SS9055B: Generalized Linear Models
Section 1: Linear Regression

Objectives

By the end of the lecture you should be able to:

1. Describe the assumptions of linear regression and the least squares method of estimation.
2. Fit linear regression models in `R` and summarize the results.
3. Conduct hypothesis tests and compute information criterion to compare model fit.
4. Construct graphical and numeric diagnostics to assess assumptions and goodness-of-fit.

1 Introduction

You have probably guessed that the name generalized linear models refers to the fact that the models we will study in this course are generalization of the linear regression model. In particular, the framework of generalized linear models includes models that extend the assumptions of linear regression by first allowing the response to have a distribution other than the normal and second defining a relationship between the mean and the linear predictor (the linear combination of coefficients and predictors often denoted by $\boldsymbol{x}_i'\boldsymbol{\beta}$) to be something other than the identity function. Don't worry if that doesn't make sense yet. It will soon.

To understand the work we do this semester you will need to be familiar with the basic methods of linear regression. This reading provides a quick, and by no means complete, summary of the essential topics of linear regression and introduces many of the basic functions in `R` that we will use throughout the semester. Some of the material may be new, and some of my opinions may be different from what you have been taught before. However, you should already be familiar with most of the material.

The `R` code for reproducing the output in this document is included in the accompanying file `1_linear_regression.R`. Note that I have become a devotee of the packages in the tidyverse (https://www.tidyverse.org/) including `dplyr` and `ggplot`. Some of the syntax may not be familiar if you use base `R`, like the use of the pipe operator `|>`. However, this should not be a problem if you have used `R` before. I find the syntax of the tidyverse functions is quite intuitive, and I am happy to help if you have not used the tidyverse before. There are also lots of resources on the web.

2 Assumptions

The objective of linear regression is to predict the distribution of a scalar response variable conditional on a fixed number, p, of observed predictors variables. I will let n denote the number of observations, Y_i the response variable for observation i,and $x_{i1}, \ldots, x_{ip}$ the corresponding values of the predictor variables. The basic assumptions of the linear regression model are that:

1. $Y_1, \ldots, Y_n$ are independent conditional on the values of the predictors.

2. The expected value of Y_i, μ_i, is a linear function of the predictors:

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

3. All Y_i have the same variance, $\text{Var}(Y_i) = \sigma^2$, $i = 1, \ldots, n$.

4. Each Y_i is normally distributed.

The first three assumptions are key to the model, but the assumption of normality is sometimes omitted. The assumption of normality allows us to show directly that the sampling distributions of the estimators are also normal. However, it turns out that the sampling distributions are likely to be normal, or approximately normal even if the response itself is not normally distributed. Least squares estimates of the coefficients, $\beta_0, \ldots, \beta_p$, can be computed regardless of the distribution of the error terms, $\epsilon_i = Y_i - \mu_i$, $i = 1, \ldots, n$. Moreover, least squares estimators are weighted averages of the observations and so their normality can be justified through the central limit theorem if n is large even if the distribution of the error term is not normal or is unknown.

These assumptions can be written more succinctly by saying that $Y_1, \ldots, Y_n$ are independent random variables such that

$$Y_i | x_{i1}, \ldots, x_{ip} \sim \text{Normal}(\mu_i, \sigma^2)$$

where

$$\mu_i = \boldsymbol{x}_i' \boldsymbol{\beta}$$

with $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$. We can write this even more compactly with matrix notation and the multivariate normal distribution by stating that

$$\boldsymbol{Y} | X \sim N_n(\boldsymbol{\mu}, \sigma^2 I_n)$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, X is the $n \times (p+1)$ design matrix with $\boldsymbol{x}_i'$ in the i^{th} row, and $\boldsymbol{\mu} = X\boldsymbol{\beta}$. The notation $N_n(\boldsymbol{\mu}, \Sigma)$ represents the n-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix Σ, and I_n is the identity matrix of dimension n. Note that the assumption of independence is hidden in the fact that the variance matrix is diagonal. Normal random variables are independent if and only if they have zero covariance/correlation (this isn't true for other distributions).

The simple linear regression model can also be extended by allowing the variance covariance matrix to be other than the identity. First, we might assume that the variance matrix remains diagonal but is no longer a multiple of the identity matrix so that the variance of the error is not the same for all observations. That is, we can assume that

$$\boldsymbol{Y} | X \sim N_n(\boldsymbol{\mu}, \sigma^2 W)$$

where $W = \text{diag}(w_1, \ldots, w_n)'$ for some positive weights $w_i > 0$, $i = 1, \ldots, n$. Even more generally, we might assume that

$$\boldsymbol{Y} | X \sim N_n(\boldsymbol{\mu}, \sigma^2 \Sigma)$$

for some symmetric, positive-definite matrix Σ. It is common to assume that W or Σ are known so that only σ^2 must be estimated for the variance matrix to be defined.

3 Least Squares Estimation

I don't want to dwell on theory too much in this section, but I will introduce least squares estimation for completeness. As we will see in later sections, the least squares estimators are equivalent to the maximum likelihood estimators given the assumption that the response variables are normally distributed.

The ordinary least squares estimators of the coefficients, denoted $\hat{\boldsymbol{\beta}}$, are found by minimizing the sum of the squares of the differences between observed values and their fitted means (i.e., the sum of squared residuals):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mu_i)^2 = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - X\boldsymbol{\beta})'(\boldsymbol{Y} - X\boldsymbol{\beta}).$$

This equation can be minimized mathematically with the result that

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{Y}.$$

It is also possible to show that under the assumptions of the linear regression model the variance of $\hat{\boldsymbol{\beta}}$ is

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}.$$

Note that I have written $\hat{\boldsymbol{\beta}}$ as a function of the random variable $\boldsymbol{Y}$ instead of a single set of observed values, $\boldsymbol{y}$, to highlight the fact that the estimator, $\hat{\boldsymbol{\beta}}$, is itself a random variable. Unfortunately, the common notation does not distinguish the estimator, $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{Y}$, from estimate corresponding to a specific set of observed values, $\boldsymbol{y}$, which is usually also denoted by $\hat{\boldsymbol{\beta}} = (X'X)X'\boldsymbol{y}$.

If the response variables do not have the same variance or are not independent then we can still obtain estimates by minimizing the weighted or generalized sums of squares. The resulting estimators are respectively

$$\hat{\boldsymbol{\beta}} = (X'W^{-1}X)^{-1}X'W^{-1}\boldsymbol{Y} \tag{1}$$

with

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'W^{-1}X)^{-1},$$

if $\text{Var}(\boldsymbol{Y}) = \sigma^2 W$ and

$$\hat{\boldsymbol{\beta}} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\boldsymbol{Y} \tag{2}$$

with

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(X'\Sigma^{-1}X)^{-1}.$$

if $\text{Var}(\boldsymbol{Y}) = \sigma^2\Sigma$. These are referred to as the weighted least squares estimators (equation (1)) and generalized least squares estimators (equation (2)). Note that the use of "generalized" here is not connected with the framework of generalized linear modelling. It will be good to keep these expressions in mind as we will see them in the later readings that introduce inference for generalized linear models.

4 Example

The remainder of this reading focuses on the more applied aspects of linear regression and its implementation in `R`. As an example I will consider the `mtcars` data set which is included in

the `datasets` package contained in base R and can be loaded with the command `data(mtcars)`. According to the help file (which can be opened with the command `help(mtcars)`), the data was extracted from the 1974 issues of *Motor Trend* magazine and was originally published in Henderson and Velleman (1981). The data contains observations of 11 different variables for 32 different cars. For simplicity we will focus on 5 of the variables:

- mpg – Miles/(US) gallon
- cyl – Number of cylinders
- hp – Gross horsepower
- wt – Weight (1000 lbs)
- am – Transmission (0 = automatic, 1 = manual)

We will treat the efficiency of the cars in miles/gallon as the response variable and try to predict this with the other four variables. We will treat the number of cylinders and the type of transmission as categorical predictors and the horsepower and weight as continuous predictors. The plot in Figure 1 illustrates the pairwise relationships between all 5 variables.

5 Main Effects

As a first example, we will consider the model including all of the main effects (i.e., modelling the efficiency as a linear function of the 4 predictor variables). Output from fitting this model using the function `lm()` is provided in Listing 1 and contains the following pieces:

- Call: simply repeats the call to `lm()`. This is useful to keep track of what you are doing if you are fitting many models.
- Residuals: summary statistics for the residuals.
- Coefficients: a table summarizing the estimated values of the coefficients and providing the results of simple tests of significance for each.
- Unlabelled: overall summaries of the fitted model.

Although it is tempting to start at the top of the summary the information at the bottom is what should be considered first. The F-test considers whether there is evidence that any of the predictors are important (i.e., that any linear combination of the predictors is better at predicting the response values than setting the mean to the same value for all of the observations). In this case, the F-test provides very, very strong evidence that at least one of the predictor variables is linearly related to the cars' efficiency ($F = 33.57$ on 5 and 26 DF, $p < .0001$). The model explains almost 87% of the total variability in the efficiency, and the estimate of the residual standard error was just 2.41 mpg.

Moving up we reach the table of coefficients. However, we really need to do one more thing before we can interpret the values in this table. Anytime you quote an estimate in your results you need to provide a measure of uncertainty, and the best measures of uncertainty are confidence intervals. 95% confidence intervals for the coefficients computed with the function `confint()` are provided in Listing 2.

Combining the results from the table of coefficients and the confidence intervals, we see that the intercept is estimated to be 33.71 mpg with 95% confidence interval (28.35, 39.06). We interpret this to mean that if our model is correct then the average automatic car (the reference value) with 4 cylinders (the reference value), 0 horsepower, and 0 weight would drive 33.71 miles on 1 gallon of gas. This is not a very useful interpretation because a car with no horsepower and no weight doesn't actually exist and doesn't go anywhere, ever. A much more reasonable thing is to provide the estimate of the response for some representative values of the predictors, like the mean of each predictor value or the values for one specific observation. This can be done by computing the fitted value which is discussed in Section 11.

The coefficients of the explanatory variables in a regression model are always interpreted in terms of how the mean of the response would be affected if we were to change that one variable by one unit while all of the other predictors remain the same. We say that we are considering the effect of one variable while adjusting for the effect of all other variables.

Let's start with the categorical variables: the number of cylinders and the type of transmission. Categorical variables are incorporated into the linear predictor by encoding the categories in terms of a number of indicator or dummy variables. If a predictor has m levels then we need to include the intercept and $m-1$ dummy variables to separate the levels. There are many ways in which this can be done and we need to know the specific encoding that has been implemented before we can interpret the coefficients. For example, the number of cylinders has three levels, 4, 6, and 8, and so the same model can be fit by defining two dummy variables

$$\begin{aligned} x_{i1} &= 1(\text{car } i \text{ has 6 cylinders}) \\ x_{i2} &= 1(\text{car } i \text{ has 8 cylinders}) \end{aligned}$$

or

$$\begin{aligned} x_{i1} &= 1(\text{car } i \text{ has 6 cylinders}) \\ x_{i2} &= 1(\text{car } i \text{ has 6 or 8 cylinders}) \end{aligned}$$

or

$$\begin{aligned} x_{i1} &= 1(\text{car } i \text{ has 6 cylinders}) \\ x_{i2} &= 1(\text{car } i \text{ has 4 cylinders}) \end{aligned}$$

where the function $1(\cdot)$ is the indicator function such that $1(S) = 1$ if statement S is true and 0 if S is false. By default, the design matrix in `R` is constructed using the first encoding so that the first level alphabetically is treated as the reference level. Each dummy variable then models the difference in the mean response between the reference level and one of the other levels (when all other variables are held fixed).

In the case of the number of cylinders, 4 cylinders is taken as the reference value, the first dummy variable models the difference in the mean efficiency between cars with 4 and 6 cylinders (and the same horsepower, weight, and transmission type) and the second models the difference in the mean efficiency between cars with 4 and 8 cylinders (and the same horsepower, weight, and transmission type). We would need to consider the difference between these parameters if we wanted to know the difference between cars with 6 an 8 cylinders (and the same horsepower, weight, and transmission type).

For example, the point estimate of the effect of 6 vs 4 cylinders is -3.03 with 95% confidence interval (-5.92,-.14) and p-value .041. We interpret this to mean that there is reasonable evidence that cars with 6 cylinders are less efficient than cars with 4 cylinders, driving on average 3.03 mpg less (95%CI=.14,5.92) given that they are the same otherwise (i.e., they have the same horsepower, weight, and transmission type).

The coefficients of continuous predictors are easier to interpret because coding is not an issue. We interpret the coefficient as the change in the mean response when the predictor increases by one unit while the other predictors remain fixed. For example, the point estimate for the coefficient of weight is -2.50 mpg/1000 lbs with 95% confidence interval (-4.32,-.68) and p-value .009. This provides very strong evidence that the efficiency of the cars is affected by weight while the other variables are fixed. All else (number of cylinders, horespower, and type of transmission) being equal, we estimate that the mean efficiency decreases by 2.50 miles per gallon of gas for each increase in weight of 1000 lbs and we are 95% confident that this difference is between .68 and 4.32 miles.

Finally, the `summary()` function provides the five-number summary for the residuals. This is seldom important as it does not provide enough information to really judge the distribution of the residuals, and we need to look at more sophisticated diagnostics, discussed in Section 10.

6 Testing and p-values

You are going to find out that I am very opinionated, and there are two things in the summary of the linear regression model that I don't like. The first is that it provides results for the t-tests comparing all coefficients to zero. The second is that it prints those pesky dots and stars to indicate whether the t-tests for each coefficients are between certain levels of significance.

The reason that I don't like the tests are provided for all coefficients is that only a few of these tests are relevant. In some cases, some of the tests may be misleading, but in other cases, the tests may be incorrect. In this case, it makes no sense to test whether or not the intercept is equal to zero. As noted above, this is asking whether the average fuel efficiency of an automatic car (the reference value) with 4 cylinders (the reference value), 0 horsepower, and 0 weight is 0 gallons/mile. This is a non-sensical question. More generally, `summary()` prints the test statistic and corresponding p-value for each predictor in the model without asking if they are needed, and these tests are often irrelevant or shouldn't be conducted in the first place (as when we discuss interactions below). However, our eyes get attracted to the highly significant p-values, and it's hard to ignore them once you've seen them.

If you are looking at the p-values for all of the different coefficients then it probably means that you haven't thought enough before you started your analysis. It's likely that specific predictors are of interest as they relate to the research hypotheses while others are included in the model simply to ensure that the assumptions are satisfied. It is always good to limit the number of hypothesis tests you conduct in any analysis to reduce the chance of a Type I Error, and you must train yourself to consider only the tests that are important. To be frank, I would much rather that summary included confidence intervals and not the tests and that specific tests could be chosen by the user. However, I don't get to decide these things.

I don't like the stars because it perpetuates the myth that there are some universal levels of significance that are more important than others (the almighty .05 being the most common). The p-value for a test is a continuous value that falls between 0 and 1 and should be interpreted as such. A p-value close to 0 provides strong evidence that the null hypothesis (in this case that the

corresponding coefficient is equal to 0) is incorrect and as the p-value increases the evidence gets weaker and weaker. A p-value like .001 or less can always be interpreted as very strong evidence against the null and in most cases any p-value above .25 can safely be interpreted to mean that the value of the test statistic could easily have occurred by chance if the null hypothesis is true and so there is very little evidence against the null. However, interpreting p-values in between depends on many factors including the design of the experiment and the consequences of making a wrong decision either way. We will try to interpret p-values along this continuum instead of comparing p-values to a single threshold and declaring that a test is either significant or not. I highly encourage you to read the American Statistical Association's view on p-values provided in Wasserstein and Lazar (2016). I will discuss significance testing more in Section 9 on model selection.

7 Interactions

In my experience, interaction terms are often misinterpreted by students and applied researchers who think that they model the how one of the predictors variables change as the value of another predictor changes. This is not the case. Instead, interaction terms examine how the effect that one explanatory variable has on the mean response changes as we change the value of one or more of the other predictors. For example, we might expect that the effect weight on efficiency is bigger for cars with more cylinders (at least, someone with mechanical knowledge might think that). We'll examine these effects through a series of models including just two of the predictors and their interactions.

Mathematically, let's suppose that a model contains two categorical predictor variables, x_1 and x_2, each with two levels. The linear predictor including the main effects and interaction (we always include the main effects if there is an interaction) is

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}.$$

The following table displays the values of the mean for all four possible combinations of x_1 and x_2

	x_2	
x_1	0	1
0	β_0	$\beta_0 + \beta_2$
1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Now consider the effect of the first categorical variable, x_1, i.e., the difference in μ_i when $x_1 = 0$ and when $x_1 = 1$. When $x_2 = 0$ the difference is

$$(\beta_0 + \beta_1) - (\beta_0) = \beta_1.$$

However, when $x_2 = 1$ the difference is

$$(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) = \beta_1 + \beta_3.$$

The effect of x_1, the difference in the mean when x_1 increase from 0 to 1, depend on the value of x_2. If $x_2 = 0$ then the difference is just β_1, but if $x_2 = 1$ then the difference is $\beta_1 + \beta_3$. The interaction term, β_3, determines this difference. If $\beta_3 > 0$ then the effect of x_1 is more positive when $x_2 = 1$ and if $\beta_3 < 0$ then the effect of x_1 is more negative when $x_2 = 1$. If $\beta_3 = 0$ then

the effect of x_1 is the same whether $x_2 = 0$ or $x_2 = 1$. Note that the effect of the interaction term is symmetric: β_3 also models how the effect of x_2 depends on the value of x_1 in exactly the same way.

Consider the case of modelling the fuel efficiency as a function of both the number of cylinders and the type of transmission. There are three levels for the number of cylinders and two types of transmission, so the number of predictors required to add the interaction term in the model $(3-1) \times (2-1) = 2$. The output from this model is provided in Listing 3 and the two interaction terms are labelled as `cyl6:am1` and `cyl8:am1` in the table of coefficients. We can interpret these either as describing the difference in the effect of changing the number of cylinders from 4 to 6 to 8 for manual vs automatic cars or as the difference in the effect of changing the type of transmission for cars with 6 and 8 cylinders in comparison to cars with only 4 cylinders. The two are equivalent. For example, the estimate of the first term is -3.73 with 95% confidence interval (-10.09,2.62). We can interpret this to mean that the estimated difference in mean efficiency between cars with 4 and 6 cylinders is 3.73 mpg lower for manual cars (`am=1`) than for automatic cars (`am=0`). Alternatively, we can say that the estimated difference in mean efficiency between manual and automatic cars is 3.73 mpg lower for cars with 6 cylinders than for cars with 4 cylinders. Note, however, that the 95% confidence interval easily covers 0 and the p-value, .131, is quite large indicating that there is little evidence that this interaction is real. The difference we observed could easily have occurred by chance if the effect of changing from 4 to 6 cylinders is the same for automatic and manual cars (or the effect of changing from automatic to manual cars is the same for cars with 4 and 6 cylinders).

Next consider a model including one categorical predictor, one continuous predictor, and their interaction. Models like this are most often interpreted by considering that the interaction term describes how the slope relating the mean response and the continuous predictor varies across the levels of the categorical predictor. Results for the model including transmission type, weight, and their interaction are provided in Listing 4. Based on this model, we would estimate that the efficiency decreases by 5.30 mpg more per 1000 lb increase in weight for manual cars (`am=1`) than for automatic cars (`am=0`). I.e., if two cars differ in weight by 1000 lb then we would expect the difference in their efficiency to be 5.30 mpg lower if the cars are both manual than if they are both automatic. The 95% confidence interval for the interaction is (-8.28,-2.34) which is bounded well below zero, and the p-value is .001, both of which provide strong evidence that the interaction term is not equal to zero.

Finally, we consider the interaction between two continuous predictor. In this case, the interaction models how the effect of one predictor changes when the value of the other predictor is increased by one unit. As in the case of two categorical variables, the interaction term is symmetric in that it can be interpreted equally as the change in the effect of the first predictor as the second is varied or vice versa. Listing 5 provides the output from the model including the weight, horsepower, and their interaction as predictors of the efficiency. The estimate of the interaction of weight and horsepower is only .03 mpg, but the confidence interval, (.01,.04) , is very narrow and the p-value is only .001. We interpret this to mean that there is strong evidence of an interaction. We can say that we estimate that the effect of weight becomes more positive by .03 mpg/1000 lbs when horsepower increases by one unit. Equivalently, we also estimate that the mean effect of a one unit increase in horsepower is more positive by .03 mpg for cars weighing 1000 lbs more.

8 Model Comparison

Model comparison generally refers to the process of comparing the fit of two models with different sets of predictors. For any model there are two competing issues to consider. We want to construct the best possible predictions for the response, but we also want to avoid including predictors that are not really important. We can always construct perfect predictions of the observed data simply by setting $\mu_i = y_i$. One way this can be done by including a categorical covariate with a separate value for each observation. However, this model presents an extreme case of overfitting and is practically useless. It does not allow us to make conclusions about the general trends in the data that answer research questions and cannot be used to compute predictions for new observations since this would introduce new values of the covariate into the model for which the coefficients are unknown.

There are two general strategies for comparing models to identify the predictors that provide the best description of the data while avoiding overfitting. The first is to compare nested models via a hypothesis test with the null hypothesis that some of the coefficients are equal to fixed values (usually 0 which effectively removes the associated predictors from the model). The second is to apply some sort of information criterion like the AIC.

8.1 F-tests and ANOVA

The analysis of variance (ANOVA) compares nested regression models. Two regression models are said to be nested if the expression for the linear predictor in one model (the reduced model) can be obtained by fixing the values of some coefficients in the other model (the full model) [1]. Most often the reduced model is constructed by setting the values of some of the coefficients in the full model to 0 which effectively removes these predictors from the model (i.e., the reduced model assumes that the mean response is not affected by these predictors). Two nested models can be compared via an F-test. The degrees of freedom for the numerator of the test is equal to the number of parameters in the full model that are fixed to obtain the reduced model. If the degrees of freedom in the numerator is one then the F-test is equivalent to a t-test. In fact, the t-tests presented in the model summaries produced by `R` are equivalent to F-tests comparing the fitted model with the reduced models formed by removing each coefficient in the model one-at-a-time.

Consider the model of the `mtcars` data including the main effects of transmission type and the number of cylinders, and their interaction. This model has two more terms than the model including only the main effects and so two extra t-tests are included in the summary table. This is not a substantial problem. However, if the number of categories for these two variables were larger then we would need to conduct many t-tests to test each of the coefficients associated with this interaction term. This is inefficient, and it also introduces the multiple testing problem. The more tests we conduct, the higher the chances of committing at least one Type I Error and including a predictor that does not actually affect the mean of the response. A better strategy would be to assess first whether there is any evidence of an interaction at any levels of the two predictors. This can be examined with an F-test implemented in `R` with the `anova()` function. If there is sufficient evidence of an effect at some level then we can use t-tests to identify where.

The ANOVA table for a regression model is constructed by conducting the F-tests removing each term in the model (i.e., setting the corresponding coefficients equal to 0) one-at-a-time. A

[1] Technically, two models are nested if the columns of the design matrix of the reduced model belong to the vector space formed by the columns of the design matrix of the full model. However, we won't need to go into such detail.

term is either a main effect or an interaction and may include multiple coefficients associated with a categorical variable. The column labeled `Df` identifies the number of coefficients that are removed (i.e., the degrees of freedom of the numerator of the F-test statistic).

Listing 6 contains the ANOVA table for the model including the number of cylinders, the transmission type, and their interaction. The F-test for the interaction term has 2 degrees of freedom (because there are two coefficients included in this term of the model) and the p-value is .269. This indicates that there is little evidence at all for the interaction overall and hence there is no point in conducting the t-tests for the individual coefficients included within this term.

The ANOVA table also illustrates the problem with `R` conducting too many tests. The table includes the tests for each of the main effects, `cyl` and `am`, as well as the test for the interaction term, `cyl:am`. However, these tests of the main effects are irrelevant. It does not make sense to test for the main effect of a predictor if that predictor is part of a significant interaction. In this example, it does not make sense to test the main effect of the number of cylinders if there is an interaction between the number of cylinders and the type of transmission. The whole point of the interaction is that the effect of the number of cylinders depends on the type transmission, and so the question about the main effect is irrelevant. These tests can only be performed if the interaction term is deemed to be not significant and the model is refit with only the main effects. Unfortunately, our eye is immediately drawn to these tests – especially when they are highlighted with multiple asterisks. I would prefer that `R` did not present the results from the tests of the main effects or interactions that are part of higher order interactions.

8.2 Model Comparison Criterion

The alternative to hypothesis testing is to compare models based on some criterion that measures the fit of the model to the data while accounting for the number of parameters in the linear predictor. The simplest criterion is the adjusted R-squared which discounts the R-squared for the number of parameters in the model. R-squared measures the proportion of variance in the response explained by the predictors in the model and always increases as more predictors are added. In comparison, the adjusted R-squared may increase or decrease depending on how much the residual standard error is reduced. In fact, the adjusted R-squared can be negative indicating that the model is no better than the model with constant mean.

Several other criterion exist and the well known are the AIC (Akaike's information criterion) and the BIC (Bayesian information criterion). Both are formulated as minus twice the log-likelihood of the model, a term that decreases as the fit of the model to the data improves, plus a penalty term determined by the number of parameters. Hence, models with smaller values of the criteria are preferred. Generally, a difference of 2 in the AIC or BIC of two models is considered to represent strong evidence for preferring one model over another and anything less than this is interpreted as meaning that neither model can be chosen over the other. However, as with p-values, the difference in AIC or BIC should be interpreted on a continuous scale and not compared to specific thresholds. If a model has K parameters then the penalty for the AIC is $2K$ and the penalty for the BIC is $K\log(n)$. A simple linear regression model with an intercept, a slope, and an unknown residual variance would have $K = 3$ parameters so that the penalty for the AIC would be 6 and the penalty for the BIC would be $3\log(n)$. A model with a single, categorical predictor with 5 categories and an unknown residual variance would have $K = 6$ parameters so that the penalty for the AIC would be 12 and the penalty for the BIC would be $6\log(n)$. In most cases, the sample size is much larger than the number of parameters in the model ($n >> K$) and so the penalty for the BIC is larger ($\log(n) > K$). This means that the BIC will tend to favour models with fewer parameters than

the AIC. This is true whenever $\log(n) > 2$ or $n > e^2 = 7.4$.

One key advantage of these criteria is that they can be applied to compare models that are not nested. In terms of linear regression, we can compare two models based on any set of predictors provided that the response variables are the same.

Listings 7 and 8 provide the AIC and BIC values for the four models compared by the ANOVA table in the previous section (note that although the ANOVA table contains only three tests it compares four models since each test represents a comparison between two models in the sequence). The AIC provides strong evidence that the three models including `cyl` are all preferred to the model including only `am`, which means essentially that `cyl` is an important predictor. There is also fairly strong evidence in favour of the model including only the main effects of the two predictors, but I would say that this is not completely definitive. Both the model including the interaction and the model including only `cyl` have AIC values that are about 2 units higher. In contrast, the BIC, which favours simpler models, provides strong evidence to prefer the model with only the main effects or the model with `cyl` alone, but does not distinguish between these. Which of these criterion to use is a matter of preference and debate, which we won't get into. What's important for this course is to know the difference and to be able to use these criterion properly.

9 Model Selection

The term model selection generally refers to choosing which predictors should be included in the linear predictor and, perhaps, whether these predictors should be transformed in anyway (e.g., by including interactions, polynomial terms, etc). There are two main strategies for selecting the predictors in a model. The first strategy is to use some form of stepwise selection. This includes backward selection which starts with a large number of predictors and removes terms one at a time until all remaining predictors are important, forward selection which starts with the model with only the intercept and adds predictors one at a time until none of the remaining predictors improves the model in a meaningful way, and what is often called stepwise selection which allows predictors both to be added and removed in alternate steps. The individual comparisons between models may be conducted either with tests of significance or with model comparison criterion like the AIC and BIC.

As an example, we might start with the model containing all four predictors and their pairwise interactions and then select a reduced model through stepwise selection based on the AIC using the `step()` function in `R`. Results are provided in Listing 9. Starting from the initial model which includes the intercept, 5 coefficients associated with the main effects, and 9 coefficients related to the interactions, the final model includes only the main effects plus coefficients related to the interactions between the number of cylinders and the horsepower (`cly6:hp` and `cyl8:hp`) and between the weight and type of transmission (`wt:am1`). One thing to note is that selection is performed with respect to the entire term not the individual coefficients in the model. For example, the interaction between the number of cylinders and the horsepower is represented by two coefficients (`cly6:hp` and `cyl8:hp`), but these coefficients are never assessed individually. Instead, we either keep both or remove both.

The second model selection strategy is to start with some candidate set of models which correspond to different hypotheses about the system being studied and then to compare these models. It is natural to consider information criterion for model selection in this strategy because the models representing different hypotheses will rarely be nested. For example, an expert with a good understanding of automechanics (not me) might hypothesize that the efficiency of cars is best

predicted by either: 1) the number of cylinders, horsepower, and their interactions, 2) the weight, type of transmission, and their interaction, or 3) the number of cylinders, transmission type, and their interaction. Listing 10 provides the AIC values for these 3 competing models. A difference in AIC of two or more is generally considered to be large, and so this provides clear evidence that the second model is preferred to the other three.

I strongly believe in the second strategy for a variety of reasons:

1. It limits the number of tests that are performed. As noted above, every extra test increases the chance of making an error and erroneously including/excluding a variable.

2. It limits the number of models that we fit. If many predictors are available then the stepwise selection may fit many different models. This will lead to overfitting and we are likely to choose models that are not plausible based on an expert understanding of the system under study.

3. It forces us to focus primarily on the science. Any statistical analysis is conducted to answer specific questions. Our analysis should answer these questions directly and should not go beyond these questions (unless this is explicitly stated). Applying a stepwise algorithm may often lead us to condcut tests outside the realm of scientific interest.

4. This approach leads us to consider multiple models that may support different hypotheses rather than selecting one. Stepwise selection leads us to choose one "best" model and to ignore all others. Unfortunately, science doesn't work this way and it is rare for the data to support only one model. The second strategy allows us to consider multiple models. In the example there is one clear best model, but in most real analyses several models will have AIC values that are very close together. Reporting our uncertainty about which model is best supported by the data is crucial to presenting a transparent analysis. This also leads naturally to model averaging which allows estimates and predictions from different, competing models to be combined in a coherent way.

I will admit that I often employ a mix of the two approaches. Following the second strategy, I will start with a candidate set of models and compare these via AIC. However, I will often perform backward selection on each of the models that competes for the best to see if they can be simplified (especially if the selected models contains complex, higher-order interactions). I feel that this strikes a balance between limiting the number of models that are fit and removing unnecessary predictors to try and simplify the results. The key is to be clear about exactly how model selection was performed, which models were selected for initial comparison and which were preferred, and what further *post hoc* comparisons were performed to simplify the final model(s).

10 Diagnostics

Regression diagnostics serve two purposes. First, they assess whether the assumptions of the model are likely to be satisfied (or are close enough to being satisfied that the results can be believed). Second, they assess how robust the results are by which I mean how sensitive the estimates are to the specific observed data and how much they might change if the data were manipulated (e.g., if observations were changed or removed).

The assumptions of normality and equal variance are best assessed through plots of the residuals. Normality is assessed with a normal quantile-quantile plot. The assumption of constant

variance is assessed by plotting the residuals versus the predictors and/or versus the fitted values. R includes built in functions to produce these plots, but I prefer to generate them myself so that I have more control. Figure 2 displays plots of the residuals for the model of mpg including the main effects of both cyl and am. The QQ-plot is based on the raw residuals whereas the remaining plots are constructed from the standardized residuals which should have a distribution with mean zero and variance close to 1 regardless of the true residual standard error (if the assumptions of the model hold). I have used boxplots to show the distribution of the residuals versus cyl and am since both are categorical variables. The results suggest that the residuals are close to normally distributed but raise concerns regarding the assumption of constant of variance. In particular, there seems to be much lower variance in the efficiency of cars with 6 cylinders in comparison to cars with 4 or 8 cylinders. This would warrant further exploration. The plots of the residuals versus the predictors and the fitted values can also be used to check for outliers. There do not appear to be any extreme outliers in this analysis.

The assumption of independence is harder to assess because we need to have some idea of how the observations might be related to know what form of dependence to look for. If the observations are collected in time or over some spatial domain then it may make sense to check for dependence between close-by values. In the case of the cars data we might expect there to be dependence between the cars that are produced by the same manufacturer. For example, all of the cars made by Mazda might be more efficient than expected simply based on the values of cyl and am. This could be assessed by extracting the brand from the names of each car and then plotting the residuals versus the brand to look for patterns that might reveal some form of dependence.

Alternative diagnostics can be used to assess how much the results depend on each of the observations in the data set. Common diagnostics are the leverage, Cook's distance, the DFBetas, and the DFFit. The leverage assesses where the vector of predictors for each observation fits into the overall distribution of the predictors. An observation has high leverage if its vector of covariates is far from the centre of the distribution. In one dimension, a point has high leverage if the value of the predictor is far above or far below the mean. Heuristically, the leverage measures the potential for a point to have an undue influence on the results, but it does not tell you whether this potential is realized. The actual influence is measured by Cook's distance which depends on both the leverage and the residual for the observations. A single point has a high Cook's distance if it has both high leverage and a high residual (i.e., it is an outlier). Such a point is said to be influential because the fit of the regression model will change significantly if the observation is removed. In fact, Cook's distance can be computed by removing each point from the data, refitting the model, and assessing the overall change in the fitted values. The DFFit is similar except that it only considers the change in the fitted value of the data point that has been removed. The DFBeta considers the change in each of the regression coefficients when each data point is removed. This leads to $n \times (p+1)$ different values – one for each combination of observation and coefficient.

Figure 3 contains plots of the Leverage, Cook's distance, DFFit, and DFBeta for the intercept term for the model including the main effects of cyl and am. The DFBetas could also be plotted for the coefficients of the two predictors, but I have not included these simply to save space. There are guidelines for determining what values constitute significant deviations from the assumptions, but these are based on specific assumptions which may or may not be satisfied. I recommend looking at the plots to see if there are points whose values are extreme compared with the remaining observations. The plots show that the leverage is similar for all of the observations, but there is one observation with an unusually high Cook's distance (#20) and one with an extreme value of DFBeta for the intercept (#21). This suggests that the fitted model is sensitive to these two data points. If we remove either data point then estimates of the coefficients and the fitted values

change significantly. It would be worth investigating this further to see if there may be problems with the data (perhaps a value was recorded incorrectly) or if there are other predictors that need to be included to improve the fit of the model to these observations.

11 Fitted Values and Prediction

Finally, I will consider the problems of computing fitted values and predicted values. These two problems are very related, but there is a subtle difference. Computing the fitted value for given values of the predictors, say $\boldsymbol{x}_{\text{new}}$ asks: What is the value of the mean response at $\boldsymbol{x}_{\text{new}}$? Predicting the response asks: What is the value of a single new observation at $\boldsymbol{x}_{\text{new}}$? If one is interested only in a point estimate then the best answers to both questions are the same. We simply multiply the predictor values by the estimated coefficients, $\boldsymbol{x}'_{\text{new}}\hat{\boldsymbol{\beta}}$. The resulting value is both the estimate of the mean observation, $\hat{\mu}_{\text{new}}$, and the predicted value of the single observation, $\hat{y}_{\text{new}}$. The difference arises when we compute estimates of uncertainty, standard errors or confidence intervals, for these values. Considering that

$$\begin{aligned}\text{Var}(\hat{\mu}_{\text{new}}) &= \text{Var}(\boldsymbol{x}'_{\text{new}}\hat{\boldsymbol{\beta}}) \\ &= \boldsymbol{x}'_{\text{new}}\text{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_{\text{new}} \\ &= \sigma^2\boldsymbol{x}'_{\text{new}}(X'X)^{-1}x_{\text{new}}\end{aligned}$$

the standard error for the fitted value (i.e., the estimated mean) is

$$\text{SE}(\hat{\mu}_{\text{new}}) = \hat{\sigma}\sqrt{\boldsymbol{x}'_{\text{new}}(X'X)^{-1}\boldsymbol{x}_{\text{new}}}.$$

On the other hand

$$\begin{aligned}\text{Var}(\hat{Y}_{\text{new}}) &= \text{Var}(\mu_{\text{new}} + \epsilon_{\text{new}}) \\ &= \text{Var}(\mu_{\text{new}}) + \text{Var}(\epsilon_{\text{new}}) \\ &= \sigma^2\boldsymbol{x}'_{\text{new}}(X'X)^{-1}x_{\text{new}} + \sigma^2\end{aligned}$$

so that the standard error of the predicted value (the prediction error) is

$$\text{SE}(y_{\text{new}}) = \hat{\sigma}\sqrt{\boldsymbol{x}'_{\text{new}}(X'X)^{-1}\boldsymbol{x}_{\text{new}} + 1}.$$

Heuristically, the prediction error accounts for the variability of the single observation about its mean as well as the uncertainty in the mean itself. Whereas the standard error of the fitted value decreases to zero as the sample size increases the prediction error decreases toward σ, the residual error. If the model we have fitted is correct (i.e., it actually generated the data), then we will learn more and more about the mean response at any point as we collect more data. However, there is uncertainty in predicting the value of a single new observation that can never be overcome.

Somewhat confusingly, both fitted values and predicted values are computed in base `R` with the `predict()` function, and the two are distinguished by setting the `type` argument. If `type='confidence'` then confidence intervals for the mean are computed. If `type='prediction'` then prediction intervals are computed instead. Listing 11 compares the fitted values with 95% confidence intervals and predicted values with 95% prediction intervals for the first 5 observations in the cars data. Note that I have used the function `augment()` from the `broom` package which is really just a wrapper around `predict()` (and some of the other functions that produce output from fitted linear

models). Whereas the `predict()` function returns a list of results that must be manipulated to produce further output (e.g., to make plots etc.), `augment()` converts the output to a matrix and combines it with the original data which makes it simpler to manipulate later (e.g., by plotting with `ggplot()`). It is immediately obvious that the prediction intervals are wider than the corresponding confidence intervals.

12 Conclusion

There is much more that can be said on the topic of linear regression, but this covers the basic information that we will need in our study of generalized linear models. As I mentioned in the introduction, the framework of generalized linear models extends linear regression in two basic ways: 1) by allowing for the response to be other than normally distributed and 2) by allowing a relationship between the mean and the linear predictor that is not linear. However, we need to cover a little more background before we can begin our study of generalized linear models in earnest. In particular, we will take time over the next couple of weeks to look at maximum likelihood estimation – an alternative for least squares estimation that we can apply to our more general models.

One final note on style. I have included the raw output from `R` in order that you can see what it looks like and can identify the different pieces of the analysis. Please do not include raw output in your own reports. Part of your job as an analyst is to select and format the appropriate output that you want to present to you audience (me, your supervisor, or a future boss or client).

References

Henderson, H. V. and Velleman, P. F. (1981). Building multiple regression models interactively. *Biometrics*, 37(2):391–411.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133.

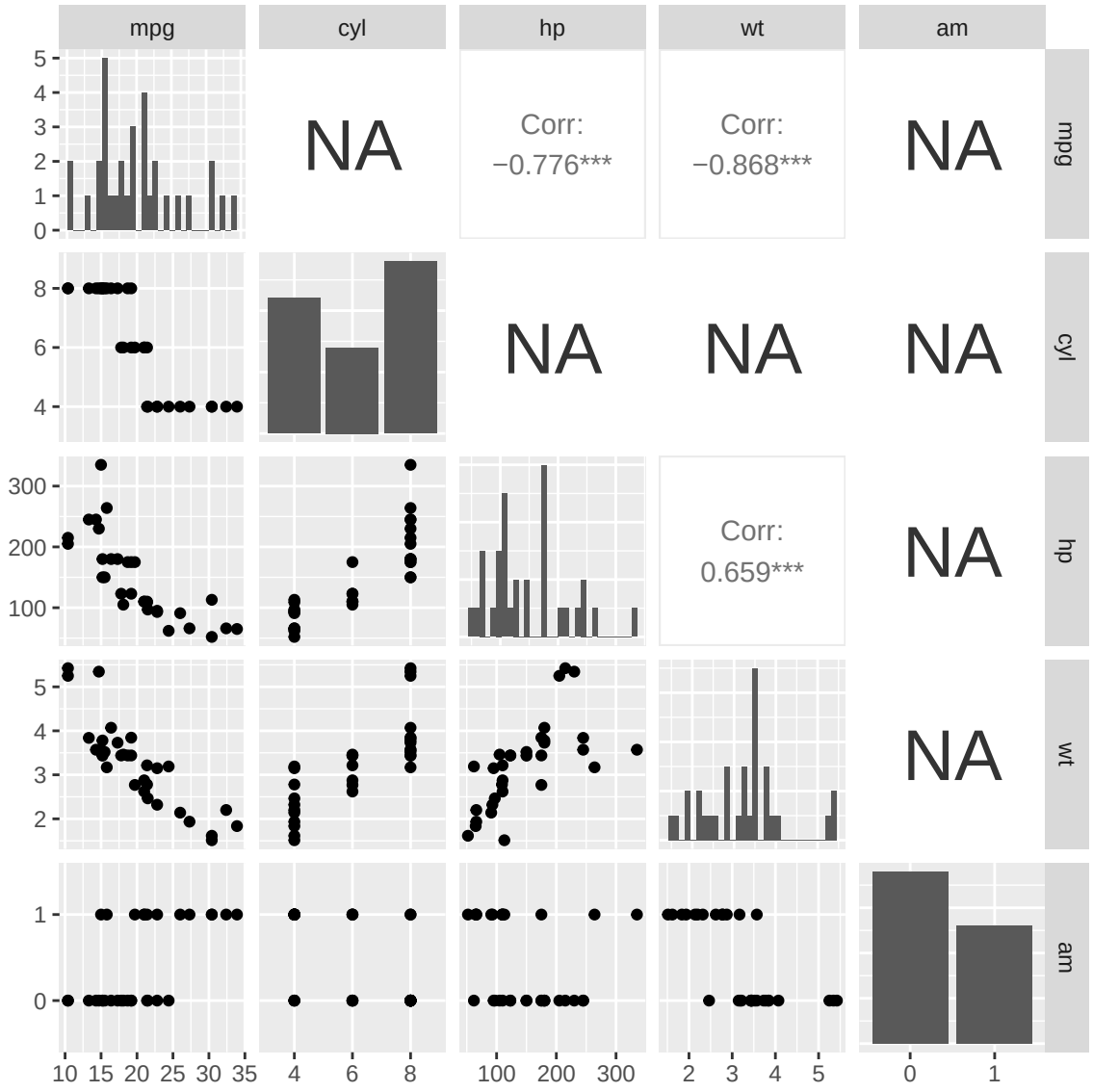

Figure 1: Pairs plot of the 5 variables selected from the `mtcars` data. The plots on the diagonal illustrate the density of each variable. The numbers above the diagonal describe the correlations when both variables are contiuous.

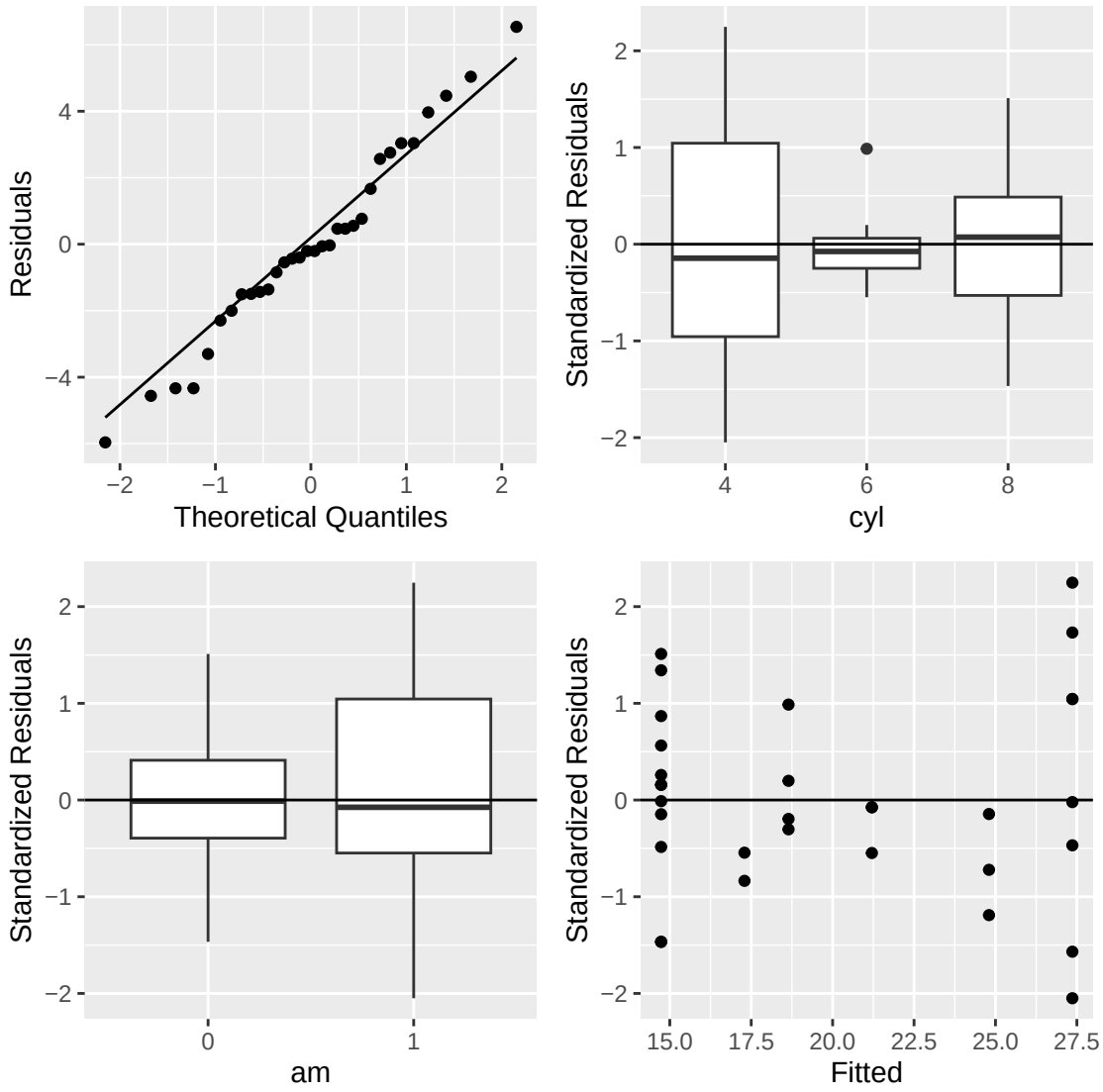

Figure 2: Residuals plots for the model of extttmpg including the main effects of `cyl` and `am`. These include the normal QQ-plot of the residuals (top left), and plots of the standardized residuals versus `cyl` (top right), `am` (bottom left), and the fitted values (bottom right).

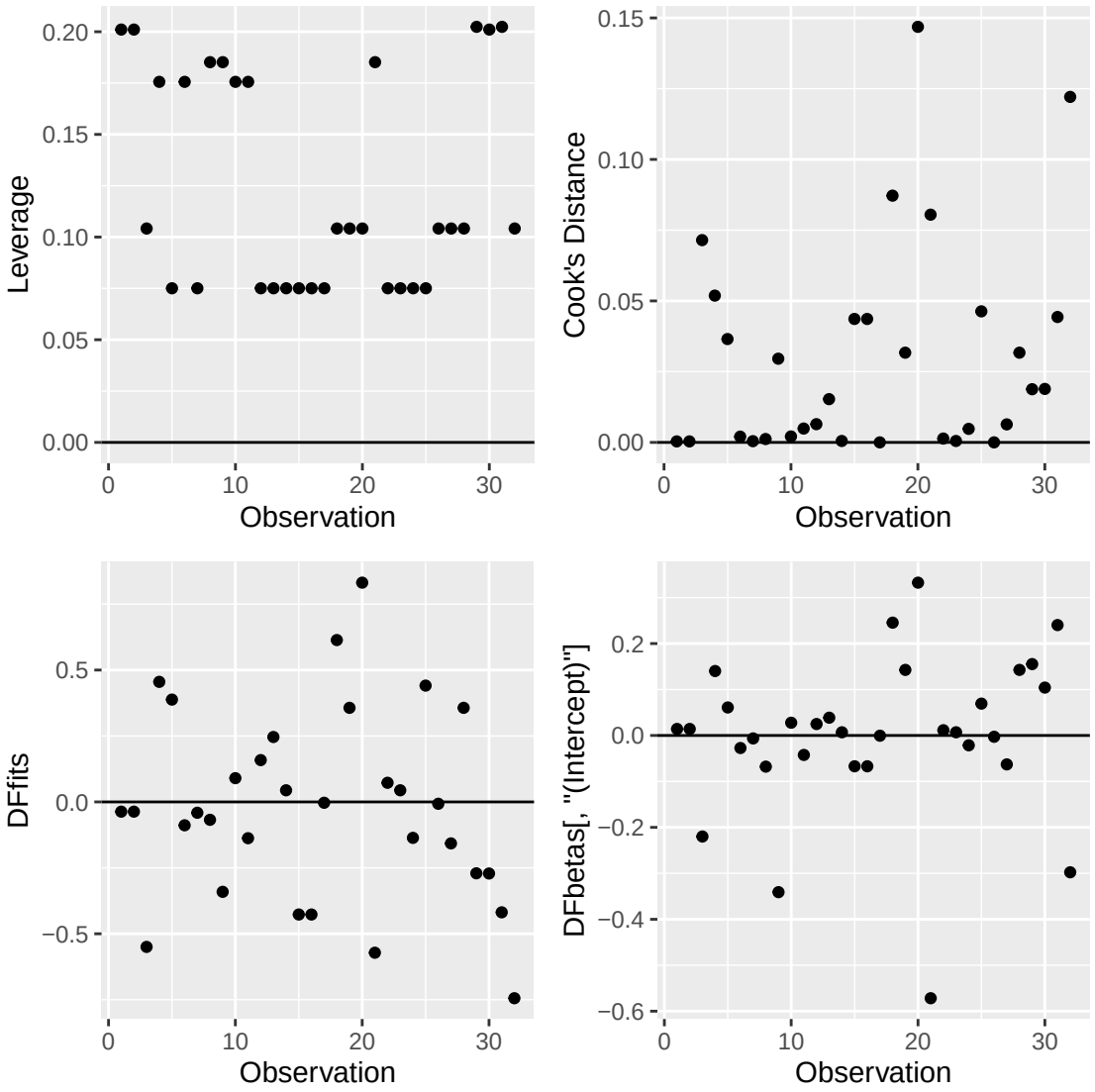

Figure 3: Further diagnostics plots for the model of extttmpg including the main effects of `cyl` and `am`. These include plots of the leverage (top left), Cook's distance (top right), DFFit (bottom left), and the and DFBeta for the intercept (bottom right).

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mycars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Listing 1: Summary of the multiple linear regression model predicting the efficiency of the 32 cars as a function of the main effect of the 4 explanatory variables.

```
##                   2.5 %       97.5 %
## (Intercept) 28.35390366 39.062744138
## cyl6        -5.92405718 -0.138631806
## cyl8        -6.85902199  2.531671342
## hp          -0.06025492 -0.003963941
## wt          -4.31718120 -0.676477640
## am1         -1.06093363  4.679356394
```

Listing 2: 95% confidence intervals for the coefficients of the main effects model.

```
##
## Call:
## lm(formula = mpg ~ cyl * am, data = mycars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6750 -1.1000  0.1125  1.6875  5.8250
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   22.900      1.751  13.081 6.06e-13 ***
## cyl6          -3.775      2.316  -1.630 0.115155
## cyl8          -7.850      1.957  -4.011 0.000455 ***
## am1            5.175      2.053   2.521 0.018176 *
## cyl6:am1      -3.733      3.095  -1.206 0.238553
## cyl8:am1      -4.825      3.095  -1.559 0.131069
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.032 on 26 degrees of freedom
## Multiple R-squared:  0.7877, Adjusted R-squared:  0.7469
## F-statistic: 19.29 on 5 and 26 DF,  p-value: 5.179e-08
##                   2.5 %     97.5 %
## (Intercept)  19.3014371 26.4985629
## cyl6         -8.5354513  0.9854513
## cyl8        -11.8733157 -3.8266843
## am1           0.9553109  9.3946891
## cyl6:am1    -10.0947539  2.6280873
## cyl8:am1    -11.1864206  1.5364206
```

Listing 3: Summary of the multiple linear regression model predicting the efficiency of the 32 cars as a function of the number of cylinders, the type of transmission, and their interaction.

```
##
## Call:
## lm(formula = mpg ~ wt * am, data = mycars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6004 -1.5446 -0.5325  0.9012  6.0909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
## wt           -3.7859     0.7856  -4.819 4.55e-05 ***
## am1          14.8784     4.2640   3.489  0.00162 **
## wt:am1       -5.2984     1.4447  -3.667  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 28 degrees of freedom
## Multiple R-squared:  0.833, Adjusted R-squared:  0.8151
## F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
##                 2.5 %    97.5 %
## (Intercept) 25.229642 37.602469
## wt          -5.395234 -2.176581
## am1          6.143928 23.612917
## wt:am1      -8.257693 -2.339028
```

Listing 4: Summary of the multiple linear regression model predicting the efficiency of the 32 cars as a function of transmission type, weight, and their interaction.

```
##
## Call:
## lm(formula = mpg ~ wt * hp, data = mycars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0632 -1.6491 -0.7362  1.4211  4.5513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
## wt          -8.21662    1.26971  -6.471 5.20e-07 ***
## hp          -0.12010    0.02470  -4.863 4.04e-05 ***
## wt:hp        0.02785    0.00742   3.753 0.000811 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.153 on 28 degrees of freedom
## Multiple R-squared:  0.8848, Adjusted R-squared:  0.8724
## F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
##                    2.5 %      97.5 %
## (Intercept)  42.42359654 57.19325031
## wt          -10.81750352 -5.61574508
## hp           -0.17069436 -0.06950982
## wt:hp         0.01264983  0.04304647
```

Listing 5: Summary of the multiple linear regression model predicting the efficiency of the 32 cars as a function of horsepower, weight, and their interaction.

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 44.8517 3.725e-09 ***
## am         1  36.77   36.77  3.9988   0.05608 .
## cyl:am     2  25.44   12.72  1.3832   0.26861
## Residuals 26 239.06    9.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Listing 6: ANOVA table for the model including the number of cylinders, transmission type, and their interaction.

```
##      df      AIC
## lm2   7 169.1633
## lm2a  5 168.3989
## lm2b  4 170.5640
## lm2c  3 196.4844
```

Listing 7: AIC values for comparing the models including the interaction between `cyl` and `am` (lm2), the main effects of both `cyl` and `am` (lm2a), the main effect of `cyl` alone (lm2b), and the main effect of `am` alone (lm2c).

```
##      df      BIC
## lm2   7 179.4234
## lm2a  5 175.7276
## lm2b  4 176.4269
## lm2c  3 200.8816
```

Listing 8: BIC values for comparing the models including the interaction between `cyl` and `am` (lm2), the main effects of both `cyl` and `am` (lm2a), the main effect of `cyl` alone (lm2b), and the main effect of `am` alone (lm2c).

```
## Start:  AIC=67.24
## mpg ~ (cyl + hp + wt + am)^2
##
##          Df Sum of Sq    RSS    AIC
## - cyl:wt  2    0.6926 103.16 63.457
## - cyl:am  2    1.8176 104.28 63.804
## - cyl:hp  2    5.8623 108.33 65.022
## - hp:wt   1    0.0052 102.47 65.243
## - hp:am   1    3.4193 105.89 66.292
## - wt:am   1    6.1169 108.58 67.097
## <none>                102.47 67.241
##
## Step:  AIC=63.46
## mpg ~ cyl + hp + wt + am + cyl:hp + cyl:am + hp:wt + hp:am +
##     wt:am
##
##          Df Sum of Sq    RSS    AIC
## - cyl:am  2    1.3945 104.55 59.886
## - hp:wt   1    0.0841 103.24 61.483
## - hp:am   1    3.6818 106.84 62.579
## - cyl:hp  2   13.1830 116.34 63.305
## <none>                103.16 63.457
## - wt:am   1    9.9355 113.09 64.399
##
## ...
##
## Step:  AIC=57.08
## mpg ~ cyl + hp + wt + am + cyl:hp + wt:am
##
##          Df Sum of Sq    RSS    AIC
## <none>                108.53 57.082
## - cyl:hp  2    21.939 130.47 58.973
## - wt:am   1    19.049 127.58 60.257
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am + cyl:hp + wt:am, data = mycars)
##
## Coefficients:
## (Intercept)         cyl6         cyl8           hp           wt          am1
##    36.65881     -7.19711    -10.82118     -0.08268     -2.31293      9.14282
##     cyl6:hp      cyl8:hp       wt:am1
##     0.05954      0.07634     -3.04685
```

Listing 9: Truncated output from stepwise selection starting with the model with all two way interactions. Full output can be obtained by running the code in R.

```
##        df      AIC
## Model1  7 169.0836
## Model2  5 157.4760
## Model3  7 169.1633
```

Listing 10: Comparison of models with AIC.

```
## # A tibble: 32 x 4
##     mpg .fitted .lower .upper
##   <dbl>   <dbl>  <dbl>  <dbl>
## 1  21      21.2   18.4   24.0
## 2  21      21.2   18.4   24.0
## 3  22.8    27.4   25.3   29.4
## 4  21.4    18.6   16.0   21.3
## 5  18.7    14.7   13.0   16.5
## # i 27 more rows
## # A tibble: 32 x 4
##     mpg .fitted .lower .upper
##   <dbl>   <dbl>  <dbl>  <dbl>
## 1  21      21.2  14.3    28.1
## 2  21      21.2  14.3    28.1
## 3  22.8    27.4  20.7    34.0
## 4  21.4    18.6  11.8    25.5
## 5  18.7    14.7   8.21   21.3
## # i 27 more rows
```

Listing 11: Fitted (top) and predicted values (bottom) for the first 5 observations in the cars data computed from the model including the main effects of `cyl` and `am`.