Crash Course in Linear Regression

Haziq Jamil

2021-09-07

Contents

Introduction 1

Example 2

The linear regression model 4

Assumptions 5

Estimation 5

Residuals 6

Properties of estimator 6

Inference 7

Example 10

Exercise 1 14

Exercise 2 17

Read more 18

Introduction

Regression analysis is one of the **most frequently used** statistical techniques. It aims to build up an explicit relationship between one *response variable*, often denoted as y, and one or several *explanatory variables*, often denoted as $x_1, \dots, x_p$.

Alternative nomenclature for the variables:

y variable	x variable
Response variable	Explanatory variables
Dependent variable	Independent variable

y variable	x variable
Output variable	Input variable
	Covariates
	Regressors

GOAL:

- To understand how y depends on $x_1, \ldots, x_p$ (inference)
- To predict unobserved y value based on observed $x_1, \ldots, x_p$

Example

Look at this data set. It contains information about various cars and their related variables (a data frame with 32 observations on 11 (numeric) variables).

- [, 1] `mpg` Miles/(US) gallon
- [, 2] `cyl` Number of cylinders
- [, 3] `disp` Displacement (cu.in.)
- [, 4] `hp` Gross horsepower
- [, 5] `drat` Rear axle ratio
- [, 6] `wt` Weight (1000 lbs)
- [, 7] `qsec` 1/4 mile time
- [, 8] `vs` Engine (0 = V-shaped, 1 = straight)
- [, 9] `am` Transmission (0 = automatic, 1 = manual)
- [,10] `gear` Number of forward gears
- [,11] `carb` Number of carburetors

```
mtcars
```

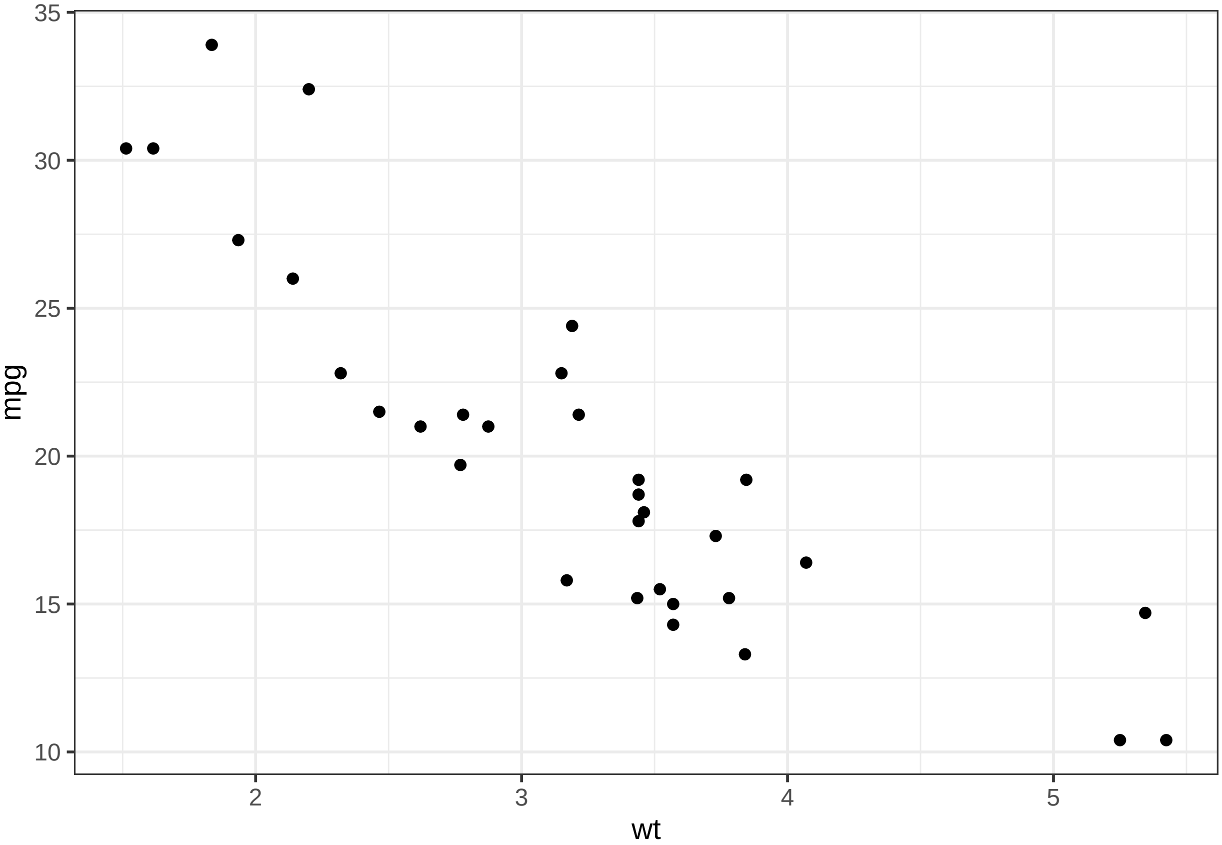
```
##                      mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Let's concentrate on two variables: `mpg` and `wt`. Plot them:

```
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  # geom_abline(slope = -6, intercept = 42, linetype = "dashed", col = "red",
  #             alpha = 0.8) +
  # geom_abline(slope = -7, intercept = 39, linetype = "dashed", col = "blue",
  #             alpha = 0.8) +
  theme_bw()
```



Questions of interest:

1. What is the general trend of `mpg` against `wt`?
2. How can we draw the line of best fit through the data points?
3. How do we explain the points that do not lie exactly on the line?

The linear regression model

For a set of response variables $\mathbf{y} = \{y_1, \dots, y_n\}$ and corresponding explanatory variables $\mathbf{x}_k = \{x_{1k}, \dots, x_{nk}\}$ for $k = 1, \dots, p$, the linear regression model is given by

$$\begin{gathered} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \\ \epsilon_i \sim N(0, \sigma^2) \text{ (iid)} \\ \text{for } i = 1, \dots, n \end{gathered}$$

Using matrix notation, we can write this as

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

or simply

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times (p+1)$ matrix of observations (sometimes called the *design matrix*), $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors.

The errors essentially measure the random deviation or random noise that the observation makes from the 'true' model.

Assumptions

- $\mathrm{E}(\epsilon_i) = 0, \forall i$.
- $\mathrm{Var}(\epsilon_i) = \sigma^2, \forall i$.
- $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$.
- We assume that the errors are normally distributed.
- We assume that the explanatory variables are fixed (non-random).

Note that The model is **linear** in the parameters $\beta_0, \dots, \beta_p$. This means we cannot have a model with β_1^2, or β_k^q or the like. However, it is fine to have the covariates x_k^2, x_k^3 or any transformation of them.

Estimation

In order to proceed with either inference or prediction, we need to *estimate the model.* This means estimating the unknown values of the parameters (collectively called θ) of the model, which are

$$\theta = \{\beta_0, \beta_1, \dots, \beta_p, \sigma^2\}.$$

There are several methods available for estimating the linear regression model. Among them is the *least squares method.* Essentially, the line that fits the best through all the data points should have the smallest total error.

Let $\hat{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\beta}$. Note that we use hats to represent estimates of coefficients/parameters. Consider the sum of the squared errors

$$\begin{aligned}\sum_{i=1}^{n} \epsilon_i^2 &= \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\end{aligned}$$

By definition, the *least squares estimator (LSE)* for $\boldsymbol{\beta}$ minimises the sum of squared errors, i.e.

$$\hat{\boldsymbol{\beta}} = \arg\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

To solve for $\boldsymbol{\beta}$, consider

$$\begin{aligned}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 &= \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\|^2 \\ &= \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 + 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\end{aligned}$$

Now if we choose $\boldsymbol{\beta}$ such that

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

then we have

$$\begin{aligned}\|\mathbf{y}-\mathbf{X}\beta\|^2 &= \|\mathbf{y}-\mathbf{X}\hat{\beta}\|^2 + \|\mathbf{X}(\hat{\beta}-\beta)\|^2 \\ &\geq \|\mathbf{y}-\mathbf{X}\hat{\beta}\|^2\end{aligned}$$

for any $\boldsymbol{\beta}$, which satisfies the least squares condition. Therefore, the LSE for $\boldsymbol{\beta}$ must satisfy

$$\begin{aligned}\mathbf{X}^\top(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) &= 0 \\ \mathbf{X}^\top\mathbf{y}-\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} &= 0 \\ \mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}^\top\mathbf{y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

provided $\mathbf{X}^\top\mathbf{X}$ is not singular, i.e. it is invertible.

With the knowledge of the LSE for the coefficients, and assuming that the errors are zero-meaned variables, an estimate of the variance of the errors is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

but usually we will use another estimator for σ^2, as we will see later.

Residuals

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$, the **predicted value** for the i'th observation using estimates $\hat{\boldsymbol{\beta}}$. Residuals are defined to be the difference between the observed value and the predicted value,

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

If the model is correct, then the residuals should behave like random noise!

Properties of estimator

1. The LSE for $\boldsymbol{\beta}$ turns out to be the **maximum likelihood** estimator for $\boldsymbol{\beta}$ too. This is easy to see: The likelihood under the normal linear model is

   $$L(\boldsymbol{\beta},\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\beta\|^2\right\}$$

   and since under the LSE, $\|\mathbf{y}-\mathbf{X}\beta\|^2 \geq \|\mathbf{y}-\mathbf{X}\hat{\beta}\|^2$, which means that the likelihood is maximised at the LSE as well.

2. The estimator for $\hat{\boldsymbol{\beta}}$ is normally distributed with mean and variance given by

   $$\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$$

   which implies that the estimators for the coefficients are **unbiased**.

3. The unbiased estimator for σ^2 is given by the formula

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n-p-1}$$

   which is distributed according to a χ^2 distribution with $n-p-1$ degrees of freedom.

4. $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are independent of each other.

5. The residuals have mean zero

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\epsilon}}) &= \mathrm{E}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathrm{E}(\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{0} \end{aligned}$$

Inference

Interpretation of coefficients

The coefficients β_j represents the 'strength or influence' of the variable x_j on y. It is the effect on y of changing x_j by a single unit, holding the other covariates fixed.

Consider the effect of the coefficient β_1. Let $y^{(0)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, and also let $y^{(1)} = \beta_0 + \beta_1(x_1 + 1) + \cdots + \beta_p x_p$. Since everything else cancels out, the difference between $y^{(1)}$ and $y^{(0)}$ is simply

$$y^{(1)} - y^{(0)} = \beta_1(x_1 + 1) - \beta_1 x_1 = \beta_1$$

As we can see, β_1 represents the change in the response variable when the variable x_1 is increased by **one unit**, and **keeping all other variables fixed**. A positive value for the coefficient imparts a change in the positive direction in the response, and vice versa. Of course, the logic is the same for all of the other variables $j = 2, 3, \ldots, p$.

Standard errors of coefficients

Recall that the **standard deviation** for each $\hat{\beta}_j$ (from the normal distribution in property 2) is given by

$$\mathrm{SD}(\hat{\beta}_j) = \sqrt{\sigma^2 v_{jj}}$$

where $v_j j$ is the $(j+1, j+1)$th element of the matrix $(\mathbf{X}^\top \mathbf{X})^{-1}$. Notice that we do not know the true value of σ and therefore, the standard deviation of the coefficients as well.

The **standard error** for each $\hat{\beta}_j$ is given by

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_{jj}}$$

where $\hat{\sigma}^2$ is the unbiased estimator for σ^2 given in property 3 above. Essentially, by replacing the unknown value σ^2 by its estimator, we get a "estimate" of the SD which we call the standard error.

Coefficient of determination

Introduce the decomposition

$$\overbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}^{\text{Total SS}}=\overbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}^{\text{Regression SS}}+\overbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}^{\text{Residual SS}}$$

The term on the LHS is the *Total Sum of Squares*, which represents the total variation in the data (responses). The first term on the RHS is called the *Regression Sum of Squares*, which represents the variation in the regression model. The second term on the RHS we have seen before, called the *Residual Sum of Squares*, which measures variability between predicted and observed values.

Define the **coefficient of determination**, as

$$R^2=\frac{\text{Regression SS}}{\text{Total SS}}=1-\frac{\text{Residual SS}}{\text{Total SS}}$$

R^2 takes values between 0 and 1. $100R^2$ is the percentage of the total variation in $\{y_i\}$ explained by all the regressors. Therefore, the closer R^2 is to 1, the better the model agreement is.

Sometimes, the **adjusted** R^2 value is used instead, because it has nice distributional properties (for hypothesis testing)

$$R^2_{\text{adj}}=1-\frac{\text{Residual SS}/(n-p-1)}{\text{Total SS}/(n-1)}$$

Roughly speaking, both should give about similar values.

Tests for single coefficients

For each coefficient β_j, where $j=0,1,\ldots,p$,

$$\frac{\hat{\beta}_j-\beta_j}{\text{SE}(\hat{\beta}_j)}\sim t_{n-p-1}$$

We can then use this fact to test the hypothesis

$$\begin{aligned}\text{H}_0&:\beta_j=b_j\\ \text{H}_1&:\beta_j\neq b_j\end{aligned}$$

Let

$$T=\frac{\hat{\beta}_j-b_j}{\text{SE}(\hat{\beta}_j)}$$

We reject the null hypothesis at the level α against the alternative hypothesis if $|T|>t_{n-p-1}(\alpha/2)$, where $t_k(\alpha)$ is the top α-point of the t_k distribution.

Note we can also test H_0 against the alternatives

- $\text{H}_1:\beta_j>b_j$, and we reject the null hypothesis if $T>t_{n-p-1}(\alpha)$; and
- $\text{H}_1:\beta_j<b_j$, and we reject the null hypothesis if $T<-t_{n-p-1}(\alpha)$.

The $(1-\alpha)$ confidence interval for β_j is

$$\hat{\beta}_j\pm t_{n-p-1}(\alpha/2)\cdot\text{SE}(\hat{\beta}_j)$$

Remarks:

- When n is large, then the t-test becomes approximately equivalent to the Z-test / Wald test (using the normal distribution).
- We are not usually interested in testing the intercept. In most practical data applications, the intercept is not likely to be zero as it represents the 'average value' of the response when all covariates are zero.

Tests for all zero-regression coefficients

Consider the hypothesis

$$\begin{aligned} \mathrm{H}_0 &: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ \mathrm{H}_1 &: \text{Not all } \beta_1, \ldots, \beta_p \text{ are } 0 \end{aligned}$$

Let

$$\begin{aligned} T = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2/p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2/(n-p-1)} &= \frac{n-p-1}{p} \cdot \frac{\text{Regression SS}}{\text{Residual SS}} \\ &= \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \end{aligned}$$

We reject the null hypothesis at the α significance level if $T > F_{p,n-p-1}(\alpha)$.

This has links to the ANOVA table which you may be familiar with:

Source	d.f.	Sum of squares	Mean SS	F-statistic
Regressors	p	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p}$	$\frac{p \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{(n-p-1) \sum_{i=1}^n (y_i - \hat{y}_i)^2}$
Residual	$n-p-1$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1}$	
Total	$n-1$	$\sum_{i=1}^n (y_i - \bar{y})^2$		

Standardized residuals

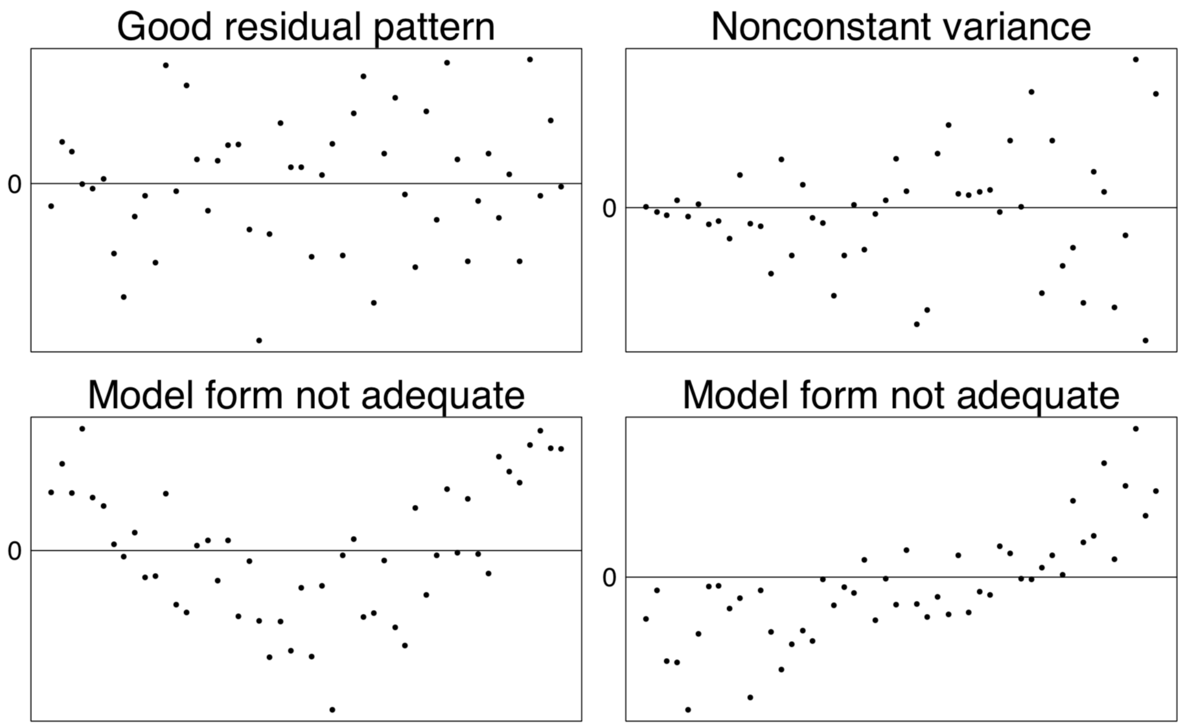
Let $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$, an $n \times n$ matrix. This matrix is called the **projection matrix** or the **hat matrix** for linear regression. It has several interesting and useful properties, which we will not go into detail now.

For $i = 1, \ldots, n$, define the **standardized residuals** as

$$\hat{e}_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

where h_{ii} is the (i,i)th entry of the matrix $\mathbf{H}$. The standardized residuals then have mean zero and variance one.

We use the standardized residuals as a means of diagnosing the fit of the linear model, and to test assumptions of our linear model. For instance, we can plot the standardized residuals - in a qq-plot (to test for normality) - against covariates (to test for homo/heteroscedasticity) - against predicted values (to test for homo/heteroscedasticity)



Example

Let's fit a simple linear regression model to the `mtcars` data set. In `R`, the command to fit linear regression model is `lm()`.

```
mod <- lm(formula = mpg ~ wt, data = mtcars)
summary(mod)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Let's go through the results one by one.

Call

```
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
```

This tells us that the following regression had been fitted:

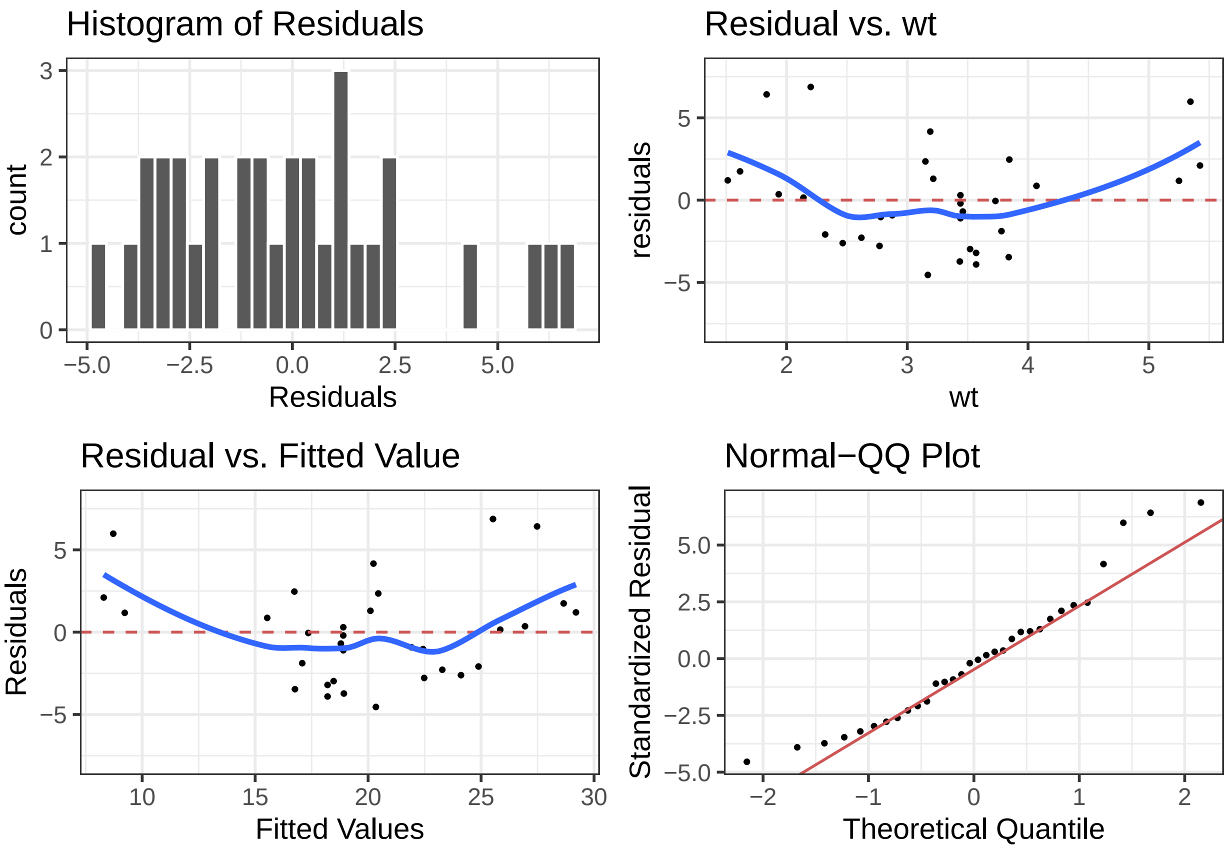
$$\texttt{mpg} = \beta_0 + \beta_1 \cdot \texttt{wt} + \epsilon$$

Residuals

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
```

This gives us a 5-point summary of the distribution of the residuals $\hat{\epsilon}_i$. The intention is to tell us how well the model fits the data, but it will also be easier to diagnose model fit using residual plots.

```
diag.plots <- lindia::gg_diagnose(mod, plot.all = FALSE)
diag.plots <- lapply(diag.plots, function(x) x + theme_bw())
diag.plots[[2]] <- diag.plots[[2]] + geom_smooth(se = FALSE)
diag.plots[[3]] <- diag.plots[[3]] + geom_smooth(se = FALSE)
lindia::plot_all(diag.plots[1:4])
```



In this case, we see some kind of curved relationship between residuals and fitted values / `wt`.

Coefficients

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

This piece of information tells us that the model has been estimated to be

$$\texttt{mpg} = 37.3 - 5.3\,\texttt{wt}$$

The table also shows individual hypothesis tests of significance (i.e. testing that the coefficient is non-zero) conducted for each of the coefficients. We find that both β_0 and β_1 are statistically significant, with both tests strongly rejecting the null hypothesis that $\beta_0 = 0$ (p-value $= < 2e16$) and $\beta_1 = 0$ (p-value $= 1.29e-10$).

What is the interpretation of the coefficient for `wt`? Firstly, notice that it is negative, which means that there is an inverse relationship between the explanatory variable `wt` and the response variable `mpg`. This makes sense, because the heavier the car is, the less fuel efficient it is. For every unit increase in `wt` (given in 1000 lbs), we see on average a decrease of -5.3 miles per gallon.

Residual standard error

```
## Residual standard error: 3.046 on 30 degrees of freedom
```

This gives the estimate for $\hat{\sigma}$, otherwise known as residual standard error. To understand why this is called what it is, take a look at the formula. As we mentioned above, this follows a χ^2 distribution with $n-p-1 = 30$ degrees of freedom.

Multiple R-squared

```
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
```

As the name states, this gives the estimate for R^2 and R^2_{adj} respectively.

F-statistic

```
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

This is the piece of information which relates to testing all coefficients are non-zero simultaneously. The F-statistic, $T = 91.38$ is calculated using the formula given above, and it is compared against the $F_{1,30}$ distribution. We see that it gives a p-value of $1.294e - 10$ which implies strongly that the null hypothesis is rejected, concluding that at least one of the coefficients is non-zero.

Line of best fit

Using the information from the linear regression, one can produce the following table:

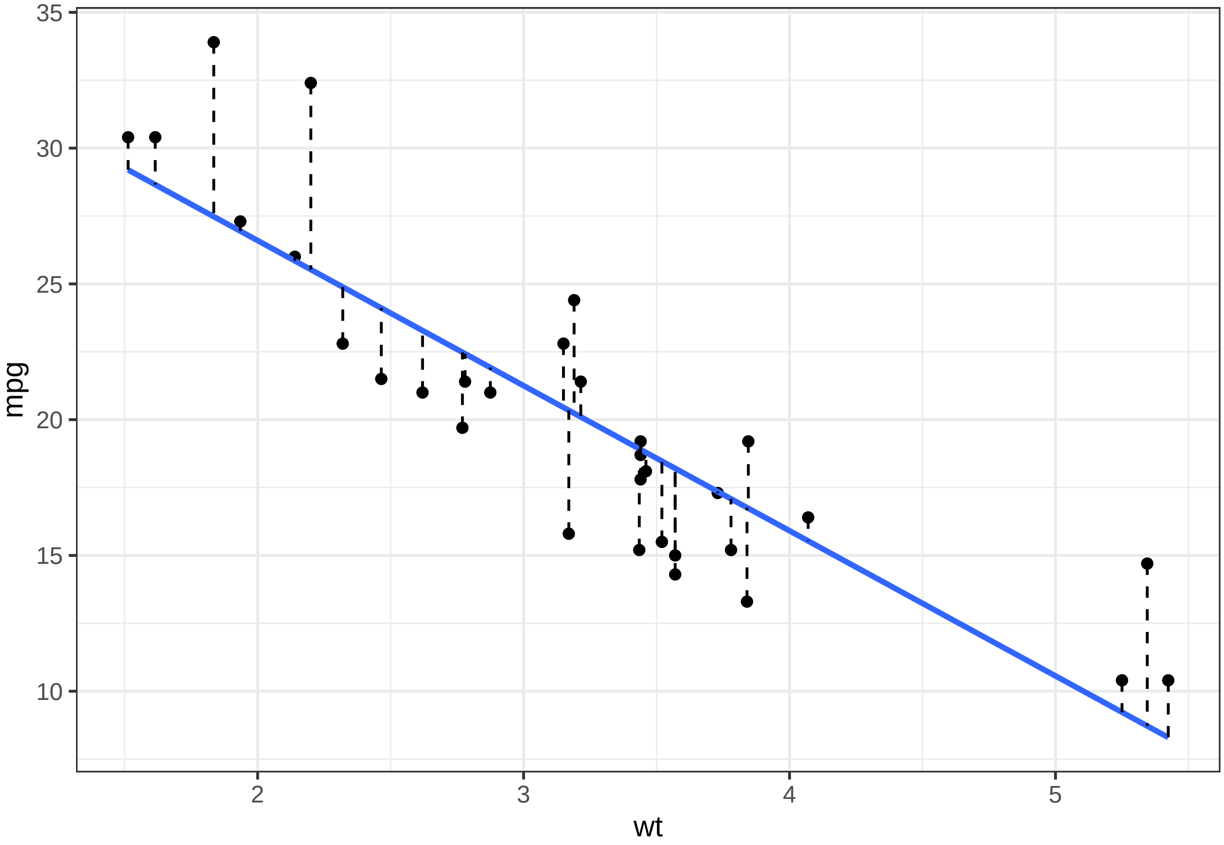
```r
knitr::kable(fortify(mod)[, -c(3:5)])
```

	mpg	wt	.fitted	.resid	.stdresid
Mazda RX4	21.0	2.620	23.282611	-2.2826106	-0.7661677
Mazda RX4 Wag	21.0	2.875	21.919770	-0.9197704	-0.3074305
Datsun 710	22.8	2.320	24.885952	-2.0859521	-0.7057525
Hornet 4 Drive	21.4	3.215	20.102650	1.2973499	0.4327511
Hornet Sportabout	18.7	3.440	18.900144	-0.2001440	-0.0668188
Valiant	18.1	3.460	18.793254	-0.6932545	-0.2314831
Duster 360	14.3	3.570	18.205363	-3.9053627	-1.3055222
Merc 240D	24.4	3.190	20.236262	4.1637381	1.3888971
Merc 230	22.8	3.150	20.450041	2.3499593	0.7839269
Merc 280	19.2	3.440	18.900144	0.2998560	0.1001080
Merc 280C	17.8	3.440	18.900144	-1.1001440	-0.3672871
Merc 450SE	16.4	4.070	15.533127	0.8668731	0.2928865
Merc 450SL	17.3	3.730	17.350247	-0.0502472	-0.0168379
Merc 450SLC	15.2	3.780	17.083024	-1.8830236	-0.6315997
Cadillac Fleetwood	10.4	5.250	9.226650	1.1733496	0.4229607
Lincoln Continental	10.4	5.424	8.296712	2.1032876	0.7697987
Chrysler Imperial	14.7	5.345	8.718926	5.9810744	2.1735331
Fiat 128	32.4	2.200	25.527289	6.8727113	2.3349021
Honda Civic	30.4	1.615	28.653805	1.7461954	0.6103569
Toyota Corolla	33.9	1.835	27.478021	6.4219792	2.2170827
Toyota Corona	21.5	2.465	24.111004	-2.6110037	-0.8796401
Dodge Challenger	15.5	3.520	18.472586	-2.9725862	-0.9931363
AMC Javelin	15.2	3.435	18.926866	-3.7268663	-1.2441801
Camaro Z28	13.3	3.840	16.762355	-3.4623553	-1.1627910
Pontiac Firebird	19.2	3.845	16.735633	2.4643670	0.8277197
Fiat X1-9	27.3	1.935	26.943574	0.3564263	0.1224441
Porsche 914-2	26.0	2.140	25.847957	0.1520430	0.0517719
Lotus Europa	30.4	1.513	29.198941	1.2010593	0.4225427
Ford Pantera L	15.8	3.170	20.343151	-4.5431513	-1.5154971
Ferrari Dino	19.7	2.770	22.480940	-2.7809399	-0.9308693
Maserati Bora	15.0	3.570	18.205363	-3.2053627	-1.0715194
Volvo 142E	21.4	2.780	22.427495	-1.0274952	-0.3438822

We can then plot the line of best fit as follows

```r
ggplot(mtcars, aes(x = wt, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_segment(data = fortify(mod),
               aes(x = wt, xend = wt, y = mpg, yend = .fitted),
               linetype = "dashed") +
  theme_bw()
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Exercise 1

Consider the following data set, which has observations on four variables

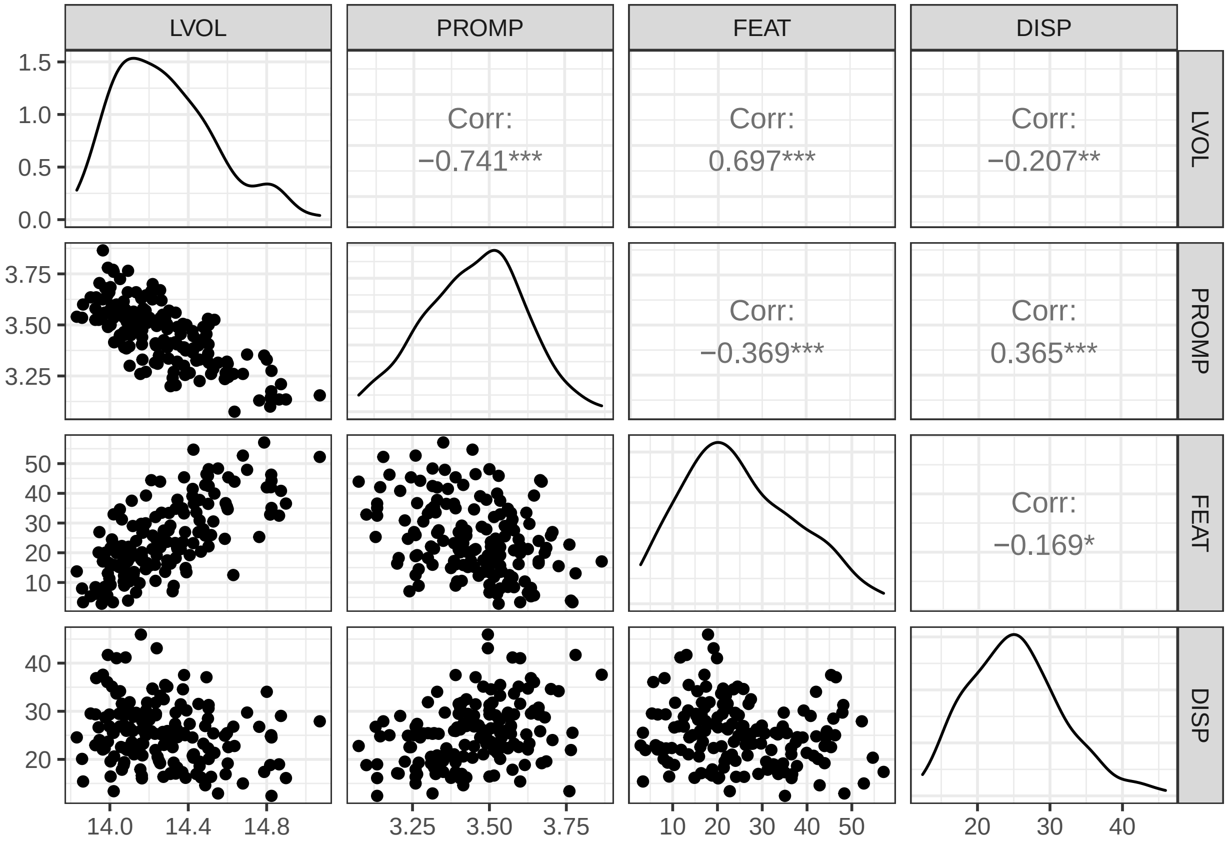
- LVOL logarithms of weekly sales volume
- PROMP promotion price
- FEAT feature advertising
- DISP display

```
foods <- as_tibble(read.table("foods.txt", header = TRUE))
foods
```

```
## # A tibble: 156 x 4
##     LVOL PROMP  FEAT  DISP
##    <dbl> <dbl> <dbl> <dbl>
##  1  14.5  3.52  39.9  21.4
##  2  14.2  3.7   25.8  34.6
##  3  14.3  3.42  23.3  27.4
##  4  14.3  3.55  25.5  25.7
##  5  14.2  3.64  39.2  30.2
##  6  14.0  3.78  13.1  41.7
##  7  14.0  3.86  17.1  37.6
##  8  14.0  3.60  24.5  35.1
##  9  14.1  3.72  15.5  34.2
## 10   14.2   3.52   24.8   29.2
## # ... with 146 more rows
```

Here's an exploratory plot of the all the variables

```
GGally::ggpairs(foods) + theme_bw()
```



Here are the results of the regression model fitted on the data set

```
mod0 <- lm(LVOL ~ 1, foods)
mod1 <- lm(LVOL ~ PROMP, foods)
mod2 <- lm(LVOL ~ PROMP + FEAT, foods)
mod3 <- lm(LVOL ~ PROMP + FEAT + DISP, foods)
mtable("Model 1" = mod1, "Model 2" = mod2, "Model 3" = mod3,
       summary.stats = c("sigma", "R-squared", "F", "p", "N",
                         "Log-likelihood", "Deviance" , "AIC", "BIC"))
```

```
##
## Calls:
## Model 1: lm(formula = LVOL ~ PROMP, data = foods)
## Model 2: lm(formula = LVOL ~ PROMP + FEAT, data = foods)
## Model 3: lm(formula = LVOL ~ PROMP + FEAT + DISP, data = foods)
##
## ==========================================================
##                         Model 1      Model 2      Model 3
## ----------------------------------------------------------
##   (Intercept)       18.409***    17.150***    17.237***
##                     (0.302)      (0.249)      (0.249)
##   PROMP             -1.197***    -0.904***    -0.956***
##                     (0.087)      (0.069)      (0.073)
##   FEAT                            0.010***     0.010***
##                                  (0.001)      (0.001)
##   DISP                                         0.004*
##                                               (0.002)
## ---------------------------------------------------------
##   sigma              0.172        0.127        0.125
##   R-squared          0.549        0.756        0.763
##   F                187.104      236.997      163.425
##   p                  0.000        0.000        0.000
##   N                156          156          156
##   Log-likelihood    54.371      102.362      104.752
##   Deviance           4.549        2.459        2.385
##   AIC             -102.743     -196.724     -199.504
##   BIC              -93.593     -184.525     -184.254
## =========================================================
##   Significance: *** = p < 0.001; ** = p < 0.01;
##                 * = p < 0.05
```

As a side note,

- Log-likelihood refers to the log-likelihood value of the normal regression model, obtained using the estimated values of the parameters.
- AIC refers to *Akaike's information criterion*, an estimator of the relative quality of statistical models for a given set of data. It is given by the formula $\text{AIC} = -2 \text{ log-likelihood} + 2k$, where k is the number of parameters of the model.
- BIC refers to the *Bayesian information criterion*, another criterion for model selection. It is given by the formula $\text{BIC} = -2 \text{ log-likelihood} + k \log n$, where k is the number of parameters of the model and n is the number of data points.
- There are other information criterion out there, but these are probably the two most commonly used ones.
- The model with the lowest information criterion is preferred.

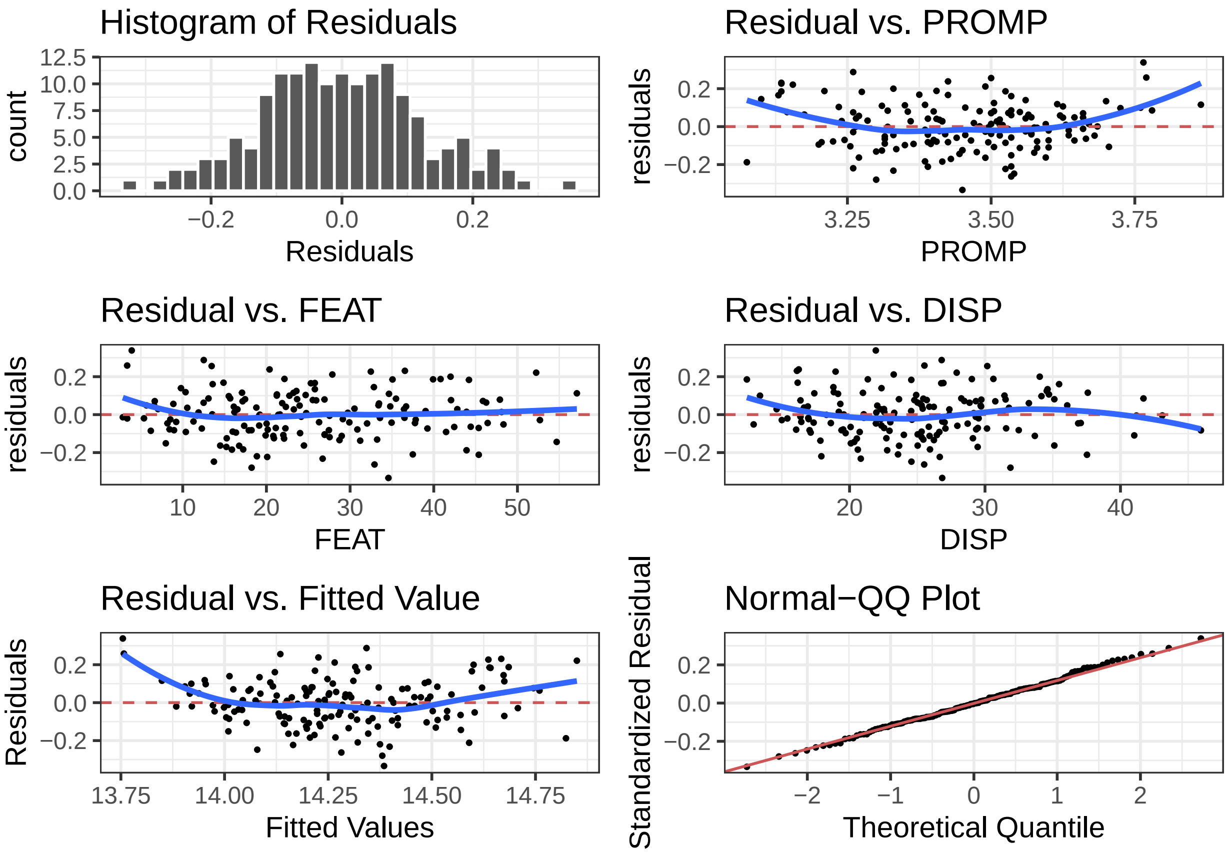
Consider the ANOVA table for Model 3

```
knitr::kable(anova(mod3),  digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PROMP	1	5.527	5.527	352.305	0.000
FEAT	1	2.090	2.090	133.243	0.000
DISP	1	0.074	0.074	4.729	0.031
Residuals	152	2.385	0.016		

And here are the diagnostic plots

```
diag.plots <- lindia::gg_diagnose(mod3, plot.all = FALSE)
diag.plots <- lapply(diag.plots, function(x) x + theme_bw())
diag.plots[[2]] <- diag.plots[[2]] + geom_smooth(se = FALSE)
diag.plots[[3]] <- diag.plots[[3]] + geom_smooth(se = FALSE)
diag.plots[[4]] <- diag.plots[[4]] + geom_smooth(se = FALSE)
diag.plots[[5]] <- diag.plots[[5]] + geom_smooth(se = FALSE)
lindia::plot_all(diag.plots[1:6])
```



QUESTIONS

1. What are the regression equations for Models 1-3?
2. Which model should we choose? Comment on model selection based on various criterion such as R^2, Log-likelihood, AIC, BIC and possibly others.
3. Consider Model 3. Are all the coefficients statistically significant?
4. For Model 3, state the 95% confidence interval for the three coefficients.
5. For Model 3, give an interpretation of the coefficients of these models.
6. In the ANOVA table above, which value represents the estimate for σ^2?
7. In the ANOVA table above, how do you obtain the F-value of of 163.425 as given in the regression table results?
8. What is the contribution of the variable `DISP` to the regression model? In other words, what is the percentage of `DISP`'s regression sum of squares in consideration of all regressors' sum of squares?
9. Comment on the diagnostic plots.
10. Can we plot Model 3 fitted regression line?

Exercise 2

Read the article Kobina & Abledu, "Multiple Regression Analysis of the Impact of Senior Secondary School Certificate Examination (SSCE) Scores on the final Cumulative Grade Point Average (CGPA) of Students of Tertiary Institutions in Ghana." (link here: https://pdfs.semanticscholar.org/68ba/ad469ae2f59f194f6e97f1c3d262fc2c6375.pdf).

1. In your own words, state the authors' research question. Are they trying to answer a causal question?
2. What is their independent variable? What are their dependent variables?
3. What is the main finding from the paper? In other words, how do the authors answer their research question? What is the main evidence they use to make this claim? Put this in your own words.

Read more

1. When we have data that is non-continous (e.g. binary data, categorical data, count data, etc.), certain assumptions of the linear model are violated (which ones?). We have to fit a generalised linear model in such cases. Examples include logistic regression, Poisson regression, etc. Find some introductory material on these kinds of regression models.
2. What happens when we have explanatory variables which are categorical in nature? E.g. Sex, Likert scale responses, political party, etc. Read up on the use of "dummy variables" in regression analysis.
3. When we have a lot of predictor variables, the model may not adequately fit the data. How do we perform model/variable selection? One way is to do forward/backward selection or otherwise known as stepwise addition/delection scheme. Read up on this method.
4. When we have too many variables, specifically when $p > n$, the linear regression model becomes mathematically impossible to fit. Why is this?