IEHC0046 BASIC STATISTICS FOR MEDICAL SCIENCES

Linear Regression: Practical III

11 November, 2020

In this practical session, we will learn how to use interaction terms in a regression model. We will be using the dataset you have seen before in previous sessions. We will also learn how to fit a non-linear relationship in the regression models, as well as learn some methods of data transformation to achieve linearity and normality assumptions. We’ll use the tidyverse, summarytools, multcomp, car, and describedata packages. You will need to install the describedata package.

# install.packages("describedata") # Uncomment to install  
library(describedata)  
library(car)  
library(multcomp)  
library(tidyverse)  
library(summarytools)  
load("elsa.Rdata")

The three sections to this practical are as follows:

1. Linear regression with interaction terms
2. Linear regression with non-linear relationships
3. Data transformation

# 1. Linear regression with interaction terms

## 1.1. Interaction between sex and age

The term interaction is used to describe situations in which the relationship between and differs according to the level of one or more of the other variables. We will run a multiple regression with sbp as the outcome, and age and sex as the predictor variables. We know from the lecture that the relationship between sbp and age differs by sex, so we will test for an interaction term between sex and age.

First, run the multiple regression without an interaction.

mod\_1 <- lm(sbp ~ age + sex, elsa)  
summary(mod\_1)

##   
## Call:  
## lm(formula = sbp ~ age + sex, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53.901 -12.861 -1.667 10.368 87.778   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 99.62065 2.27674 43.76 <2e-16 \*\*\*  
## age 0.69246 0.03685 18.79 <2e-16 \*\*\*  
## sexfemale -1.71432 0.71722 -2.39 0.0169 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.47 on 2689 degrees of freedom  
## (437 observations deleted due to missingness)  
## Multiple R-squared: 0.1173, Adjusted R-squared: 0.1166   
## F-statistic: 178.6 on 2 and 2689 DF, p-value: < 2.2e-16

To add an interaction between sex and age, we have two options:

lm(sbp ~ age\*sex, elsa)  
lm(sbp ~ age + sex + age:sex, elsa)

Both are equivalent (run to confirm this). age\*sex includes the main effects (age and sex) and the interaction term. age + sex + age:sex programs this explicitly.

mod\_2 <- lm(sbp ~ age\*sex, elsa)  
summary(mod\_2)

##   
## Call:  
## lm(formula = sbp ~ age \* sex, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53.09 -12.88 -1.87 10.65 89.47   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 114.12324 3.40772 33.490 < 2e-16 \*\*\*  
## age 0.45092 0.05606 8.044 1.29e-15 \*\*\*  
## sexfemale -27.09959 4.51549 -6.001 2.22e-09 \*\*\*  
## age:sexfemale 0.42166 0.07406 5.693 1.38e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.36 on 2688 degrees of freedom  
## (437 observations deleted due to missingness)  
## Multiple R-squared: 0.1278, Adjusted R-squared: 0.1268   
## F-statistic: 131.3 on 3 and 2688 DF, p-value: < 2.2e-16

confint(mod\_2)

## 2.5 % 97.5 %  
## (Intercept) 107.4412273 120.8052615  
## age 0.3410066 0.5608383  
## sexfemale -35.9537634 -18.2454077  
## age:sexfemale 0.2764362 0.5668912

**Q: Can you write down the equation for the predicted sbp?**

where is a dummy variable = 0 if male and = 1 if female.

**Q: Can you interpret the results? What are the sbp intercepts (i.e. when age = 0) and sbp slopes (the increase in sbp for each year increase in age) for men and women?**

Men is the reference group (coded as 0 in female), so the sbp intercept for men is 114.12 and its 95% CI is from 107.44 to 120.81. The intercept of sbp for women is 87.02 (calculated as 114.12 - 27.10 x 1).

The interaction term is 0.42 (95% CI 0.28 to 0.57). Its p value is very low (< 0.001), so we have enough statistical evidence that the relationship between sbp and age differs by sex. In other words, the slopes of sbp are different for men and women. In men, the increase in sbp for each year increase in age (i.e., the slope) is 0.45 (calculated as 0.45 x 1) and its 95% CI is from 0.34 to 0.56). In women, the slope is 0.87 (calculated as 0.45 + 0.42).

To get the 95% CI for the women’s sbp intercept and age slope, use glht().

glht(mod\_2, "`(Intercept)` + `sexfemale` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age \* sex, data = elsa)  
##   
## Quantile = 1.9608  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## `(Intercept)` + sexfemale == 0 87.0237 81.2144 92.8329

glht(mod\_2, "`age` + `age:sexfemale` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age \* sex, data = elsa)  
##   
## Quantile = 1.9608  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## age + `age:sexfemale` == 0 0.8726 0.7777 0.9675

**Q: What is the systolic blood pressure for women aged 60?**

glht(mod\_2,  
 "`(Intercept)` + `sexfemale` +  
 60\*`age` + 60\*`age:sexfemale` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age \* sex, data = elsa)  
##   
## Quantile = 1.9608  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate  
## `(Intercept)` + sexfemale + 60 \* age + 60 \* `age:sexfemale` == 0 139.3788  
## lwr   
## `(Intercept)` + sexfemale + 60 \* age + 60 \* `age:sexfemale` == 0 138.4509  
## upr   
## `(Intercept)` + sexfemale + 60 \* age + 60 \* `age:sexfemale` == 0 140.3068

## 1.2. Interaction between sex and wealth

Next, we will run a multiple regression again with sbp as the outcome, and age, female and quintile of wealth (wealth5) as the predictor variables, and also an interaction term between female and wealth5. (Note that this model is just an example of how to interpret the results of the interaction term, but maybe it is too simple for real research.) We’ll need to convert wealth5 to a factor so R interprets it as a categorical variable.

elsa <- elsa %>%  
 mutate(wealth5 = factor(wealth5))

We can add an interaction between two factor variables using the same syntax as above.

mod\_3 <- lm(sbp ~ age + sex\*wealth5, elsa)  
summary(mod\_3)

##   
## Call:  
## lm(formula = sbp ~ age + sex \* wealth5, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.50 -12.92 -1.66 10.27 86.69   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 103.94033 2.61674 39.721 <2e-16 \*\*\*  
## age 0.66293 0.03779 17.543 <2e-16 \*\*\*  
## sexfemale -3.77539 1.78737 -2.112 0.0348 \*   
## wealth52 -2.28066 1.92400 -1.185 0.2360   
## wealth53 -3.18107 1.83954 -1.729 0.0839 .   
## wealth54 -3.71459 1.76088 -2.110 0.0350 \*   
## wealth55 -2.63963 1.70147 -1.551 0.1209   
## sexfemale:wealth52 5.11727 2.49619 2.050 0.0405 \*   
## sexfemale:wealth53 3.51886 2.43930 1.443 0.1493   
## sexfemale:wealth54 2.46276 2.36375 1.042 0.2976   
## sexfemale:wealth55 -0.87389 2.29291 -0.381 0.7031   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.45 on 2639 degrees of freedom  
## (479 observations deleted due to missingness)  
## Multiple R-squared: 0.1227, Adjusted R-squared: 0.1194   
## F-statistic: 36.92 on 10 and 2639 DF, p-value: < 2.2e-16

**Q: Can you interpret the results?**

In this model, males in the lowest wealth group (group 1) are the references categories. The intercept is 103.94 (95%CI 98.81 to 109.07), which is predicted sbp for males aged 0 in the lowest wealth group. The intercept of sbp for women in the lowest wealth group is 103.94 + (-3.78)= 100.16. The slope of sbp for men and women is 0.66.

The interaction terms between female and wealth are 5.12, 3.52, 2.46, and -0.87, respectively. There is some statistical evidence that there is an interaction between the 2nd wealth group and female (p=0.040). For other groups, there is limited evidence to reject the null hypothesis that there is no interaction between wealth groups and sex.

The (non-interaction) coefficients of wealth groups are all negative, suggesting that, for men, those in higher wealth groups had lower sbp levels than men in the lowest wealth group, although only the differences between the 4th wealth group and the lowest wealth group reached conventional criteria of p<0.05.

**Q: What is the predicted systolic blood pressure of women aged 55 and in the 2nd wealth group? What about women aged 55 and in the lowest (1st) wealth group? What about men aged 60 and in the highest (5th) wealth group?**

glht(mod\_3, "`(Intercept)` + `sexfemale` + `wealth52` +  
 `sexfemale:wealth52` + 55\*`age` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age + sex \* wealth5, data = elsa)  
##   
## Quantile = 1.9609  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate  
## `(Intercept)` + sexfemale + wealth52 + `sexfemale:wealth52` + 55 \* age == 0 139.4625  
## lwr   
## `(Intercept)` + sexfemale + wealth52 + `sexfemale:wealth52` + 55 \* age == 0 137.2440  
## upr   
## `(Intercept)` + sexfemale + wealth52 + `sexfemale:wealth52` + 55 \* age == 0 141.6809

glht(mod\_3, "`(Intercept)` + `sexfemale` + 55\*`age` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age + sex \* wealth5, data = elsa)  
##   
## Quantile = 1.9609  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## `(Intercept)` + sexfemale + 55 \* age == 0 136.6258 134.3140 138.9377

glht(mod\_3, "`(Intercept)` + `wealth55` + 60\*`age` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age + sex \* wealth5, data = elsa)  
##   
## Quantile = 1.9609  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## `(Intercept)` + wealth55 + 60 \* age == 0 141.0762 139.0439 143.1085

**Q: Use the linearHypothesis() function to test whether including the interaction term between female and wealth5 improves model fit significantly?**

lincoms <- paste0("sexfemale:wealth5", 2:5, " = 0")  
linearHypothesis(mod\_3, lincoms)

## Linear hypothesis test  
##   
## Hypothesis:  
## sexfemale:wealth52 = 0  
## sexfemale:wealth53 = 0  
## sexfemale:wealth54 = 0  
## sexfemale:wealth55 = 0  
##   
## Model 1: restricted model  
## Model 2: sbp ~ age + sex \* wealth5  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2643 901694   
## 2 2639 898515 4 3179 2.3342 0.05349 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Results show that there is some statistical evidence (p=0.054) against the null hypothesis that all of the interaction coefficients are 0. So we may want to keep the interaction in the model.

# 2. Linear regression with non-linear relationships

## 2.1. Summary statistics

We want to run a multiple regression with total cholesterol (chol) as the outcome, and age, female, and bmi as the predictor variables. First, we want to examine the summary statistics for chol and bmi these two new variables. (We checked the summary statistics for age and female in the last session.)

elsa %>%  
 select(chol, bmi) %>%  
 descr()

## Descriptive Statistics   
## elsa$Total cholesterol in mmol/l   
## Label: Total cholesterol in mmol/l   
## N: 3129   
##   
## bmi chol  
## ----------------- --------- ---------  
## Mean 27.53 5.98  
## Std.Dev 4.49 1.10  
## Min 14.81 2.50  
## Q1 24.48 5.20  
## Median 27.03 5.90  
## Q3 29.95 6.60  
## Max 52.47 12.00  
## MAD 4.01 1.04  
## IQR 5.47 1.40  
## CV 0.16 0.18  
## Skewness 0.83 0.51  
## SE.Skewness 0.05 0.05  
## Kurtosis 1.43 1.00  
## N.Valid 2930.00 2475.00  
## Pct.Valid 93.64 79.10

\*\*Q. Show the histogram and kernel density plot of the outcome variable (chol). Is chol normally distributed?

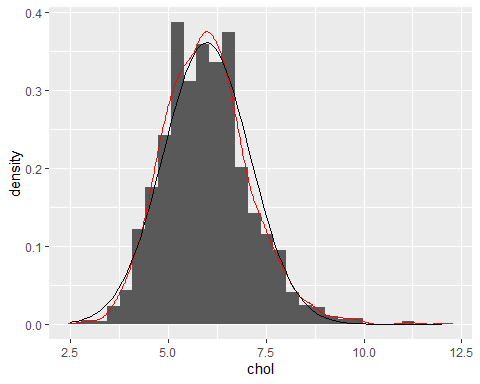
We can plot the histogram and kernel density simultaneously thanks to ggplot’s *layered* approach to plot construction.

ggplot(elsa) +  
 aes(x = chol) +  
 geom\_histogram(aes(y = ..density..)) +  
 geom\_density(color = "red") +  
 stat\_function(fun = dnorm, args = list(mean = mean(elsa$chol, na.rm = TRUE),  
 sd = sd(elsa$chol, na.rm = TRUE)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 654 rows containing non-finite values (stat\_bin).

## Warning: Removed 654 rows containing non-finite values (stat\_density).

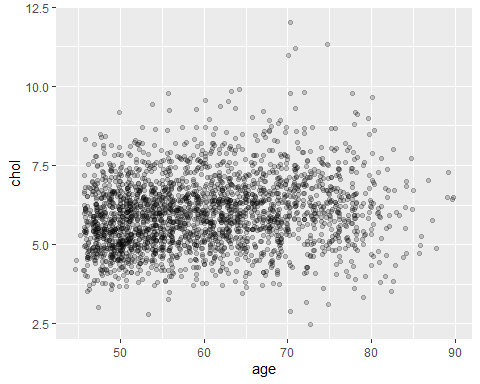


chol is almost normally distributed.

Next, we will be will create scatter plots to check the linearity assumption between chol and age and between chol and bmi.

ggplot(elsa) +  
 aes(x = age, y = chol) +  
 geom\_jitter(alpha = 0.2)

## Warning: Removed 654 rows containing missing values (geom\_point).

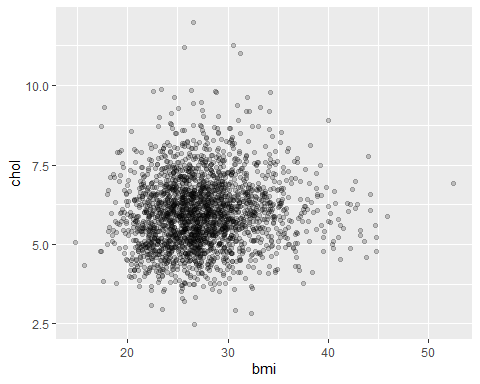


**Q: Is there evidence of a linear relationship between chol and age?**

From the scatter plot, higher values of age in general correspond with slightly higher values of total cholesterol, so there is a positive linear relationship suggested, although this is quite weak.

ggplot(elsa) +  
 aes(x = bmi, y = chol) +  
 geom\_jitter(alpha = 0.2)

## Warning: Removed 786 rows containing missing values (geom\_point).



**Q: What can you tell from the sctter plot? Is there evidence of a non-linear relationship between chol and bmi?**

There is some evidence of a non-linear relationship between chol and bmi. As BMI increases, total cholesterol at first increases; then as BMI continues to increase, total cholesterol decreases.

## 2.2. Specifying non-linear relationships

It is often found that the relationship between the outcome variable and an independent variable is non-linear. We will try two possible ways of incorporating such an independent variable in the multiple regression.

First, let us run a multiple regression with chol as the outcome, and age, female, and bmi as the predictor variables. This is will be the basic model to be compared with.

mod\_4 <- lm(chol ~ age + sex + bmi, elsa)  
summary(mod\_4)

##   
## Call:  
## lm(formula = chol ~ age + sex + bmi, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5217 -0.7138 -0.0636 0.6678 5.6679   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.999395 0.193550 20.663 < 2e-16 \*\*\*  
## age 0.019195 0.002295 8.364 < 2e-16 \*\*\*  
## sexfemale 0.413181 0.044378 9.310 < 2e-16 \*\*\*  
## bmi 0.021681 0.004979 4.354 1.39e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.068 on 2339 degrees of freedom  
## (786 observations deleted due to missingness)  
## Multiple R-squared: 0.0715, Adjusted R-squared: 0.0703   
## F-statistic: 60.03 on 3 and 2339 DF, p-value: < 2.2e-16

One possibility of handling non-linear relationship is to find an algebraic description of the relationship. For example, it may be quadratic, in which case both the variable (bmi) and its square would be included in the model.

We could do this by creating a new variable bmi\_sq equal to bmi^2 and adding this in the lm() functional call. An issue with this is that R won’t know that bmi\_sq is a function of bmi. Instead we can use the I() syntax to make this explicit. The advantage of this is that some postestimation commands will know bmi and I(bmi^2) are related. We can put an number of functions in I() includes functions of two or more variables.

mod\_5 <- lm(chol ~ age + sex + bmi + I(bmi^2), elsa)  
summary(mod\_5)

##   
## Call:  
## lm(formula = chol ~ age + sex + bmi + I(bmi^2), data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5770 -0.7158 -0.0628 0.6509 5.6244   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.9671689 0.5943821 3.310 0.000948 \*\*\*  
## age 0.0185041 0.0022969 8.056 1.24e-15 \*\*\*  
## sexfemale 0.4383695 0.0448092 9.783 < 2e-16 \*\*\*  
## bmi 0.1665896 0.0403914 4.124 3.85e-05 \*\*\*  
## I(bmi^2) -0.0024807 0.0006862 -3.615 0.000307 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.065 on 2338 degrees of freedom  
## (786 observations deleted due to missingness)  
## Multiple R-squared: 0.07666, Adjusted R-squared: 0.07508   
## F-statistic: 48.53 on 4 and 2338 DF, p-value: < 2.2e-16

**Q: Can you interpret the results of bmi and I(bmi^2)?**

The p-values of bmi and I(bmi^2) are very small (p<0.001) and suggests low compatibility with the null hypothesis that the coefficients are zero. The coefficient on bmi is greater than zero > 0. The coefficient on I(bmi^2) is less than zero. This means that total cholesterol is increasing at first, but will eventually turn around and be decreasing (an inverted-U).

**Q: By comparing the models with and without squared BMI, what do their R-squared and Root MSE tell us?**

in the current model is higher than the model without BMI squared. However, we need caution in interpreting R2 to justify the utility of a variable added into the model, as R2 increases even when the new variables have no real predictive capability. So we need to look at the values of Root MSE.

sqrt(mean(mod\_4$residuals^2))

## [1] 1.067043

sqrt(mean(mod\_5$residuals^2))

## [1] 1.064073

Lower values of Root MSE indicate better fit. Root MSE is also lower than before. So we can conclude that the model with BMI squared is a better fitting model.

The second possibility is to transform BMI into a categorical variable. For most purposes, a subdivision into 3–5 groups - depending on the sample size - is adequate to investigate non-linearity of the relationship. There is a categorical BMI variable already available in your data (bmi4).

freq(elsa$bmi4)

## Frequencies   
## elsa$bmi4   
## Label: BMI grouped: <20,20-25,25-30,30+   
## Type: Factor   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ---------------- ------ --------- -------------- --------- --------------  
## Under 20 63 2.15 2.15 2.01 2.01  
## Over 20-25 814 27.78 29.93 26.01 28.03  
## Over 25-30 1336 45.60 75.53 42.70 70.73  
## Over 30 717 24.47 100.00 22.91 93.64  
## <NA> 199 6.36 100.00  
## Total 3129 100.00 100.00 100.00 100.00

The ‘under 20’ category is very small. In this is practical, we will combine the ‘under 20’ and the ‘over 20-25’ categories. (There are pros and cons to doing this and it is important to think about what this might mean for your research.)

elsa <- elsa %>%  
 mutate(bmi3 = fct\_recode(bmi4, "Under 25" = "Under 20",  
 "Under 25" = "Over 20-25"))  
table(elsa$bmi3, elsa$bmi4, useNA = "ifany")

##   
## Under 20 Over 20-25 Over 25-30 Over 30 <NA>  
## Under 25 63 814 0 0 0  
## Over 25-30 0 0 1336 0 0  
## Over 30 0 0 0 717 0  
## <NA> 0 0 0 0 199

Let’s run the model using bmi3.

mod\_6 <- lm(chol ~ age + sex + bmi3, elsa)  
summary(mod\_6)

##   
## Call:  
## lm(formula = chol ~ age + sex + bmi3, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5824 -0.7176 -0.0722 0.6513 5.6196   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.464198 0.143303 31.152 < 2e-16 \*\*\*  
## age 0.018824 0.002297 8.194 4.11e-16 \*\*\*  
## sexfemale 0.421212 0.044746 9.413 < 2e-16 \*\*\*  
## bmi3Over 25-30 0.177368 0.051986 3.412 0.000656 \*\*\*  
## bmi3Over 30 0.279310 0.060644 4.606 4.33e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.067 on 2338 degrees of freedom  
## (786 observations deleted due to missingness)  
## Multiple R-squared: 0.07298, Adjusted R-squared: 0.07139   
## F-statistic: 46.01 on 4 and 2338 DF, p-value: < 2.2e-16

**Q: Can you interpret the results of bmi3?**

The coefficient for 25-30 BMI group is 0.18 meaning that people in the 25-30 BMI group have 0.18 mmol/l higher total cholesterol than people in the <25 BMI group. The test statistic for this coefficient is t = 3.41 and the p value = 0.001, so we are more confident that we should reject the null hypothesis that there is no difference between the 25-30 BMI group and the <25 BMI group. People with a BMI between 25 and 30 have 0.28 mmol/l higher total cholesterol than people in the <25 BMI group. Its p value is very low (<0.001) so there is enough statistical evidence to reject the null hypothesis that there is no difference between the 30+ BMI group and the <25 BMI group.

Notice how this model caputres the non-linear relationship: the move from BMI < 25 to 25-30 is association with 0.18 higher SBP, while the move from 25-30 to 30+ is associated with 0.10 higher SBP (0.28 - 0.18). We can use glht() to see if the difference is statistically significant.

glht(mod\_6, "`bmi3Over 30` - `bmi3Over 25-30` = 0") %>%  
 summary()

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Fit: lm(formula = chol ~ age + sex + bmi3, data = elsa)  
##   
## Linear Hypotheses:  
## Estimate Std. Error t value Pr(>|t|)   
## `bmi3Over 30` - `bmi3Over 25-30` == 0 0.10194 0.05614 1.816 0.0695 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## (Adjusted p values reported -- single-step method)

The p-value is low (0.0695) so there is some evidence of a difference between the two highest BMI groups on SBP levels (after adjusting for ager and sex).

**Q: By comparing this bmi3 model with the BMI-squared mode and the basic model (bmi), what do their R-squared and Root MSE tell us?**

sqrt(mean(mod\_4$residuals^2))

## [1] 1.067043

sqrt(mean(mod\_5$residuals^2))

## [1] 1.064073

sqrt(mean(mod\_6$residuals^2))

## [1] 1.066191

in the bmi3 model is lower (0.0714) than that in the bmisq model (0.0774), and Root MSE in the bmi3 model is slightly higher (1.066) than that in the BMI-squared model (1.064). These suggest that the BMI-squared is a slightly better model than the bmi3 model. But the bmi3 model is still better than the basic model, as its Root MSE (1.066) is lower than the basic model (1.067).

# 3. Data transformation

The assumption of normality will not always be satisfied by a particular set of data. Next, we see how such problems can often be overcome by transforming the data to a different scale of measurement. It is recommended to transform the dependent variable first and then if necessary you can transform the independent variables.

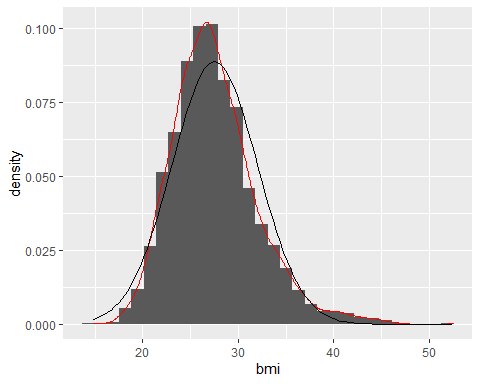
**Q: Check the histogram and kernel density of BMI**

ggplot(elsa) +  
 aes(x = bmi) +  
 geom\_histogram(aes(y = ..density..)) +  
 geom\_density(color = "red") +  
 stat\_function(fun = dnorm,   
 args = list(mean = mean(elsa$bmi, na.rm = TRUE),  
 sd = sd(elsa$bmi, na.rm = TRUE)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 199 rows containing non-finite values (stat\_bin).

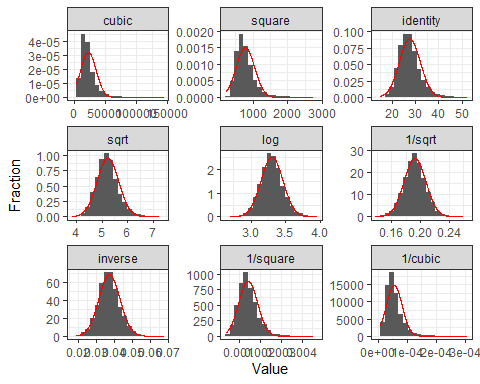
## Warning: Removed 199 rows containing non-finite values (stat\_density).



BMI is slightly positively skewed, but not so much.

We can use the gladder() function from describedata to explore possible transformations.

gladder(elsa$bmi)



To look at test statistics for which transformation is best, use ladder() from describedata.

ladder(elsa$bmi)

## # A tibble: 9 x 3  
## Transformation statistic p.value  
## <fct> <dbl> <dbl>  
## 1 cubic 0.822 3.32e-49  
## 2 square 0.907 4.19e-39  
## 3 identity 0.966 8.23e-26  
## 4 sqrt 0.984 1.82e-17  
## 5 log 0.995 4.15e- 8  
## 6 1/sqrt 0.998 6.71e- 3  
## 7 inverse 0.994 1.83e- 9  
## 8 1/square 0.964 1.83e-26  
## 9 1/cubic 0.907 4.87e-39

Smaller (chi-2) statistics indicates better transformation (alternatively, higher p-values). So it suggests that 1/sqrt is a better transformation.