Homework 1 Regression Models and Generalized Linear Models (GLMs)

Lynnstacy Kegeshi

2025-03-15

Contents

Question 1

Distinguish between the best regression models when considering the professional salary dataset.

- Linear
- Non-linear; Polynomial degree 2 (Quadratic)
- Polynomial; degree 3 trend (Cubic)

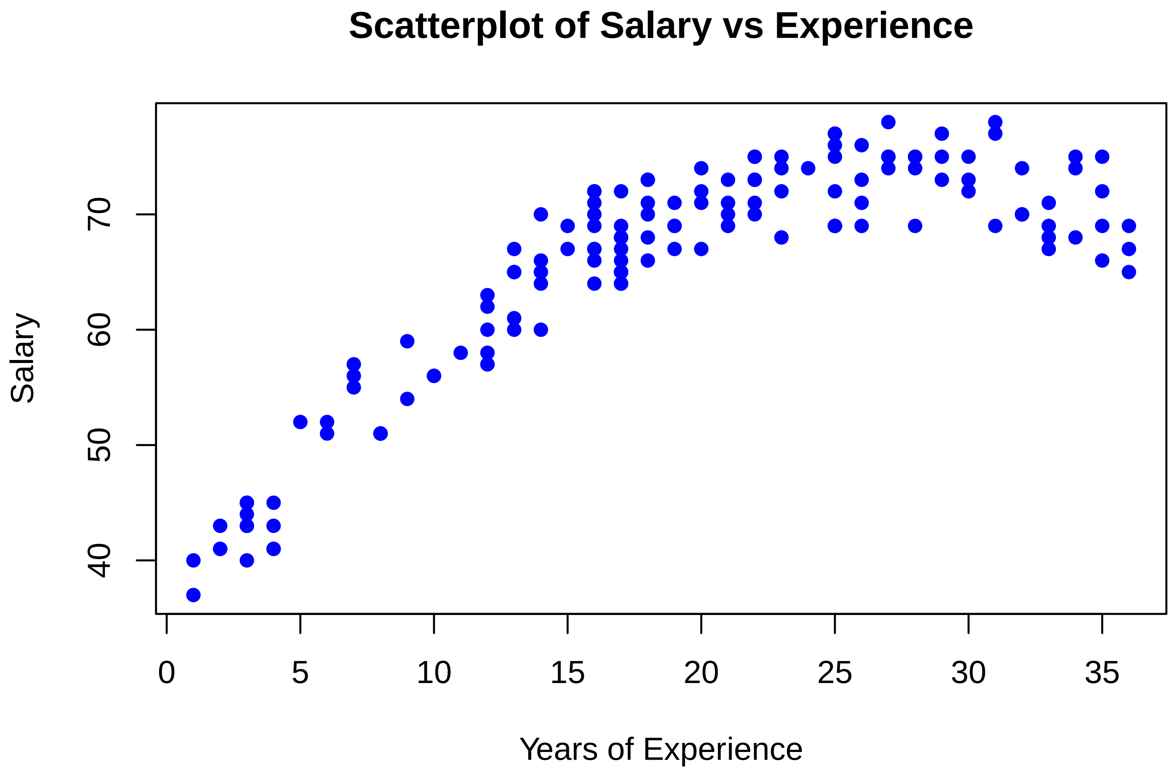
First step in solving this question is loading the data into R and making the column of the data set accessible as variables without needing the $.

```r
#Loading the data into R
prof_salary <- read.table("profsalary.txt",header = T)

#Attaching data to environment
attach(prof_salary)
```

We want to model the relationship between salary and experience. We first visualize the relationship between experience and salary to observe the trend. This will help us decide whether a linear, quadratic, or cubic model might be appropriate.

```r
# Scatterplot and linear regression in one chunk
sal_prof <- plot(Experience, Salary,
     main = "Scatterplot of Salary vs Experience",
     xlab = "Years of Experience", ylab = "Salary",
     pch = 16, col = "blue")
```



From the scatter plot, we see a positive correlation between salary and experience. On average, salaries tend to increase as experience grows. However, we observe some variability and a slight curve in the scatterplot, indicating that a quadratic model might provide a better fit. Thus, we proceed to test all three models to determine the most accurate representation of the data.

Linear Model

A linear regression model assumes a straight-line relationship between Experience and Salary.

```r
# Fit a linear regression model
ModLinear <- lm(Salary ~ Experience)
```

Quadratic Regression Model (Polynomial Degree 2)

A quadratic model captures curvature in the relationship we had earlier observed from the plot.

```r
# Fit a quadratic regression model
ModQuadratic <- lm(Salary ~ Experience + I(Experience^2))
```

Cubic Regression Model (Polynomial Degree 3)

A cubic model captures more complex patterns but sometimes can overfit the data

```r
# Fit a cubic regression model
ModCubic <- lm(Salary ~ Experience + I(Experience^2) + I(Experience^3))
```

Comparing the models

To determine the best model, we compare R^2 and AIC values:

```r
# Calculate R² and AIC for each model
r2_linear <- summary(ModLinear)$r.squared
r2_quadratic <- summary(ModQuadratic)$r.squared
r2_cubic <- summary(ModCubic)$r.squared

aic_linear <- AIC(ModLinear)
aic_quadratic <- AIC(ModQuadratic)
aic_cubic <- AIC(ModCubic)

# Creating a dataframe
model_comparison <- data.frame(
  Model = c("Linear", "Quadratic", "Cubic"),
  R_Squared = c(r2_linear, r2_quadratic, r2_cubic),
  AIC = c(aic_linear, aic_quadratic, aic_cubic)
)

# Display the dataframe
model_comparison
```

```
##       Model R_Squared      AIC
## 1    Linear 0.6753623 913.9240
## 2 Quadratic 0.9246756 707.0146
## 3     Cubic 0.9256789 707.0971
```

Based on the above table, we see clear differences in how well the three models explain salary variations based on years of experience.

The linear model suggests a general upward trend, but with an R-squared of 0.675 and the highest AIC (913.92), it does not capture the full complexity of the relationship.

The quadratic model, however, provides a much better fit, with an R-squared of 0.925 and a significantly lower AIC (707.01), indicating that salary growth is not strictly linear but follows a curved pattern.

The cubic model improves the R-squared slightly to 0.926, but its AIC (707.10) is nearly identical to the quadratic model, meaning the added complexity does not offer a meaningful improvement.

Given these results, the quadratic model is the most suitable choice for modeling salary trends which also agrees with our initial scatter plot visualization.

Question 2

Describe the Generalized Linear Model framework and its application by enumerating the assumptions and scenarios where it differs from the typical linear regression framework.

Generalized Linear Model (GLM) Framework

The **Generalized Linear Model (GLM)** framework is an extension of the traditional linear regression model that allows us to model relationships between variables when the assumptions of linear regression (such as normally distributed residuals) don't hold. The GLM framework is particularly useful for handling different types of response variables and distributional assumptions. It's widely used when we can't rely on the assumption that the residuals are normally distributed, and it accommodates response variables from different families of distributions.

The GLM consists of three main components:

1. **Random Component**: This is the distribution of the response variable Y. Instead of assuming that the response follows a normal distribution (as in linear regression), GLMs allow for various distributions, such as Binomial, Poisson, and Gamma, depending on the nature of the data.
2. **Systematic Component**: This is the linear predictor η, which is just a weighted sum of the predictor variables. It's similar to the linear predictor in traditional linear regression:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

3. **Link Function**: The link function $g(\mu)$ connects the linear predictor η to the expected value of the response variable μ:

$$g(\mu) = \eta$$

   The link function allows us to transform the expected value of the response variable so that we can model non-linear relationships between predictors and the response.

How GLM Differs from Linear Regression

The key differences between GLM and linear regression are:

1. **Response Distribution**: Linear regression assumes that the response variable follows a normal distribution. In contrast, GLMs can handle a variety of distributions, such as Binomial for binary data, Poisson for count data, and Gamma for continuous, skewed data.
2. **Link Function**: In GLM, the relationship between the predictors and the response variable is modeled through a link function, which transforms the expected value of the response. Linear regression assumes that the response and predictors are related linearly.
3. **Variance Structure**: While linear regression assumes that the variance of the residuals is constant (homoscedasticity), GLMs allow the variance of the response variable to depend on its mean. This makes GLMs more flexible in handling data with different variance structures.

Applications

1. **Medical Sciences**
   GLMs are widely used in medical fields, especially in areas like epidemiology and clinical research. Logistic regression, for instance, helps predict the probability of a binary outcome, such as the presence or absence of a disease. Poisson regression is often applied to count data, like hospital visits or disease occurrences, aiding in understanding patterns and making predictions.

2. **Social Sciences and Economics**
   In social sciences, logistic regression models binary outcomes, like voting behavior or engagement in certain actions. Poisson regression models count data, such as crime rates or accidents. These models help analyze relationships between variables, making them valuable for policy analysis and decision-making.

3. **Insurance and Actuarial Sciences**
   In insurance, GLMs are used to model claims data. Gamma and inverse Gaussian distributions help model claim amounts, while Poisson regression can model the frequency of claims. These models are essential for setting fair premiums and predicting risks, enabling insurers to adjust pricing based on claim probabilities.

4. **Finance**
   In finance, GLMs help model events like loan defaults or bankruptcies. Poisson regression is used for modeling count data, such as the frequency of defaults. Quasi-binomial and quasi-Poisson models handle overdispersed data, helping financial institutions assess risk and make data-driven portfolio decisions.

5. **Environmental Sciences**
   GLMs are important in environmental sciences, where they model data like pollutant levels or rainfall amounts. The Gamma and inverse Gaussian distributions are well-suited for skewed environmental data, helping researchers predict future trends and understand the impact of various factors on the environment.

6. **Marketing and Customer Research**
   In marketing, GLMs predict consumer behavior, like whether a customer will purchase a product. Logistic regression models the likelihood of a purchase based on demographics and past behavior. These insights allow businesses to target specific customer segments and improve marketing strategies.

7. **Epidemiology**
   GLMs are key in epidemiology for modeling disease spread and risk factors. Logistic regression predicts whether an individual has a disease, while Poisson regression models the rate of disease occurrence. These models are crucial for understanding public health issues and guiding intervention strategies.

Assumptions with Different Link Functions in GLM

Describe the assumptions with adopting different link functions in a GLM framework.

The GLM framework uses different link functions based on the type of response variable.

1. Binomial (logit link = "logit"):

- **Response Variable**: Binary outcome (e.g., success/failure, 1/0).
- **Link Function**: The logit function transforms probabilities into log-odds:

$$\log\left(\frac{\mu}{1-\mu}\right) = \eta$$

- **Assumptions**:
  - The response variable follows a Binomial distribution (for binary outcomes).
  - The outcome is a probability that is transformed using the logit link to model the log-odds.
  - The expected value μ lies between 0 and 1 (since it represents a probability).

2. Gaussian (identity link = "identity"):

- **Response Variable**: Continuous and normally distributed outcome (e.g., height, salary).
- **Link Function**: The identity function does not transform the mean and simply models the response variable directly:

$$\mu = \eta$$

- **Assumptions**:
  - The response variable follows a Normal distribution with a constant variance.
  - The relationship between predictors and the response is linear.
  - No transformation of the response is needed, so the model directly predicts the expected value of the response.

3. Gamma (inverse link = "inverse"):

- **Response Variable**: Continuous positive values (e.g., time, money, rate of occurrence).
- **Link Function**: The inverse function transforms the mean to model the response as an inverse of the linear predictor:

$$\frac{1}{\mu} = \eta$$

- **Assumptions**:
  - The response variable follows a Gamma distribution (typically used for modeling positive continuous data, such as waiting times or costs).
  - The variance increases as the mean increases (heteroscedasticity).
  - The inverse link ensures that the predicted values remain positive.

4. Inverse Gaussian (link = "
$frac1mu^2$**")**

- **Response Variable**: Continuous positive values.
- **Link Function**: The link function is of the form $\frac{1}{\mu^2} = \eta$, where μ represents the expected value.

$$\frac{1}{\mu^2} = \eta$$

- **Assumptions**:
  - The response variable follows an Inverse Gaussian distribution, often used for modeling positively skewed data.
  - The variance depends on the square of the mean, making it useful for modeling highly skewed distributions.

5. Poisson (log link = "log"):

- **Response Variable**: Count data (e.g., number of events occurring in a time period).
- **Link Function**: The log function transforms the expected count into the linear predictor:

$$\log(\mu) = \eta$$

- **Assumptions**:
  - The response variable follows a Poisson distribution, which is suitable for modeling count data with events that occur independently.
  - The variance is equal to the mean, which is a characteristic property of the Poisson distribution.
  - The log link ensures that the predicted values are positive, which is essential since counts cannot be negative.

6. **Quasi (identity link = "identity", variance = "constant"):**

- **Response Variable**: Continuous response.
- **Link Function**: The identity link function is used, similar to the Gaussian model:

$$\mu = \eta$$

- **Assumptions**:
  - The variance is assumed to be constant, similar to linear regression (homoscedasticity).
  - The quasi-likelihood model is used when the response variable does not conform to a specific distribution (e.g., Normal) but still exhibits constant variance.
  - Suitable when the data exhibits overdispersion (variance greater than expected from the assumed distribution).

7. **Quasibinomial (logit link = "logit"):**

- **Response Variable**: Binary outcome with overdispersion (e.g., success/failure rates with more variability than the binomial distribution assumes).
- **Link Function**: The logit link is used to model binary outcomes, similar to the Binomial model:

$$\log\left(\frac{\mu}{1-\mu}\right) = \eta$$

- **Assumptions**:
  - The response variable is binary, but there is overdispersion, meaning the variance is greater than expected under the Binomial model.
  - The logit link is still appropriate for transforming the binary probabilities.

8. **Quasipoisson (log link = "log"):**

- **Response Variable**: Count data with overdispersion (i.e., the variance is greater than the mean).
- **Link Function**: The log link function is used, similar to the Poisson model:

$$\log(\mu) = \eta$$

- **Assumptions**:
  - The response variable is count data, but with overdispersion (variance greater than the mean).
  - The log link is used to model the relationship between predictors and the expected count, while accounting for overdispersion in the data.