Applied Geodata Science 1

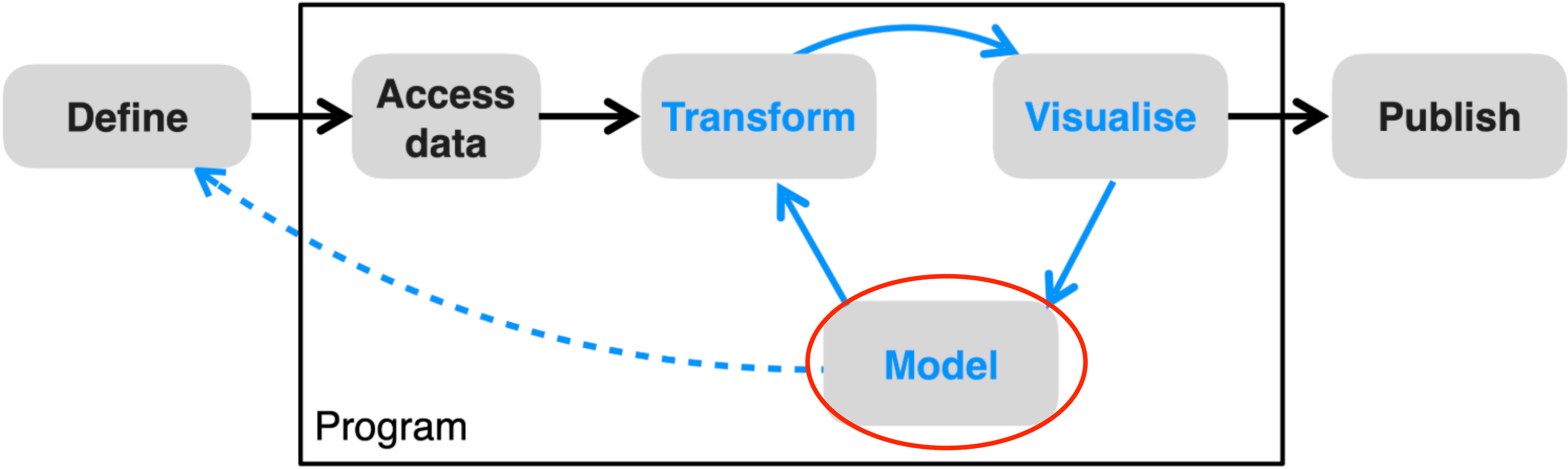
Session 9

Prof. Dr. Benjamin Stocker
14.04.2025

GECO

UNIVERSITÄT BERN

Chapter 9

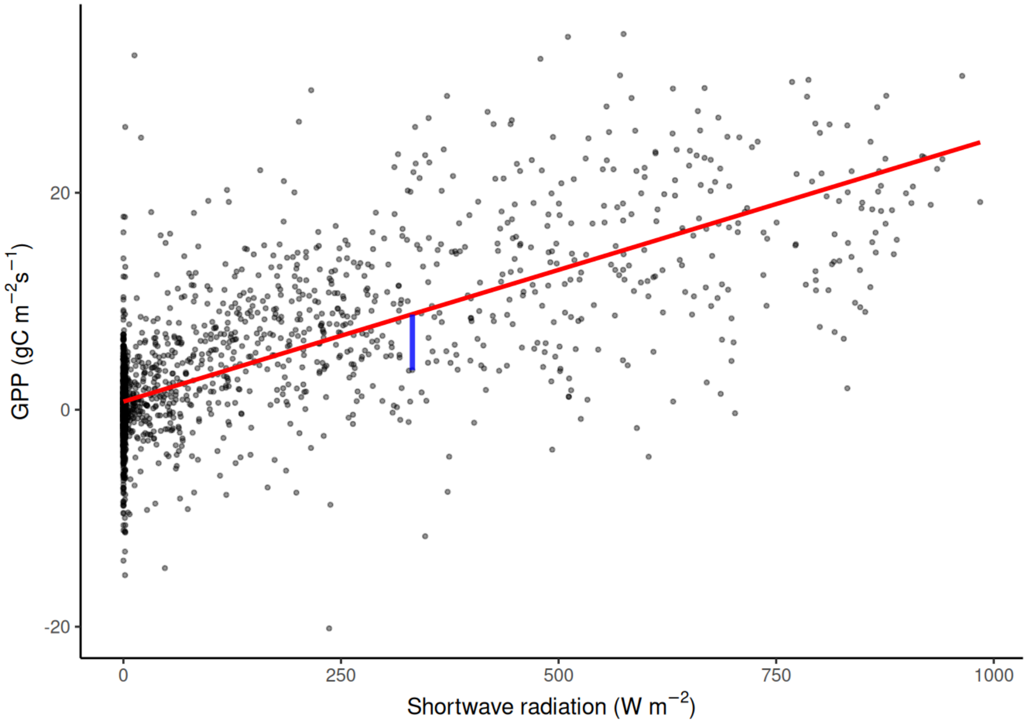


Define
Access data
Transform
Visualise
Publish
Model
Program

Regression vs. classification

	Regression	Classification
Target variable	Continuous	Categorical
Common models	Linear regression, polynomial regression, KNN, tree-based regression	Logistic regression, KNN, SVM, tree classifiers
Metrics	RMSE, R^2, adjusted R^2, AIC, BIC	Accuracy, precision, AUC, F1

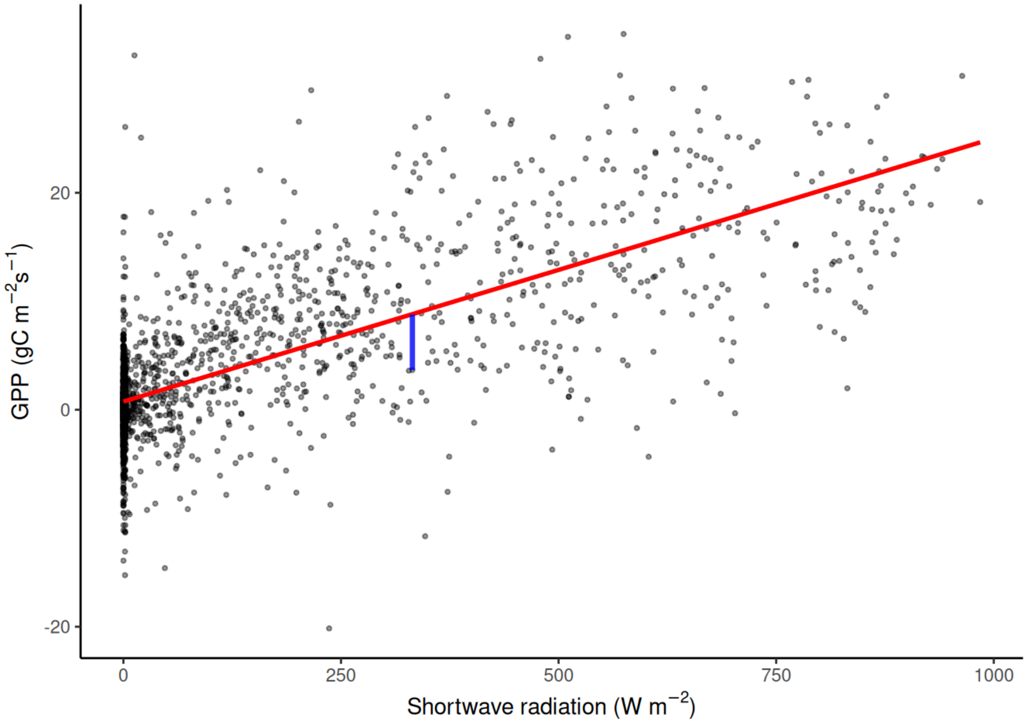
Linear regression



$$Y_i \sim \beta_0 + \beta_1 X_i, \quad i = 1, 2, \dots n \,,$$

$$\min_{\beta_0, \beta_1} \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2.$$

Linear regression

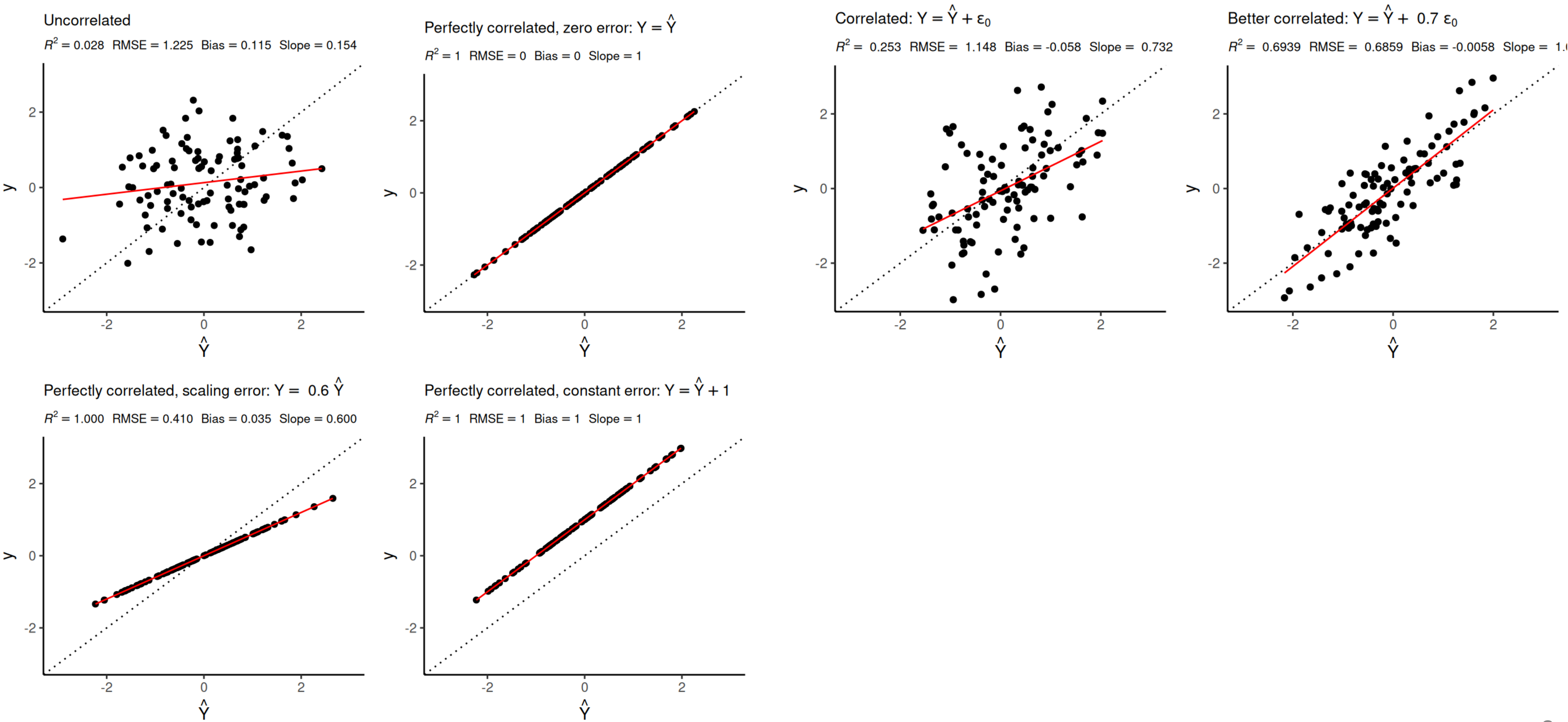



```r
# fit univariate linear regression
linmod1 <- lm(GPP_NT_VUT_REF ~ SW_IN_F, data = df)
```

```r
summary(linmod1)
```

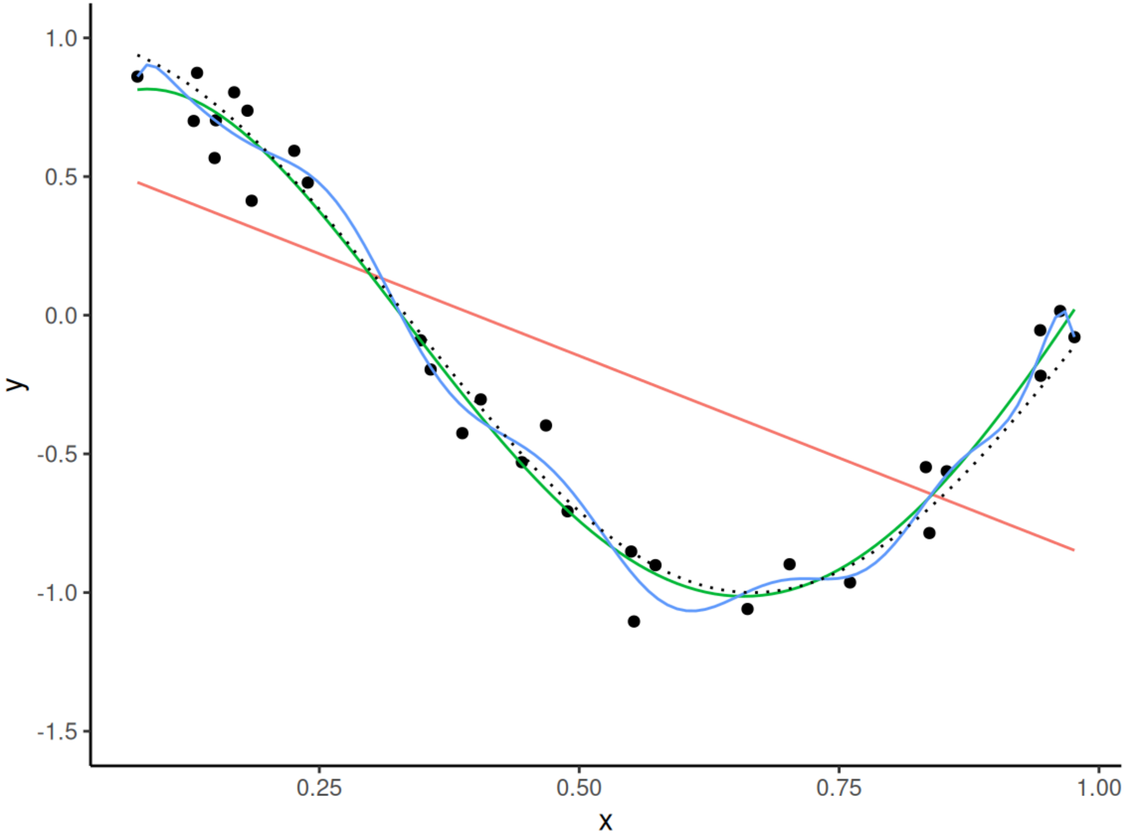
```
Call:
lm(formula = GPP_NT_VUT_REF ~ SW_IN_F, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.699  -2.092  -0.406   1.893  35.153 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.8732273  0.0285896   30.54   <2e-16 ***
SW_IN_F     0.0255041  0.0001129  225.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.007 on 41299 degrees of freedom
Multiple R-squared:  0.5525,	Adjusted R-squared:  0.5525 
F-statistic: 5.099e+04 on 1 and 41299 DF,  p-value: < 2.2e-16
```

Metrics



Uncorrelated
R^2 = 0.028 RMSE = 1.225 Bias = 0.115 Slope = 0.154
Perfectly correlated, zero error: $Y = \hat{Y}$
R^2 = 1 RMSE = 0 Bias = 0 Slope = 1
Correlated: $Y = \hat{Y} + \varepsilon_0$
R^2 = 0.253 RMSE = 1.148 Bias = -0.058 Slope = 0.732
Better correlated: $Y = \hat{Y} + 0.7\ \varepsilon_0$
R^2 = 0.6939 RMSE = 0.6859 Bias = -0.0058 Slope =
Perfectly correlated, scaling error: $Y = 0.6\ \hat{Y}$
R^2 = 1.000 RMSE = 0.410 Bias = 0.035 Slope = 0.600
Perfectly correlated, constant error: $Y = \hat{Y} + 1$
R^2 = 1 RMSE = 1 Bias = 1 Slope = 1
y
$\hat{Y}$
-2
0
2

Model selection



- Increasing model complexity always increases the explained variance (R^2) on the data used for model fitting (training).
- Increasing model complexity means increasing the number of parameters (e.g., by increasing the number of predictors in a multivariate regression model).
- The R^2 evaluated on the data used for model fitting is not a reliable estimate for R^2 on new data.

Model selection

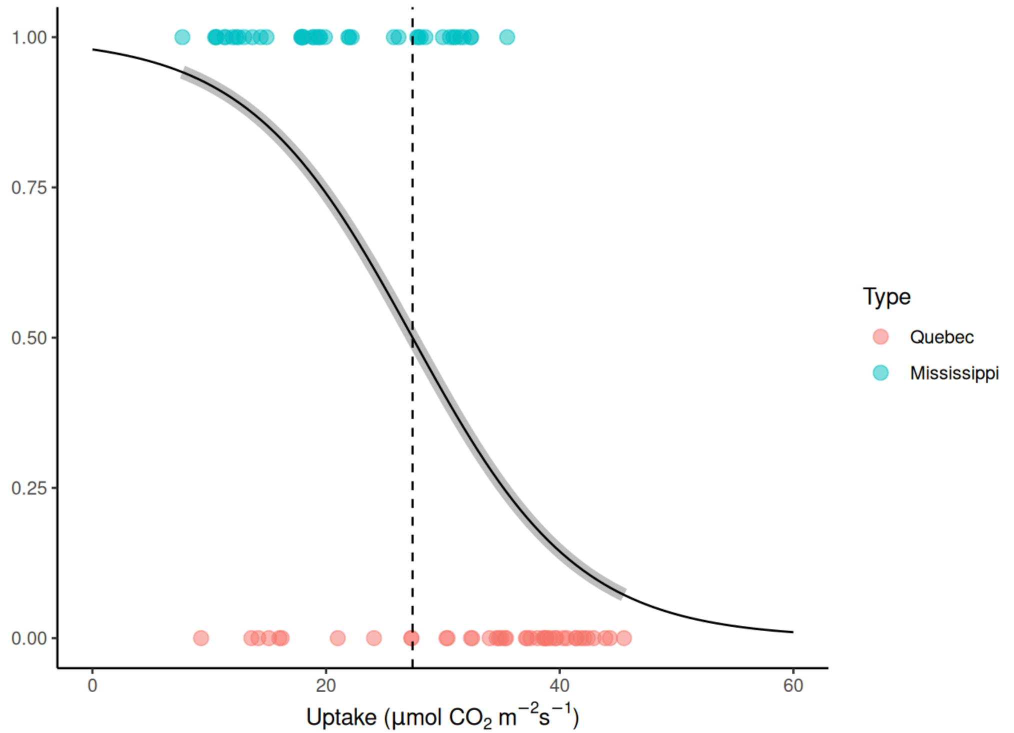
- Compare only R^2 from alternative **models with the same level of complexity** (e.g., number of predictors).
- For comparing alternative models with differing complexity (e.g., number of predictors), consider a metric that **penalises model complexity**.

Minimise:

$$\mathrm{AIC} = n \log \left(\frac{\mathrm{SSE}}{n} \right) + 2(p+2)$$

Logistic regression

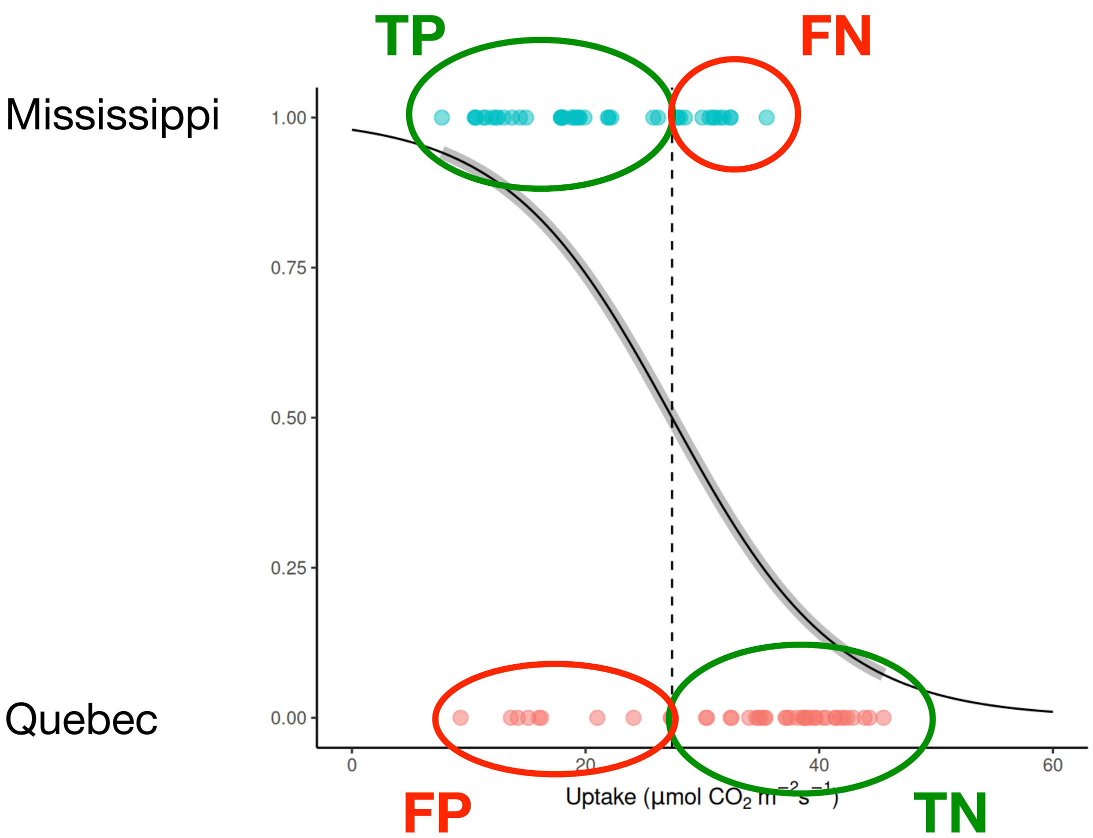
```
> library(datasets)
> CO2
Grouped Data: uptake ~ conc | Plant
   Plant        Type  Treatment conc uptake
1    Qn1      Quebec nonchilled   95   16.0
2    Qn1      Quebec nonchilled  175   30.4
3    Qn1      Quebec nonchilled  250   34.8
4    Qn1      Quebec nonchilled  350   37.2
5    Qn1      Quebec nonchilled  500   35.3
6    Qn1      Quebec nonchilled  675   39.2
```



$$\text{logit}(z) = \frac{\exp(z)}{1+\exp(z)}.$$

$$f(X,\beta) = \text{logit}(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1+\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}.$$

Logistic regression



	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	True positives (TP)	False positives (FP)
$\hat{Y} = 0$	False negatives (FN)	True negatives (TN)

Type
Quebec
Mississippi

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{N},$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$