Diagnostic Plots and Transformations of Data

Rafael Romero

2025-3-13

Contents

```
## Loading required package: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

Question 1

1. A.) Set up the design matrix X and B vector for each of the following regression models (assume i = 1, 2, 3, 4)

Model #1: Design Matrix

$$\begin{bmatrix} 1 & X11 & X11X12 \\ 1 & X21 & X21X22 \\ 1 & X31 & X31X32 \\ 1 & X41 & X41X42 \end{bmatrix}$$

Coefficient Vector

$$\begin{bmatrix} B0 \\ B1 \\ B2 \end{bmatrix}$$

Model #2: Design Matrix

$$\begin{bmatrix} 1 & X11 & X12 \\ 1 & X21 & X22 \\ 1 & X31 & X32 \\ 1 & X41 & X42 \end{bmatrix}$$

Coefficient Vector

$$\begin{bmatrix} B0 \\ B1 \\ B2 \end{bmatrix}$$

Model #3: Design Matrix

$$\begin{bmatrix} X11 & X12 & X^2, 11 \\ X21 & X22 & X^2, 21 \\ X31 & X32 & X^2, 31 \\ X41 & X42 & X^2, 41 \end{bmatrix}$$

Coefficient Vector

$$\begin{bmatrix} B1 \\ B2 \\ B3 \end{bmatrix}$$

Model #4: Design Matrix

$$\begin{bmatrix} 1 & X11 & log10X12 \\ 1 & X21 & log10X22 \\ 1 & X31 & log10X32 \\ 1 & X41 & log10X42 \end{bmatrix}$$

Coefficient Vector

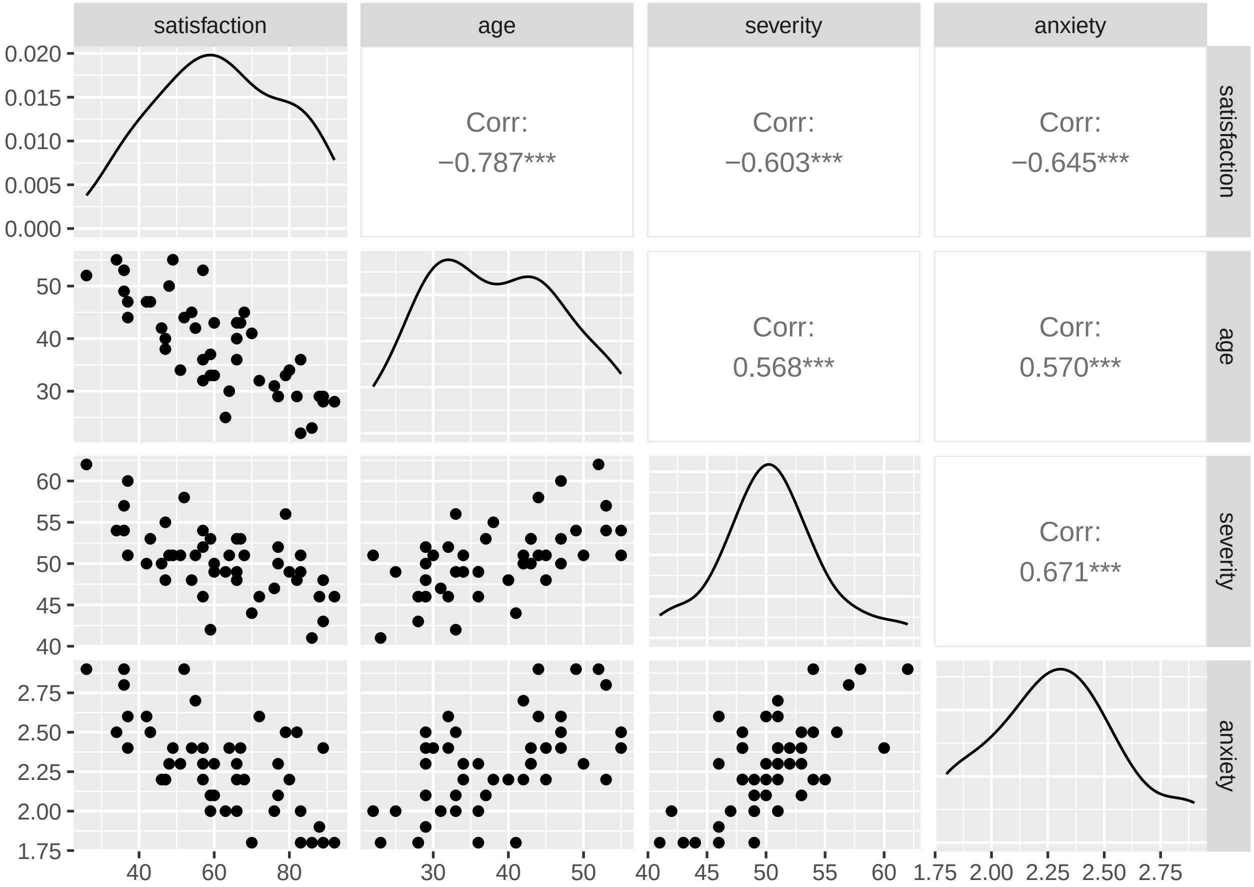
$$\begin{bmatrix} B0 \\ B1 \\ B2 \end{bmatrix}$$

Question 2

```
patients <- read.csv("data/patient_satisfaction.csv")
```

2. A.) Prepare pairs plot for all variables. Do these plots reveal any noteworthy features?

```
ggpairs(patients)
```



Observations The plots show strong correlation among the variables due to the ***. Also notice that severity and anxiety are more normally distributed. Satisfaction is right skewed which means there were more high satisfaction responses. Age is bimodal aswell.

2. B.) Fit a regression model with all three predictor variables to the data and state the estimated regression function. Interpret each estimated parameter.

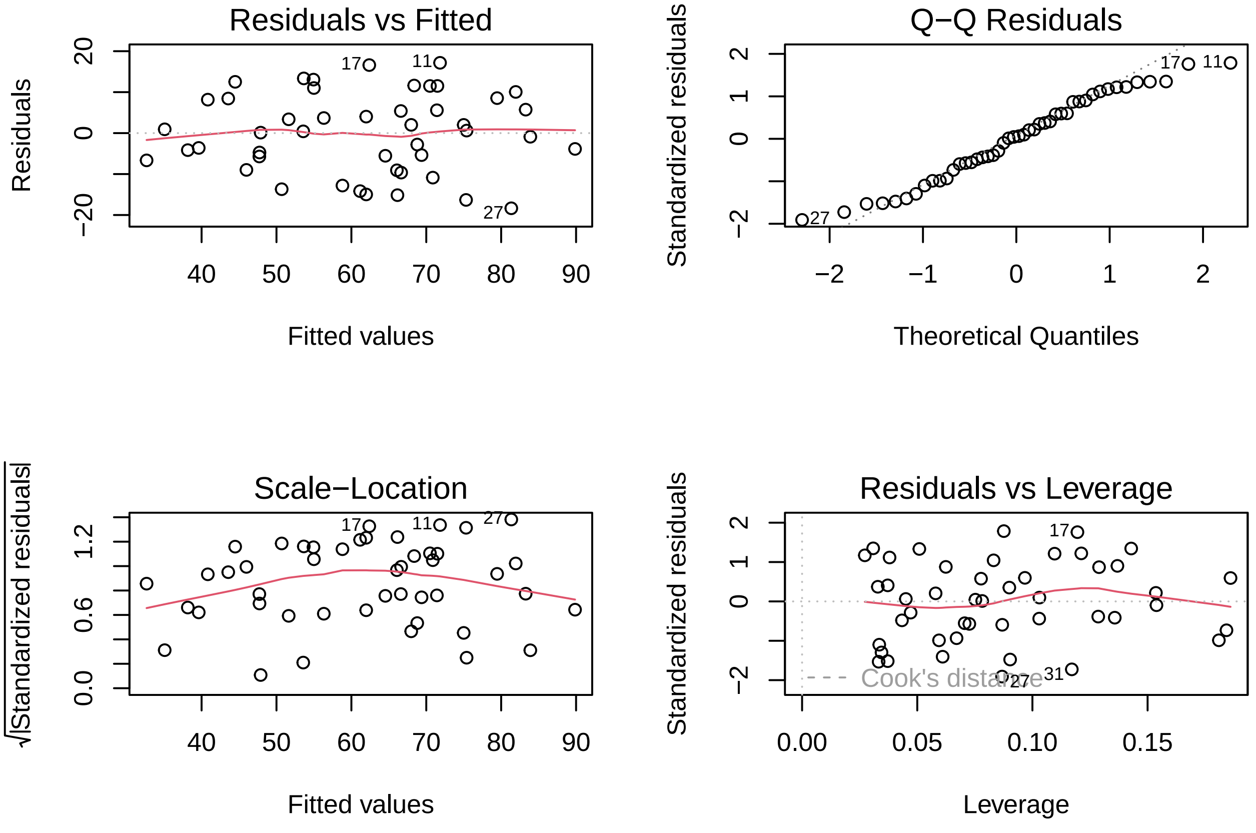
```
model <- lm(satisfaction ~ age + severity +anxiety, data = patients)
summary(model)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + severity + anxiety, data = patients)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.35  -6.42   0.52   8.37  17.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  158.491     18.126    8.74  5.3e-11 ***
## age           -1.142      0.215   -5.31  3.8e-06 ***
## severity      -0.442      0.492   -0.90    0.374
## anxiety      -13.470      7.100   -1.90    0.065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.1 on 42 degrees of freedom
## Multiple R-squared:  0.682,  Adjusted R-squared:  0.659
## F-statistic: 30.1 on 3 and 42 DF,  p-value: 1.54e-10
```

Interpretation: Each parameter is representing the one unit increase in either age, severity, or anxiety while keeping the others constant.

2. C.) Construct diagnostic plots and comment on your findings

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



Comments: Based on the residuals vs fitted plot the residuals are randomly distributed which indicate homoscedasticity. The Q-Q residuals plot shows that the residuals follow the line closely which indicates that they are normally distributed.

2. D.) Report R^2 and comment on your findings.

```
summary(model)$adj.r.squared
```

```
## [1] 0.6595
```

Comment: Since the R-squared is relatively close to one then the model does an adequate job at explaining the proportion of variance by the model.

2. E.) Assuming the fitted model is appropriate, state estimates of all B parameters with 95% confidence intervals.

```
confint(model, level = 0.95)
```

```
##                2.5 %   97.5 %
## (Intercept) 121.912 195.0708
## age          -1.575  -0.7081
## severity     -1.435   0.5508
## anxiety     -27.798   0.8575
```

2. F.) Obtain an estimate of the mean satisfaction with 90% confidence interval when age = 35,severity of illness = 45, and anxiety level = 2.2.

.

```
predict(model, newdata = data.frame(age=35,severity = 45,anxiety = 2.2), interval = "confidence", level
```

```
##     fit   lwr   upr
## 1 69.01 64.53 73.49
```

2. G.) Obtain a 90% prediction interval for a new patient's satisfaction when age = 35, severity of illness = 45, and anxiety level = 2.2.

```
predict(model,  newdata = data.frame(age=35,severity = 45,anxiety = 2.2), interval = "prediction", level
```

```
##     fit   lwr   upr
## 1 69.01 51.51 86.51
```

2. H.) Indicate which subset of predictor variables you would recommend as best for predicting patient satisfaction according to the stepwise procedure using AIC.

```
step_model <- stepAIC(model, direction = "both")
```

```
## Start:  AIC=216.2
## satisfaction ~ age + severity + anxiety
##
##            Df Sum of Sq  RSS AIC
## - severity  1        82 4330 215
## <none>                  4249 216
## - anxiety   1       364 4613 218
## - age       1      2858 7106 238
##
## Step:  AIC=215.1
## satisfaction ~ age + anxiety
##
```

```
##             Df Sum of Sq  RSS AIC
## <none>                   4330 215
## + severity  1        82 4249 216
## - anxiety   1       763 5094 220
## - age       1      3484 7814 240
```

```
summary(step_model)
```

```
##
## Call:
## lm(formula = satisfaction ~ age + anxiety, data = patients)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.445  -7.328   0.673   8.513  18.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  145.941     11.525   12.66  4.2e-16 ***
## age           -1.200      0.204   -5.88  5.4e-07 ***
## anxiety      -16.742      6.081   -2.75   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10 on 43 degrees of freedom
## Multiple R-squared:  0.676,  Adjusted R-squared:  0.661
## F-statistic: 44.9 on 2 and 43 DF,  p-value: 2.98e-11
```

Conclusion: The best subset of variables are Age and Anxiety.

2. I.) Does the stepwise procedure in (h) identify the same best subset by exhaustive search? Does this always happen?

```
b_subset <- regsubsets(satisfaction ~ age + severity + anxiety, data = patients, method = "exhaustive")
summary(b_subset)
```

```
## Subset selection object
## Call: regsubsets.formula(satisfaction ~ age + severity + anxiety, data = patients,
##     method = "exhaustive")
## 3 Variables  (and intercept)
##          Forced in Forced out
## age          FALSE      FALSE
## severity     FALSE      FALSE
## anxiety      FALSE      FALSE
## 1 subsets of each size up to 3
## Selection Algorithm: exhaustive
##          age severity anxiety
## 1  ( 1 ) "*" " "      " "
## 2  ( 1 ) "*" " "      "*"
## 3  ( 1 ) "*" "*"      "*"
```

Conclusion: Yes the stepwise procedure matches the same best subset by exhaustive search which are Age and Anxiety, but the exhaustive search still considers other models with Severity. Therefore, stepwise may not always be the same as exhuastive search.

2. J.) Would the stepwise procedure have any advantages as a screening procedure over the exhaustive search?

Answer: Yes, stepwise procedure would have advantages such as it being faster since it does not need to evaluate all the subsets of predictors. This means it may remove predictors which can reduce the chance of over fitting. This can help us eliminate variables from the start which can be greet as a screening procedure when dealing with large data sets.

Question 3

3. A.) Provide comment on the preliminary plots.

Comments We notice positive relationships between Rental prices with expenses and footage. Also a positive relationship between expenses and footage. There is also a negative relationship between rental and age. Vacancy is randomly and widely scattered which can indicate that it may not be a good predictor of rental prices.

3. B.) For the saved R object fit, write down the model that is being fitted, including assumptions.

Answer: Model:

rental = B0 + B1(age)+B2(expenses)+B3(vacancy)+B4(footage)+e

Assumptions: Linearity, Independence, homoscedasticity, residuals are normally distributed

3. C.) Explain the purpose of the statement, Provide details about the criterion and the option direction. Discuss the outcom eof each step and final model.

The statement is using stepwise forward selection with starts with an empty model and adds the best significant predictor based on AIC at each step.

Process: Step 1: It adds footage which lowered the AIC from 88.81 to 63.47 Step 2: It adds age which lowered the AIC to 42.11 Step 3: It adds expenses which drops the AIC to 23.97 Step 4: Checks vacancy, but it does not significant improve the model

final model: rental = 12.37 + (8.178 + 10ˆ-6) x (footage) - 0.1442 x (age) + 0.2672 x (expenses) **vacancy was excluded since it did not benefit the model**

3. D.) Suppose the final model fit1 is adequate. State your conclusions.

Conclusion: The significant predictors are footage, age, and expenses. Footage have higher rental prices. Older properties tend to have lower rental prices. Higher expenses correlate with higher rent.

The adjusted Rˆ2 is 0.5667 which means that the model does a good job at explaining the variance in rental prices.

Therefore, expenses, age, and footage are the most significant factors in determining rental prices.

3. E.) Comment on Residual Plots

Residuals vs Fitted The residuals seem to be randomly scatterd which indicates that the linear model is a good fit.

QQ-plot of residuals The residuals follow the line relatively close which indicates that the residuals most likely have a normal distribution.

Conclusion: I am relatively satisfied with the model, but some outliers can be further investigated as well as the residuals that deviate from the 45 degree line which can indicate skewdness to our distribution. Overall the model is still a relatively good fit.