error of our prediction is the difference between the “true” value of y^* for x^*, and our predicted $\hat{y}^*$:

$$(B_0 + B_1 x^*) - \left(\hat{B}_0 + \hat{B}_1 x^*\right)$$

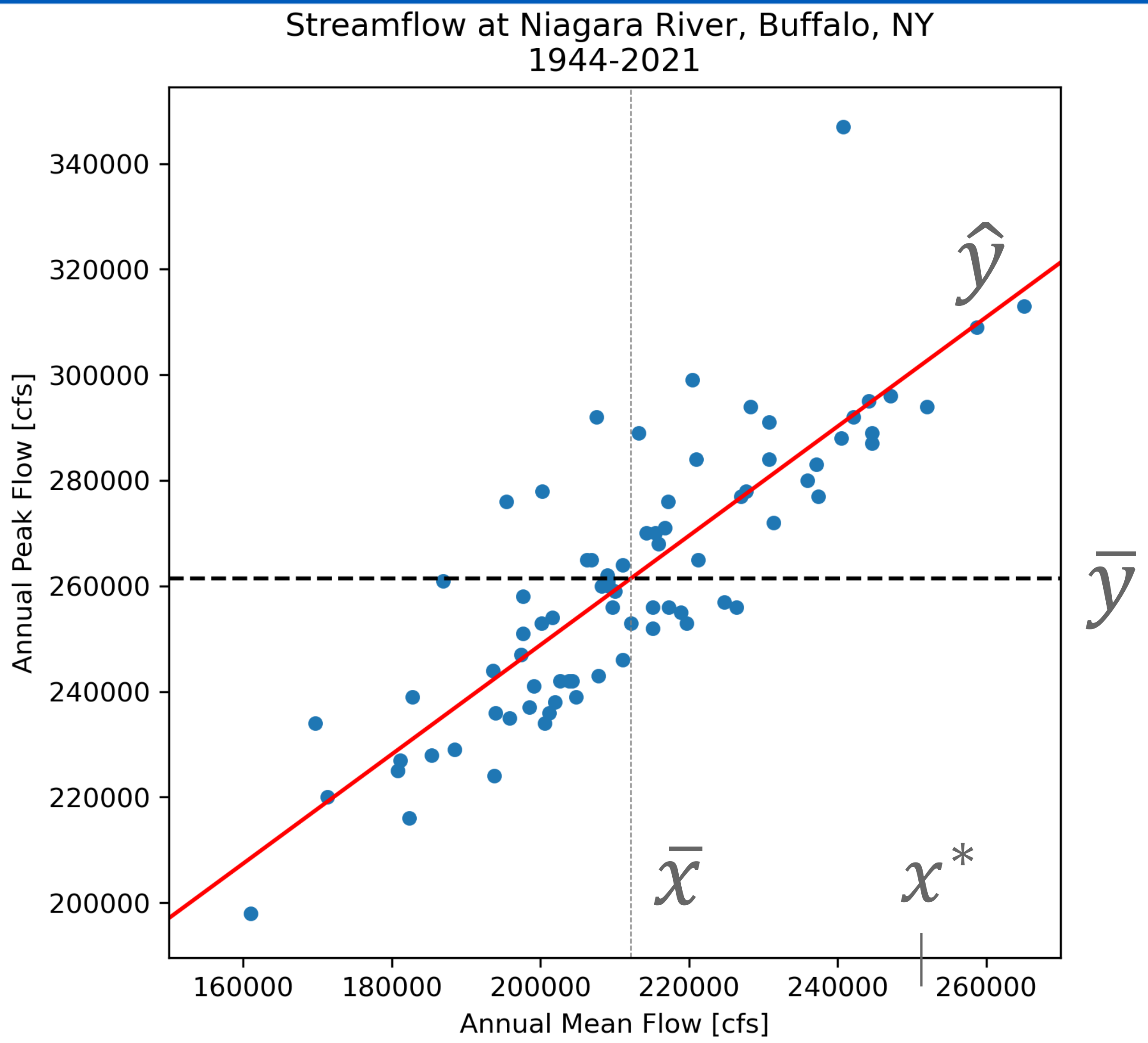
The variance of this prediction error ($\sigma^2_{E_P}$) will help define our predicted intervals,

$$\sigma^2_{E_P}(x^*) = s^2 \left[1 + \frac{1}{n} + \frac{n(x^* - \bar{x})^2}{n\sum x_i^2 + (\sum x_i)^2}\right]$$

$$\sigma^2_{E_P}(x^*) = s^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SST_x}\right]$$

(see Regression_conf_intervals_Devore_p483.pdf or https://www.ma.utexas.edu/users/mks/statmistakes/CIvsPI.html)

Constructing an Interval for the Predicted Values of Y



Constructing an Interval for the Predicted Values of Y

The combined variance of the error of prediction at x^* can be shown to be:

$$\sigma_{E_P}^2(x^*) = var(y - y^*) = s^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SST_x}\right]$$

Note: $\bar{x}$ and x_i refer to the ORIGINAL data used to make the model.
s is the original standard error.

And the statistic:

$$T = \frac{(y - y^*)}{\sigma_{E_P}(x^*)}$$

has a t-distribution with n-2 degrees of freedom.

Constructing an Interval for the Predicted Values of Y

Thus a (1 – α) prediction interval for

y at an arbitrary value of x^* is:

$$y^* \pm t_{\frac{\alpha}{2},n-2} \cdot \sigma_{E_P}(x^*)$$

Note that the uncertainty is a function of x^* and the farther away from $\bar{x}$ we find ourselves the larger the uncertainty in the prediction of y!
(Key thing to remember, these are not constant and vary with the location you want to predict.)

That’s it for today!