Minilab 6a Worksheet

Linear Regression: Model Diagnostics

We have seen how to fit a *linear model* to a dataset using the R function lm() and we have seen that the intuitive idea behind linear regression is to find the "line of best fit", i.e., the line that **minimises the sum of squared residuals** (sometimes called "least squares" regression).

In this minilab, we consider whether a fitted linear regression model is a *valid model* for the data, i.e., does the model explain the data well. Also, when there are several valid models for a given dataset, we need a way of comparing them to determine which model that has the most explanatory power.

1. Does a fitted model explain the dataset well?

Key question:

How can we check that the fitted model explains the dataset well, i.e., that it is a valid model?

(1) First we have another look at the marketing data that shows the impact of three different advertising media (youtube, facebook, and newspaper) on sales.
The values in the sales column of the marketing tibble are called the observed values of *sales*, i.e., the values in the data.

```
library(tidyverse)
library(datarium)
marketing = as_tibble(marketing)
observed_sales = marketing$sales
observed_sales
```

(2) Suppose we fit the linear model "sales~youtube".

```
model = lm(sales~youtube, data=marketing)
summary(model)
```

The output from summary() shows that the fitted model is:

$$sales = (8.439112) + (0.047537) \times youtube$$

(3) What value of *sales* does our fitted model **predict** for each observed value of *youtube*? These predicted values are often known as "fitted" values.

```
a = model$coefficients[1]
b = model$coefficients[2]
predicted_sales = a + b*marketing$youtube
predicted_sales
```

(4) The *residual sales* are the "leftovers" (positive or negative) when the *predicted sales* are subtracted from the *observed sales*. (You could think of predicted as the budget and observed as the actual, so the residual is the surplus or deficit.)

```
residual_sales = observed_sales - predicted_sales
residual_sales
```

(5) We are most interested in whether there is any pattern in the residuals. To do this, one important plot to consider is the *residuals vs predicted values*.

```
ggplot(NULL, aes(x=predicted_sales, y=residual_sales)) +
  geom_point()
```

This plot shows the pattern remaining in the data when the pattern explained by the model has been removed.

(6) Another way to quickly get the same information is to use the function augment() (from the R tidyverse package broom) to calculate the predicted (fitted) values and the residuals and then plot a scatterplot of the special variables with names beginning with dot, i.e., .fitted and .resid.


```
model = lm(sales~youtube, data=marketing)
library(broom)
augment(model) %>%
  ggplot(aes(x=.fitted,y=.resid)) +
    geom_point()
```

Therefore, we are hoping to see **no further pattern** in the plot. Check that the mean of the residuals is approximately zero. Also check that there is no kind of funnel shape or noticeable curve in this scatterplot.

2. Checking the Assumptions of a Linear Model

Checking the *residuals* vs *fitted* plot is the first step in scrutinising the validity of a fitted linear model. In order to see what else needs to be examined, we need to look into the theory behind the linear model.

In general, we think of the *line* $y = a + bx$ more formally as a **linear model**

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

where the ϵ_i are independent random errors (residuals) which follow the same normal distribution with mean zero and common variance σ^2.

Here α and β are the population parameters (i.e. the intercept and slope of the linear relationship in the population). We *estimate* the unknown α and β (from the population) by a and b calculated from the sample using the formulas:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

As long as the points on the scatterplot do not form an exact vertical line, it is always possible to calculate these best-fit values for a and b. *The question is whether the fitted model is of any use to explain the relationship between the variables.*

In fitting a linear model, we make several assumptions about the relationship between the variables in the population.

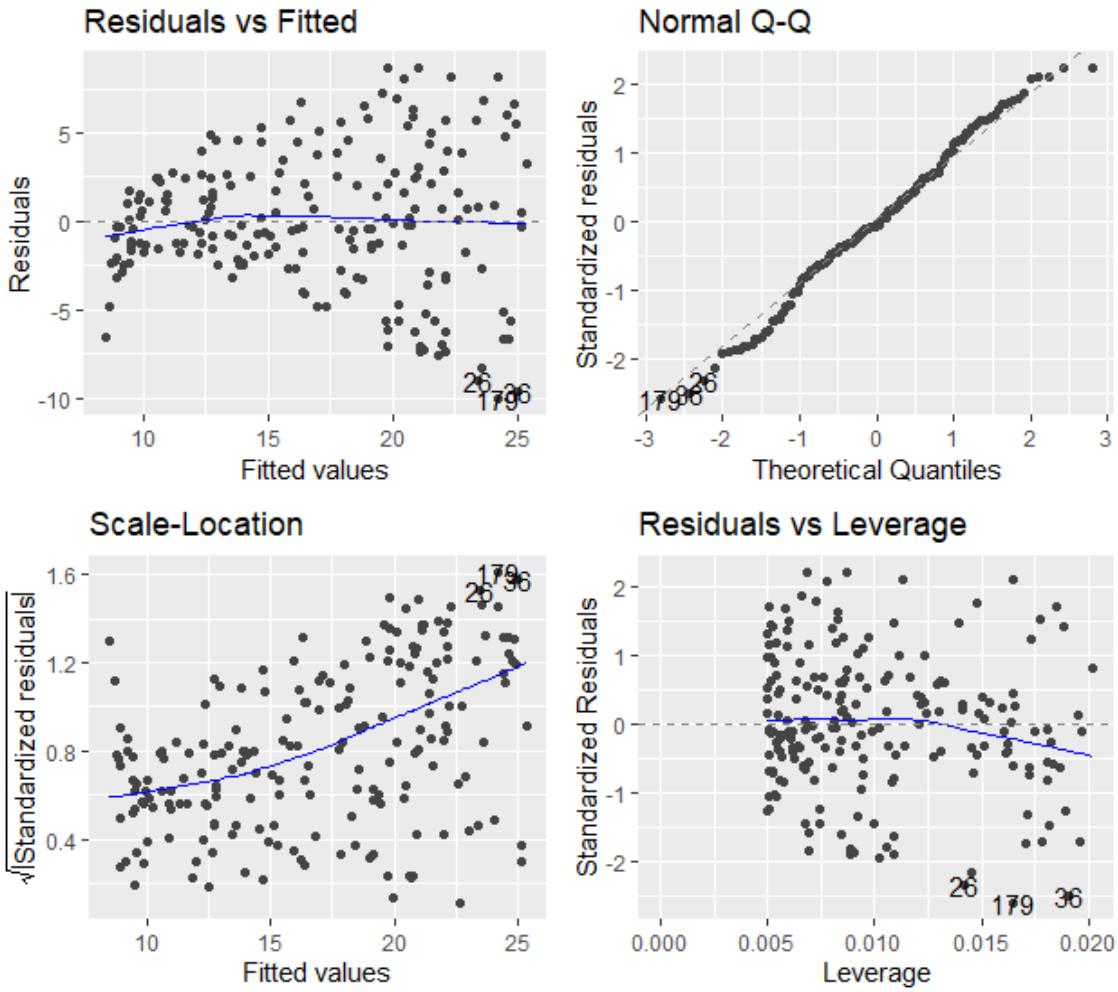
- **Linearity.** The relationship between the predictor X and the response Y is assumed to be linear.
- **Normality of residuals.** The residuals are assumed to be normally distributed.
- **Homogeneity of residual variances.** The residuals are assumed to have a constant (shared) variance (called "homoscedasticity" in statistical language).
- **Independence of residuals.** The residual at one value of the predictor value is statistically independent of the residuals at all other predictor values.

All these assumptions and potential problems can be checked by producing four diagnostic plots visualising the residuals

To construct these diagnostic plots, we use the autoplot() function from the R package "ggfortify". The diagnostic plots show residuals in four different ways.

```
install.packages("ggfortify")  # run only once
library(ggfortify)
model = lm(sales~youtube,data=marketing)
autoplot(model)
```

Hopefully, you will recognise the first one.



(1) The "Residuals vs Fitted" plot (top left) is used to check the linear relationship assumption (this is the same plot we made earlier). Ideally, the plot will show no fitted pattern, i.e., the blue line should be approximately horizontal at zero. The presence of a pattern may indicate a problem with some aspect of the linear model.

In this example, there is no pattern in the residual plot, which suggests that we can assume a linear relationship between the predictor and the response variables. GOOD

(2) The "Normal Q-Q" plot (top right) is used to visually check whether the residuals are approximately normally distributed. The points should approximately follow a straight line.

In this example, all the points fall approximately along the dashed line so we can assume normality of residuals. GOOD

(3) The "Scale-Location" plot (bottom left) is used to check the homogeneity of variance of the residuals (homoscedasticity). This plot checks whether residuals are spread roughly equally along the ranges of predictors. It is good if you see a horizontal line with equally spread points.

In this example, the variability of the residual points increases with the value of the fitted response variable, suggesting nonconstant variances in the residuals (or heteroscedasticity). BAD

(4) The "Residuals vs Leverage" plot (bottom right) is used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. These would indicate lack of independence of residuals.

- An outlier is a point that has an extreme response value. The presence of outliers may affect the interpretation of the model. Outliers can be identified by examining the *standardised residual*. Observations whose standardised residuals are greater than 3 or less than -3 are possible outliers.
- A data point has high leverage if it has extreme predictor value. This can be detected by examining the leverage statistic.

In this example, there do not appear to be any outliers or points of high leverage. GOOD

Note that the four plots show the three most extreme data points labelled with the row numbers of the data in the data set (rows 26, 36 and 179). They might be problematic. You might want to take a closer look at them individually to check if there is anything special for the subject or if it could be simply data entry errors.

These four diagnostic plots ONLY use the *residuals* and *fitted values* to assess how well a linear model explains/fits a dataset. Therefore, we can also use these diagnostic plots for a multiple regression model, i.e., where there are two or more predictor varaibles.

Exercise. Consider the linear model "sales~youtube+facebook".

```
model = lm(sales~youtube+facebook, data=marketing)
autoplot(model)
```

By looking at the four diagnostic plots, do you have any concerns over the regression assumptions or concerns regarding any particular points in this dataset?

3. Comparing Models

We have seen that we may have several different models that predict the same response variable (y).

If several models approximately satisfy the assumptions of linear regression, which one should we choose?

Image from: https://twitter.com/dsfraley/status/982725556356747266

A good starting point for comparing models is the

Coefficient of Determination

which is defined as the square of the correlation between the *observed* (original) and *fitted* (predicted by the model) values of the response variable, i.e.,

$$R^2 = \big(cor(y, \hat{y})\big)^2$$

This is known as the "R squared" value (or "Multiple R-squared" value).

It can be interpreted as the *proportion of the variance in the response variable that is predictable from the predictor variables.*

It provides a measure of how well observed responses are replicated by the model, based on the proportion of total variation of responses explained by the model.

Note that **only** in the case of simple linear regression (one quantitative predictor variable) we have $R^2 = r^2$ when r is the Pearson correlation coefficient, i.e., $r = cor(x, y)$.

(1) Consider the linear model "sales~youtube".

```
modelA = lm(sales~youtube,data=marketing)
summary(modelA)
```

This gives $R^2 = 0.6119$, i.e., 61% of the variance in sales is predictable (explained) by the linear relationship with youtube.

Exercise. Show that the Pearson correlation coefficient between sales and youtube is $r = 0.7822244$. What is r^2 in this case?

(2) Consider the linear model "sales~youtube+facebook".

```
modelB = lm(sales~youtube+facebook,data=marketing)
summary(modelB)
```

This gives $R^2 = 0.8972$, i.e., 90% of the variance in sales is predictable (explained) by the linear relationship with youtube and facebook together. We would therefore regard this as a better model that "sales~youtube" since $0.8972 > 0.6119$.

Note that "sales~youtube+facebook" is the *form* of the linear model. It means that we want to find the best-fit coefficients a, b, and c for the linear model

$$sales = a + b \times youtube + 0.18799 \times facebook$$

We never actually add the youtube column and the facebook column of the marketing tibble.

Exercise. Consider the linear model "sales~facebook+newspaper". Would you consider that this model is a "better" model than "sales~youtube"?

Summary

In this minilab, we have visually checked the regression assumptions (linearity, normality of residuals, homogeneity of residual variances, and independence of residuals) using four **diagnostic plots**. These diagnostic plots only use *residuals* and *fitted* values, so this check applies to both simple linear regression and multiple regression. We have also seen that the R^2 value (coefficient of determination) is useful to make a comparison of the explanatory power of different models.