# An analysis of the union wages data: GLM's, GAM's and JAGS

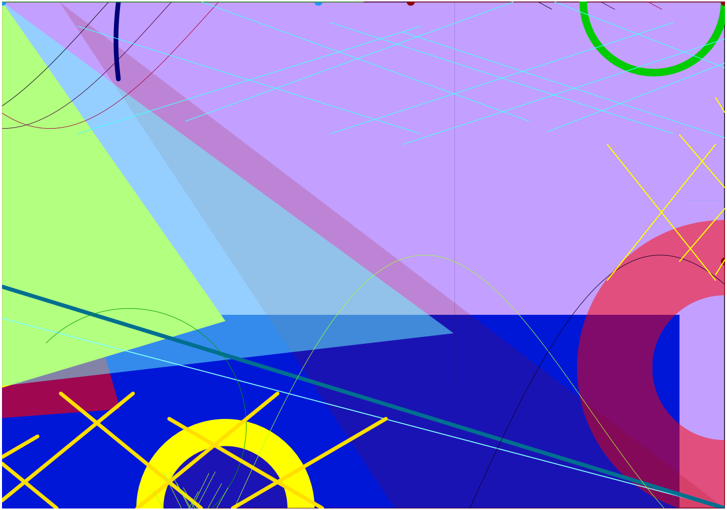
Henrique Laureano
mynameislaure.github.io

STAT 260: Nonparametric Statistics

# On the Agenda

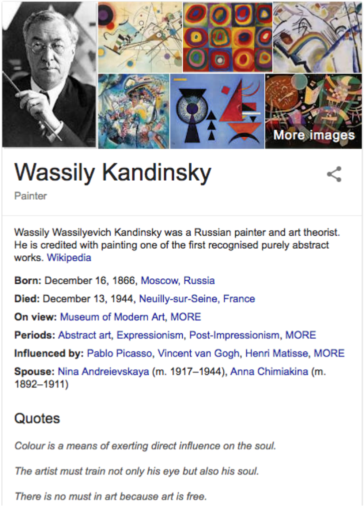
1. Data
2. GLM
3. GAM
4. JAGS

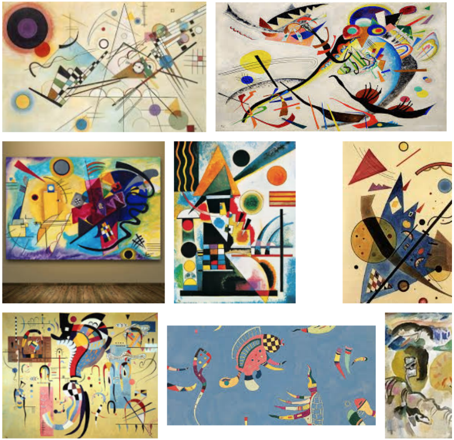
# Turning the dataset into a Kandinsky* painting



(github.com/gsimchoni/kandinsky)

# * Kandinsky



(screenshots from Google)

## Trade union data

Data on 534 U.S. workers with eleven variables (`SemiPar::trade.union`).

## Trade union data

Data on 534 U.S. workers with eleven variables (`SemiPar::trade.union`).

Variables:

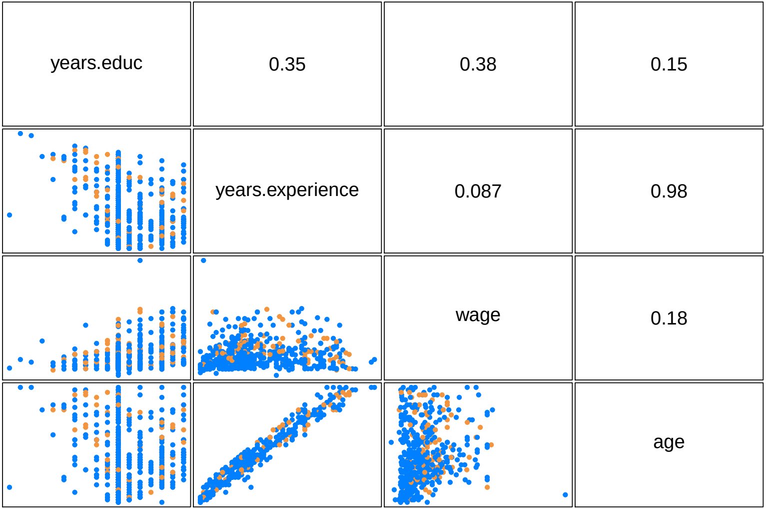
## Trade union data

Data on 534 U.S. workers with eleven variables (`SemiPar::trade.union`).

Variables:

- `union.member` (yes or no)
- `years.educ`
- `years.experience`
- `wage` (dollars per hour)
- `age`
- `female` (yes or no)
- `south` (living or not in southern region of U.S.)
- `race` (black, hispanic or white)
- `occupation` (six categories)
- `sector` (three categories)
- `married` (yes or no)

# Quantitative variables:




(colors by `union.member` status)

# On the Agenda

1. Data
2. GLM
3. GAM
4. JAGS

# Fitting Generalized Linear Models

# Fitting Generalized Linear Models

- Let $p_i$ be the probability of trade union membership;
- Using a logistic regression model

$$\begin{aligned} \mathrm{logit}(p_i) &= \beta_0 + \beta_1 \texttt{years.educ}_i + ... + \beta_{10} \texttt{married}_i, \\ \texttt{union.member}_i &\sim \mathrm{Bernoulli}(p_i), \quad i = 1, \ldots 534. \end{aligned}$$

(b/c we have 10 variables, as previously shown)

# Fitting Generalized Linear Models

- Let $p_i$ be the probability of trade union membership;
- Using a logistic regression model

$$\text{logit}(p_i) = \beta_0 + \beta_1 \texttt{years.educ}_i + \ldots + \beta_{10} \texttt{married}_i,$$
$$\texttt{union.member}_i \sim \text{Bernoulli}(p_i), \quad i = 1, \ldots 534.$$

(b/c we have 10 variables, as previously shown)

```
formula <- union.member ~
  years.educ + years.experience + wage + age + female + south +
  as.factor(race) + as.factor(occupation) + sector + married

union.glm <- glm(formula, family = binomial, trade.union)
```

Using the AIC as criterion we have ...

```r
union.glm$formula
```

```
union.member ~ wage + age + female + south + as.factor(race) +
    as.factor(occupation) + married
```

we *finish* with seven variables, two quantitatives.

```
union.glm$formula
```

```
union.member ~ wage + age + female + south + as.factor(race) +
    as.factor(occupation) + married
```

we *finish* with seven variables, two quantitatives.

... and the residues?

```r
union.glm$formula
```

```
union.member ~ wage + age + female + south + as.factor(race) +
    as.factor(occupation) + married
```

we *finish* with seven variables, two quantitatives.

... and the residues? ... and the goodness-of-fit?

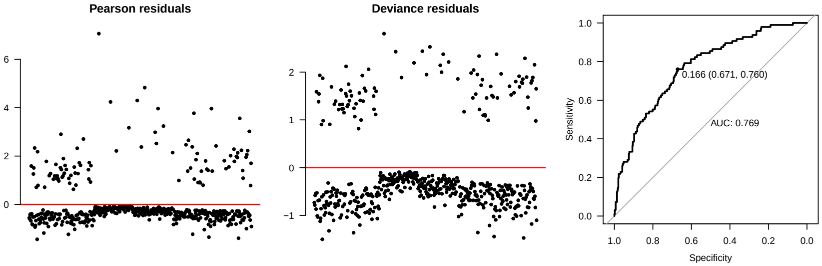
```
union.glm$formula
```

```
union.member ~ wage + age + female + south + as.factor(race) +
    as.factor(occupation) + married
```

we *finish* with seven variables, two quantitatives.

... and the residues? ... and the goodness-of-fit?

```
pearson <- residuals(union.glm, type = "pearson")
devi <- residuals(union.glm, type = "deviance")
rocurve <- pROC::roc(trade.union$union.member, fitted(union.glm))
```



# Coefficients

# Coefficients

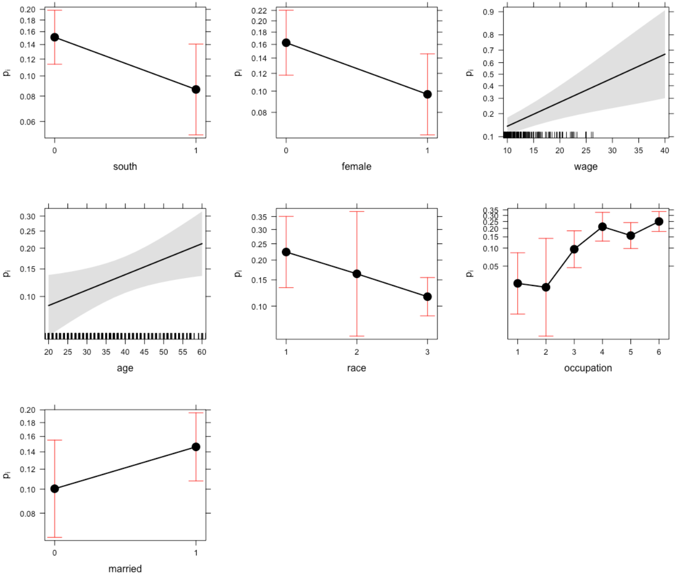
```r
round( summary(union.glm)$coeff, 5)
```

```
                         Estimate Std. Error  z value Pr(>|z|)
(Intercept)              -4.53060    0.90223 -5.02153  0.00000
wage                      0.08442    0.02592  3.25714  0.00113
age                       0.02597    0.01095  2.37229  0.01768
female                   -0.59555    0.29111 -2.04579  0.04078
south                    -0.63577    0.29703 -2.14043  0.03232
as.factor(race)2         -0.38396    0.62853 -0.61089  0.54127
as.factor(race)3         -0.78680    0.34220 -2.29922  0.02149
as.factor(occupation)2 -0.16020    1.20540 -0.13291  0.89427
as.factor(occupation)3  1.41211    0.76008  1.85784  0.06319
as.factor(occupation)4  2.34356    0.72099  3.25049  0.00115
as.factor(occupation)5  1.97851    0.66585  2.97139  0.00296
as.factor(occupation)6  2.56000    0.67209  3.80900  0.00014
married                   0.42817    0.28264  1.51489  0.12980
```

```r
                    # null.deviance: 503.0841, deviance: 426.8709
```

# Effects

# Effects



# On the Agenda

# Fitting Generalized Additive Models

# Fitting Generalized Additive Models

Logistic regression model

$$\begin{aligned}\operatorname{logit}(p_i) = \beta_0 + f_1(\texttt{years.educ}_i) + ... + f_4(\texttt{age}_i) \\ + \beta_1 \texttt{female}_i + ... + \beta_6 \texttt{married}_i,\end{aligned}$$

$$\texttt{union.member}_i \sim \text{Bernoulli}(p_i), \quad i = 1, \ldots 534.$$

(4 quantitative variables, thus 4 smooth functions/splines, and 6, remaining, qualitative variables.)

# Fitting Generalized Additive Models

Logistic regression model

$$\begin{aligned}\text{logit}(p_i) = \beta_0 + f_1(\texttt{years.educ}_i) + \ldots + f_4(\texttt{age}_i) \\ + \beta_1\texttt{female}_i + \ldots + \beta_6\texttt{married}_i, \\ \texttt{union.member}_i \sim \text{Bernoulli}(p_i), \quad i = 1, \ldots 534.\end{aligned}$$

(4 quantitative variables, thus 4 smooth functions/splines, and 6, remaining, qualitative variables.)

```
formula <- union.member ~
  s(years.educ) + s(years.experience, k = 20) + s(wage, k = 20) +
  s(age, k = 20) + female + south + race + occupation + sector +
  married

union.gam <- mgcv::gam(formula, family = binomial, trade.union)
```

# Fitting Generalized Additive Models

Logistic regression model

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + f_1(\texttt{years.educ}_i) + ... + f_4(\texttt{age}_i) \\ &\quad + \beta_1\texttt{female}_i + ... + \beta_6\texttt{married}_i, \\ \texttt{union.member}_i &\sim \text{Bernoulli}(p_i), \quad i = 1, \ldots 534. \end{aligned}$$

(4 quantitative variables, thus 4 smooth functions/splines, and 6, remaining, qualitative variables.)

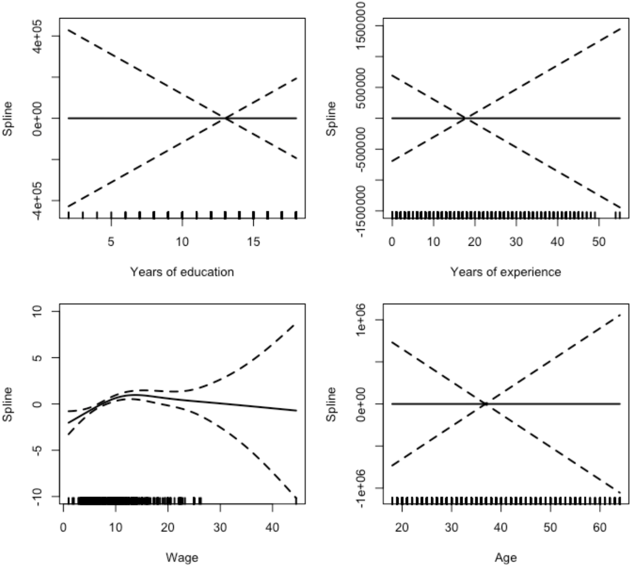
```
formula <- union.member ~
  s(years.educ) + s(years.experience, k = 20) + s(wage, k = 20) +
  s(age, k = 20) + female + south + race + occupation + sector +
  married

union.gam <- mgcv::gam(formula, family = binomial, trade.union)
```

Selecting a model looking to trade off between degree of freedom and RSS ...

Doing variable selection in qualitative features and looking to the qualitative ones . . .

Doing variable selection in qualitative features and looking to the qualitative ones . . .



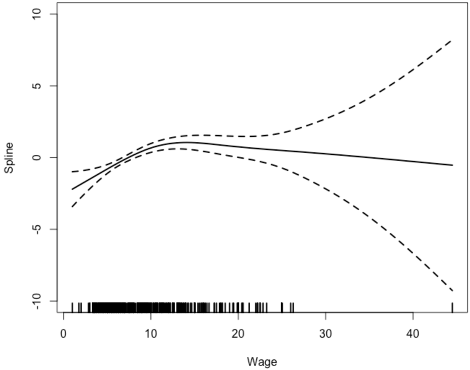
```r
round( anova(union.gam)$s.table, 5)
```

```
                          edf  Ref.df   Chi.sq p-value
s(years.educ)         1.06247 1.12229  0.00616 0.95612
s(years.experience) 1.00006 1.00000  0.00000 0.99978
s(wage)               2.73401 3.49695 23.78311 0.00008
s(age)                1.00005 1.00000  0.00000 0.99978
```

```r
round( anova(union.gam)$s.table, 5)
```

```
                         edf  Ref.df   Chi.sq p-value
s(years.educ)        1.06247 1.12229  0.00616 0.95612
s(years.experience) 1.00006 1.00000  0.00000 0.99978
s(wage)              2.73401 3.49695 23.78311 0.00008
s(age)               1.00005 1.00000  0.00000 0.99978
```

Doing variable selection in the qualitative's...

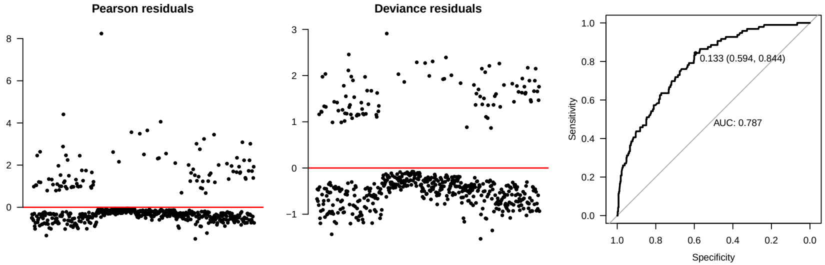


# Residues

```
union.gam$formula
```

```
union.member ~ s(wage, k = 20) + female + south + as.factor(race) +
    as.factor(occupation)
```

```
pearson <- residuals(union.gam, type = "pearson")
devi <- residuals(union.gam, type = "deviance")
rocurve <- roc(trade.union$union.member, fitted(union.gam))
```



# Coefficients

```r
round( summary(union.gam)$p.table, 5)
```

```
                           Estimate Std. Error  z value Pr(>|z|)
(Intercept)                -2.57646    0.68418 -3.76578  0.00017
female                     -0.39142    0.29565 -1.32392  0.18553
south                      -0.43371    0.30044 -1.44360  0.14885
as.factor(race)2           -0.01837    0.62644 -0.02932  0.97661
as.factor(race)3           -0.78659    0.34918 -2.25271  0.02428
as.factor(occupation)2 -0.22424    1.19651 -0.18741  0.85134
as.factor(occupation)3  1.08200    0.73580  1.47051  0.14142
as.factor(occupation)4  2.33738    0.69628  3.35695  0.00079
as.factor(occupation)5  1.73543    0.65435  2.65214  0.00800
as.factor(occupation)6  2.37213    0.64999  3.64951  0.00026
```

```r
summary(union.gam)$s.table
```

```
            edf   Ref.df   Chi.sq      p-value
s(wage) 2.82771 3.641582 29.64349 5.95743e-06
```

# On the Agenda

# Logistic regression model

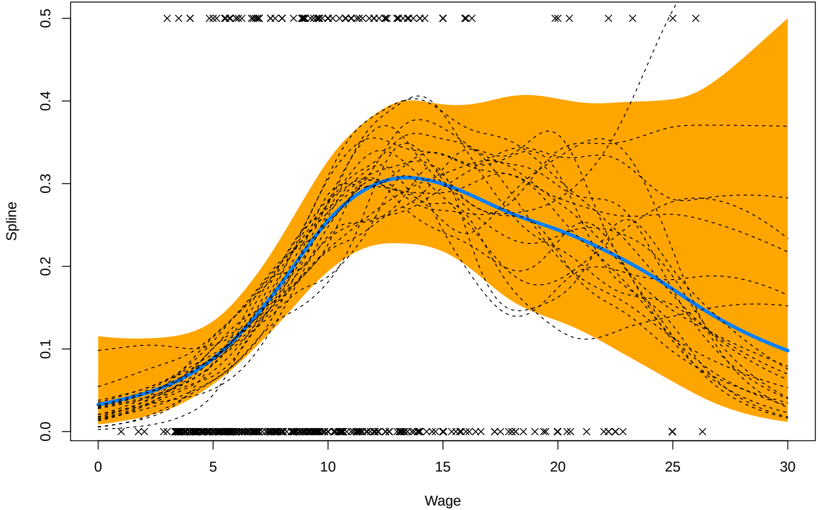
$$\text{logit}(p_i) = f(\texttt{wage}_i), \quad \texttt{union.member}_i \sim \text{Bernoulli}(p_i), \quad i = 1, \ldots 534.$$

JAGS model specification file

```
model {
  eta <- X %*% b
  for (i in 1:n) { mu[i] <-  ilogit(eta[i]) } # expected response
  for (i in 1:n) { y[i] ~ dbin(mu[i], w[i])              # response
  for (i in 1:1) { b[i] ~ dnorm(0, .018) }            # tau=1/7.5**2
                                                  # prior for s(wage)
  K1 <- S1[1:19, 1:19] * lambda[1]  + S1[1:19, 20:38] * lambda[2]
  b[2:20] ~ dmnorm(zero[2:20] ,K1)
                                          # smoothing parameter priors
  for (i in 1:2) {
    lambda[i] ~ dgamma(.05, .005)
    rho[i] <- log(lambda[i])
  }
}
```

# Results

Simulating from the model and addying a sample of 20 curves from the posterior.



# and is this. . .

thank you!

henrique.laureano@kaust.edu.sa