HW 1 Math 7450

Wenjuan Bian

2023-09-03

1. Assume that the time to divorce follows a normal model with the mean of 7 years and standard deviation of 4 years. If a study only recruits couples who have been married for at least four years, what is the model of the data?

Answer:

Assume that the least time to divorce is 0 and the longest divorce time is 100 years, the normal model for the population is a normal model with both ends truncated. The lower bound is 0 and the upper bound is 100. We can write the model for population divorce time is:

$$f(x|0 < x < 100) = \frac{\mathcal{N}(\mu, \sigma)}{F(\mu, \sigma|x < 100) - F(\mu, \sigma|x < 0))}$$

Thus,

$$E(x) = \frac{\int_0^{100} \frac{x}{\sqrt{2\pi \cdot \sigma^2}} e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}} dx}{F(\mu, \sigma|x < 100) - F(\mu, \sigma|x < 0))}$$

Let $t = \dfrac{x - \mu}{\sigma}$, we get

$$E(x) = \frac{\int_a^b \frac{t\sigma + \mu}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}{\Phi(b) - \Phi(a))}$$

Where:

$$a = \frac{\mu}{\sigma}$$
$$b = \frac{100 - \mu}{\sigma}$$

and, Φ is the cdf of standard normal distribution.

$$\begin{aligned} E(x) &= \frac{\int_a^b \frac{t\sigma}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}{\Phi(b) - \Phi(a))} + \frac{\int_a^b \frac{\mu}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}{\Phi(b) - \Phi(a))} \\ &= \frac{\int_a^b \frac{t\sigma}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}{\Phi(b) - \Phi(a))} + \mu \end{aligned}$$

Computer $\int_a^b \frac{t\sigma}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$:

$$\begin{aligned}
&\int_a^b \frac{t\sigma}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt \\
&= \sigma \int_a^b \frac{-1}{\sqrt{2\pi}} de^{-\frac{t^2}{2}} \\
&= \frac{\sigma}{\sqrt{2\pi}}[-e^{-\frac{t^2}{2}}]|_a^b \\
&\approx \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}}
\end{aligned}$$

Thus, we have:

$$E(x) \approx \frac{\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}[\Phi(b)-\Phi(a)]} + \mu = 7$$

Similarly,

$$\begin{aligned}
E(x^2) &= \frac{\int_0^{100} \frac{x^2}{\sqrt{2\pi\cdot\sigma^2}} e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}\,dx}{F(\mu,\sigma|x<100) - F(\mu,\sigma|x<0))} \\
&= \frac{\int_a^b \frac{(t\sigma+\mu)^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt}{\Phi(b)-\Phi(a))} \\
&= \frac{\int_a^b \frac{t^2\sigma^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt}{\Phi(b)-\Phi(a))} + \frac{\int_a^b \frac{2t\mu\sigma}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt}{\Phi(b)-\Phi(a))} + \frac{\int_a^b \frac{\mu^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt}{\Phi(b)-\Phi(a))} \\
&= \frac{\int_a^b \frac{t^2\sigma^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt}{\Phi(b)-\Phi(a))} + \frac{2\mu\sigma}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}} + \mu^2
\end{aligned}$$

compute $\int_a^b \frac{t^2\sigma^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt$:

$$
\begin{aligned}
&\int_a^b \frac{t^2\sigma^2}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt \\
&\qquad = \int_a^b \frac{-t\sigma^2}{\sqrt{2\pi}} de^{-\frac{t^2}{2}} \\
&\qquad = \left[-t\sigma^2 \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\right]_a^b + \sigma^2 \int_a^b \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}} dt \\
&\qquad = \left[-t\sigma^2 \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\right]_a^b + \sigma^2(\Phi(b) - \Phi(a)) \\
&\qquad \approx -\mu\sigma \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}} + \sigma^2(\Phi(b) - \Phi(a))
\end{aligned}
$$

Therefore,

$$E(x^2) = \sigma^2 + \mu^2 + \frac{\mu\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}\,[\Phi(b) - \Phi(a)]}$$

We can compute variance for x as:

$$Var(x) = E(x^2) - [E(x)]^2)$$

$$= \sigma^2 + \mu^2 + \frac{\mu\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}\,[\Phi(b) - \Phi(a)]} - \left(\frac{\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}\,[\Phi(b) - \Phi(a)]} + \mu\right)^2$$

$$= \sigma^2 - \frac{\mu\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}\,[\Phi(b) - \Phi(a)]} - \frac{\sigma^2 e^{-\frac{\mu^2}{\sigma^2}}}{2\pi[\Phi(b) - \Phi(a)]^2} = 16$$

Combine E(x) and Var(x), we get

$$
\begin{cases}
\sigma^2 - \dfrac{\mu\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}\,[\Phi(\frac{100-\mu}{\sigma}) - \Phi(\frac{-\mu}{\sigma})]} - \dfrac{\sigma^2 e^{-\frac{\mu^2}{\sigma^2}}}{2\pi[\Phi(\frac{100-\mu}{\sigma}) - \Phi(\frac{-\mu}{\sigma})]^2} = 16 \\
\dfrac{\sigma e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi}\,[\Phi(\frac{100-\mu}{\sigma}) - \Phi(\frac{-\mu}{\sigma})]} + \mu = 7
\end{cases}
$$

Solve σ and μ by above equation in R

```
library(nleqslv)

equations <- function(x) {
  mu <- x[1]
  sigma <- x[2]

  phi_term1 <- pnorm((100 - mu)/sigma) - pnorm(-mu/sigma)
  phi_term2 <- phi_term1^2

  e_term1 <- exp(-mu^2 / (2 * sigma^2))
  e_term2 <- exp(-mu^2 / sigma^2)

  equation1 <- sigma^2 - (mu * sigma * e_term1) / (sqrt(2 * pi) * phi_term1) - (sigma^2 * e_term
2) / (2 * pi * phi_term2) - 16
  equation2 <- (sigma * e_term1) / (sqrt(2 * pi) * phi_term1) + mu - 7

  c(equation1, equation2)
}

initial_guess <- c(5, 5)
solution <- nleqslv(initial_guess, equations)
solution$par
```

```
## NULL
```

```
print(solution$x)
```

```
## [1] 6.076342 4.739790
```

Therefore, we got

μ = 6.1

σ = 4.7

$b = \frac{(100-6.1)}{4.7} \approx 20.0$

$a = \frac{(0-6.1)}{4.7} \approx -1.3$

$\phi(b) - \phi(a) = 0.90$

Therefore, the density function of the truncated normal model for the population is:

$$f(x|0 < x < 100) = \frac{\mathcal{N}(6.1, 4.7)}{0.90}$$

Check whether the mean is 7 and standard deviation is 4:

```
library(truncnorm)
```

```
## Warning: package 'truncnorm' was built under R version 4.3.1
```

```
n <- 1e6

mean_p <- etruncnorm( 0, 100, 6.1, 4.7)
sample1 <- rtruncnorm(n, 0, 100, 6.1, 4.7)
sd_p <- sd(sample1)

cat("The population mean is", mean_p)
```

```
## The population mean is 6.994584
```

```
cat("The standard deviation for the population is", sd_p)
```

```
## The standard deviation for the population is 3.979589
```

If we sample among the couples who have been married for at least four years, then the lower bound for the truncated normal model is 4, the corresponding

$a = \frac{(4-6.1)}{4.7} \approx -0.45$

$\phi(b) - \phi(a) = 0.67$

Therefore, the density function of the truncated normal model for the study is:

$$f(x|4 < x < 100) = \frac{\mathcal{N}(6.1, 4.7)}{0.67}$$

2. In the truncated data regarding the time to divorce example, what is the chance that a randomly selected couple from the data set have marriage more than 5 years? What is the long term average of the time to divorce for couples in the sample?

Solution:

$$\frac{\mathcal{N}(6.1, 4.7|5 < x < 100)}{0.67} = 0.59/0.67 = 0.88$$

By question one, we get

$$E(x) \approx \frac{4.7e^{-\frac{(4-6.1)^2}{2(4.7^2)}}}{\sqrt{2\pi}(0.67)} + 6.1$$

```
mu <- 6.1
sigma <- 4.7
E_x <- (sigma * exp(-(4-mu)^2 / (2 * sigma^2)) / (sqrt(2 * pi) * 0.67)) + mu
E_x
```

```
## [1] 8.632692
```

Or

```
library(truncnorm)

n <- 1e6

mean_p <- etruncnorm( 4, 100, 6.1, 4.7)
sample1 <- rtruncnorm(n, 4, 100, 6.1, 4.7)
sd_p <- sd(sample1)

cat("The mean for the study data is", mean_p)
```

```
## The mean for the study data is 8.623301
```

```
cat("The standard deviation for the study data is", sd_p)
```

```
## The standard deviation for the study data is 3.229325
```

Therefore, The long term average of the time to divorce for couples in the sample is 8.6

3. In practice, we do not have the assumption on the underlying model of the data, how to obtain an unbiased estimator for the population mean? Explain the difference between the population mean and sample mean using the concept of unbiased estimator for the divorce data.

Solution:

If we don’t have the assumption on the underlying model, the sample mean is till the unbiased estimator for the population mean. According to law of large numbers, the sample mean converges to the population mean as the sample size increase.

For the divorce data, the entire set of all individuals who’ve ever been divorced in a country represents the population.

The true average time to divorce for everyone in that country is the population mean. If we take a random sample of, say, 1,000 divorced individuals and calculate their average time to divorce, that is the sample mean. If you were to repeatedly take different samples of 1,000 individuals and calculate the sample mean each time, we will get a distribution of sample means. An unbiased estimator means that the average of all those sample means would be equal to the population mean.

For the divorce data, the given mean of 7 is the population mean. The sample mean of mean time to divorce in the study is computed by taking the avarage of all the time to divorce data that the researchers have collected from the recruited couples who have been married for at least four years.

5. Assume that the time to death (in days) after a kidney transplantation follows a log logistic distribution with density

$$f(t) = \frac{\lambda \alpha t^{\alpha-1}}{(1+\lambda t^{\alpha})^2}$$

with and λ=0.02. Find and interpret the corresponding mean and median

Solution

The cumulative distribution function (CDF) for the log-logistic distribution is:

$$F(t) = \int_0^t \frac{\lambda \alpha t^{\alpha-1}}{(1+\lambda t^{\alpha})^2}\, dt$$

$$= 1 - \frac{1}{1+\lambda t^{\alpha}}$$

$$= \frac{\lambda t^{\alpha}}{1+\lambda t^{\alpha}}$$

$$= \frac{1}{1+\lambda^{-1} t^{-\alpha}}$$

a. The median, M, is defined by $F(t) = 0.5.$ \

Hence,

$$= \frac{1}{1+\lambda^{-1} M^{-\alpha}} = 0.5$$

$$1+\lambda^{-1} M^{-\alpha} = 2$$

$$\lambda^{-1} M^{-\alpha} = 1$$

$$M = \lambda^{\frac{-1}{\alpha}} = (0.02)^{\frac{-1}{1.2}} \approx 26.05$$

b. Mean of the Log-logistic Distribution

For the mean, we compute expected value:

$$E(X) = \int_0^{\infty} t f(t)\, dt$$

$$= \int_0^{\infty} t \left(\frac{\lambda \alpha t^{\alpha-1}}{(1+\lambda t^{\alpha})^2} \right) dt$$

$$= \int_0^{\infty} \frac{\lambda \alpha t^{\alpha}}{(1+\lambda t^{\alpha})^2}\, dt$$

Let $z = 1 + \lambda t^{\alpha}$

Then, $t = (\frac{z-1}{\lambda})^{\frac{1}{\alpha}}$

$$\int_0^\infty \frac{\lambda\alpha t^\alpha}{(1+\lambda t^\alpha)^2}\,dt = \int_1^\infty (\frac{z-1}{\lambda})^{\frac{1}{\alpha}}\ \frac{1}{z^2}\,dz$$

$$= \lambda^{\frac{-1}{\alpha}} \int_1^\infty (z-1)^{\frac{1}{\alpha}}\ d\frac{1}{z}$$

$$= \lambda^{\frac{-1}{\alpha}} \int_1^\infty (1-\frac{1}{z})^{\frac{1}{\alpha}}\ (\frac{1}{z})^{\frac{-1}{\alpha}}\ d\frac{1}{z}$$

$$= \lambda^{\frac{-1}{\alpha}} \int_1^\infty -\left(1-\frac{1}{z}\right)^{\left(\frac{1}{\alpha}+1-1\right)} \left(\frac{1}{z}\right)^{\left(\frac{-1}{\alpha}+1-1\right)} d\frac{1}{z}$$

let $x = \frac{1}{z}$

$$\lambda^{\frac{-1}{\alpha}} \int_0^1 (1-x)^{\left(\frac{1}{\alpha}+1-1\right)}\ x^{\left(\frac{-1}{\alpha}+1-1\right)}\,dx$$

$$= \lambda^{\frac{-1}{\alpha}} B[\left(\frac{1}{\alpha}+1\right), \left(\frac{-1}{\alpha}+1\right)]$$

$$= \lambda^{\frac{-1}{\alpha}} \frac{\Gamma\left(\frac{1}{\alpha}+1\right)\Gamma\left(\frac{-1}{\alpha}+1\right)}{\Gamma(2)}$$

$$= \lambda^{\frac{-1}{\alpha}}\ \frac{1}{\alpha}\,\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(\frac{-1}{\alpha}+1\right)$$

$$= \frac{\pi\lambda^{\frac{-1}{\alpha}}}{\alpha\sin\left(\frac{\pi}{a}\right)}$$

Therefore, we get:

mean = $\frac{\pi\lambda^{\frac{-1}{\alpha}}}{\alpha\sin\left(\frac{\pi}{a}\right)}$

```
alpha <- 1.2
lambda <- 0.02
mean_value <- (pi * lambda^(-1/alpha)) / (alpha * sin(pi/alpha))
print(mean_value)
```

```
## [1] 136.3977
```

Check by log-logistic function in R:

```
library(flexsurv)
```

```
## Warning: package 'flexsurv' was built under R version 4.3.1
```

```
## Loading required package: survival
```

```
sample2 <- rllogis(6e5, shape=1.2, scale=26.05)
mean(sample2)
```

```
## [1] 116.9933
```

```
median(sample2)
```

```
## [1] 26.06225
```

The Median is 26 days. This indicates that half of the kidney transplant patients have passed away by the 26th day, while the other half survive for longer than 26 days.

The Mean is 136 days, indicating that the average survival time of all kidney transplant patients is 136 days.

6. If $E(\hat{\theta}) = \theta$ for an unknown parameter θ, is $(\hat{\theta})^2$ an unbiased estimator for θ^2 ? Explain.

Answer:

If $E(\hat{\theta}) = \theta$ for an unknown parameter $$, it doesn't necessarily imply that $E(\hat{\theta}^2)$ is equal to θ^2. This is because:

$$E(\hat{\theta}^2) \neq [E(\hat{\theta})]^2$$

Given $E(\hat{\theta}) = \theta$:

$$E(\hat{\theta}^2) = \mathrm{Var}(\hat{\theta}) + [E(\hat{\theta})]^2$$
$$E(\hat{\theta}^2) = \mathrm{Var}(\hat{\theta}) + \theta^2$$

For $E(\hat{\theta}^2)$ to be an unbiased estimator of θ^2, $\mathrm{Var}(\hat{\theta})$ should be zero, which isn't the case for an estimator.

Thus, just because $\hat{\theta}$ is an unbiased estimator of θ doesn't imply that $\hat{\theta}^2$ is an unbiased estimator of θ^2. The variance of $\hat{\theta}$ plays a crucial role in determining the unbiasedness of $\hat{\theta}^2$.

7. Assume that the variation of systolic blood pressure is STD=4 for patients before taking a blood pressure medicine and STD=2 for patients after taking the BP medicine. Also assume that the correlation between the two measurements is 0.4 on the same experimental subject. Find the variation of the drug effect on systolic blood pressure. How to find an unbiased estimator for the variation of the drug effect? Explain.

Solution

Let x be the blood pressure before medicine and y be the blood pressure after medicine, then the drug effect on systolic blood pressure is y-x. Therefore the variation of drug effect is Var(y-x). By definition:

$$\begin{aligned} Var(y-x) &= E[(y-x)-E(y-x)]^2 \\ &= E[(y-E(y))-(x-E(x))]^2 \\ &= E[(y-E(y)^2+(x-E(x))^2-2(y-E(y))(x-E(x))] \\ &= E[(y-E(y))^2]+E[(x-E(x))^2]-2E[(y-E(y))(x-E(x))] \\ &= Var(y)+Var(x)-2Cov(x,y) \end{aligned}$$

$$\begin{aligned} Cov(x,y) &= \rho_{xy}\cdot\sqrt{Var(x)}\sqrt{Var(y)} \\ &= 0.4\times 2\times 4 \\ &= 3.2 \end{aligned}$$

Therefore,

$$\begin{aligned} &Var(y-x) \\ &= Var(y)+Var(x)-2Cov(x,y) \\ &= 2^2+4^2-2\times 3.2 \\ &= 13.6 \end{aligned}$$

8. The data set insurance-data-txt (available in CANVAS) contains consumer information of 6-month insurance premium and the driving experience in years.

8.1 Divide the data into two sub data sets, one for customers who drove more than two years and another for those who drove less than 2 years.

```
data8 <- read.table(file = "Math745-HW1 Insurance-data.txt", header = TRUE)
head(data8)
```

```
##   D_year premium
## 1   2.16  537.02
## 2   2.72  539.33
## 3   1.03  589.70
## 4   2.55  544.80
## 5   2.89  541.76
## 6   2.94  543.19
```

```
data_2 <- data8[data8$D_year > 2, ]

data_1 <- data8[data8$D_year <= 2, ]

head(data_1)
```

```
##    D_year premium
## 3    1.03  589.70
## 7    1.75  556.55
## 8    0.29  597.61
## 9    0.59  586.43
## 11   1.17  574.77
## 13   0.13  595.41
```

```
head(data_2)
```

```
##    D_year premium
## 1    2.16  537.02
## 2    2.72  539.33
## 4    2.55  544.80
## 5    2.89  541.76
## 6    2.94  543.19
## 10   2.01  556.77
```

```
dim(data_1)
```

```
## [1] 210   2
```

```
dim(data_2)
```

```
## [1] 1041    2
```

8.2 Use the data to fit a linear regression model $P_{remium} = a + b * (Drivingyear) + \epsilon$ for each sub data set, what assumptions do you need for the data analysis? Are the assumptions plausible?

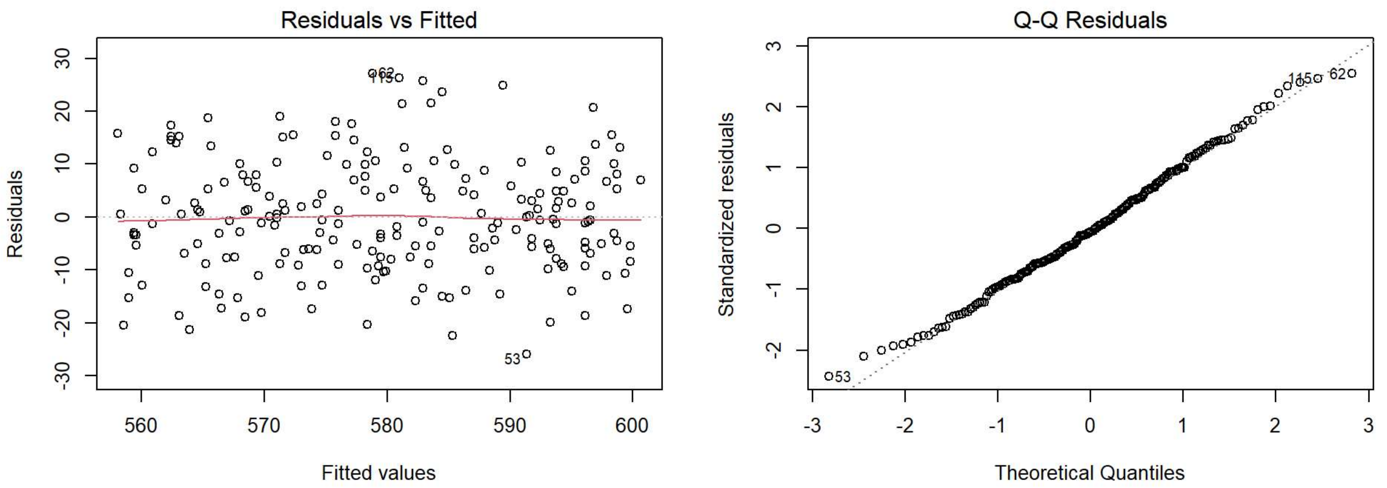
```
model1 <- lm(premium ~ D_year, data=data_1)
model2 <- lm(premium ~ D_year, data=data_2)



summary(model1)
```

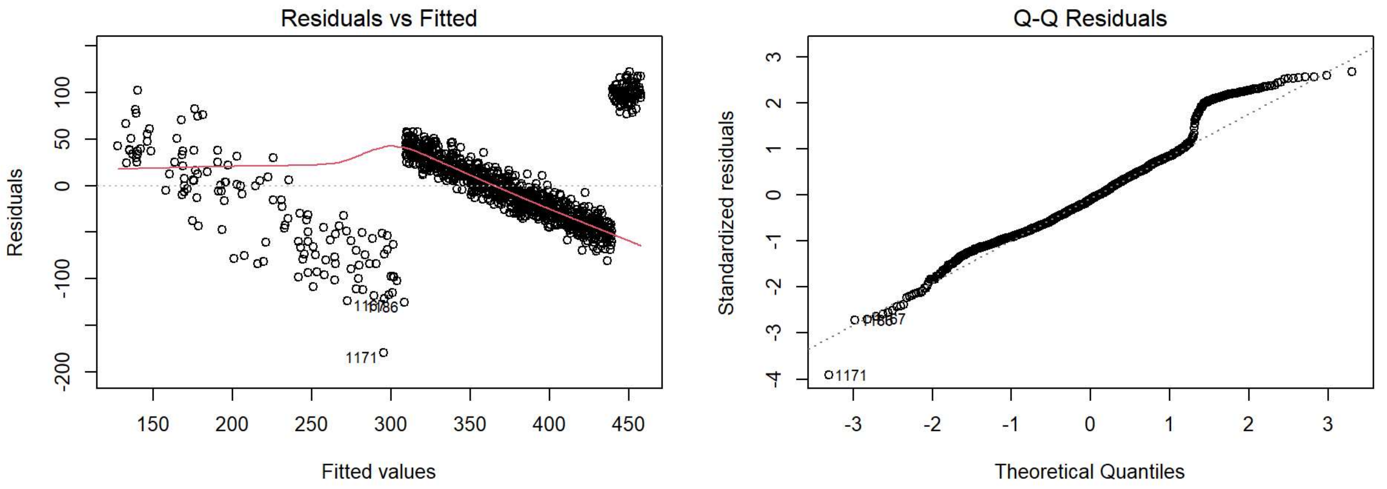
```
##
## Call:
## lm(formula = premium ~ D_year, data = data_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9507  -7.3756  -0.5714   7.1390  27.1324
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  601.319      1.496  401.91   <2e-16 ***
## D_year       -21.626      1.321  -16.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.71 on 208 degrees of freedom
## Multiple R-squared:  0.5629, Adjusted R-squared:  0.5608
## F-statistic: 267.9 on 1 and 208 DF,  p-value: < 2.2e-16
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = premium ~ D_year, data = data_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -179.365  -31.879   -4.603   24.754  122.014
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 495.0615     3.0752  160.99   <2e-16 ***
## D_year      -18.6064     0.3822  -48.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.85 on 1039 degrees of freedom
## Multiple R-squared:  0.6952, Adjusted R-squared:  0.695
## F-statistic:  2370 on 1 and 1039 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))

plot(model1, which=1)
plot(model1, which=2)


plot(model2, which=1)
plot(model2, which=2)
```





```
par(mfrow=c(1,1))
```

8.3 At 0.05 significance level, test the hypothesis that consumers with more than two years driving experience pay less premium than those who spent less than two years behind the wheel.

Solution

H_{0}: There is no difference in the mean premiums between consumers with more than two years of driving experience and those with two years or less.

H_{a}: Consumers with more than two years of driving experience pay less premium on average than those with two years or less.

```
t.test(data_2$premium, data_1$premium, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  data_2$premium and data_1$premium
## t = -77.649, df = 1248.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -213.1262
## sample estimates:
## mean of x mean of y
##  362.2853  580.0273
```

The p-value is 2.2e-16, which is much smaller that the significance level of 0.05. Therefore, there’s strong evidence to suggest that consumers with more than two years of driving experience pay significantly less premium than those with two years or less of driving experience. The mean premium for those with more than two years of driving experience is 362.28, while the mean for those with two years or less of driving experience is 580.03

8.4 Plot the data and find a data-driven model behind the data.

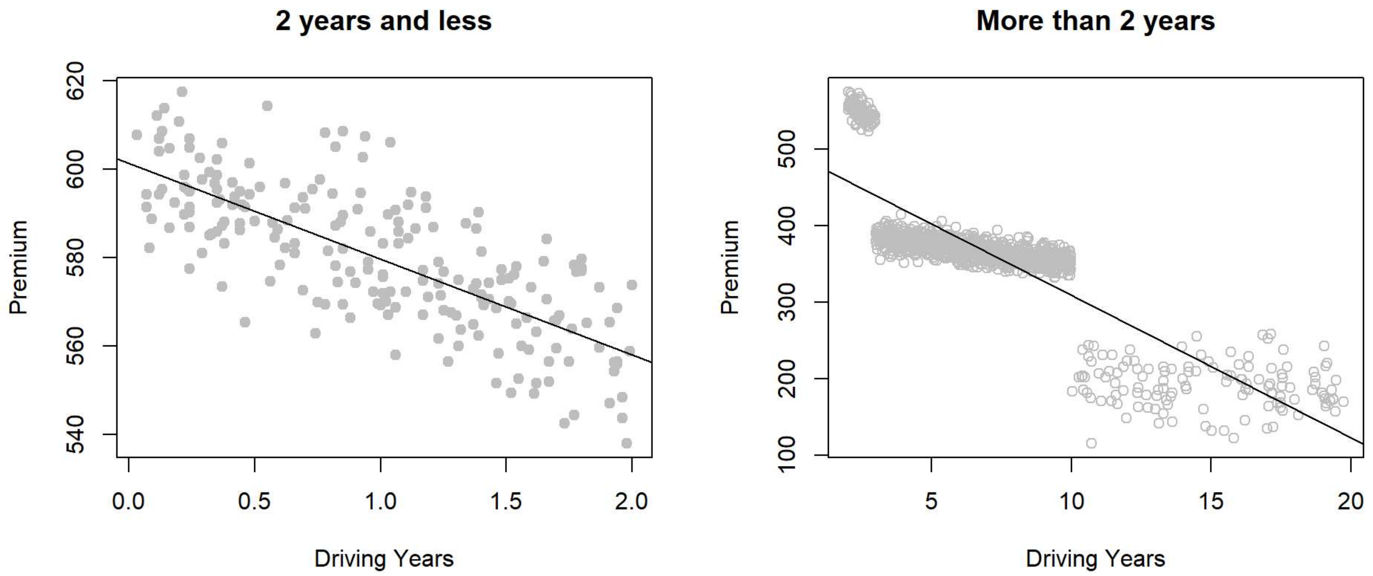
```
dim(data_1)
```

```
## [1] 210   2
```

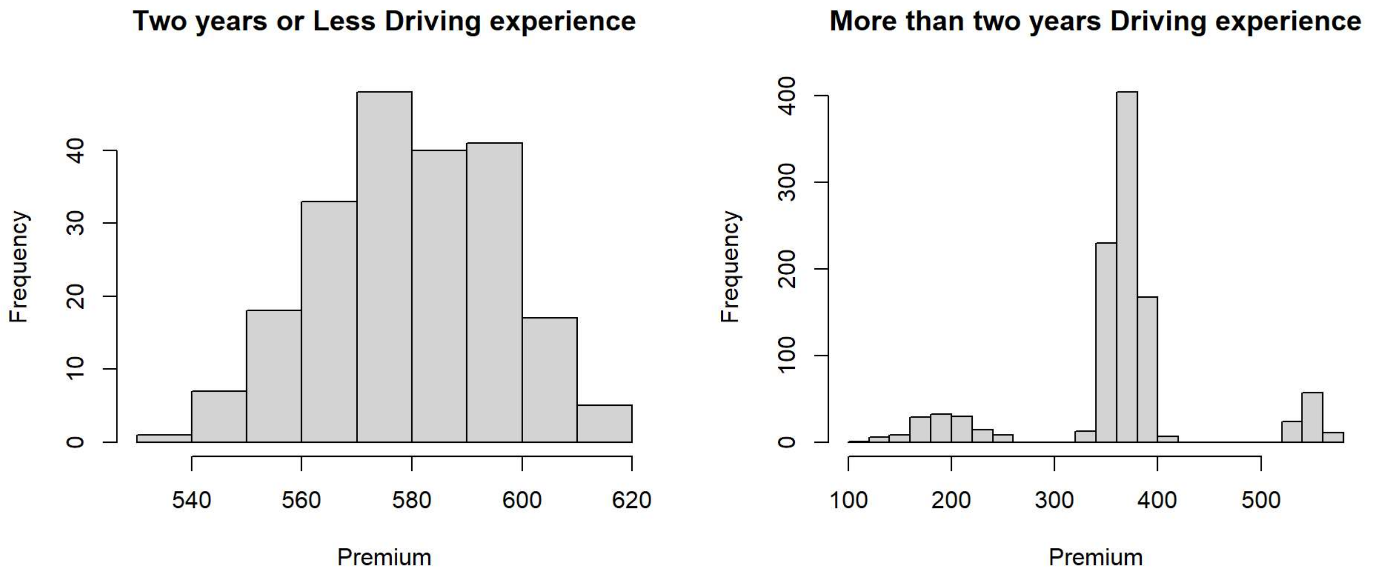
```
dim(data_2)
```

```
## [1] 1041    2
```

```
par(mfrow=c(1,2))

plot(data_1$D_year, data_1$premium, main="2 years and less",
     xlab="Driving Years", ylab="Premium", col="grey", pch=19)
abline(lm(premium ~ D_year, data=data_1), col="black")

plot(data_2$D_year, data_2$premium, main="More than 2 years",
     xlab="Driving Years", ylab="Premium", col="grey", pch=1)
abline(lm(premium ~ D_year, data=data_2), col="black")
```

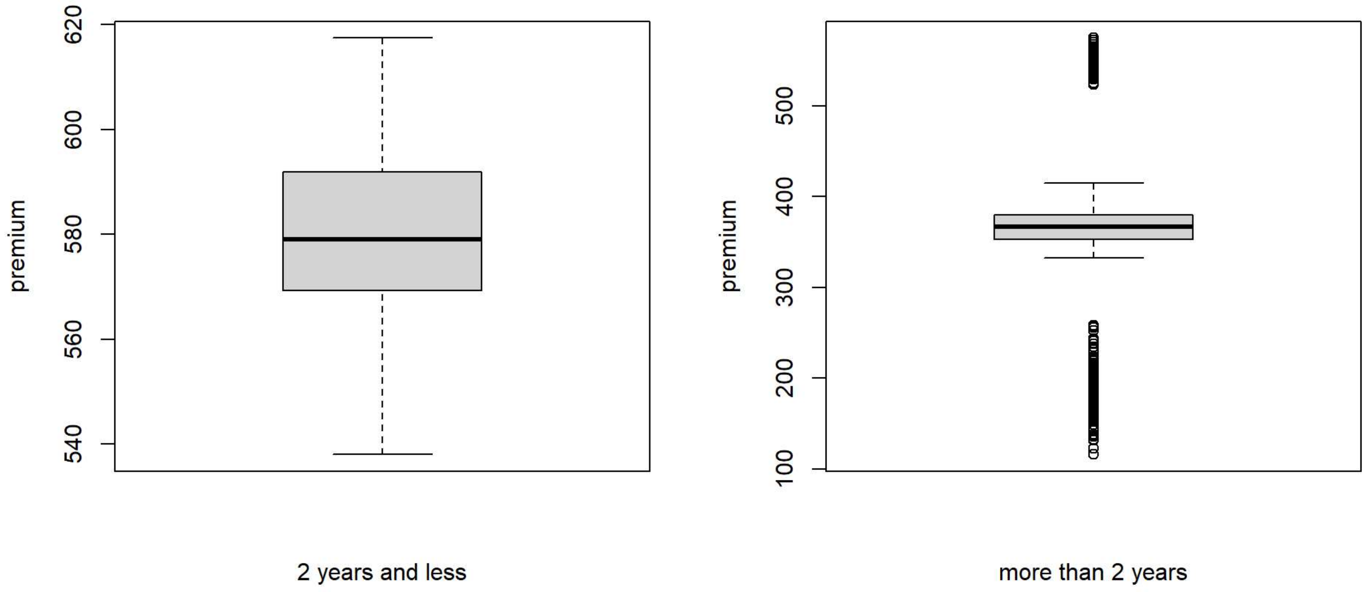


```
par(mfrow=c(1,2))

hist(data_1$premium, main="Two years or Less Driving experience", xlab="Premium", breaks=10)
hist(data_2$premium, main="More than two years Driving experience", xlab="Premium", breaks=20)
```



```
par(mfrow=c(1,1))
```

```
par(mfrow=c(1,2))
boxplot(data_1$premium, ylab = "premium", xlab= "2 years and less")
boxplot(data_2$premium, ylab = "premium", xlab= "more than 2 years")
```



```
par(mfrow=c(1,1))
```

Data_1 represents consumers with two years or less of driving experience and contains 210 observations. In contrast, Data_2 includes those with more than two years of driving experience and has 1041 observations. The scatter plot shows a strong negative linear association between driving years and premium for Data_1. However, for Data_2, the effect of driving years on the premium decreases stepwise. This indicates that the overall association is not strictly linear

From the histograms, we can observe that the premiums in both data_1 and data_2 appear to be normally distributed. The boxplots suggest the presence of some potential outliers in data_2.

Based on the data, linear model is a good fit for the data_1, but not ideal for data_2. By the visionization plots of data_2, Generalized Additive Models (GAMs) may provide a better fit to capture the complex trend.

```
library(mgcv)
```

```
## Loading required package: nlme
```

```
## This is mgcv 1.8-42. For overview type 'help("mgcv-package")'.
```

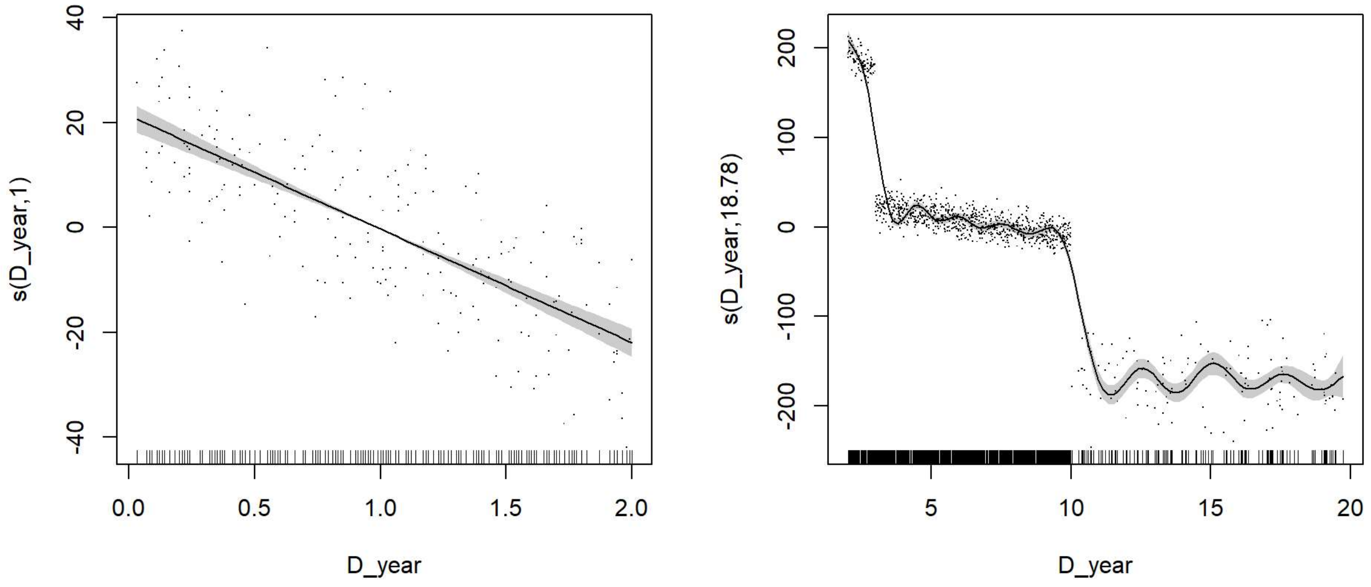
```
model3 <- gam(premium ~ s(D_year), data=data_1)
model4 <- gam(premium ~ s(D_year, k = 20), data=data_2)

summary(model3)
```

```
## 
## Family: gaussian 
## Link function: identity 
## 
## Formula:
## premium ~ s(D_year)
## 
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 580.0273     0.7391   784.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value    
## s(D_year)   1      1 267.9  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =  0.561   Deviance explained = 56.3%
## GCV = 115.82  Scale est. = 114.71    n = 210
```

```
summary(model4)
```

```
## 
## Family: gaussian 
## Link function: identity 
## 
## Formula:
## premium ~ s(D_year, k = 20)
## 
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  362.285      0.653   554.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value    
## s(D_year) 18.78  18.99 795.8  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## R-sq.(adj) =  0.936   Deviance explained = 93.7%
## GCV = 452.48  Scale est. = 443.88    n = 1041
```

```
par(mfrow=c(1,2))
plot(model3, residuals=TRUE, shade=TRUE)
plot(model4, residuals=TRUE, shade=TRUE)
```



```
par(mfrow=c(1,1))
```

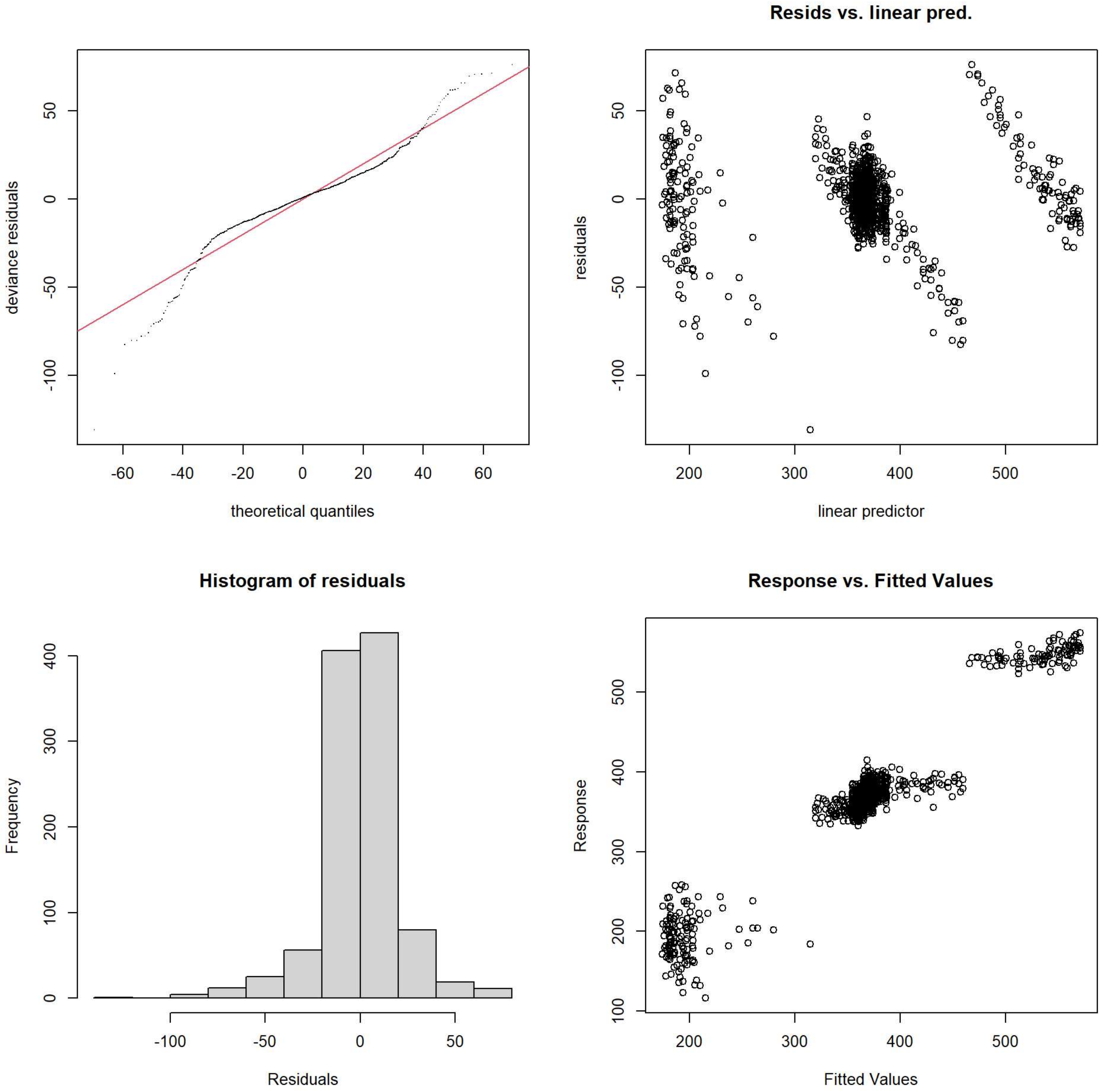
From the plots, we can observe that the GAM regression accurately captures the trend of how driving years affect premiums for both data_1 and data_2.

8.5 How do you interpret the inference method in 8.2 and the data-driven model in 8.4?

```
act1 <- data_1$premium
act2 <- data_2$premium

preds3 <- predict(model3, data_1)
abe3 <- abs(preds3 - act1)
MAE3 <- mean(abe3)
MAE3
```

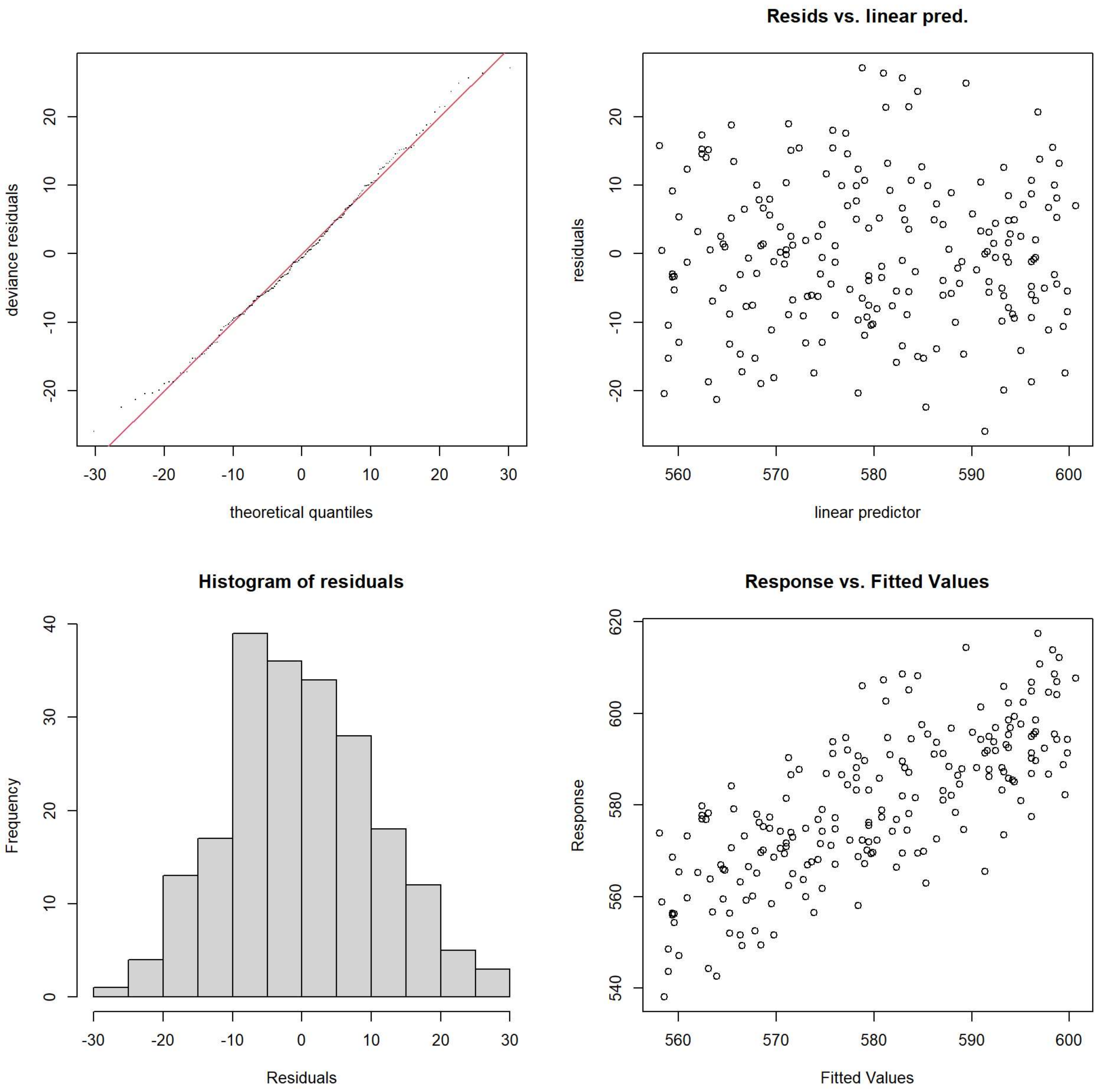
```
## [1] 8.598846
```

```
preds4 <- predict(model4, data_2)
abe4 <- abs(preds4 - act2)
MAE4 <- mean(abe4)

par(mfrow=c(2,2))
gam.check(model4)
```


deviance residuals
theoretical quantiles
-60
-40
-20
0
20
40
60
50
0
-50
-100
Resids vs. linear pred.
residuals
linear predictor
200
300
400
500
Histogram of residuals
Frequency
Residuals
-100
-50
0
50
0
100
200
300
400
Response vs. Fitted Values
Response
Fitted Values
100
200
300
400
500

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 9 iterations.
## The RMS GCV score gradient at convergence was 0.002387389 .
## The Hessian was positive definite.
## Model rank =  20 / 20
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##             k'  edf k-index p-value
## s(D_year) 19.0 18.8    0.53  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
gam.check(model3)
```


Resids vs. linear pred.

Histogram of residuals

Response vs. Fitted Values

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 11 iterations.
## The RMS GCV score gradient at convergence was 8.802332e-06 .
## The Hessian was positive definite.
## Model rank =  10 / 10
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##           k' edf k-index p-value
## s(D_year)  9   1    0.99    0.40
```

```
par(mfrow=c(1,1))

preds1 <- predict(model1, data_1)
preds2 <- predict(model2, data_2)





abe1 <- abs(preds1 - act1)
abe2 <- abs(preds2 - act2)
MAE1 <- mean(abe1)
MAE2 <- mean(abe2)


MAE1
```

```
## [1] 8.598846
```

```
MAE2
```

```
## [1] 35.59207
```

```
MAE3
```

```
## [1] 8.598846
```

```
MAE4
```

```
## [1] 14.28148
```

```
mae_t <- data.frame(
  Data = c("Data_1", "Data_2"),
  lm = c(8.598846, 35.59207),
  gam = c(8.598846, 14.28148)
)

print(mae_t)
```

```
##     Data        lm       gam
## 1 Data_1  8.598846  8.598846
## 2 Data_2 35.592070 14.281480
```

When we fit

8.6 Explain the unbiasedness of parameter estimation in models 8.2) and 8.4)

Unbiasedness refers to the accuracy of an estimation.

When the predictor variable and response variable are strong linearly associated, the Ordinary Least Squares method in lm provides unbiased parameter estimates. This means that, on average, the estimates match the true population values. Fitting data_1 with lm and glm yield pretty close MSE. This means when we switched to a more complex model, our predictions haven’t been improved. In this case, we prefer parametric model by lm for data_1.

GAMs can capture non-linear trends in data. The model provides unbiased estimates of underlying patterns when the smoothing parameters are chosen correctly. Unlike lm, GAM focuses on estimating smooth functions rather than fixed parameters. Since the premium in data_2 doesn’t exhibit a strictly linear relationship with its predictor, gam more effectively captures its non-linear trend, leading to a much smaller MAE.