TREND ANALYSIS

GLY606 Water Data Analysis & Modeling

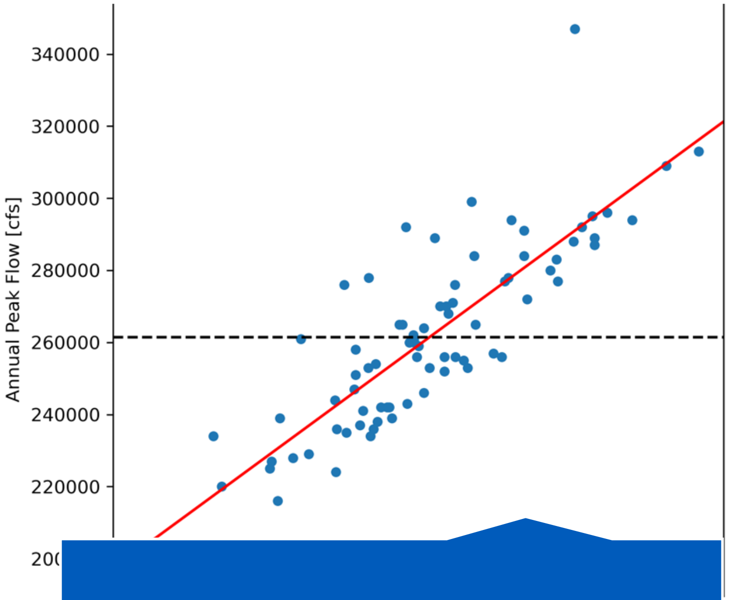
Oct 2nd 2024

University at Buffalo The State University of New York

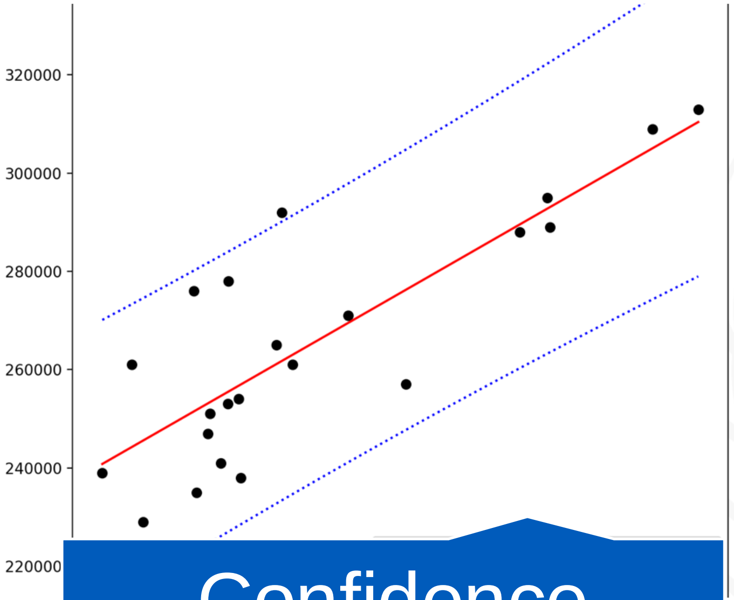
Announcements

- **No in-class lecture this Friday.**
- Instead, you will be assigned practice Jupyter notebooks, that will go over the trend analysis we will cover today.
  - The notebooks will also cover the practices for hypothesis testing.
  - You will need to finish the exercises in the notebooks, save them as HTML files, and submit them through UBLearns, just like homework.
  - Due date: 1 pm, Oct 11th 2024 (Friday)
- We will not have a separate assignment this week ☺

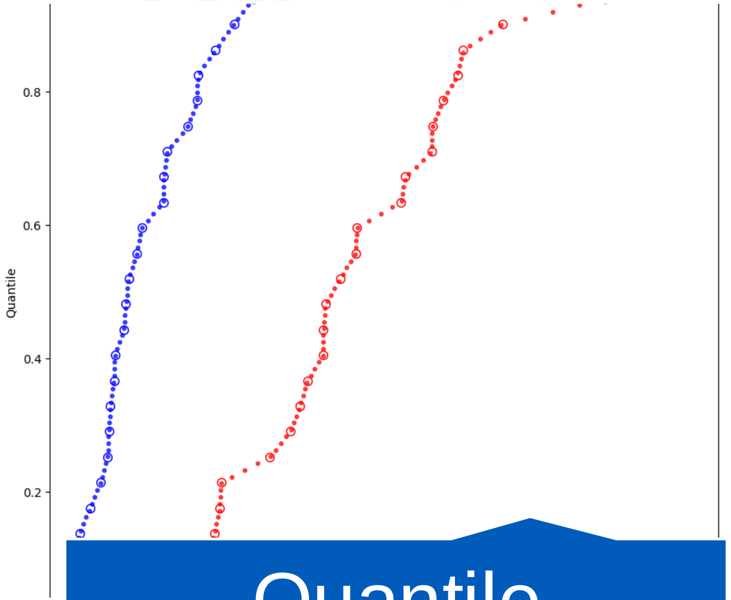
Trend analysis



Linear regression



Confidence intervals



Quantile regression

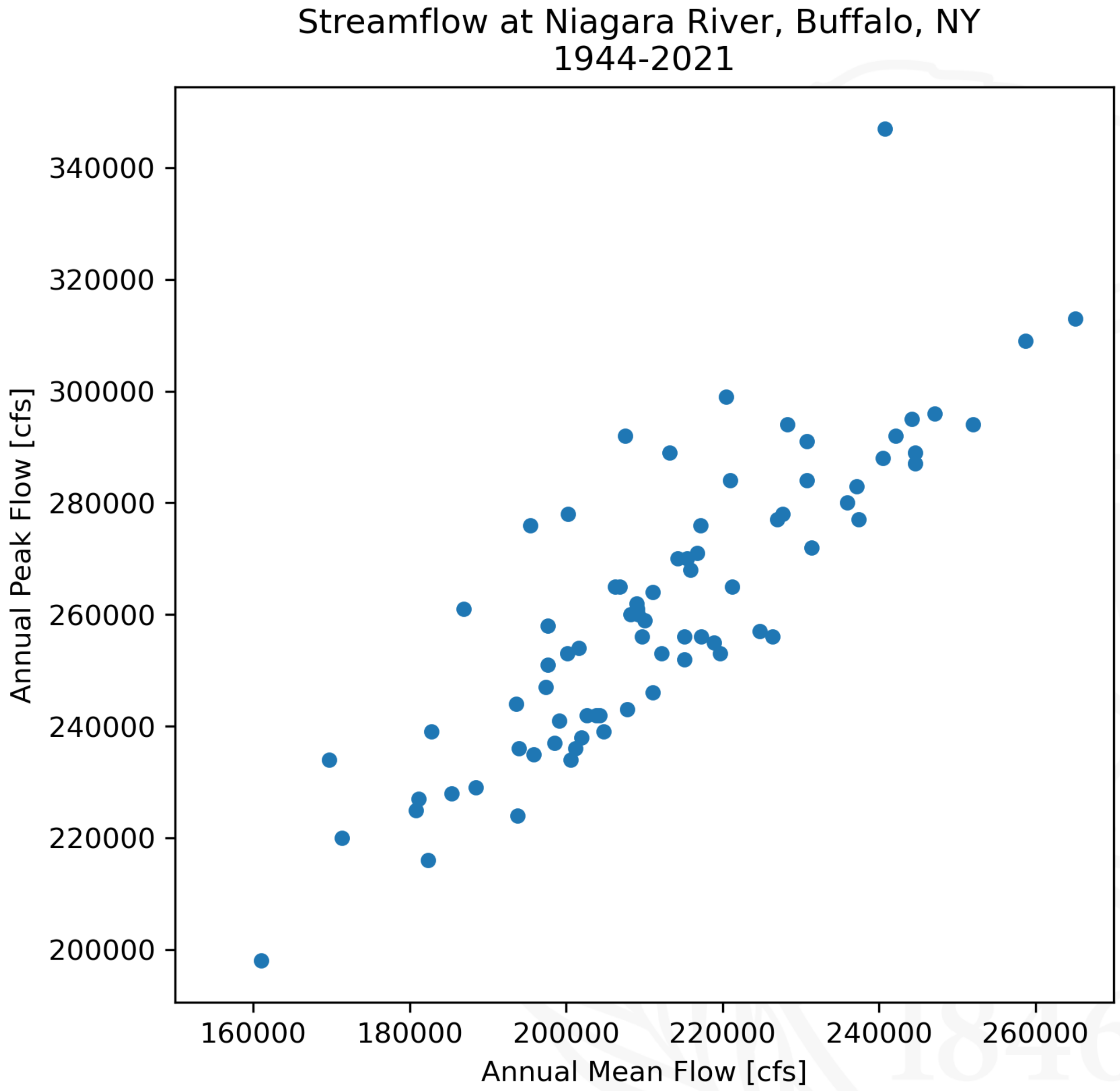
Least Squares Linear Regression

- 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

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

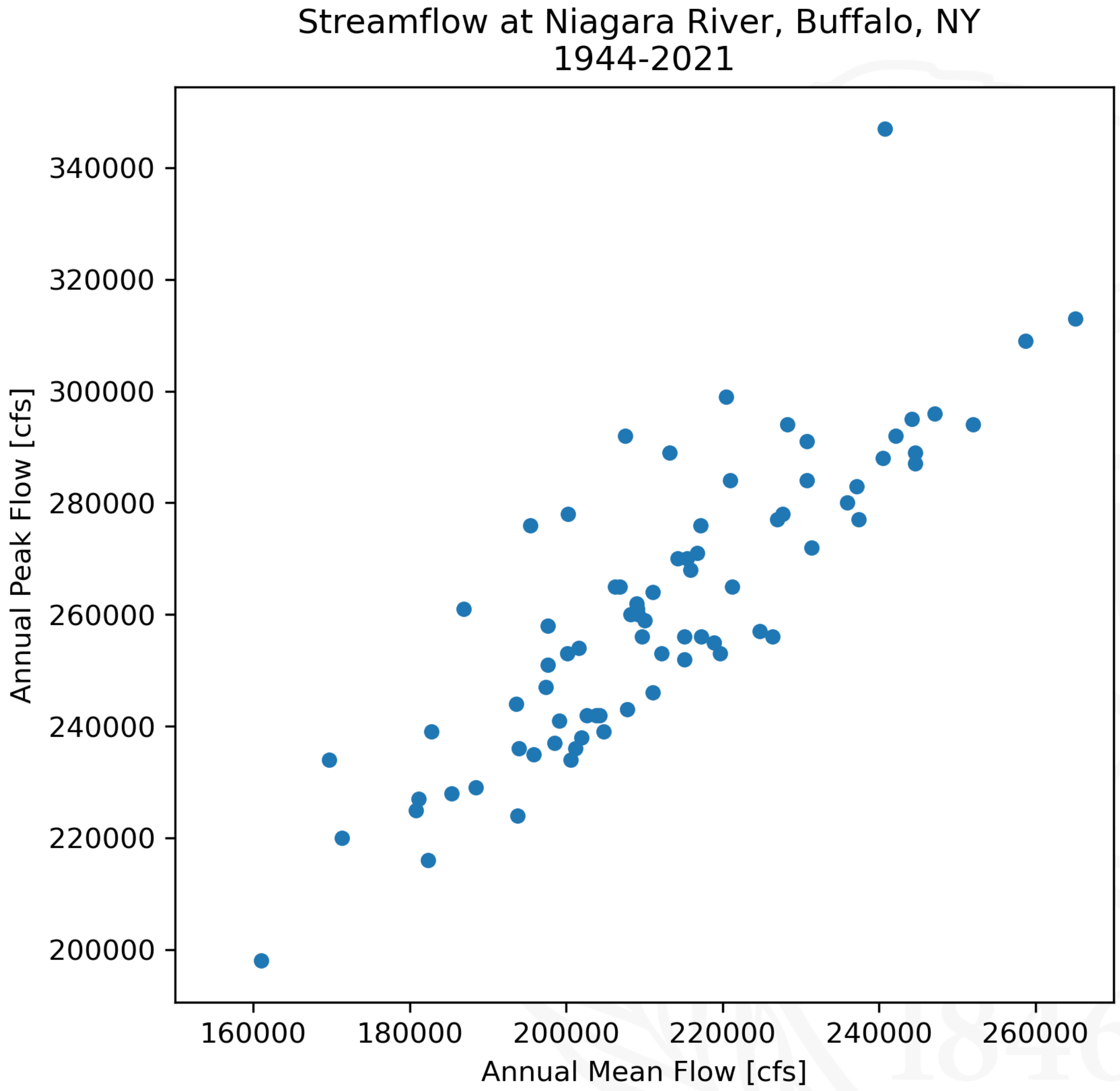
$$\sigma = \sqrt{\frac{SSE}{(n-2)}}$$

Correlation Coefficient

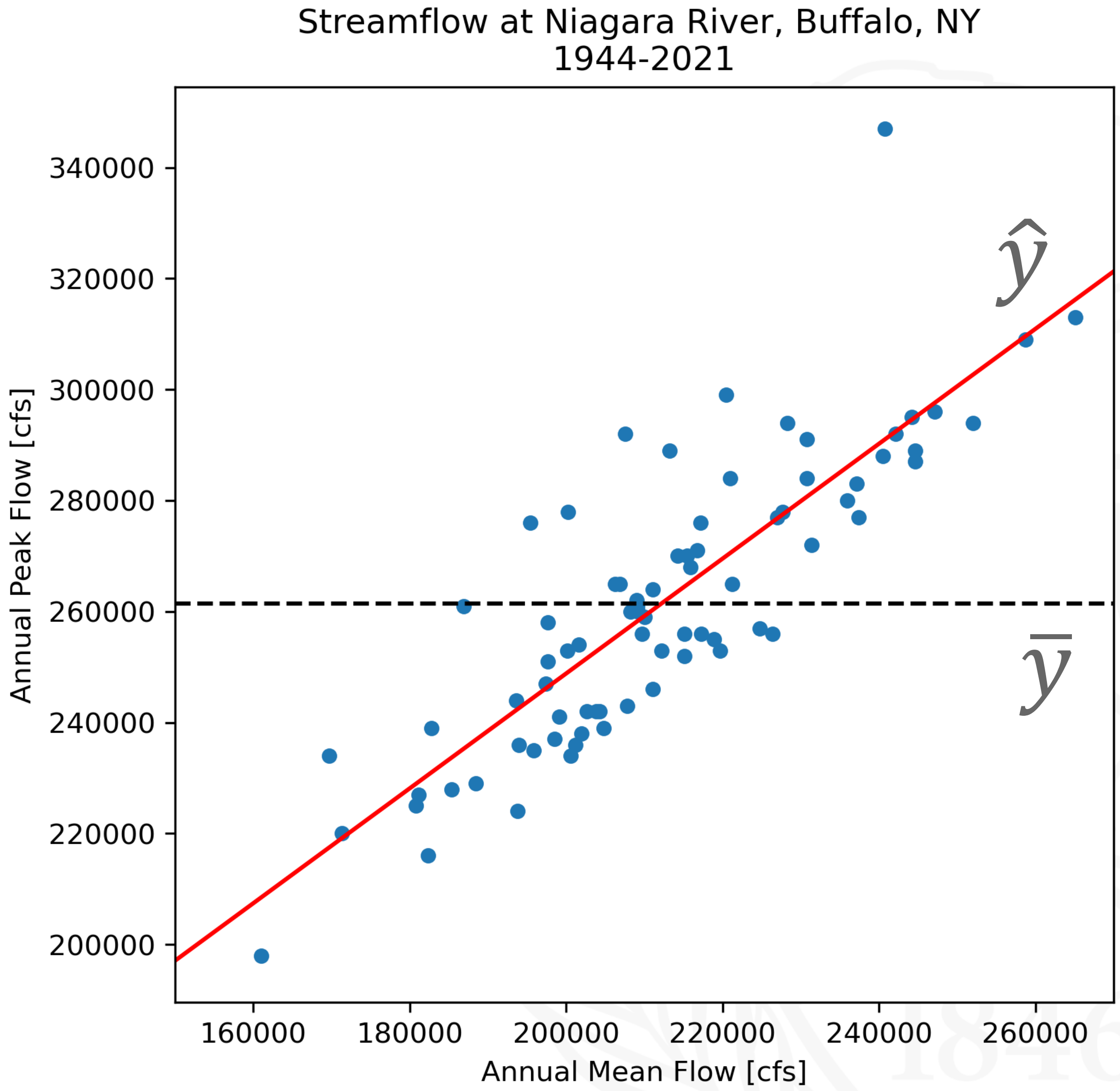
(Variance explained by the model)

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

Can we use mean annual flow to predict peak flow?

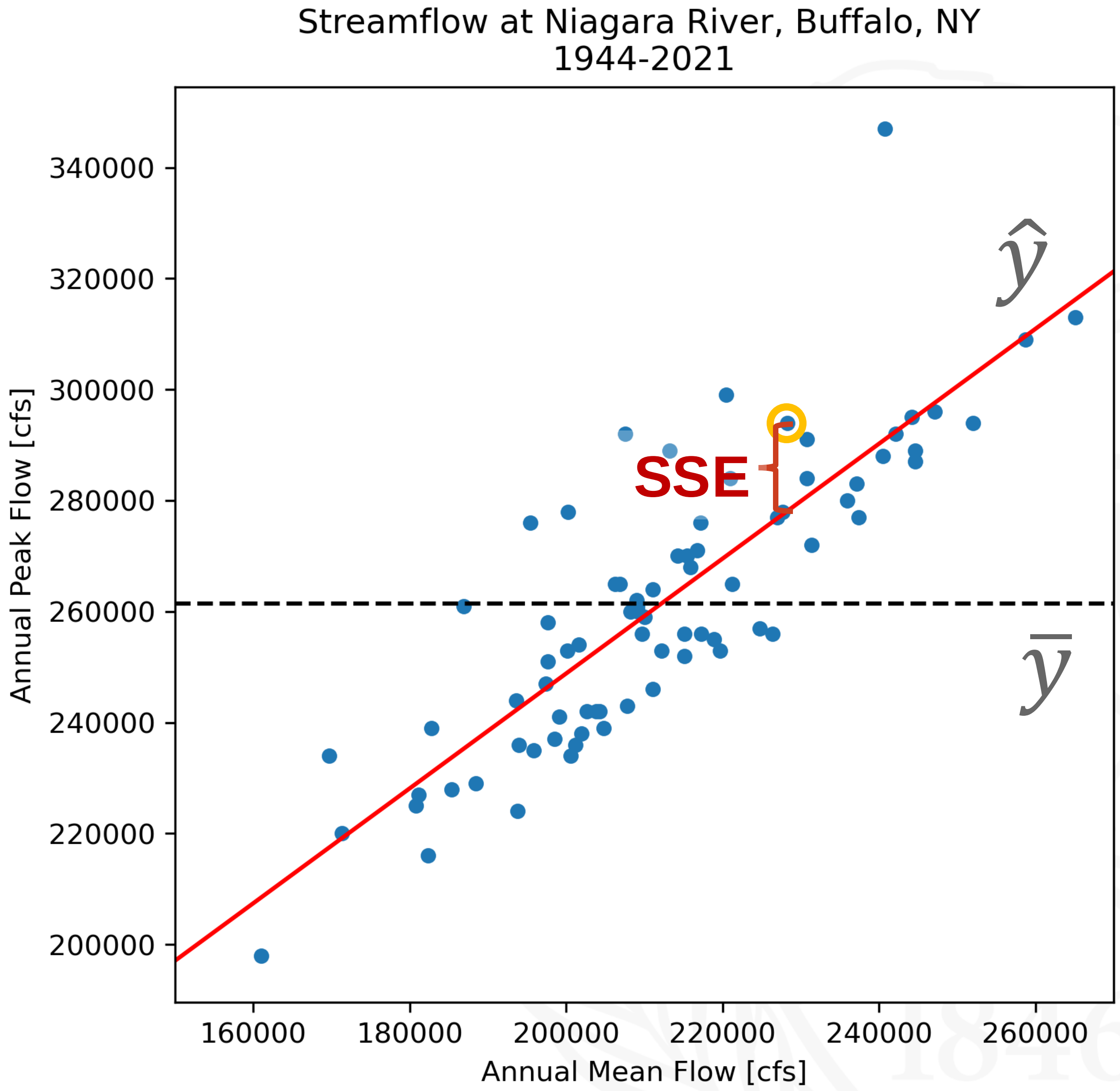


Can we use mean annual flow to predict peak flow?



Can we use mean annual flow to predict peak flow?

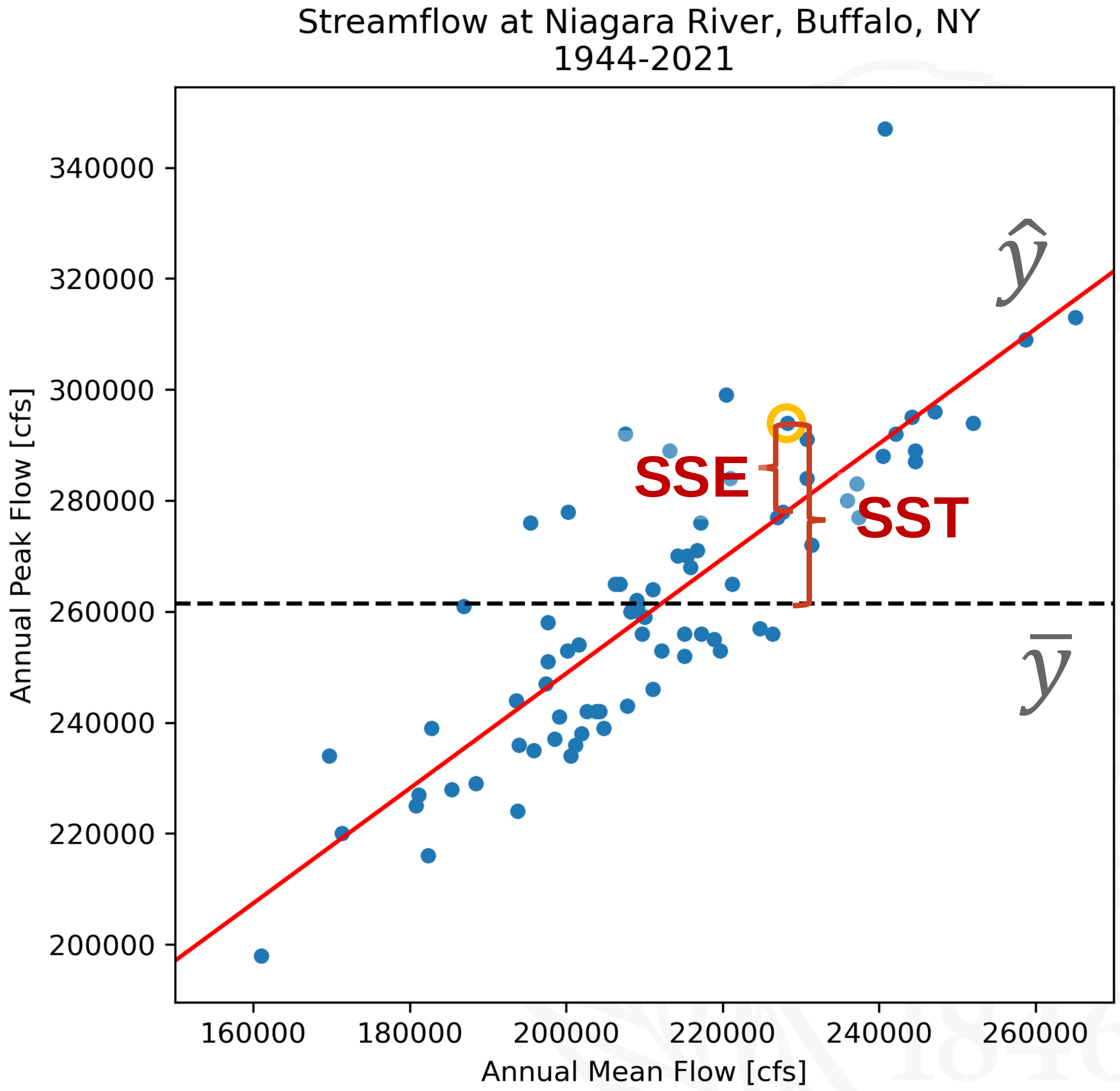
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



Can we use mean annual flow to predict peak flow?

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

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

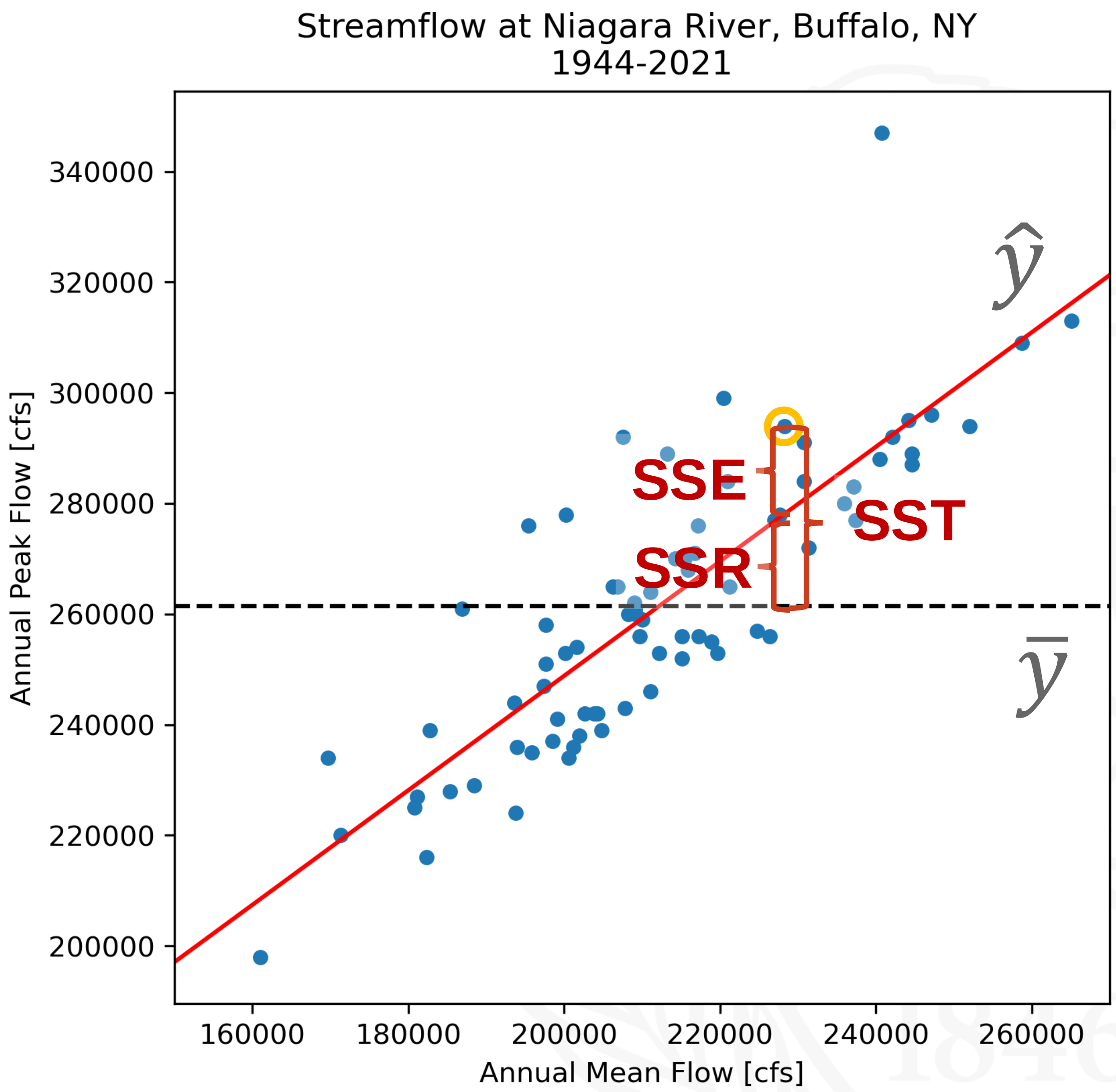


Can we use mean annual flow to predict peak flow?

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

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

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



Confidence Bounds on Regression Parameters

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.

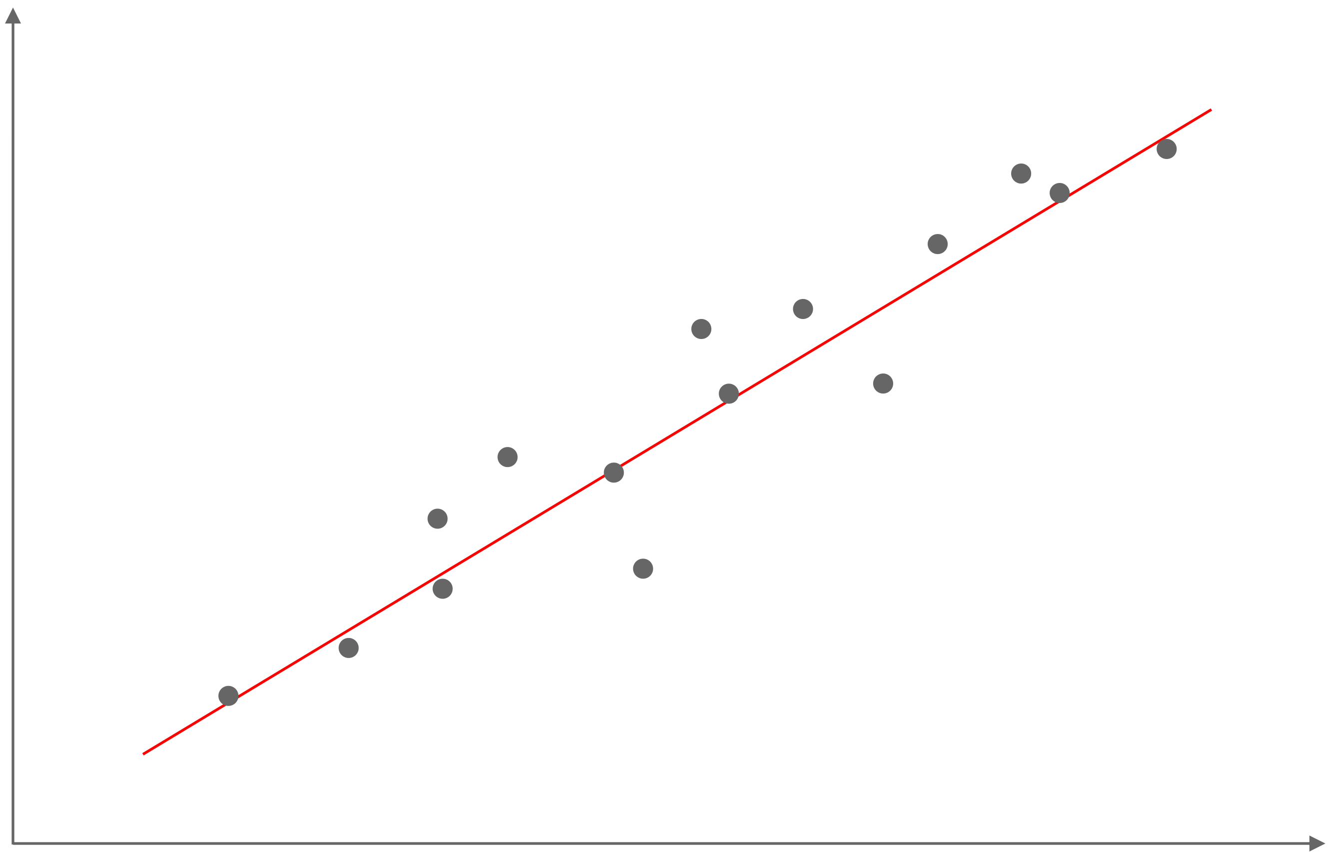
$$S_{B_1}{}^2 = \frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 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?

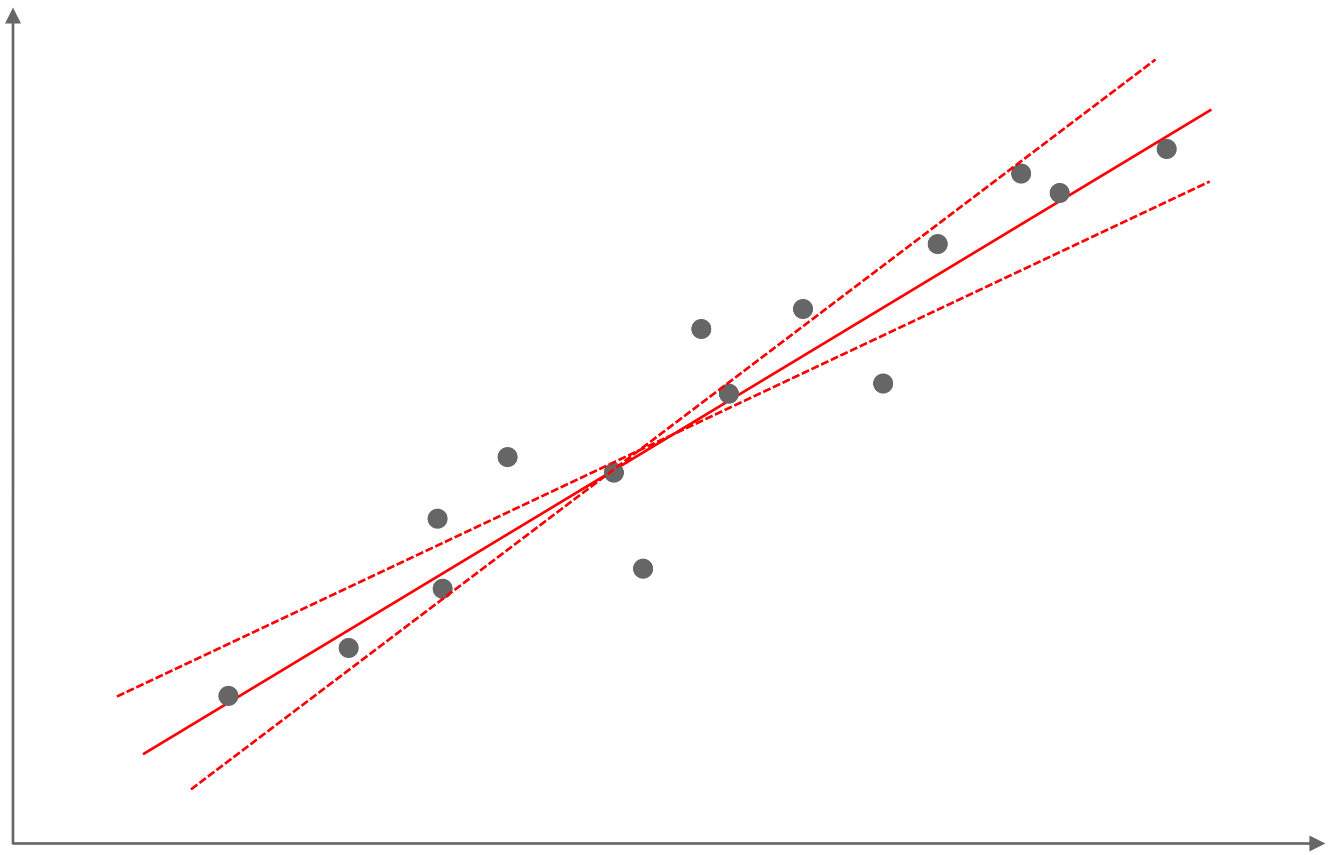
$$\hat{y} = \hat{B}_0 + \hat{B}_1 x$$



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

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

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



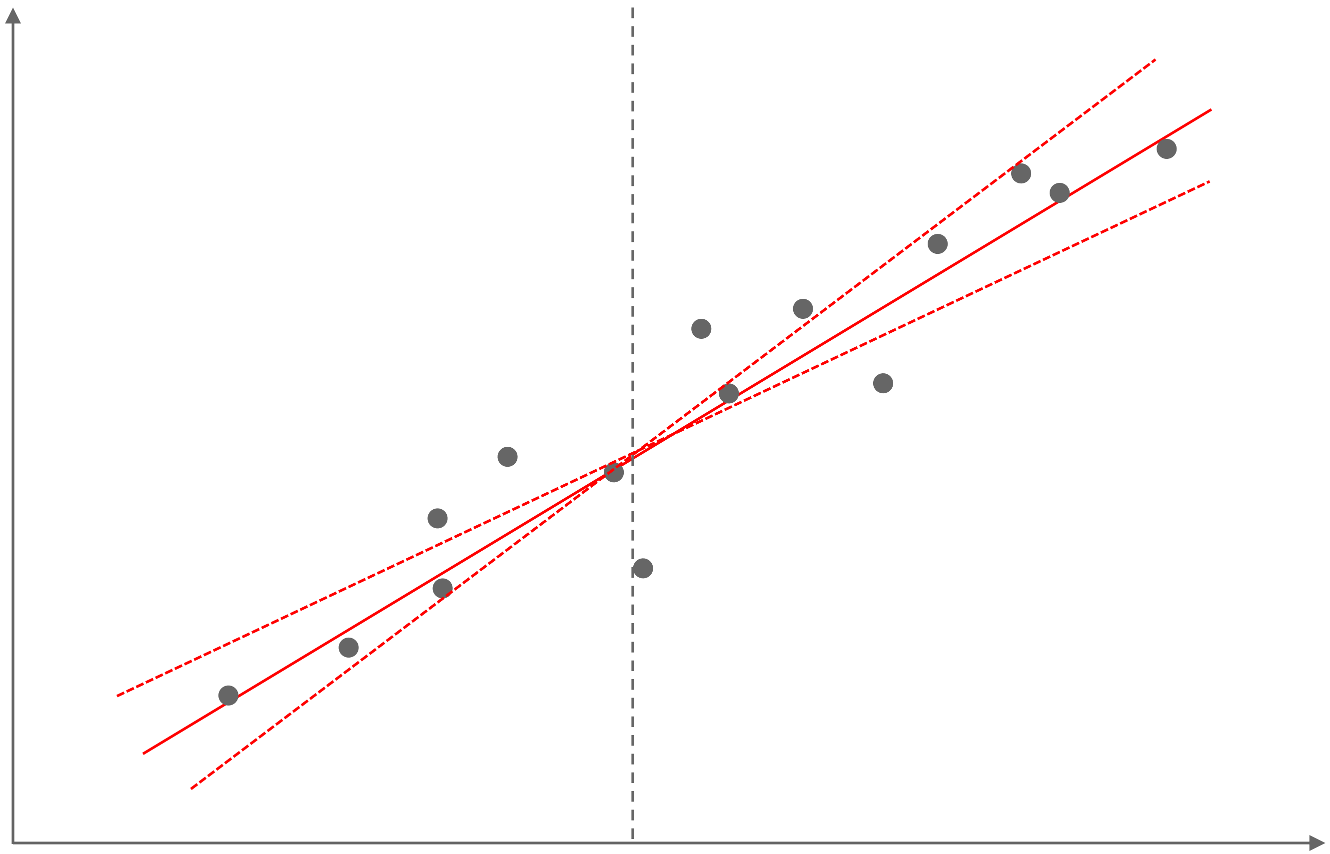
And what's this point where the B1 slope values are pivoting?

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

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

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

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



$$\bar{x}$$

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

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:

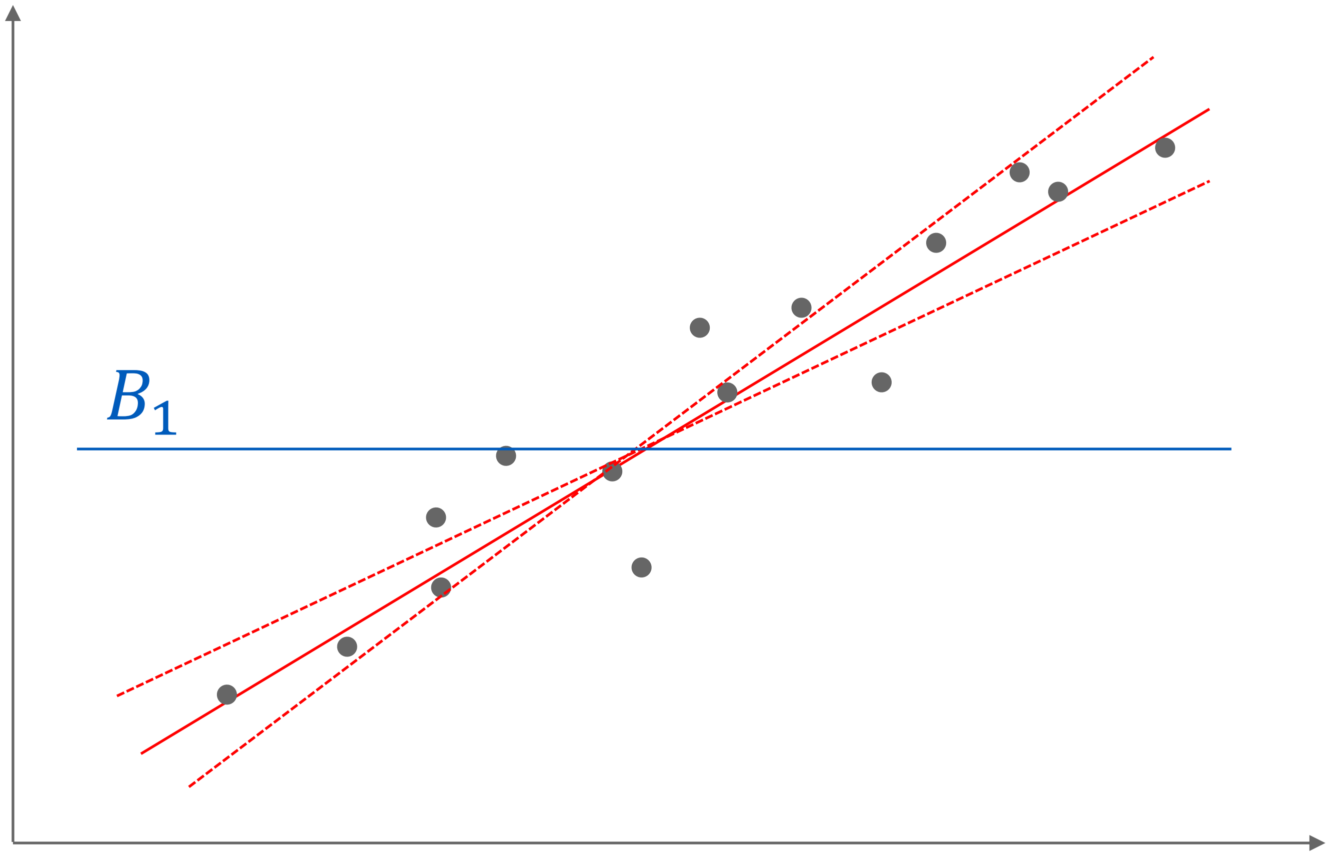
- $t \geq t_{\alpha, n-2}$
- $t \leq -t_{\alpha, n-2}$
- $t \leq -t_{\frac{\alpha}{2}, n-2}$ or $t \geq t_{\frac{\alpha}{2}, n-2}$

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



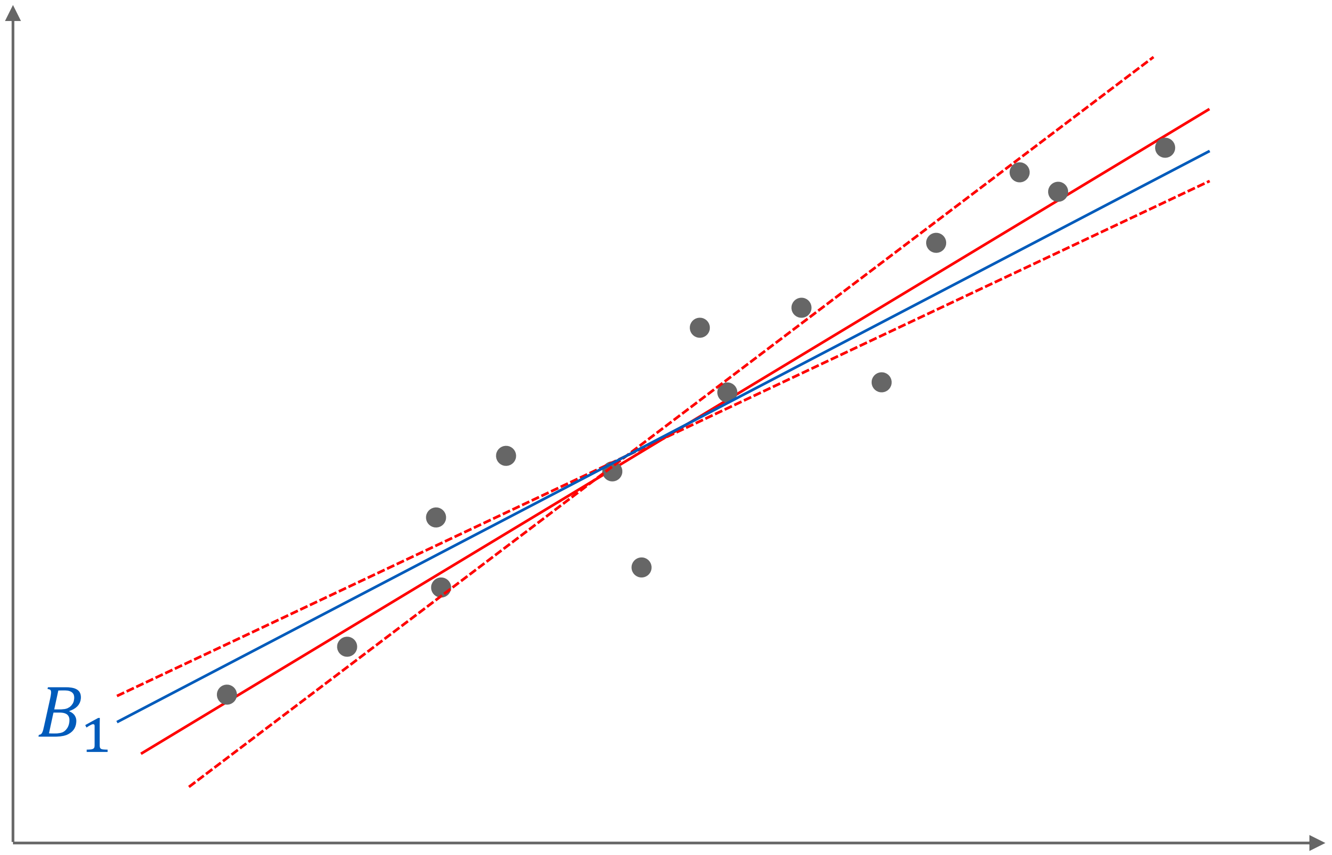
$$\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot S_{B_1}$$

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



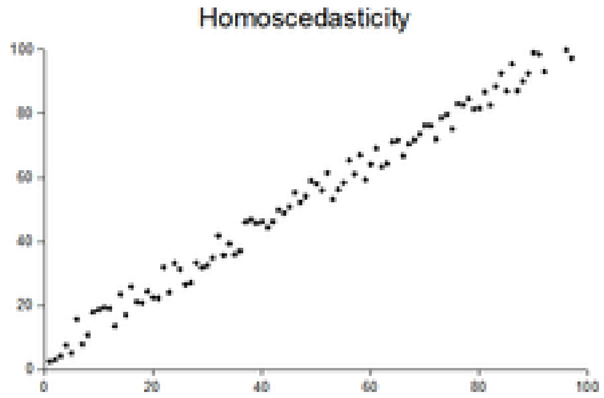
$$\hat{B}_1 \pm t_{\frac{\alpha}{2},n-2} \cdot S_{B_1}$$

Estimating the Trend Using a Least Squares Linear Model

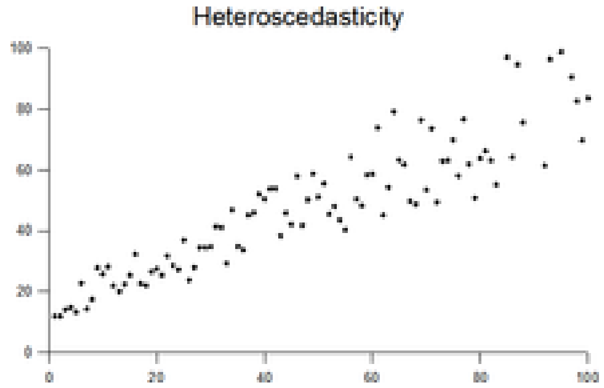
Some conditions that should be met for good results (see Helsel et al. for more)

- 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 $\hat{y}^*$:

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

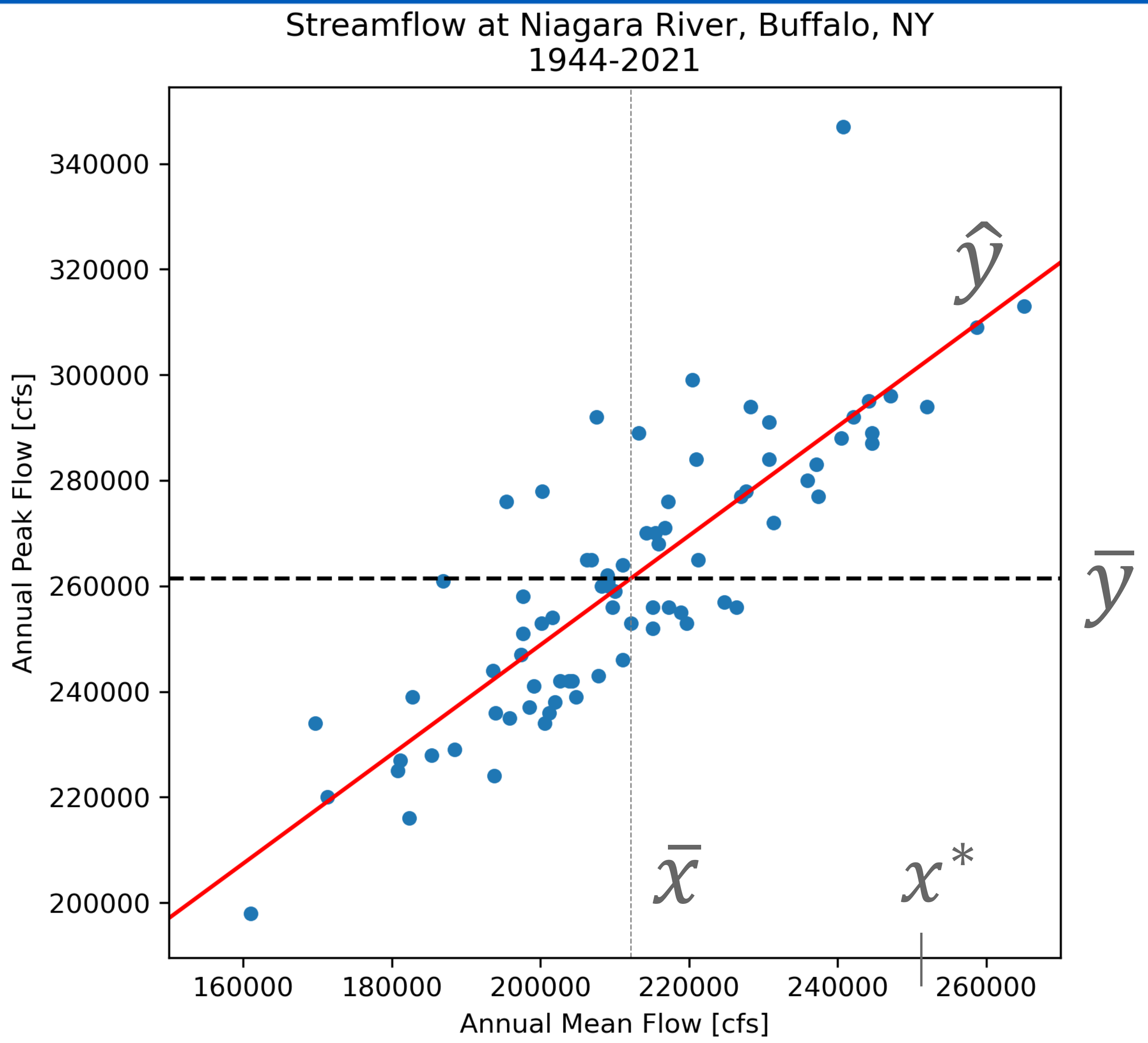
The variance of this prediction error ($\sigma_{E_P}^2$) will help define our predicted intervals,

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

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

(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 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.)

Quantile Regression

Quantile Regression

Advantage of Quantile regression

- Do not require that the underlying probability distributions are known or have any particular form.
- A linear relationship between the two variables is not required.
- The time series of the data need not be the same (or even from the same times) in the explanatory and dependent variables. That is, paired data is not required (although in many cases it is desirable).

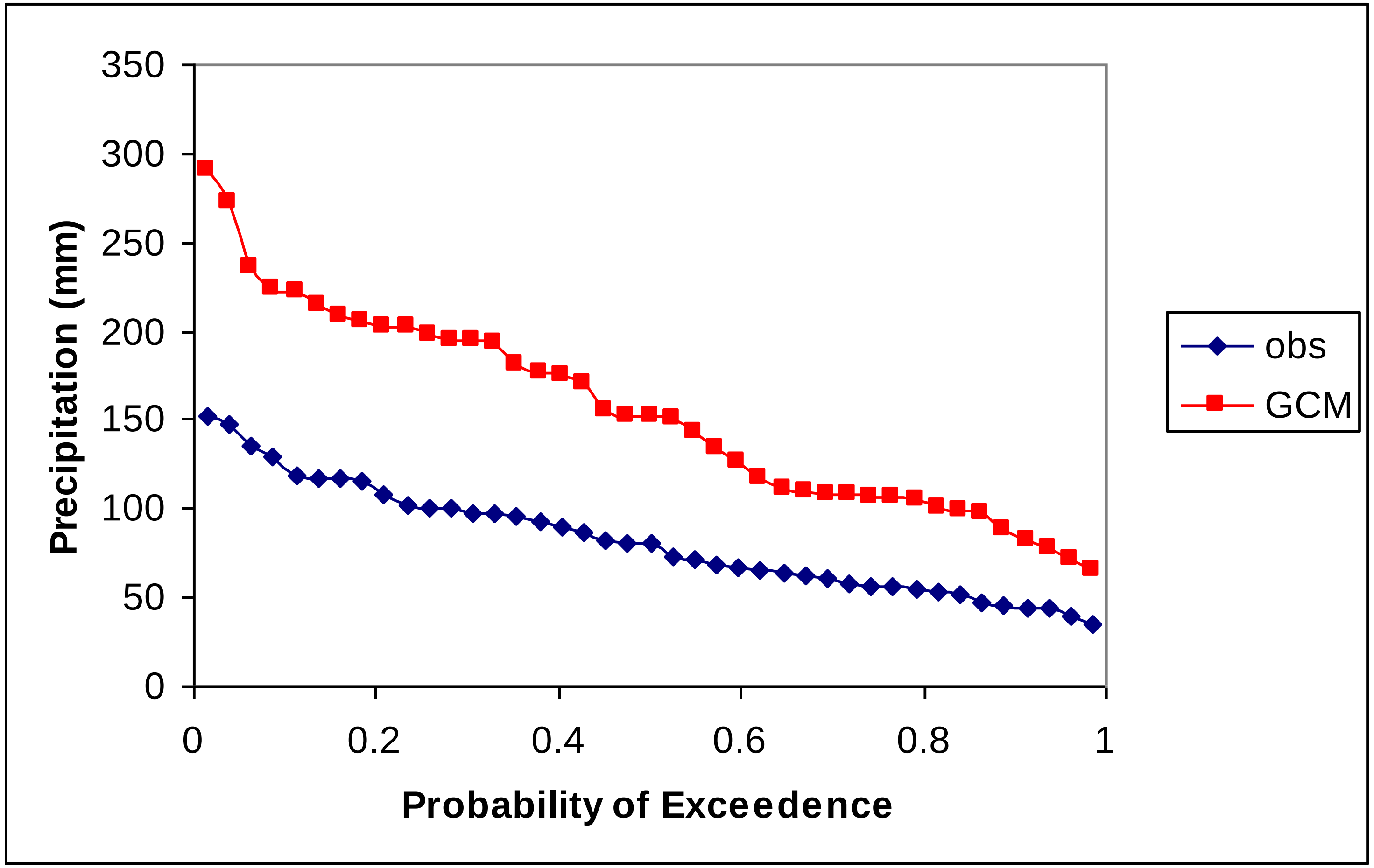
Example

Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

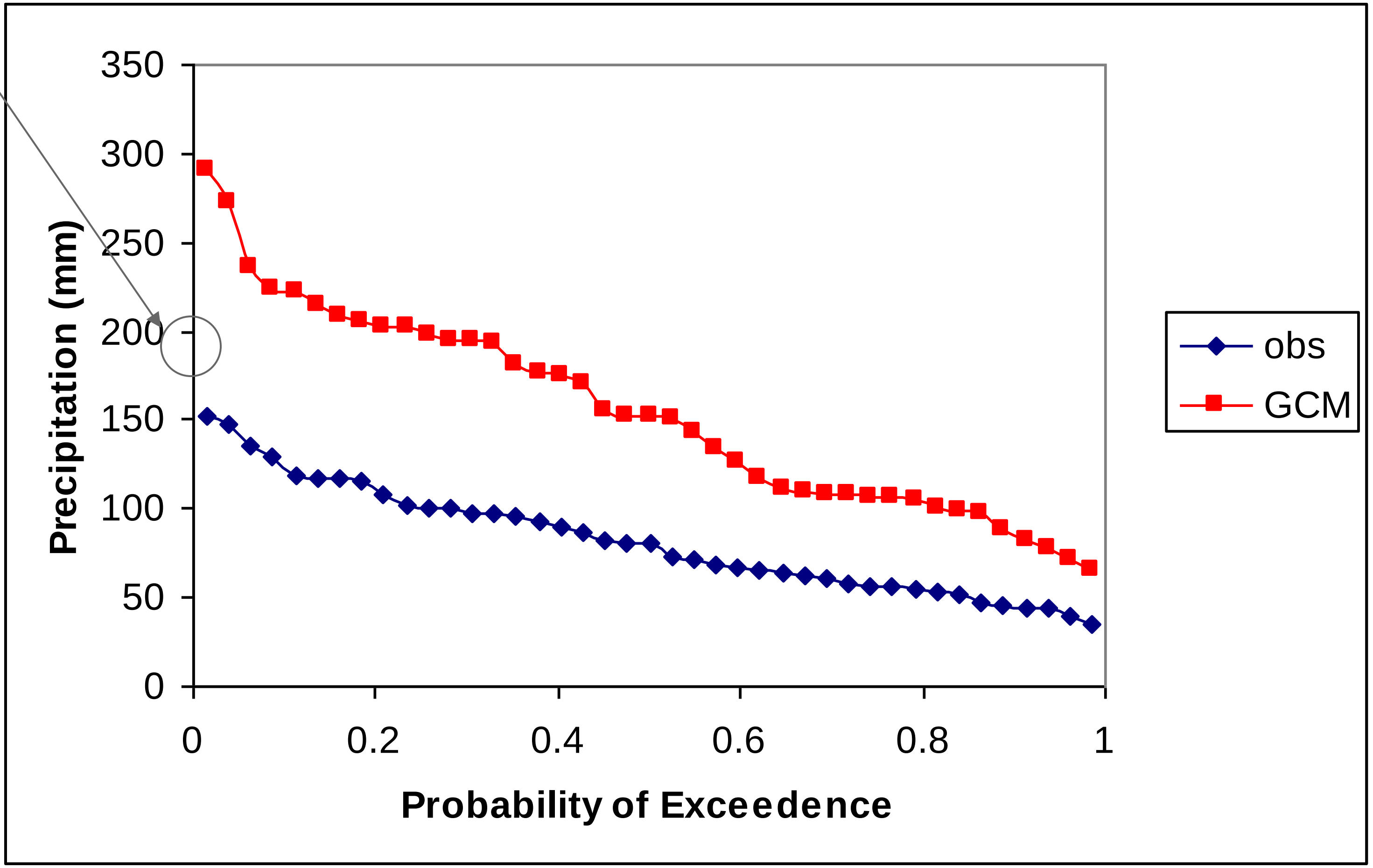
We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.



Example

GCM Input = 190

We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.



Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

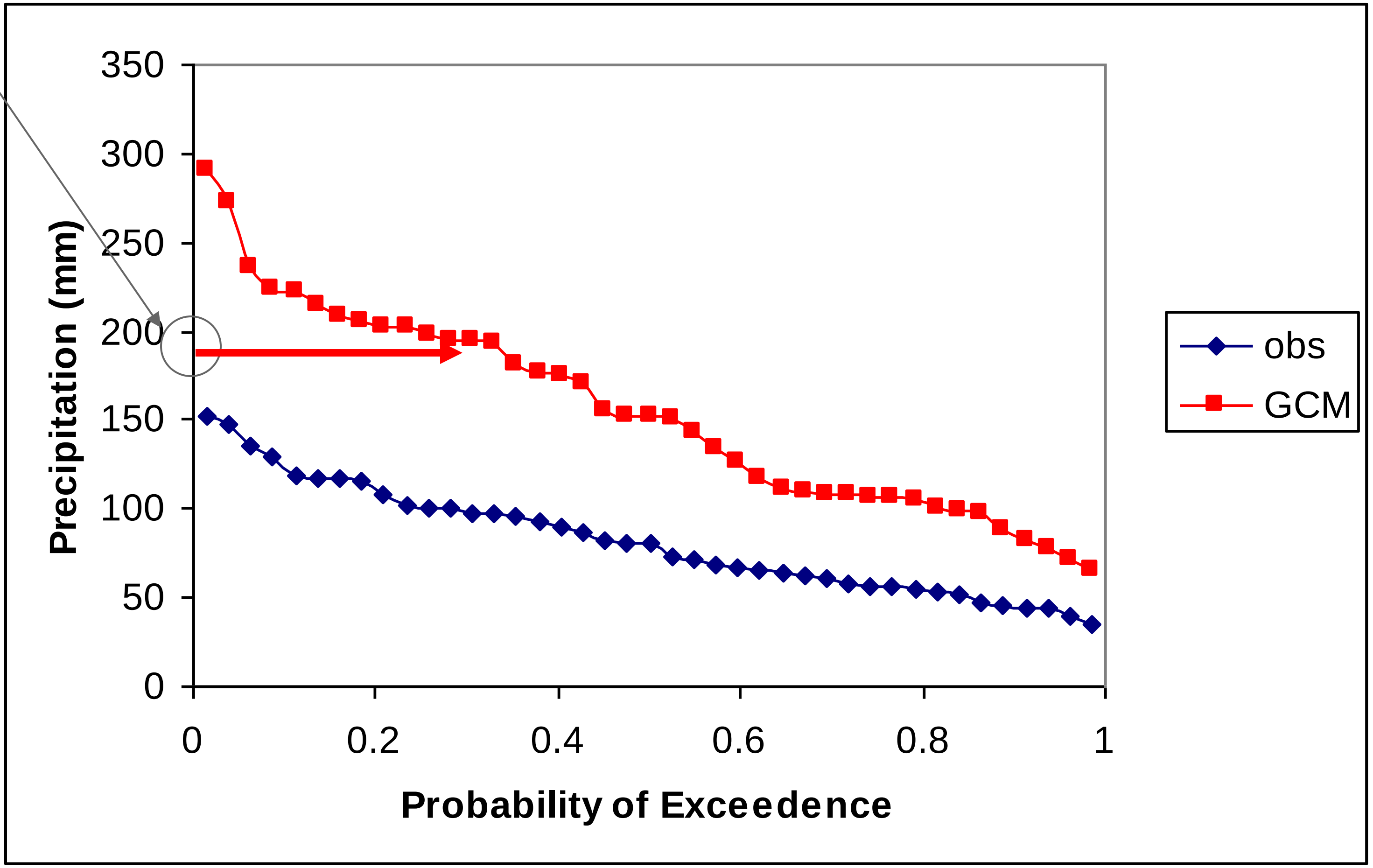
Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile

Example

GCM Input = 190

We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.



Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile

Example

GCM Input = 190

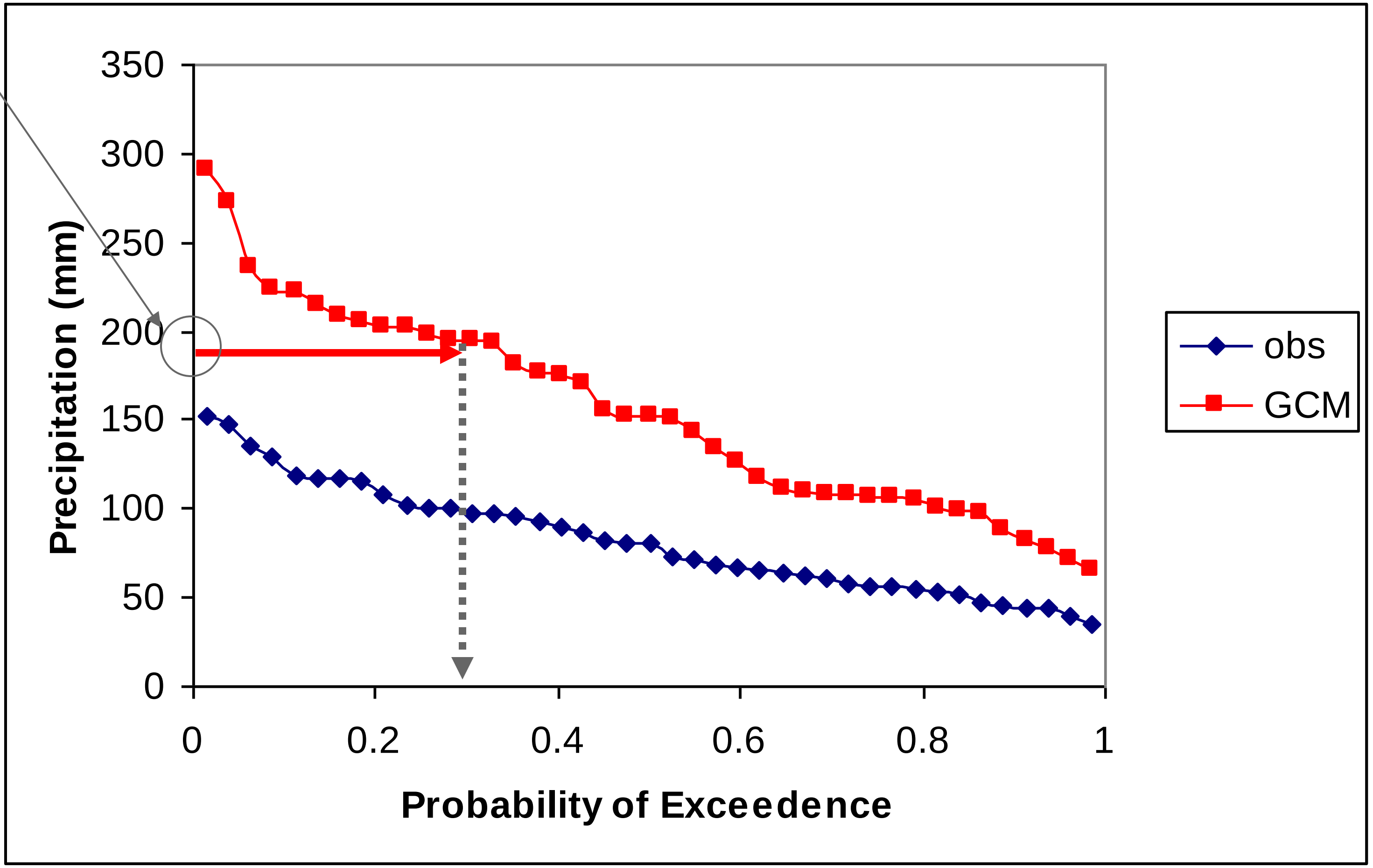
Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile

We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.



Example

GCM Input = 190

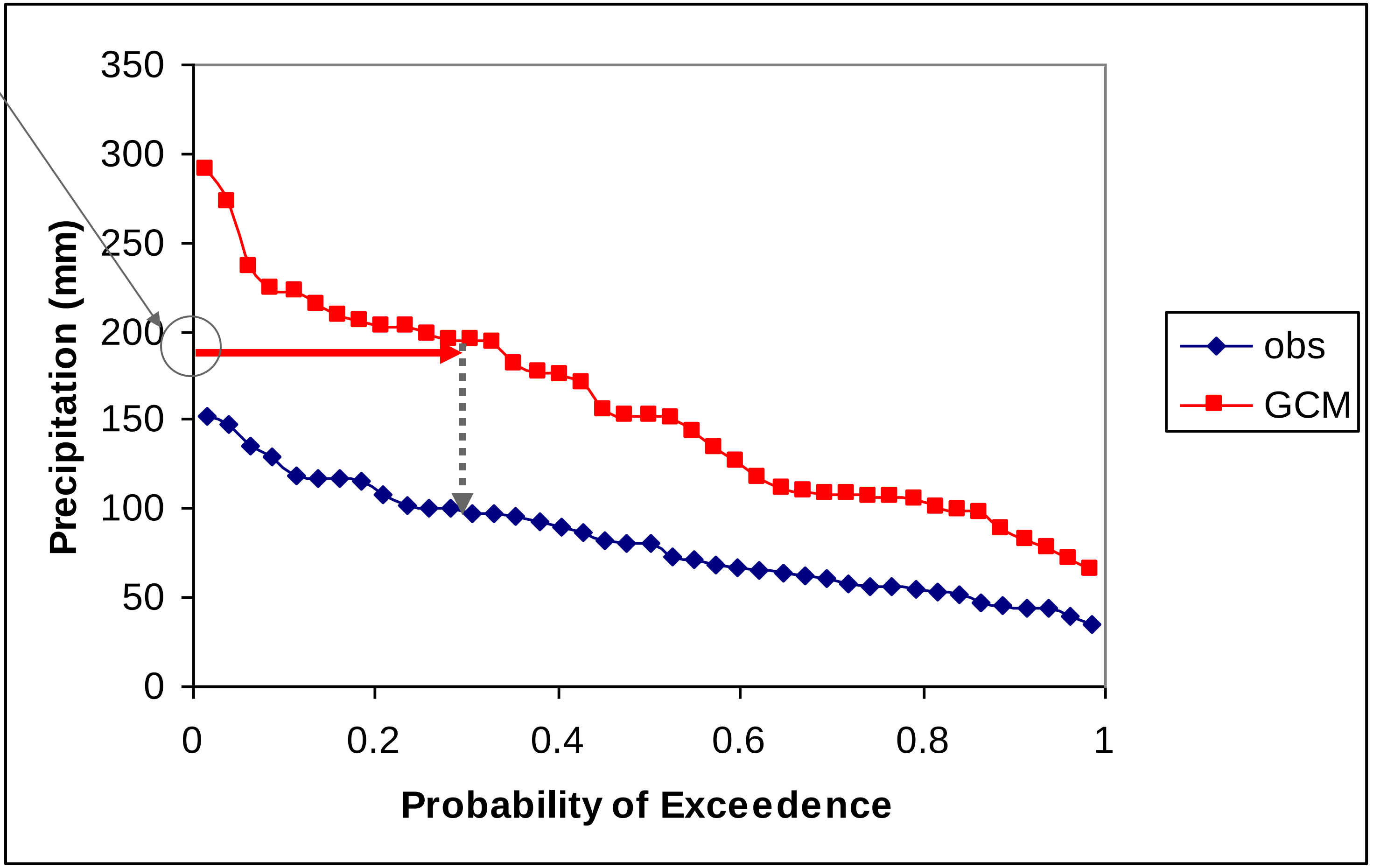
We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.

Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile



Example

GCM Input = 190

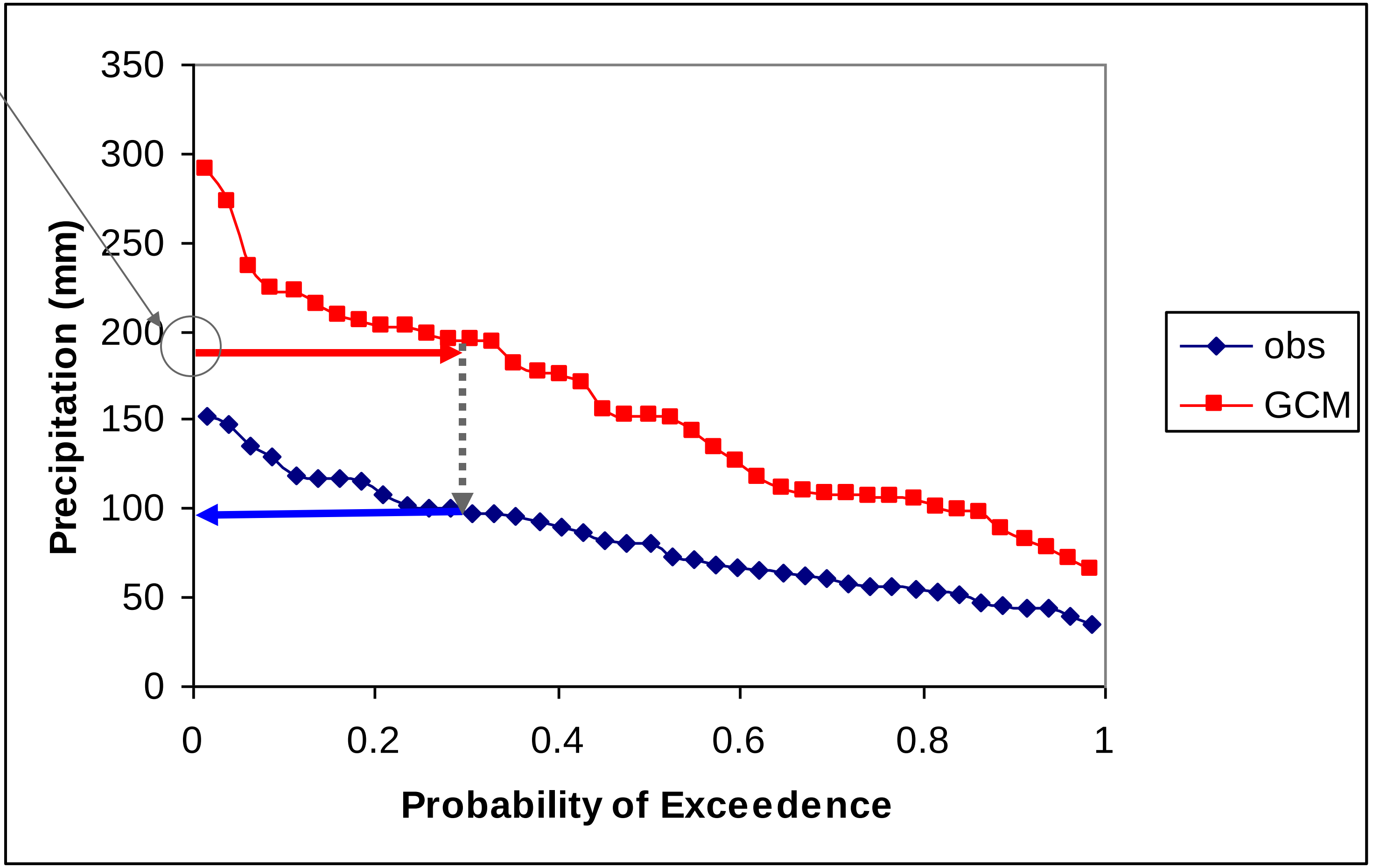
Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile

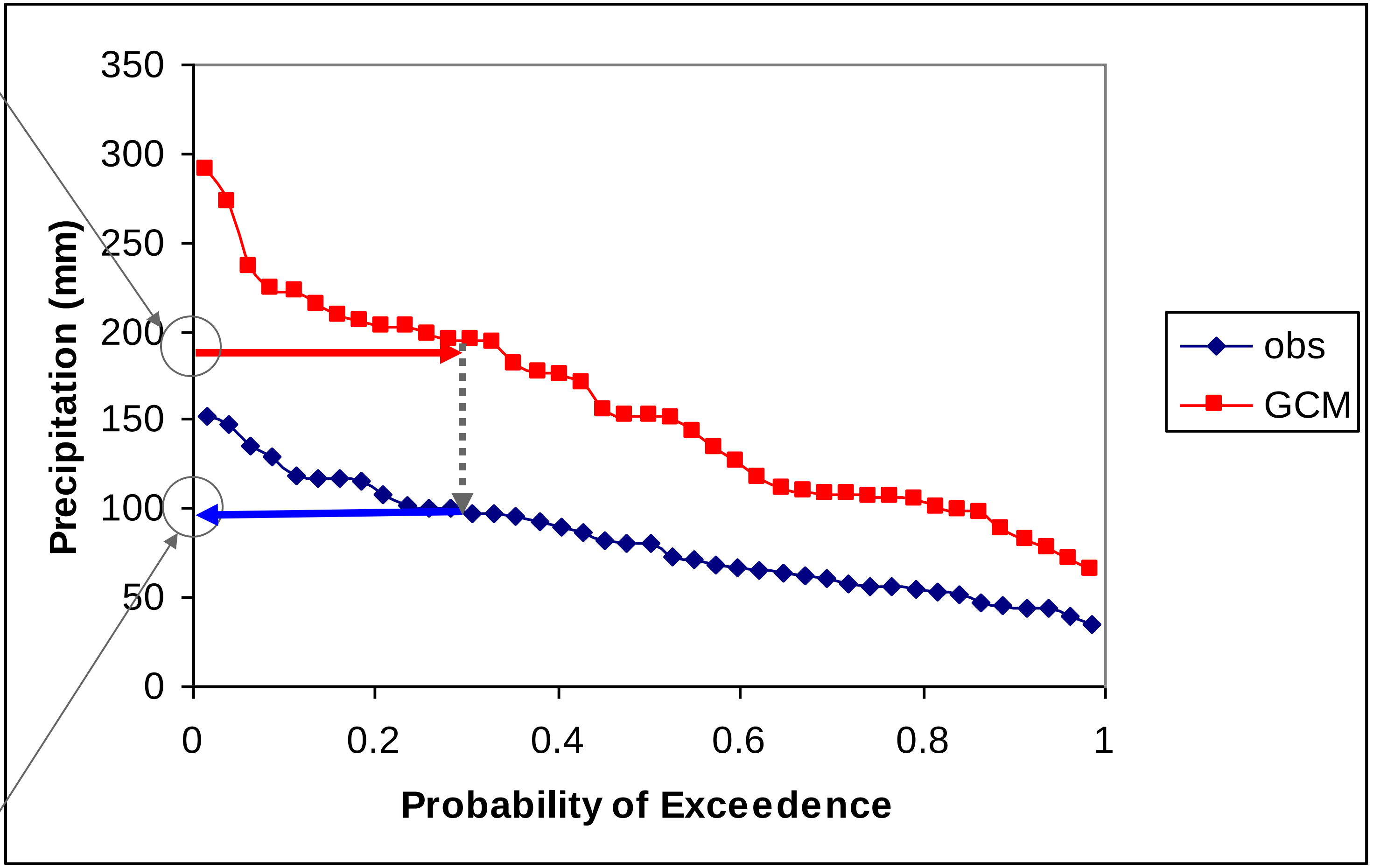
We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.



Example

GCM Input = 190

We presume that relative ranking and frequencies of events are correct, even if actual values don't match up in a linear way.



Bias-correct simulated precipitation from Global Climate Model (GCM) using local observations

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile

Bias Corrected Output = 100

Announcement (again)

- No in-class lecture this Friday.
- Instead, you will be assigned practice Jupyter notebooks, that will go over the trend analysis we will cover today.
  - The notebooks will also cover the practices for hypothesis testing.
  - You will need to finish the exercises in the notebooks, save them as HTML files, and submit them through UBLearns, just like homework.
  - Due date: 1 pm, Oct 11th 2024 (Friday)
- We will not have a separate assignment this week ☺

Least Squares Linear Regression

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

$$y = B_0 + B_1 x$$

- We can solve for the linear parameters as follows:

$$B_1 = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\left(\sum_{i=1}^{n} {x_i}^2\right) - (\sum_{i=1}^{n} x_i)^2} \qquad B_0 = \frac{(\sum_{i=1}^{n} y_i) - B_1(\sum_{i=1}^{n} x_i)}{n} = \bar{y} + B_1 \bar{x}$$

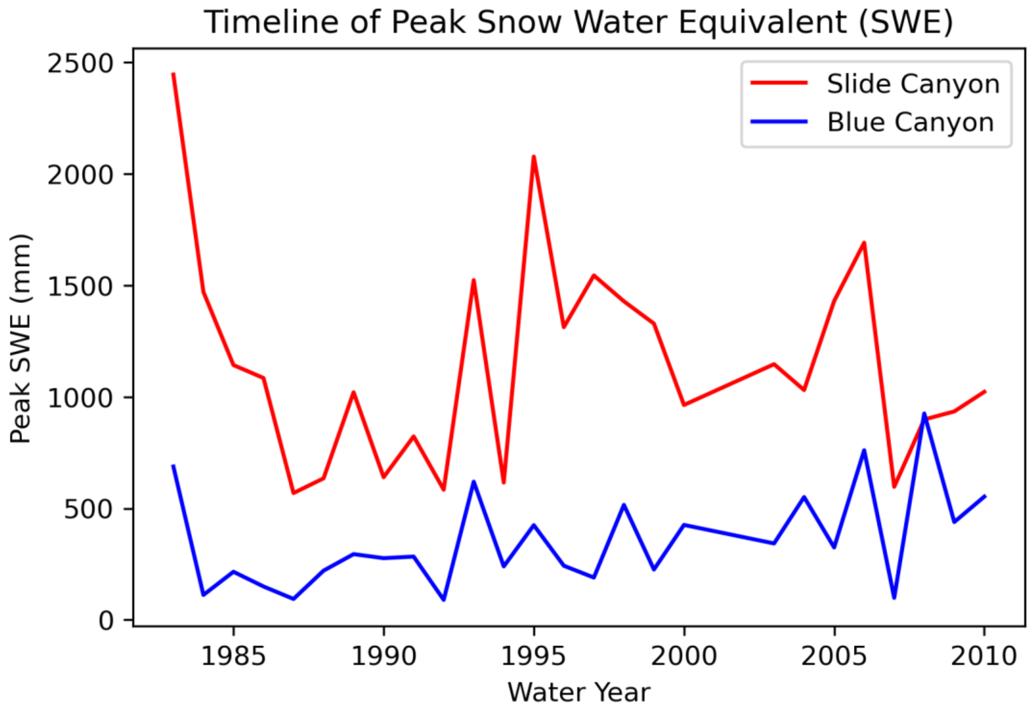
- Now we can use the linear model to make predictions:

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

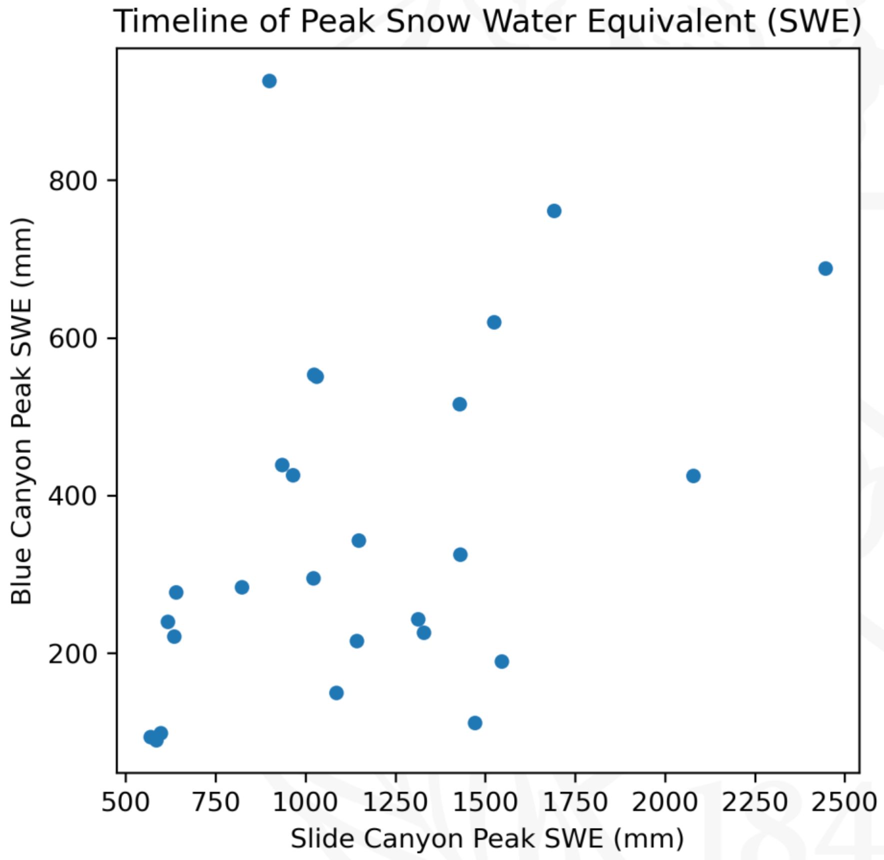
- And finally compute residuals between our data and model's predictions:

$$(y_i - \hat{y}_i)$$

Example: Use SWE in Slide Canyon to predict SWE in Blue Canyon

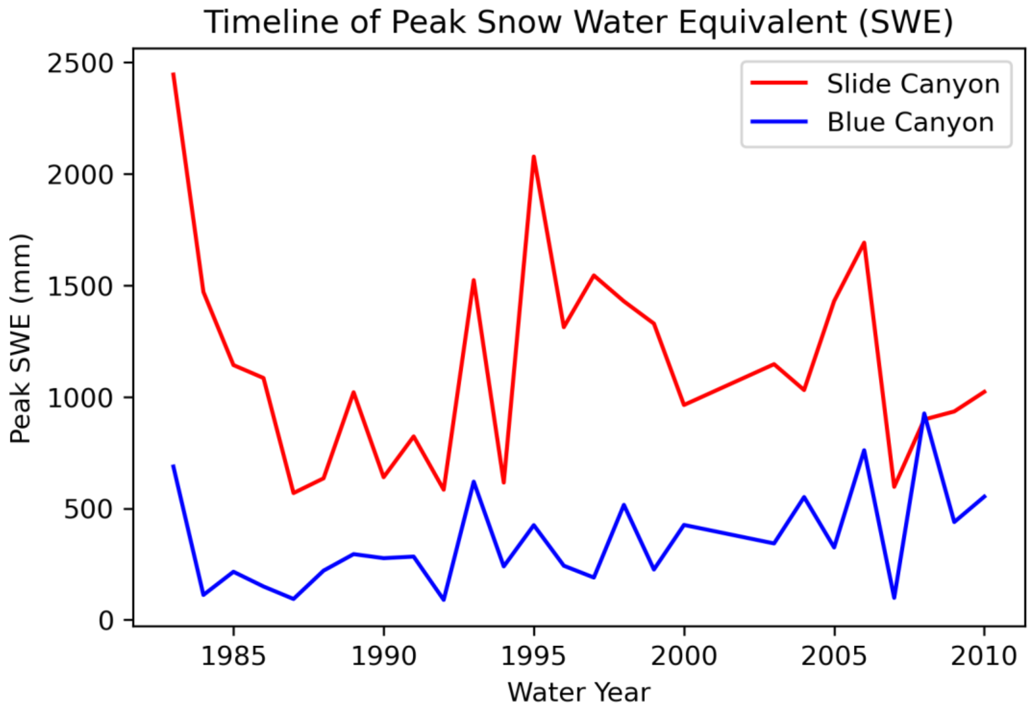


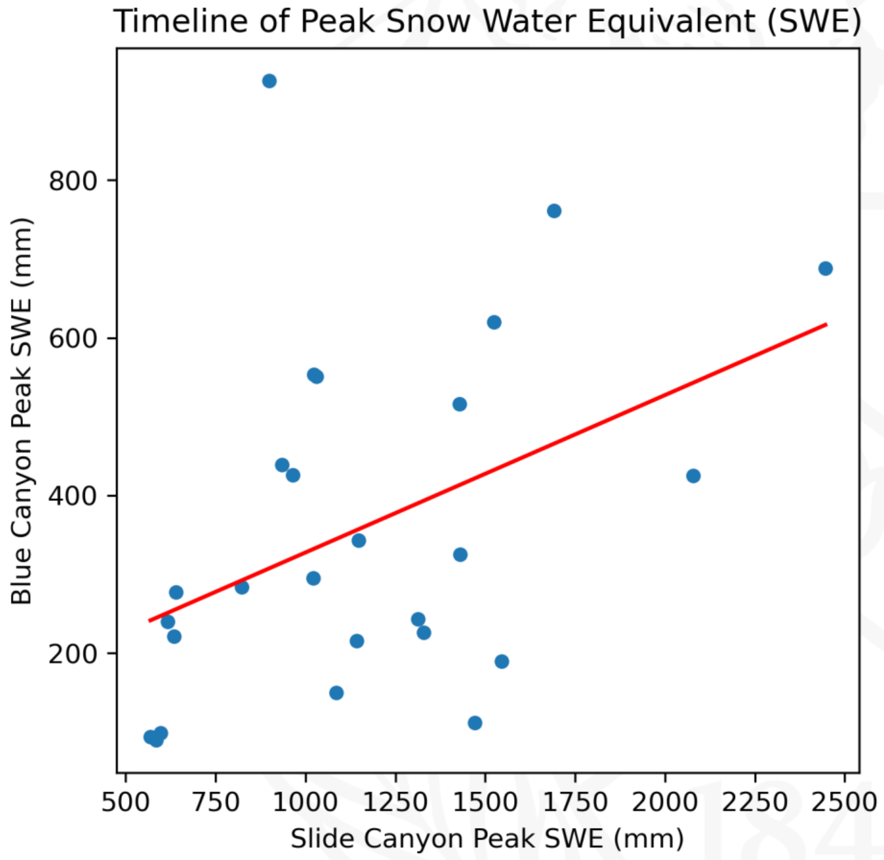
We can use the Linear Regression Model!



Example: Use SWE in Slide Canyon to predict SWE in Blue Canyon

We can use the Linear Regression Model!

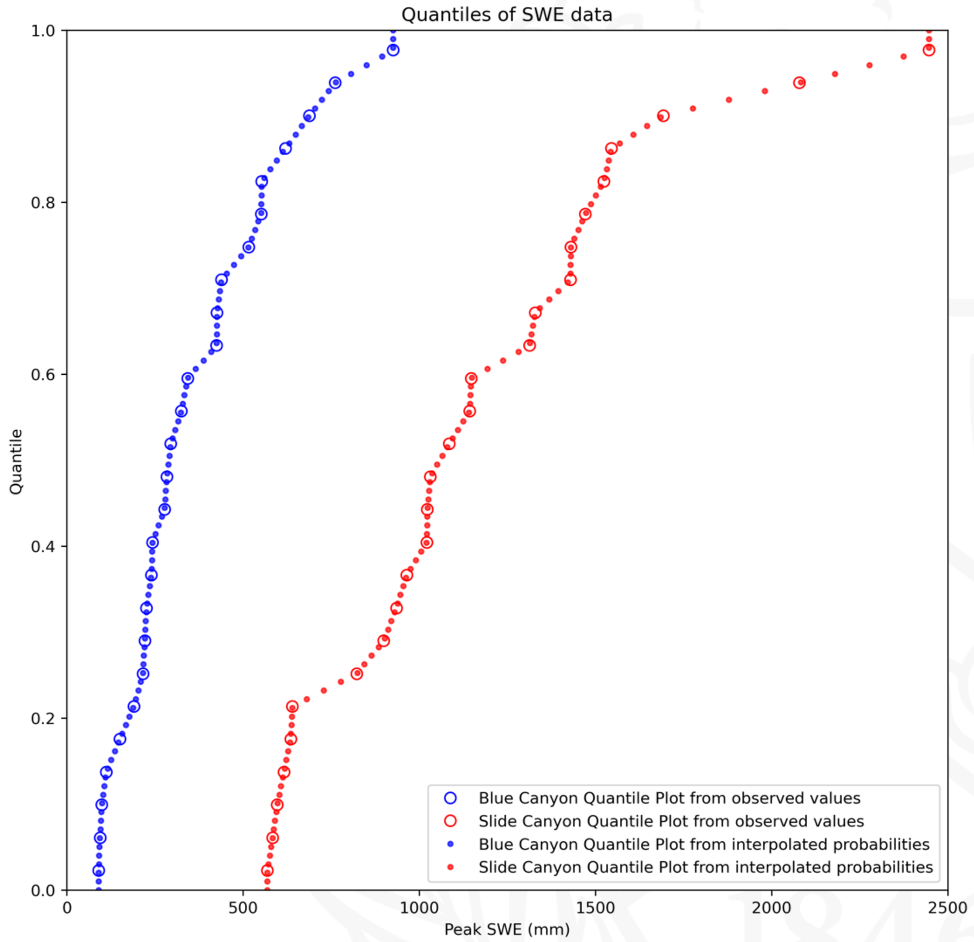




Example: Use SWE in Slide Canyon to predict SWE in Blue Canyon

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

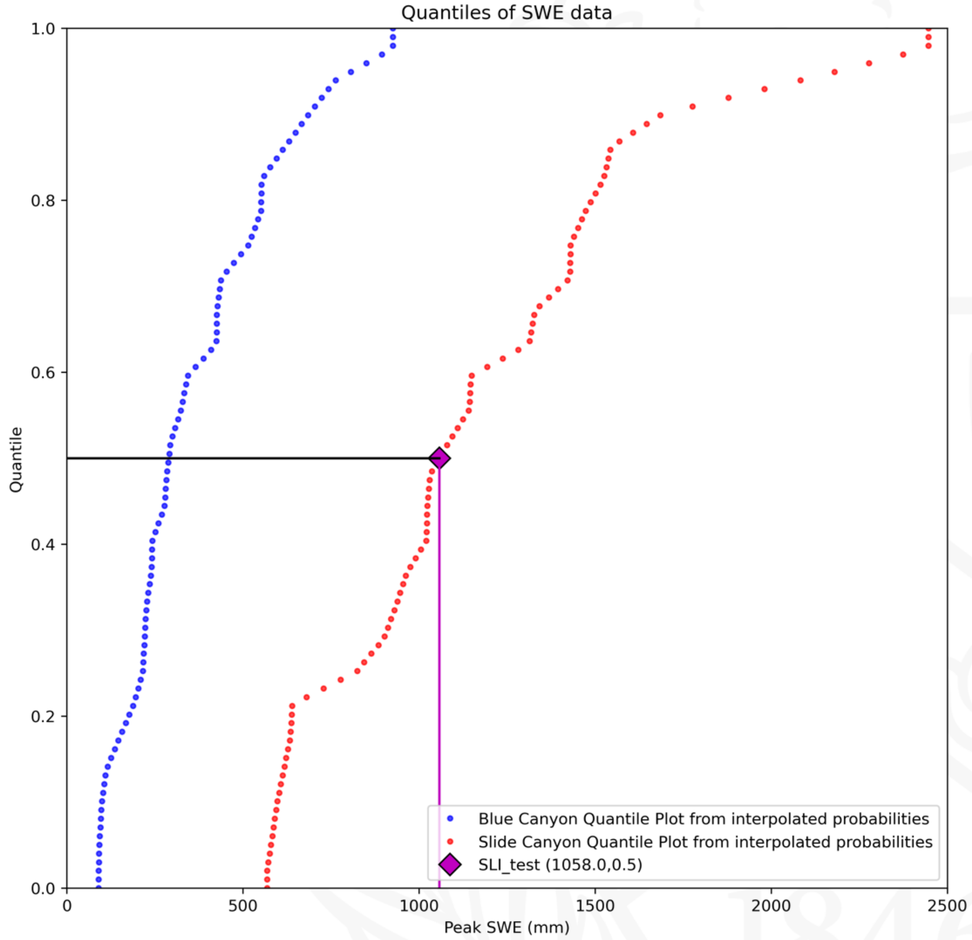


Example: Use SWE in Slide Canyon to predict SWE in Blue Canyon

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile

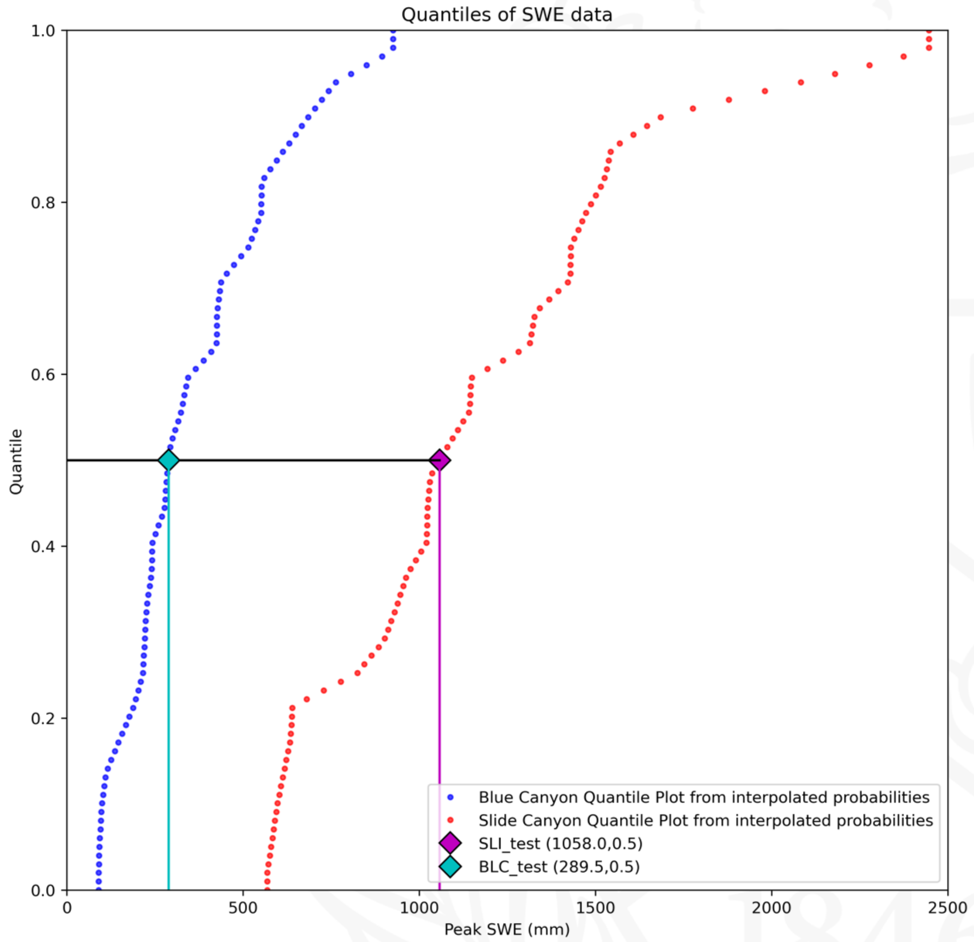


Example: Use SWE in Slide Canyon to predict SWE in Blue Canyon

How does Quantile Regression work?

Step 1: For each of your two datasets, create an empirical CDF

Step 2: Use the two empirical CDFs as a way of looking-up (or mapping) values from the predictor to the predictand, by matching which physical value corresponds to the same quantile



Example: Use SWE in Slide Canyon to predict SWE in Blue Canyon