Supplement: Statistical modeling

SSC 328M

1 Introduction

In statistics, we often want to describe how one variable (the *response* variable) depends on one or more other variables (the *predictor* variables). Consider the following example. The weight of a person will likely depend on a number of factors, such as the person's height (taller people will weigh more than shorter people, all else equal), the person's sex (men tend to be heavier than women at the same height), the person's age (people tend to gain weight as they age), and the person's activity level (we expect people who are physically active to weigh less, on average). So in this case, weight is the response variable, and height, sex, age, and activity level are predictor variables. We would apply a statistical modeling approach to answer questions such as the following:

1. Which of the predictor variables show a significant relationship with the response variable? We could easily imagine a predictor variable that should not show a significant relationship, such as the color of the person's car in our weight example, and we would expect the statistical approach to be able to differentiate between predictors that are significant and those that are not.

2. How exactly does the response change with changing predictor variables? For example, for every inch that a man is taller, how much more does he weigh?

3. How much variability in the response variable does each predictor variable explain? For example, if we know a woman's height, does that pretty much tell us her weight, or are there so many other factors that influence weight, including ones that may be unknown, that knowing a woman's height does not tell us much about her weight, even though taller women are heavier on average.

2 Defining a linear model

In the above example, we could write the relationship between predictor and response variables as follows:

```
weight ~ height + sex + age + exercise
```
(1)

(This is the way this relationship would be expressed in R). The tilde symbol (`~`) should be read as "is modeled as". So what this formula says is that weight is modeled as the sum of height, sex, age, and exercise status. Of course we don't mean to say that weight is exactly the sum of these variables, that doesn't make any sense. How would we even add height and sex? That's why we are using a tilde symbol and not an equals sign. What formula (1) is meant to say is that weight increases or decreases in proportion to each of the predictors, and that all the predictors influence weight independently of each other. (Independence is expressed by the fact that we are using the plus sign to combine the predictors.) We can turn formula (1) into an exact equation by writing

$$\text{weight} = a + (b_1 \times \text{height}) + (b_2 \times \text{sex}) + (b_3 \times \text{age}) + (b_4 \times \text{exercise}) + \epsilon. \tag{2}$$

In this equation, the variables b_1, b_2, b_3, b_4 are called regression coefficients. They describe how the response variable changes with changing predictor variables. For example, we might find $b_1 =$ 8lbs/in, which would mean that a person's weight increases, on average, by 8lbs for every inch that they are taller. Categorical variables, such as sex, can be encoded as zeros and ones (we could say a zero stands for women and a one for men), in which case the value of b_2 would tell us the number of lbs a men would be heavier than a women at the same height, and so on.

The variable a is called the intercept. It tells us the value of the response variable if all the predictor variables are zero. In this particular example, the weight of a person who is zero inches tall and eats zero calories a day may seem like a weird concept. However, in general, we need an intercept variable to construct a reasonable statistical model.

The variable ϵ measures any unidentified variation in the data. This could be simply random variation (some people may be heavier than others for no apparent reason) or systematic variation due to unidentified causes. An example of the latter could be genetic background. The genetic background of people most likely affects their weight, but our model does not explicitly account for genetic background, so variation in genetic background gets absorbed in the error term.

The variables a, b_1, b_2, b_3, b_4, and ϵ are obtained by *fitting* the model to the data. What this means is that we find a set of values for a, b_1, b_2, b_3, b_4 such that the error (ϵ) is made as small as possible. It is important to realize in this context that we need multiple observations (multiple people for which we know their weight, their height, their sex, and so on) to be able to create a meaningful fit. Hence, we don't have just one equation of the form (2), we actually have multiple such equations, one for each observational unit (person, in this case). A statistician would write these equations as follows:

$$y_i = a + b_1 x_{1,i} + b_2 x_{2,i} + b_3 x_{3,i} + b_4 x_{4,i} + \epsilon_i. \tag{3}$$

Compare this equation to (2). The y_i has taken the place of weight, $x_{1,i}$ has taken the place of height, $x_{2,i}$ the place of sex, and so on. Everything else has remained the same. What equation (3) says is that the weight of person i (that is, y_i) is given by the sum of the intercept (a) plus the product of b_1 and $x_{1,i}$ (that is, the height of the ith person times the weight increase per unit height) plus the product of b_2 and $x_{2,i}$ (that is the sex of the ith person times the weight increase for male sex) and so on, plus the error term for person i (given by ϵ_i). The quantity ϵ_i is also called the *residual*. It measures the difference between the model prediction and the actual value of the response variable for observation i.

Notice how the response variable y_i, the predictor variables $x_{1,i}$, $x_{2,i}$, $x_{3,i}$, $x_{4,i}$, and the residuals ϵ_i all depend on the person i. By contrast, the intercept a and the regression coefficients b_1, b_2, b_3, b_4 do not depend on the person i. This is because predictor and response variables correspond to individual observations, but intercept and regression coefficients capture the general relationship between these variables, i.e., intercept and regression coefficients are assumed to be the same for all observational units (people, in this case).

A statistical model that can be described by an equation such as (3) is called a *linear model*. Linear models are extremely powerful, and are used throughout modern statistics. All the specific statistical analyses discussed in this document (regression, analysis of variance, analysis of covariance) are special cases of linear models. If you know how to handle linear models, you can analyze almost any data set that may come your way.

We mentioned earlier that fitting the model to the data means finding a set of values for a, b_1, b_2, b_3, and b_4 such that the error is made as small as possible. We can now make this statement more precise. For each observational unit, the error (residual) is ϵ_i. We measure the overall error

in the model fit by summing the squares of the residuals:

$$\mathrm{SS}_{\mathrm{residuals}} = \sum_{i=1}^{n} \epsilon_i^2. \tag{4}$$

Here, "SS" stands for "sum squares", and n is the sample size (number of people in our data set). Thus, fitting the model to the data means minimizing the sum squares of the residuals $\mathrm{SS}_{\mathrm{residuals}}$.

In practice, calculating the exact values of the regression coefficients that minimize the sum squares of the residuals can be rather complicated. Fortunately, we don't have to worry about how exactly this calculation is carried out. R will do it for us. What we need to understand and remember is that the regression coefficients apply to all observations equally, and that they are chosen such that the residuals become as small as possible.

3 Hypothesis testing

We mentioned in the introduction that one of the questions we may ask in a statistical modeling context is which predictors show a significant relationship with the response variable. How do we know whether a predictor shows a significant relationship? Consider the example from the previous section. If height had no influence on weight, then we would expect the regression coefficient that converts height into weight (b_1) to be zero. If $b_1 = 0$, then no matter how tall people are, we would always expect them to have the same weight.

More formally, if we want to test the hypothesis that height is a significant predictor of weight, we would consider the following hypotheses:

H_0: weight is independent of height, $b_1 = 0$.
H_A: weight depends on height, $b_1 \neq 0$.

If we reject H_0, then height is a significant predictor of weight, and it needs to be included in our model. By contrast, if we fail to reject H_0, then height makes no difference to our model and we might just as well have left it out of the model.

Let's now consider the hypothesis that a person's sex makes a difference for that person's weight. In this case, the hypotheses would be:

H_0: weight is independent of sex, $b_2 = 0$.
H_A: weight depends on sex, $b_2 \neq 0$.

As you can see, now we're simply asking whether a different regression coefficient is significantly different from zero. In general, hypothesis testing in a modeling framework boils down to testing whether specific regression coefficients are significantly different from zero or not.

4 One-factor analysis of variance (ANOVA)

It is a common occurrence that we want to determine the influence of a categorical predictor variable on a quantitative response. For example, if we are only interested in the weight differences as explained by sex, then we have one categorical predictor (sex) and one quantitative response (weight). Earlier in the course, we would have used a t-test to analyze this scenario, but now we want to consider it in the context of a linear model.

Statisticians refer to categorical variables also as "factors". A linear model with one categorical predictor is called *one-factor analysis of variance* or just *analysis of variance* (in short, ANOVA).

Let's consider a concrete example. The data file "weights.csv" contains a fictitious data set containing weight (in kg), height (in cm), age (in years), sex (male or female), exercise status (light, moderate, or heavy), and hair color (blonde, brown, black, or red) of 300 people:

```
> weights<-read.csv("weights.csv")
> head(weights)
  weight height  sex age exercise hair.color
1  81.63  176.9 male  37 moderate      brown
2  78.98  172.4 male  38    light      brown
3  78.14  176.2 male  31 moderate     blonde
4  84.66  179.8 male  39    light      brown
5  81.42  181.6 male  32    heavy      brown
6  76.86  170.8 male  36    light      brown
```

Women seem, on average, to be lighter than men in this data set (Figure 1). A t test confirms this hypothesis:

```
> t.test(weights$weight[weights$sex=='female'],weights$weight[weights$sex=='male'])

        Welch Two Sample t-test

data:  weights$weight[weights$sex == "female"] and weights$weight[weights$sex == "male"]
t = -14.7452, df = 286.86, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
```

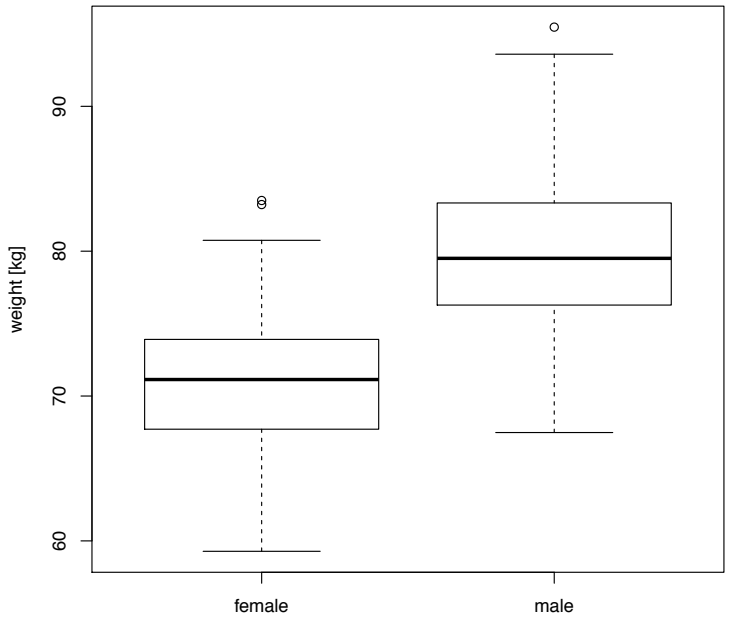

Figure 1: Boxplots of weight for female and male subjects in the example data set.

```
95 percent confidence interval:
 -10.169326  -7.774141
sample estimates:
mean of x mean of y
 70.79880  79.77053
```

We can test the same hypothesis with a linear model as follows:

```
> fit<-lm(weight~sex, data=weights)
> anova(fit)
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq F value    Pr(>F)
sex         1 6036.9  6036.9  217.42 < 2.2e-16 ***
Residuals 298 8274.2    27.8
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The `lm()` function fits the model to the data; in this case, the model simply tries to explain weight on the basis of sex, `weight~sex`. The `data=weights` argument to `lm()` tells the fitting function which data set to use. The `lm()` function produces a fitted model, which we store in the variable `fit`. To analyze the model, we use the function `anova()`, which displays the results in an ANOVA table.

The ANOVA table contains one row for each predictor variable (here, only sex) plus a row for the residuals. Each row lists the number of degrees of freedom for that variable (`Df`), the sum squares (`Sum Sq`), the mean squares (`Mean Sq`, the sum squares divided by the number of degrees of freedom), the F value (mean squares for the variable divided by the mean squares of the residuals), and the associated P value (`Pr(>F)`). In the case of sex, the number of degrees of freedom is 1 (number of groups $-$ 1), the sum squares are 6036.9, the mean squares are $6036.9/1 = 6036.9$, and $F = 6036.9/27.8 = 217.42$. The associated P value is extremely small ($< 2.2 \times 10^{-16}$), so we can reject the null hypothesis and we conclude that weight differs significantly for men and women in this data set. Thus, our modeling approach gave us the same result as the t test shown previously.

However, the modeling approach gives us more than the t test, because it also tells us how much of the variation in weight is explained by sex. The amount of variation explained is the sum squared associated with a variable divided by the sum square total. Here, we find:

$$R^2 = 6036.9/(6036.9 + 8274.2) = 42\%. \tag{5}$$

Thus, 42% of the variation in weight is explained by sex, and 58% of the variation in weight is residual (explained by factors other than weight or not explained at all).

While the t test can only compare a response variable among two groups, ANOVA can handle a larger number of groups. For example, in the same data set, the exercise status groups people into three groups, people that engage only in light exercise, people that engage in moderate exercise, and people that engage in heavy exercise. The ANOVA null hypothesis is that weight is independent of exercise status, i.e., all three groups have the same weight on average.

Let's test this hypothesis on our example data set:

```
> fit<-lm(weight~exercise, data=weights)
> anova(fit)
```

```
Analysis of Variance Table

Response: weight
           Df  Sum Sq Mean Sq F value Pr(>F)
exercise    2   180.4  90.221  1.8963 0.1519
Residuals 297 14130.7  47.578
```

According to this test, exercise status does not have a significant effect on weight ($P = 0.1519$) and explains very little variation in the data ($R^2 = 180.4/14130.7 = 1.3\%$).

Note that the ANOVA looks at all groups at the same time. So if reject the null hypothesis, we only know that some of the groups are different than others. If we want to know which group specifically is different from which other group, we have to look at pairs of groups one by one. One option is to do pairwise t tests with correction for multiple testing.

5 Multi-factor ANOVA

In the preceding section, we found that weight differed significantly by sex but not by exercise status. Since sex explained over 40% of the variation in weight, it is possible that this variation drowned out the small difference in weight caused by exercise status. If we analyzed the data separately by sex (i.e., do one analysis only for men and one only for women), we might find that exercise status does have a significant effect in these subgroups.

In statistical modeling, we handle those kinds of situations by incorporating multiple predictors into the model. For example, we can carry out an ANOVA that considers both sex and exercise status as predictor variables:

```
> fit<-lm(weight~sex+exercise, data=weights)
> anova(fit)
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq  F value  Pr(>F)
sex         1 6036.9  6036.9 222.4104 < 2e-16 ***
exercise    2  239.9   119.9   4.4187 0.01285 *
Residuals 296 8034.3    27.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

As we can see, now the ANOVA table has one row for sex and one row for exercise, each with its associated P value. And we also see that now exercise makes a significant contribution to the model. We say that exercise is significant *when controlling for sex*.

The model with sex and exercise assumes that those two factors are independent, i.e., that whatever weight reduction exercise causes is the same in men and women. This need not be the case, however. We could imagine that exercise might have a bigger effect in women than in men, for example because men who exercise a lot build more muscle and hence weigh more, relatively speaking, than men who don't exercise that much.

If the effect of one factor in a model depends on another factor, we speak of an *interaction*. The R function `interaction.plot()` allows us to visualize the data in a way that will highlight possible interactions. Figure 2 shows the output of `interaction.plot()` on this data set. As we can see, there is little difference in mean weight between people who exercise lightly or moderately,

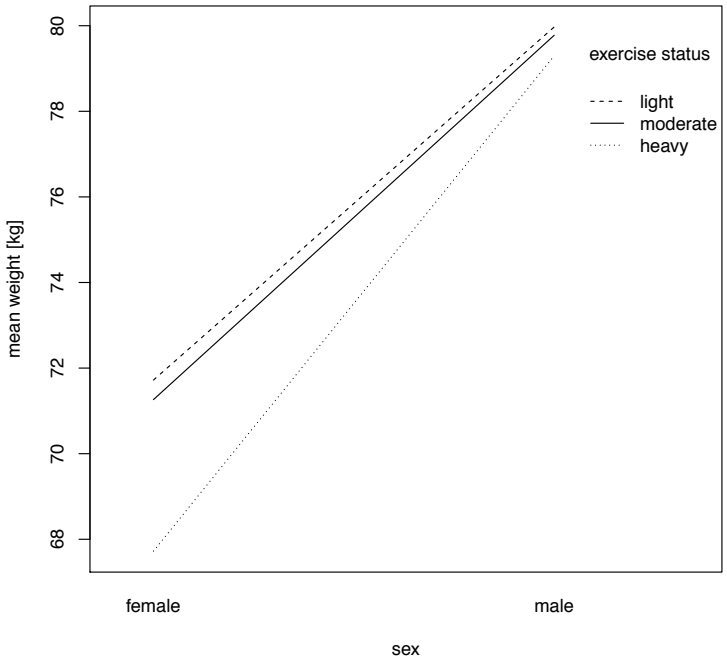

Figure 2: Interaction plot of weight explained by sex and exercise status. The y-axis shows mean weight in subgroups of the data set. The six subgroups considered here are men and women with either light, moderate, or heavy exercise status. Lines connect men and women with the same exercise status. Non-parallel lines suggest that there is an interaction in the data set.

but there is a much bigger effect for heavy exercise. However, women seem to benefit much more than men. The mean reduction in weight is approximately 3kg in women who exercise heavily compared to women who exercise lightly or moderately, but in men the comparable reduction in weight is less than 1kg.

The modeling framework allows us to test whether the interaction is significant or not. All we have to do is replace `sex+exercise` by `sex*exercise` in the model formula. The `*` sign indicates that sex and exercise could interact, whereas the `+` implies that interactions are not taken into account.

This is the result we obtain from an ANOVA with interaction:

```
> fit<-lm(weight~sex*exercise, data=weights)
> anova(fit)
Analysis of Variance Table

Response: weight
             Df Sum Sq Mean Sq  F value  Pr(>F)
sex           1 6036.9  6036.9 224.5655 < 2e-16 ***
exercise      2  239.9   119.9   4.4615 0.01234 *
sex:exercise  2  130.9    65.4   2.4340 0.08945 .
Residuals   294 7903.5    26.9
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Now the ANOVA table contains an additional row, labeled `sex:exercise`. This row represents the possible interaction between sex and exercise. As we can see, the associated P value ($P = 0.089$) is small but not < 0.05. We can conclude that while there seems to be a weak tendency for sex and exercise to interact, this interaction is not statistically significant at the 5% level. If we used $\alpha = 0.1$, however, we could reject the null hypothesis of no interaction. In this case, we could conclude that sex and exercise do interact significantly.

Important: If we find a significant interaction, we do not test for the significance of the individual factors. In other words, whether exercise alone is significant or not does not matter if the interaction between sex and exercise is significant. Interaction terms trump individual terms.

6 Simple linear regression

So far, the predictor variables we have considered were categorical. However, our data set also has quantitative predictors: height and age. The statistical modeling approach can handle both categorical and quantitative predictors. When we model a quantitative response in terms of one (or more) quantitative predictors, we say that we are doing a *linear regression analysis*.

Let's consider the simplest case, where weight is explained by height. This linear regression model can be fit in R like so:

```
> fit<-lm(weight~height, data=weights)
> summary(fit)

Call:
lm(formula = weight ~ height, data = weights)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4147  -2.4183   0.1097   2.6083  14.2361

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.82779    4.06815  -10.28   <2e-16 ***
height        0.69097    0.02397   28.82   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.561 on 298 degrees of freedom
Multiple R-squared: 0.736,      Adjusted R-squared: 0.7351
F-statistic: 830.8 on 1 and 298 DF,  p-value: < 2.2e-16
```

Note that the first line of R code again calls the `lm()` function with the appropriate model, which in this case is `weight~height`. However, in the second line we now use the `summary()` function instead of the `anova()` function. The difference between the `summary()` function and the `anova()` function is that they display different information about the fitted model. Both functions can be used on all linear models, but the output of the `anova()` function is not as useful in the regression context and the output of the `summary()` function is not as useful in the ANOVA context.

The `anova()` function shows us the sum squares, F values, and P values of each factor. The `summary()` function, on the other hand, shows us the individual regression coefficients with standard error and associated P value.

Let's go over the output of the `summary()` function in the above example. The most interesting section is the section entitled `Coefficients:`. There are two rows, one for `(Intercept)` and one for `height`. The `(Intercept)` row tells us about the intercept in the regression model. [We had previously introduced the intercept as the variable a in Eq. (3) on page 2.] The `height` row tells us about the regression coefficient for the predictor variable `height`.

Each row in `Coefficients:` section contains four numbers, titled `Estimate` (this is the coefficient as estimated by the linear model), `Std. Error` (this is the standard error of the coefficient estimate), `t value` (this is a t value, as in the standard t test), and `Pr(>|t|)` (this is the P value, testing the null hypothesis that the coefficient in question is equal to zero). Thus, in our example, the intercept is -41.8 ± 4.1kg. In other words, the regression model estimates that a person of zero height weighs -42kg. This statement is a meaningless extrapolation of the data set, but we note that the intercept is highly significant and hence necessary to build a meaningful linear model.

The coefficient for `height` is more interesting. R states the estimate as 0.69 ± 0.024. What this number means is that for every additional cm in height, people in the data set are on average 0.69kg heavier. The P value is extremely small, so this result is highly significant.

At the bottom of the output from the `summary()` function, we see that R prints two R^2 values, an F-statistic, and another P value. The R^2 values tell us the overall amount of variation in the data explained by the model, considering all the predictor variables that were used. The first of the two, `Multiple R-squared`, corresponds to the R^2 we have discussed above, and is calculated by dividing the sum of the sum squares of all predictor variable by the total sum squares of the model. The second, `Adjusted R-squared`, is calculated using a slightly modified formula. It takes into account that each predictor variable will capture some of the variance in the data by random chance. A detailed discussion of the adjusted R^2 is beyond the scope of this document.

The F-statistic and P value at the bottom of the `summary()` output test the overall hypothesis that the variance in the response variable explained by the model is not zero. This hypothesis test is usually less interesting than the tests for individual non-zero regression coefficients.

7 Multiple linear regression

As in the case of ANOVA, we can build linear regression models with multiple predictor variables. For example, we can regress weight against both height and age. We do this by simply adding `age` to the regression formula:

```
> fit<-lm(weight~height+age, data=weights)
> summary(fit)

Call:
lm(formula = weight ~ height + age, data = weights)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0328 -1.4978  0.0168  1.7410  7.8350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -90.00037    4.47016  -20.13   <2e-16 ***
height        0.90896    0.02334   38.95   <2e-16 ***
age           0.25000    0.01679   14.89   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.699 on 297 degrees of freedom
Multiple R-squared: 0.8488,     Adjusted R-squared: 0.8478
F-statistic: 833.7 on 2 and 297 DF,  p-value: < 2.2e-16
```

Note that now we have a row corresponding to `age` in the `Coefficients:` section. That row tells us about the increase in weight with increasing age. Also note that the estimates for the intercept and the coefficient for height have changed. The overall R^2 is now 85%, and it was only 74% in the previous model. Thus, the new model captures more variation in the data, and hence the regression coefficients should be considered to be more accurate. Hence, from this analysis we can conclude that weight increases, on average, by 0.91kg for every cm in height and by 0.25kg for every year in age.

8 Analysis of co-variance (ANCOVA)

So far, we have discussed ANOVA, where we have a quantitative response and categorical predictors, and linear regression, where we have both a quantitative response and quantitative predictors. Often times, however, we want to consider a categorical and a quantitative predictor at the same time. Such an analysis is called analysis of covariance (ANCOVA).

For example, when we first looked at the weight data set, we found that weight differed by sex, and men were on average almost 10kg heavier than women (Figure 1). However, men in this data set are also taller than women (not shown). Therefore, we have to ask whether the increase in weight for men is caused by their increase in height, or whether men are heavier than women *even when controlling for height*. In other words, we want to know whether a woman and a man of equal height are expected to have comparable weight.

We perform an ANCOVA by simply adding both the quantitative and the categorical predictor into the regression model, like so:

```
> fit<-lm(weight~height+sex, data=weights)
> summary(fit)

Call:
lm(formula = weight ~ height + sex, data = weights)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4057  -2.4104   0.0954   2.5924  14.2285

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.39275    5.97427  -6.929 2.64e-11 ***
height        0.68822    0.03660  18.801  < 2e-16 ***
sexmale       0.06251    0.62781   0.100    0.921
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.566 on 297 degrees of freedom
Multiple R-squared: 0.736,      Adjusted R-squared: 0.7342
F-statistic:   414 on 2 and 297 DF,  p-value: < 2.2e-16
```

The additional coefficients row, labeled `sexmale`, tells us how much heavier men than women are at equal height. As we can see, this is a rather small number (0.06kg), and it is not significant. Thus, this model would suggest to us that men and women of equal height have equal weight. However, what this analysis ignores is that age is also a significant predictor. In this data set, men are taller but women are older (not shown). Therefore, if we control only for height, we find that men and women have comparable weight, but these men and women are not of comparable age. If we want to compare men and women of comparable height and age, we have to include both predictor variables in our model:

```
> fit<-lm(weight~height+age+sex, data=weights)
> summary(fit)

Call:
lm(formula = weight ~ height + age + sex, data = weights)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1960  0.0562  0.2956  0.9679  1.1910

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.58942    2.52842  -27.13   <2e-16 ***
height        0.69186    0.01488   46.49   <2e-16 ***
age           0.48569    0.01254   38.74   <2e-16 ***
sexmale       9.60279    0.35469   27.07   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.45 on 296 degrees of freedom
Multiple R-squared: 0.9565,     Adjusted R-squared: 0.9561
F-statistic:  2170 on 3 and 296 DF,  p-value: < 2.2e-16
```

We see that now men are again almost 10kg heavier than women, if we compare men and women of similar height and age. Also, the overall R^2 is now 97%. Thus, this model fits much better than any model we previously considered.

9 Linear models with multiple categorical and quantiative predictors

As we have seen in the previous sections, linear models can handle both categorical and quantitative predictors, in arbitrary combinations. In general, one would want to add all predictor variables that

could possibly explain some variability in the response, and then use the linear model framework to test which predictors are significant and which are not.

In our weight example, the data set contains information about height, age, sex, exercise status, and hair color. We fit a model considering all these predictors with the following R command:

```
> fit<-lm(weight~height+age+sex+exercise+hair.color, data=weights)
```

Let us first run the `anova()` function on this fitted model:

```
> anova(fit)
Analysis of Variance Table

Response: weight
            Df  Sum Sq Mean Sq    F value Pr(>F)
height       1 10533.2 10533.2 15543.8322 <2e-16 ***
age          1  1614.3  1614.3  2382.2269 <2e-16 ***
sex          1  1541.2  1541.2  2274.4087 <2e-16 ***
exercise     2   424.5   212.2   313.2039 <2e-16 ***
hair.color   3     0.7     0.2     0.3497 0.7894
Residuals  291   197.2     0.7
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We see that height, age, sex, and exercise status all have a significant relationship with weight, whereas hair color does not.

To see the individual regression coefficients, we have to use the `summary()` function:

```
> summary(fit)

Call:
lm(formula = weight ~ height + age + sex + exercise + hair.color,
    data = weights)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7610 -0.4481  0.2958  0.4159  1.7701

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -72.933913   1.450465 -50.283   <2e-16 ***
height             0.701833   0.008493  82.636   <2e-16 ***
age                0.492389   0.007210  68.295   <2e-16 ***
sexmale            9.661172   0.203926  47.376   <2e-16 ***
exerciselight      2.935997   0.123909  23.695   <2e-16 ***
exercisemoderate   2.905738   0.139709  20.798   <2e-16 ***
hair.colorblonde   0.024580   0.135769   0.181    0.856
hair.colorbrown    0.039352   0.123463   0.319    0.750
hair.colorred     -0.160334   0.209804  -0.764    0.445
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 0.8232 on 291 degrees of freedom
Multiple R-squared: 0.9862,     Adjusted R-squared: 0.9858
F-statistic:  2603 on 8 and 291 DF,  p-value: < 2.2e-16
```

We can summarize the results from this regression model as follows: on average, weight increases by 0.7kg for every cm of height and by 0.5kg for every year of age. Men are on average 9.7kg heavier than women, if all else is equal. (All else means all predictor variables considered, i.e., height, age, exercise status, and hair color.) Light and moderate exercise both yield approximately 2.9kg in extra weight, on average, compared to heavy exercise. None of the hair colors make any difference. Overall, the model explains nearly all (99%) of the variation in the data.

10 Assumptions of linear models

All linear models make the following assumptions:

- Residuals are normally distributed.
- The variance in the data is similar across all predictor values.
- Predictors are independent of each other (unless we explicitly model interactions, as discussed in Section 5).
- For quantitative predictors, the response depends linearly on the predictors.

All these assumptions should be checked when building a linear model. Let's do this check for a simple ANCOVA model, where we model weight as a function of age and sex.

To test whether residuals are normally distributed, we do a q-q plot of the residuals:

```
> fit<-lm(weight~age+sex,data=weights)
> qqnorm(fit$residuals)
```

The resulting graph is shown in Figure 3, left. The points mostly fall onto a straight line, and hence the assumption of normally distributed residuals is satisfied in this case.

To test whether variance is uniform, we can plot the residuals against the fitted values:

```
> plot(fit$fitted.values,fit$residuals)
```

The resulting graph is shown in Figure 3, right. We see a uniform band of points, spread around zero. This outcome indicates that the variance is uniform. Non-uniform variance would display as a systematic trend in the data, for example a fanning out such that residuals tend to be small for small fitted values and large for large fitted values.

Both non-normality and non-uniformity of residuals can often be ameliorated by transforming the predictor and/or the response variables. Transformations in regression models work just as they do for other statistical tests; for example, we can take the logarithm of a predictor or response variable and see if the transformed data set agrees better with the modeling assumptions.

To assess whether predictors are independent of each other and whether the response variable depends linearly on quantiative predictors, we plot the response against various predictors and look at the data. For example, we can plot weight against age separately for men and women:

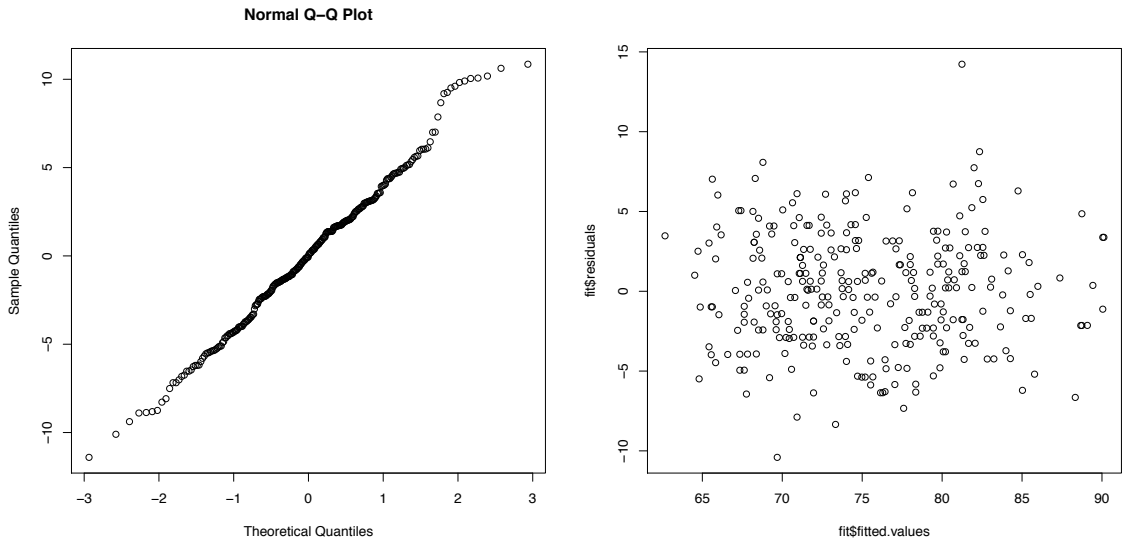

Figure 3: Diagnostic plots of a fitted model. Left: q-q plot of residuals. Right: residuals plotted against fitted values.

```
> men<-weights[weights$sex=='male',]
> women<-weights[weights$sex=='female',]
> plot(men$age,men$weight, xlim=c(20,69), ylim=c(55,95), xlab='age [years]',
ylab='weight [kg]', pch=19)
> points(women$age,women$weight)
```

The resulting plot is shown in Figure 4. We see that weight changes linearly with age, for both men and women. Hence, the assumption of linearity is satisfied. Furthermore, the relationship between weight and age seems similar for men and women (the slopes are comparable), but the weights for men are simply shifted relative to the weights for women (at the same age, men are, on average, about 18kg heavier). Therefore, the assumption of independence is satisfied as well. If age and sex were not independent, then we might see a different slope for men and women.

If the assumptions of linearity or independence are not satisfied, then one can develop more sophisticated models that allow for these issues. However, such models are beyond the scope of this document.

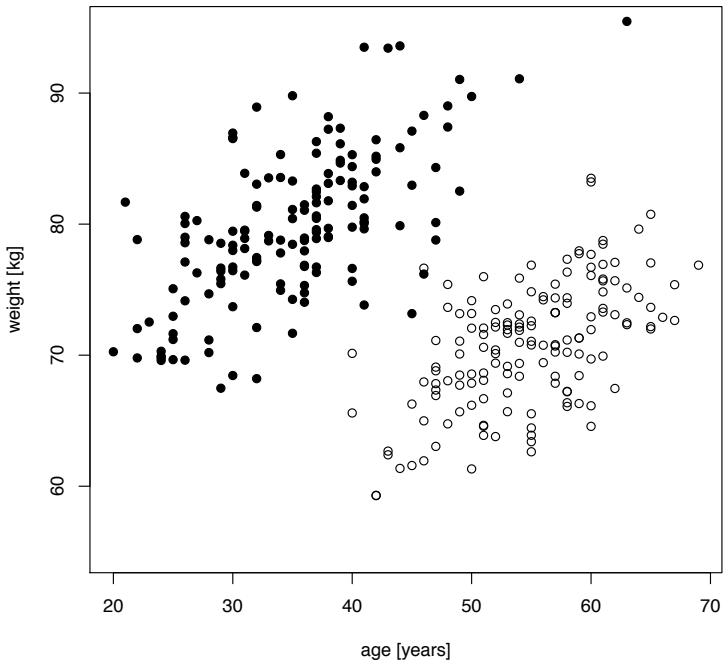

Figure 4: Weight plotted as a function of height for men (filled symbols) and women (open symbols). Weight varies linearly with age for both groups. Moreover, age and sex seem to be independent predictors, because male weight is shifted relative to female weight but shows a comparable linear relationship with comparable slope.