Ve406 Lecture 19

Jing Liu

UM-SJTU Joint Institute

July 23, 2018

- Recall GLMs assumes Y follows a certain conditional distribution with

$$\mu_i = \mathbb{E}\left[Y_i \mid \mathbf{X}_i\right]$$

where μ_i is modelled by by a linear predictor via a link function

$$g\left(\mu_i\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

- Generalised Additive Models (GAMs) are an extension of GLMs to

$$g\left(\mu_i\right) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik})$$

where f_j are often piecewise smooth functions of x_{ij}.

- GAMs are an extension of additive models (AMs) which only model

$$\mu_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_k(x_{ik})$$

- GAMs provide a powerful data-driven of models, especially, predictive models

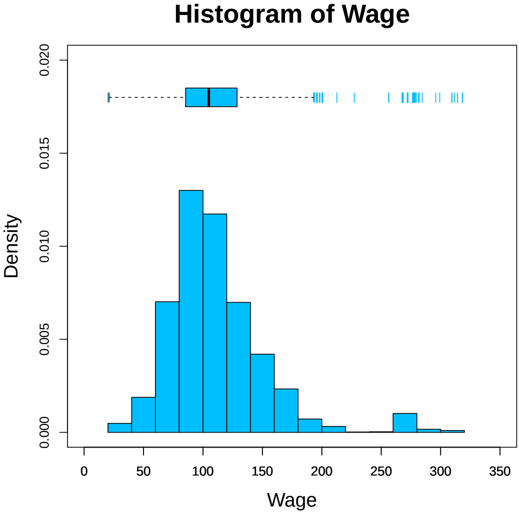
- Recall we struggled to obtain a satisfactory predictive model for the following

`wage`	Raw wage in the Mid-Atlantic region	
`age`	Age of the worker	
`year`	The year that wage information was recorded	
`education`	A factor with levels:	1. < HS Grad 2. HS Grad 3. Some College 4. College Grad 5. Advanced Degree

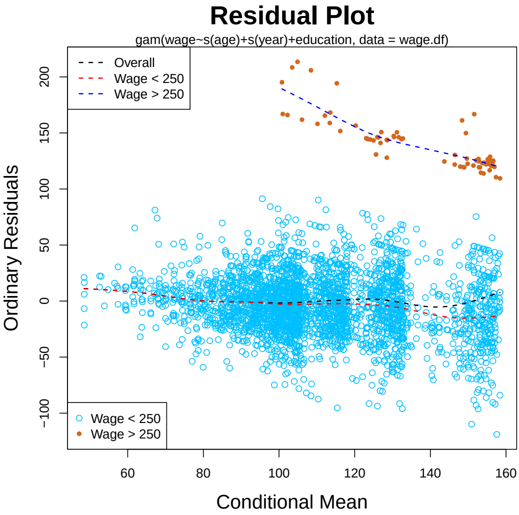
```
> library(ISLR); wage.df =
+   Wage[, c("year", "age", "education", "wage")]
> str(wage.df)
```

```
'data.frame':   3000 obs. of  4 variables:
 $ year     : int  2006 2004 2003 2003 2005 2008 2009 2008 2006 2004 ...
 $ age      : int  18 24 45 43 50 54 44 30 41 52 ...
 $ education: Factor w/ 5 levels "1. < HS Grad",..: 1 4 3 4 2 4 3 3 3 2 ...
 $ wage     : num  75 70.5 131 154.7 75 ...
```

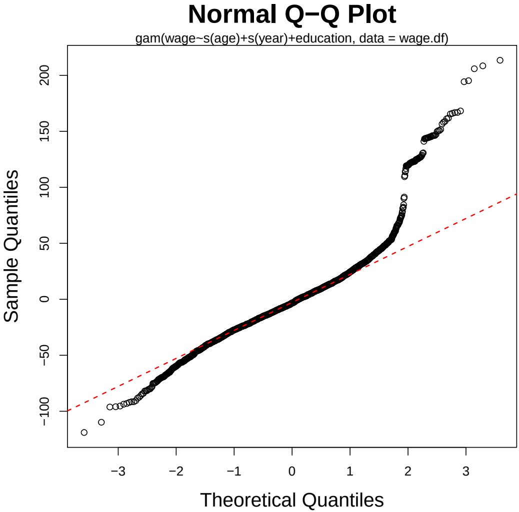
- Recall the outliers/extreme values in the wage was one of the concerns.



- An ordinary additive model is far from satisfactory.



- It is clear the errors are not normally distributed, and severally right skewed.



- The gamma distribution,

$$f_Y(y;\alpha,\theta) = \frac{\frac{1}{\theta}}{\Gamma(\alpha)}\left(\frac{y}{\theta}\right)^{\alpha-1}\exp\left(-\frac{y}{\theta}\right) \qquad \text{where} \quad \alpha>0, \theta>0$$

  is a flexible distribution that can be used to mode a response

$$y \in (0,\infty)$$

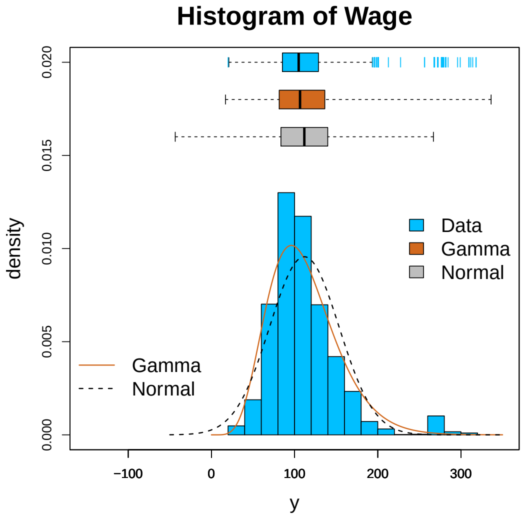
  which it makes more sense than normal for a response can only be positive.

- The mean and variance of gamma distribution is given by

$$\mu = \mathbb{E}[Y] = \alpha\theta \qquad \text{and} \qquad \sigma^2 = \mathrm{Var}[Y] = \alpha\theta^2 = \mu\theta$$

- This makes gamma useful to model increasing variability in the data.
- Having a longer right tail than normal makes gamma more suitable to model data that have a lot of unusually large values.

- Gamma is more flexible and more suitable for modelling wage.



- For known α, the log-likelihood is given by

$$\ell = -\alpha\left(\frac{y_i}{\mu} + \ln\mu\right) + c(y_i, \alpha)$$

where $c(y_i, \alpha)$ is a function of y_i and α only.

- In terms of regression, it can be done under under GLM and thus GAM using

$$g(\mu) = \mu^{-1} \qquad \text{or} \qquad g(\mu) = \ln(\mu)$$

as the link function and the shape parameter α is assumed to be a constant.

- In terms of GLM,

$$\frac{1}{\mu_i} = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \implies \text{a plot of } x_{ij} \text{ Vs } \frac{1}{\bar{y}_i} \text{ should be roughly linear}$$
$$\ln\mu_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \implies \text{a plot of } x_{ij} \text{ Vs } \ln\bar{y}_i \text{ should be roughly linear}$$

where $\bar{y}_i$ denotes the sample mean of y_i having the same value of x_{ij}.

- Of course, this requires we have repeated observations for each x_{ij},

```
> nrow(wage.df)
```

```
[1] 3000
```

```
> sort(unique(wage.df$age))
```

```
 [1] 18 19 20 21 22 23 24 25 26 27
[11] 28 29 30 31 32 33 34 35 36 37
[21] 38 39 40 41 42 43 44 45 46 47
[31] 48 49 50 51 52 53 54 55 56 57
[41] 58 59 60 61 62 63 64 65 66 67
[51] 68 69 70 71 72 73 74 75 76 77
[61] 80
```

```
> wage_age.df =
+   aggregate(list(wage = wage.df$wage),
+             by = list(age = wage.df$age), mean)
```

```
> head(wage_age.df)
```

```
  age     wage
1  18 64.49306
2  19 53.99049
3  20 69.03334
4  21 75.90695
5  22 72.25167
6  23 74.73047
```

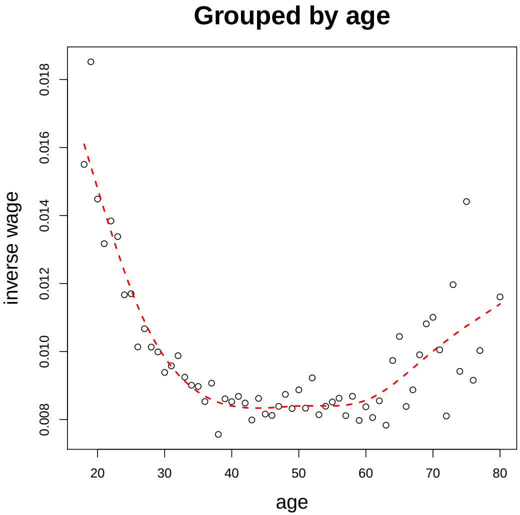
```
> mean(wage.df$wage[wage.df$age == 18])
```

```
[1] 64.49306
```

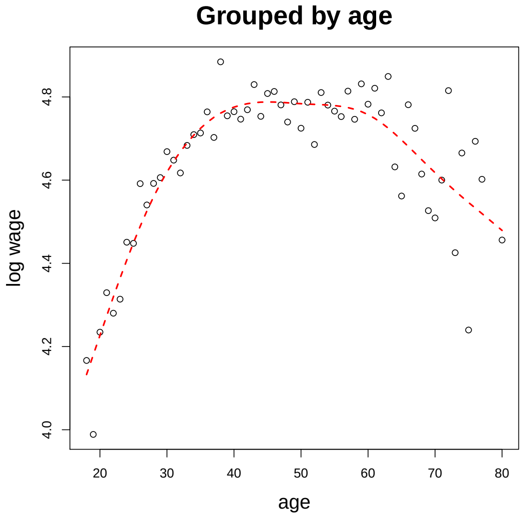
```
> mean(wage.df$wage[wage.df$age == 19])
```

```
[1] 53.99049
```

- It seems gamma regression under GLM with inverse link is not appropriate.



- It seems gamma regression under GLM with inverse link is not appropriate.



- Instead of trying various transformation on x_{ij}, let us use smoothing spline

```
> library(gam)
>
> wage.inv.GAM =
+   gam(wage~s(age)+s(year)+education,
+       family = Gamma(link = "inverse"),
+       data = wage.df)
>
> wage.log.GAM =
+   gam(wage~s(age)+s(year)+education,
+       family = Gamma(link = "log"),
+       data = wage.df)
```

- Just like GLMs, GAMs do not possess additive residuals to the predictor

$$y_i \neq \beta_0 + \hat{f}_1(x_{i1}) + \hat{f}_2(x_{i2}) + \cdots + \hat{f}_k(x_{ik}) + \hat{e}_i$$

- In general, Pearson residuals for GLM and GAM are defined as

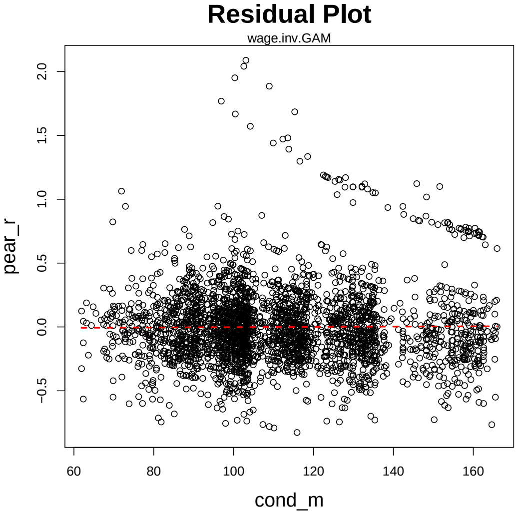
$$\hat{e}_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{Var}\left[\hat{\mu}_i\right]}}$$

which should approximately have zero mean and constant variance.

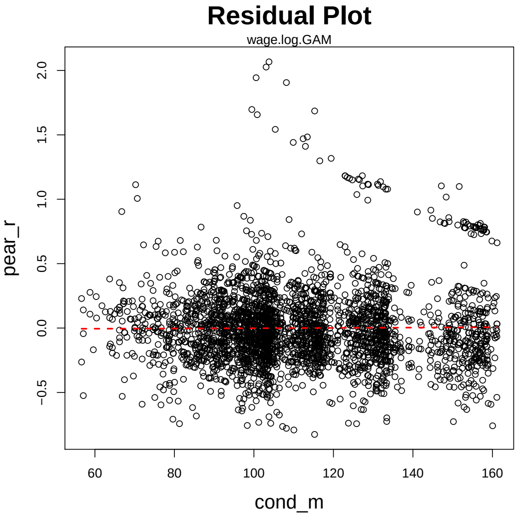
```
> res.inv.df = data.frame(
+    cond_m = fitted(wage.inv.GAM),
+    pear_r = residuals(wage.inv.GAM, type = "pearson"))

> with(res.inv.df, plot(
+    cond_m, pear_r, main = "Residual Plot",
+    cex.lab = 1.5, cex.main = 2))
>
> with(res.inv.df, lines(smooth.spline(
+    cond_m, pear_r), col = "red", lty = 2, lwd = 2))
>
> mtext("wage.inv.GAM")
```

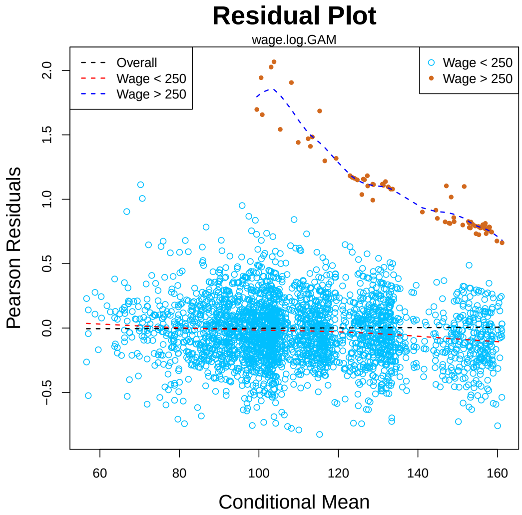
- The residual plot indicates no problem other than the outliers



- We obtain a similar residual plot, so we expect similar results from both



- You might think the residual plot didn't improve much from our first model



- However, both Gamma regression model is far better than the original model

```
> sum(residuals(wage.normal.GAM, type="pearson")^2)
```

```
[1] 3692824
```

```
> sum(residuals(wage.inv.GAM, type="pearson")^2)
```

```
[1] 260.4095
```

```
> sum(residuals(wage.log.GAM, type="pearson")^2)
```

```
[1] 261.5848
```

- Like logistic and Poisson regression, deviance can be used to check goodness of fit, but unlike logistic and Poisson, we have to use the scaled deviance

$$D^* = \frac{D}{\hat{\phi}} \sim \chi^2_{n-(k+1)} \qquad \text{where} \quad D = 2\left(\ell_{sat} - \ell_{prop}\right)$$

and $\hat{\phi}$ is MLE of dispersion parameter which is given by $\phi = \frac{1}{\alpha}$ for Gamma.

```
> summary(wage.inv.GAM)
```

```
(Dispersion Parameter for Gamma family taken to be 0.0872)

    Null Deviance: 371.6636 on 2999 degrees of freedom
Residual Deviance: 248.1586 on 2987 degrees of freedom
```

```
> 1 - pchisq(248.1586/0.0872, 2987)
```

```
[1] 0.9676302
```

```
> summary(wage.log.GAM)
```

```
(Dispersion Parameter for Gamma family taken to be 0.0876)

    Null Deviance: 371.6636 on 2999 degrees of freedom
Residual Deviance: 248.9206 on 2987 degrees of freedom
```

```
> 1 - pchisq(248.9206/0.0876, 2987)
```

```
[1] 0.9715768
```

- However, since the two models are not nested, we CANNOT use deviance based test to judge which model is a better one.
- We could consider AIC of the three models

```
> AIC(wage.normal.GAM, wage.inv.GAM, wage.log.GAM);
```

```
                 df      AIC
wage.normal.GAM  8 29888.23
wage.inv.GAM     8 29029.51
wage.log.GAM     8 29038.84
```

which seems to prefer the gamma regression with the inverse link.

- We could also use cross-validation to judge the quality of our models

```
> k = 100 # number of subsamples
> n = nrow(wage.df)
> n.test = n/k
> row.index = sample(1:n, n)
> pred.nor = matrix(0, nrow = n.test, ncol = k)
> pred.inv = matrix(0, nrow = n.test, ncol = k)
> pred.log = matrix(0, nrow = n.test, ncol = k)
```

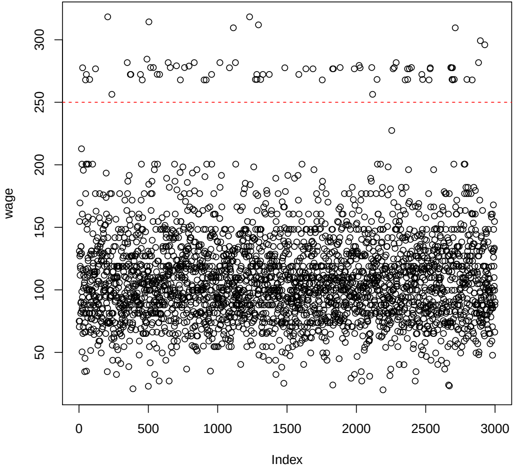
```
> for (i in 1:k){
+    start = (i-1)*n.test+1; end = i*n.test
+    index = row.index[start:end]
+    wage.inv.GAM = gam(wage~s(age)+s(year)+education,
+          data = wage.df[-index,])
+    pred.inv[, i] = predict(wage.inv.GAM,
+      wage.df[index,], type = "response")
}

> tmp = pred.nor[1:n] - wage.df[row.index, "wage"]
> mse.nor = mean((tmp)^2)
> tmp = pred.inv[1:n] - wage.df[row.index, "wage"]
> mse.inv = mean((tmp)^2)
> tmp = pred.log[1:n] - wage.df[row.index, "wage"]
> mse.log = mean((tmp)^2)
```

Q: Which one do you think is the best model in terms of MSE estimated by CV?

```
> c(mse.nor, mse.inv, mse.log)
[1] 1241.127 1231.781 1234.101
```

- Recall we didn't find any single variable that can be used to explain



- We could use logistic regression under GAM to see whether collectively

`age`, `year` and `education`

can explain this apparent separation in `wage`.

```
> wage.LG.GAM =
+   gam(I(wage > 250)~s(age)+s(year)+education,
+       data = Wage, family = binomial)

> 1 - pchisq(wage.LG.GAM$deviance,
+            wage.LG.GAM$df.residual)
```

```
[1] 1
```

which means there is no indication of lack of fit.

- This indicates that we probably should consider using mixture models, which will be covered next week, so we will revisit this dataset again!