lab2 Answers

#### ***Diala Ezzeddine***

#### ***January 22, 2019***

**Objective: Learn how to apply Multiple Linear Regression algorithm.**

For a health insurance company to make money, it needs to collect more in yearly premiums than it spends on medical care to its beneficiaries.

As a result, insurers invest a great deal of time and money in developing models that accurately **forecast medical expenses** for the insured population.

Medical expenses are difficult to estimate because the costliest conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese.

The goal of this analysis is to use patient data to estimate the average medical care expenses for such population segments. These estimates can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs.

Collecting data

For this analysis, we will use a simulated dataset containing hypothetical medical expenses for patients in the United States. This data was created using demographic statistics from the US Census Bureau, and approximately reflect real-world conditions.

The data includes **1,338 examples** of beneficiaries currently enrolled in the insurance plan, with features indicating characteristics of the patient as well as the total medical expenses charged to the plan for the calendar year.

There are 6 features:

* **age**: An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
* **sex**: The policy holder’s gender, either male or female.
* **bmi**: The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
* **children**: An integer indicating the number of children/dependents covered by the insurance plan.
* **smoker**: A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
* **region**: The beneficiary’s place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.

It is important to give some thought to how these variables may be related to billed medical expenses. For instance, we might expect that older people and smokers are at higher risk of large medical expenses.

Unlike many other machine learning methods, in regression analysis, the relationships among the features are typically specified by the user rather than being detected automatically.

Let’s explore some of these potential relationships.

Exploring and preparing the data

**1. How to load the data for analysis? Do we need to use stringsAsFactors to load the data properly? why?**

insurance <- read.csv("C:/Users/diala/OneDrive/Documents/R\_labs/Data/insurance.csv",   
 stringsAsFactors = TRUE)

We need to use stringsAsFactors = TRUE to load the data properly because the three nominal variables: sex, smokers and region are factors. So, it is appropriate to convert it to factors from the begining.

**2. Show the structure of the data.**

str(insurance)

## 'data.frame': 1338 obs. of 7 variables:  
## $ age : int 19 18 28 33 32 31 46 37 37 60 ...  
## $ sex : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...  
## $ bmi : num 27.9 33.8 33 22.7 28.9 ...  
## $ children: int 0 1 3 0 0 0 1 3 2 0 ...  
## $ smoker : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...  
## $ region : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...  
## $ expenses: num 16885 1726 4449 21984 3867 ...

**3. What feature is the dependent feature and which ones are the independent features? What is the type of each feature in the data?**

The model’s dependent variable is expenses, which measures the medical costs each person charged to the insurance plan for the year.The dependent feature is numeric.

There are 6 independent features used to study the dependent feature. Those independent features are age, sex, bmi, children, smoker and region. Three independent features are numeric: age, bmi and children. The three remaining are factors.

Prior to building a regression model, it is often helpful to check for normality. Although linear regression does not strictly require a normally distributed dependent variable, the model often fits better when this is true.

**4. How to study the dependent feature and why using this method?**

# summarize the charges variable  
summary(insurance$expenses)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1122 4740 9382 13270 16640 63770

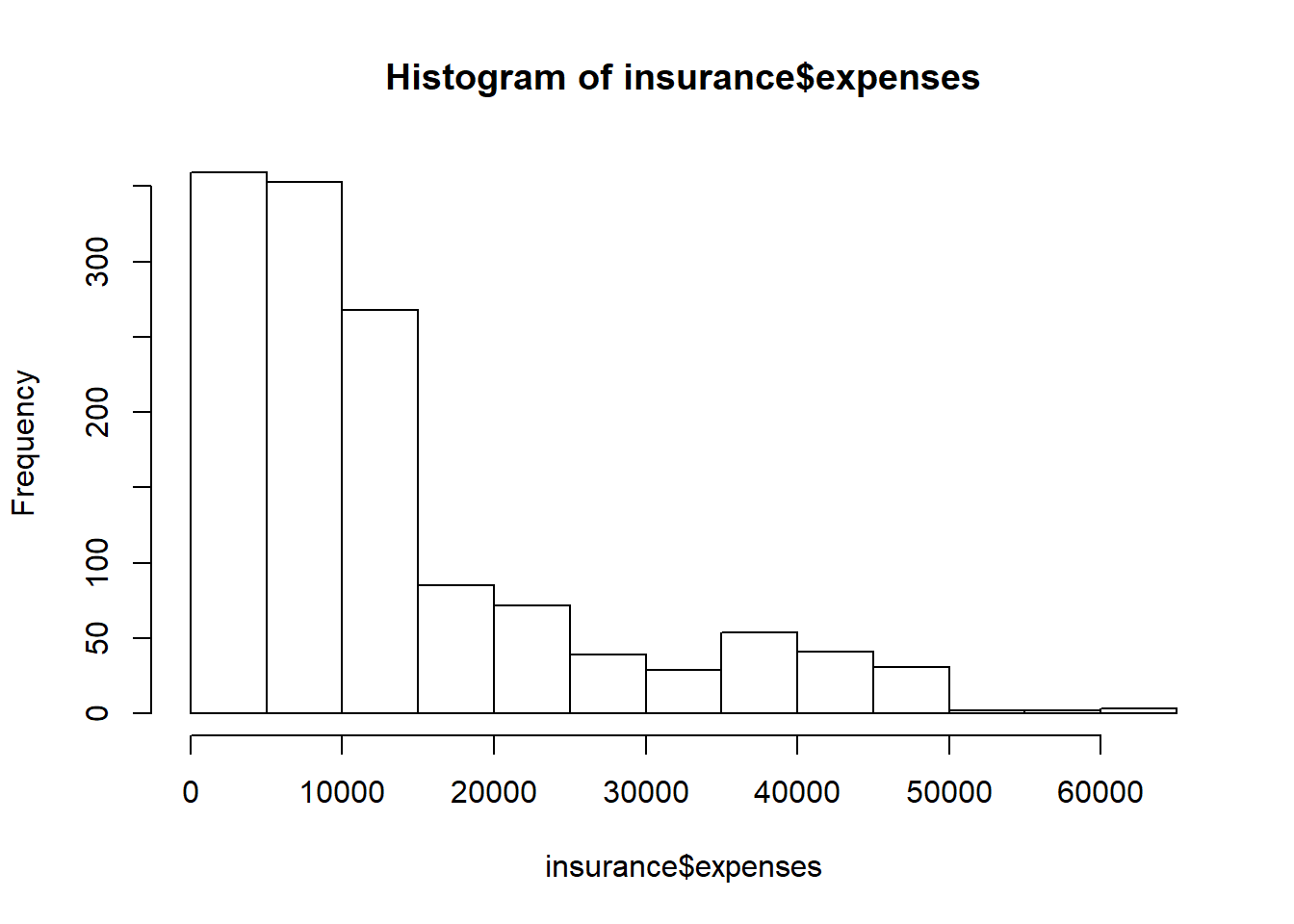
We use summary() because expenses is a numeric feature.

**5. Give two methods that describe if the dependent variable is normally distributed?Explain the distribution of the dependent feature.**

The first way is looking into the results from summary(). As the mean value is greater than the median, this implies that the distribution of insurance expenses is skewed-right.

The second way is by looking to the histogram visual.

hist(insurance$expenses)



We can conclude that insurance expenses is not normally distributed but skewed-right.

The majority of people in the data have yearly medical expenses between zero and $15,000, and those expenses can be more than $50,000 in few cases.

Let’s investigate the independent features.

**6. Which features are factors in the data? Look into the distribution of each one. Explain the results.**

The three factor features are sex, smoker and region.

To study those nominal features, we can use the table() function and the mode.

# table of sex  
table(insurance$age)

##   
## 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42   
## 69 68 29 28 28 28 28 28 28 28 28 27 27 27 26 26 26 25 25 25 25 25 27 27 27   
## 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64   
## 27 27 29 29 29 29 28 29 29 29 28 28 26 26 26 25 25 23 23 23 23 22

# table of smoker  
table(insurance$smoker)

##   
## no yes   
## 1064 274

# table of region  
table(insurance$region)

##   
## northeast northwest southeast southwest   
## 324 325 364 325

Here, we see that the data has been divided nearly evenly among four geographic regions. Most of the insured people are coming from southeast. There is more non smokers in the data than smokers and the data is divided almost evenly among the different ages from 20 years to 64 years old except that this data has more people of age 18 and 19.

Exploring relationships among features - correlation matrix

Before fitting a regression model to data, it can be useful to determine how the independent variables are related to the dependent variable and each other.

A correlation matrix provides a quick overview of these relationships.Given a set of variables, it provides a correlation for each pairwise relationship.

**7. Create a correlation matrix for the four numeric variables in the insurance data frame, use the cor() command.**

# exploring relationships among numeric features and the dependent feature: correlation matrix  
cor(insurance[c("age", "bmi", "children", "expenses")])

## age bmi children expenses  
## age 1.0000000 0.1092719 0.04246900 0.29900819  
## bmi 0.1092719 1.0000000 0.01275890 0.19834097  
## children 0.0424690 0.0127589 1.00000000 0.06799823  
## expenses 0.2990082 0.1983410 0.06799823 1.00000000

At the intersection of each row and column pair, the correlation is listed for the variables indicated by that row and column.

The diagonal is always 1.0000000 since there is always a perfect correlation between a variable and itself. The values above and below the diagonal are identical since correlations are symmetrical: cor(x, y) = cor(y, x).

**8. Is there any interesting relationship between features? Explain those associations**

There is no strong correlations in the correlation matrix, but there are some interesting associations.

1. age and bmi appear to have a weak positive correlation, meaning that as someone ages, their body mass tends to increase.
2. There is also a moderate positive correlation between age and expenses, bmi and expenses, and children and expenses. These associations imply that as age, body mass, and number of children increase, the expected cost of insurance goes up.

Visualizing relationships among features - scatterplot matrix

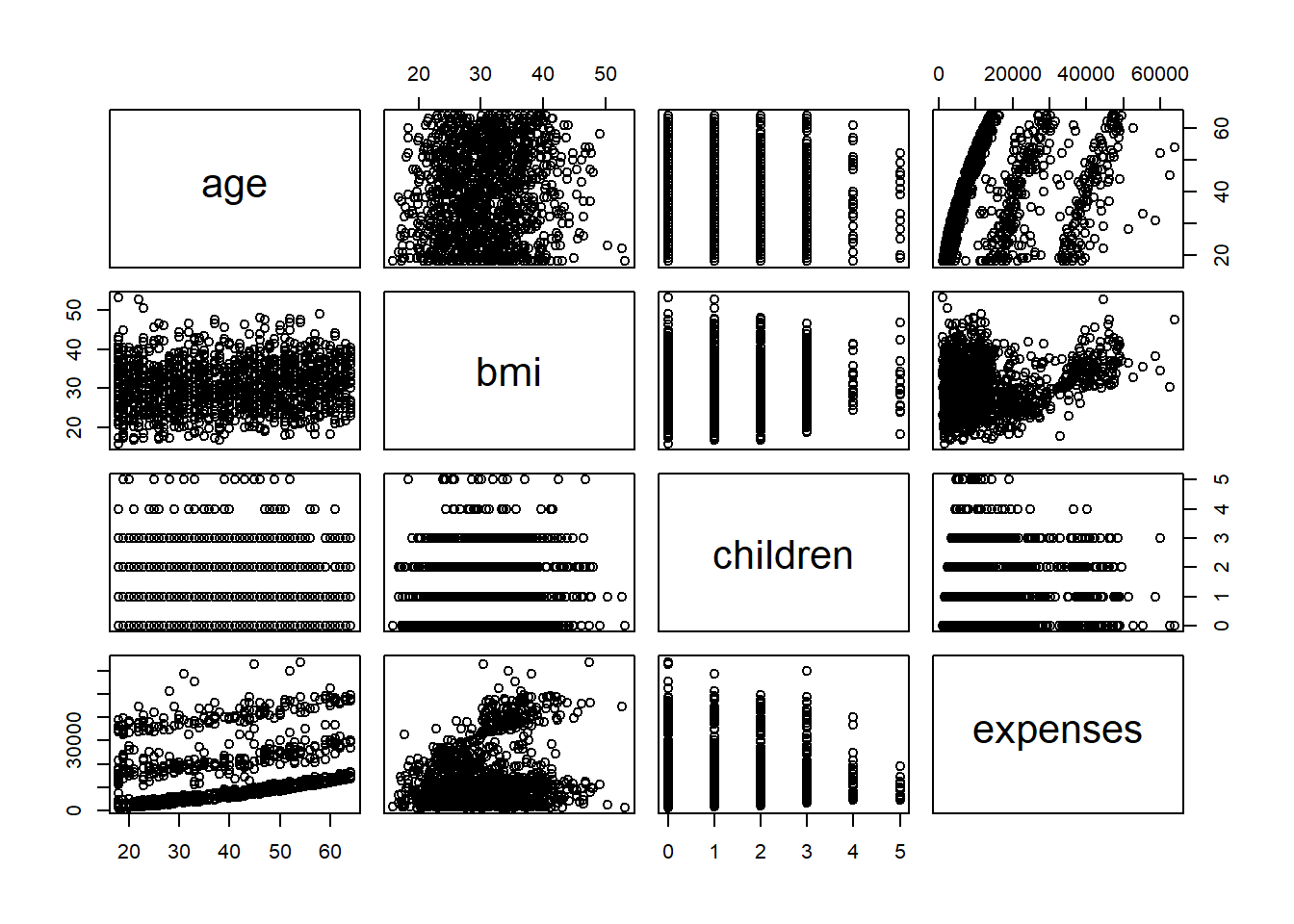
It can also be helpful to visualize the relationships among numeric features by using a scatterplot. Although we could create a scatterplot for each possible relationship, doing so for a large number of features might become hard.

An alternative is to create a scatterplot matrix, which is simply a collection of scatterplots arranged in a grid. It is used to detect patterns among three or more variables.

The scatterplot matrix is not a true multidimensional visualization because only two features are examined at a time. Still, it provides a general sense of how the data may be interrelated.

**9. Use the pairs() function to create a scatterplot matrix for the four numeric features: age, bmi, children, and expenses.**

# visualing relationships among features: scatterplot matrix  
pairs(insurance[c("age", "bmi", "children", "expenses")])



In the scatterplot matrix, the intersection of each row and column holds the scatterplot of the variables indicated by the row and column pair.

The diagrams above and below the diagonal are transpositions since the x axis and y axis have been swapped.

**10. Do you notice any patterns in these plots? Explain.**

Some patterns look like random clouds of points, a few seem to display some trends.

The relationship between age and expenses displays several relatively straight lines, while the bmi versus expenses plot has two distinct groups of points. It is difficult to detect trends in any of the other plots.

We can enhance the previous scatterplot by using the pairs.panels() function from the **psych** package. This produces a slightly more informative scatterplot matrix.

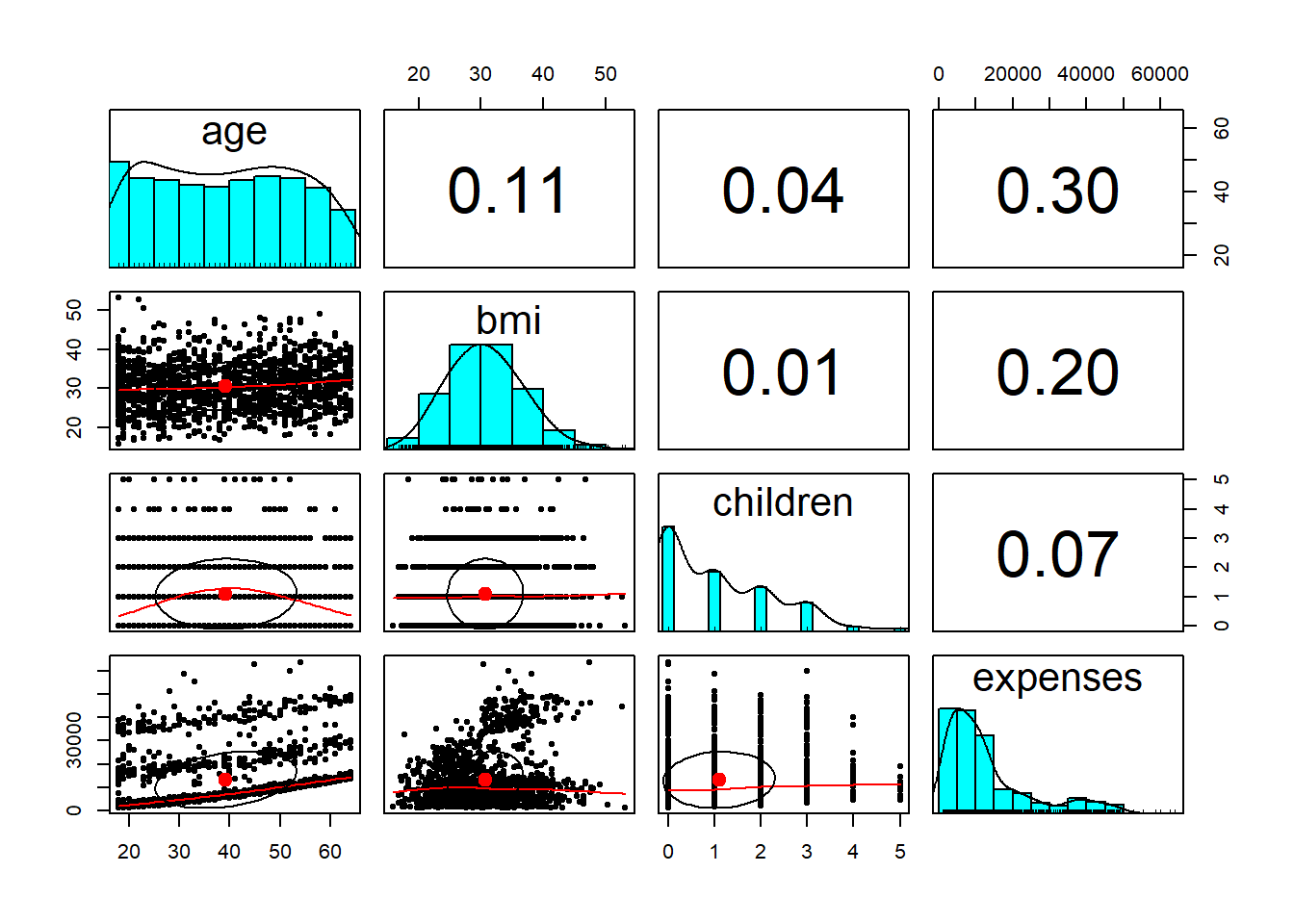
* Above the diagonal, the scatterplots have been replaced with a correlation matrix.
* On the diagonal, a histogram depicting the distribution of values for each feature is shown.
* The scatterplots below the diagonal are presented with additional visual information.

**11. To enhance the previous scatterplot matrix create the new one using the pairs.panels() function. Do you notice any new patterns in these new graphics?**

# more informative scatterplot matrix  
library(psych)

## Warning: package 'psych' was built under R version 3.5.1

pairs.panels(insurance[c("age", "bmi", "children", "expenses")])



We can notice a new pattern in these new graphics between age and bmi and between children and expenses.

To analysis this new graphic, you need to know:

* **correlation ellipse** The oval-shaped object on each scatterplot is a **correlation ellipse**. It provides a visualization of correlation strength. The dot at the center of the ellipse indicates the point at the mean values for the x and y axis variables. The correlation between the two variables is indicated by the shape of the ellipse; the more it is stretched, the stronger the correlation.
* **loess curve** The curve drawn on the scatterplot is called a **loess curve**. It indicates the general relationship between the x and y axis variables.

Training a model on the data

To fit a linear regression model to data with R, the **lm()** function can be used. This is included in the **stats package**, which should be included and loaded by default with your R installation.

**12. Use the lm() syntax provided in the chapter slides, to fit a linear regression model to this data( Use all the available features). Name it “ins\_model”.**

#Training a regression model   
  
ins\_model <- lm(expenses ~ age + children + bmi + sex + smoker + region,  
 data = insurance)  
ins\_model <- lm(expenses ~ ., data = insurance) # this is equivalent to above

**13. Show the estimated beta coefficients. Explain the results. Why there are 8 coefficients reported in addition to the intercept not 6?**

# see the estimated beta coefficients  
ins\_model

##   
## Call:  
## lm(formula = expenses ~ ., data = insurance)  
##   
## Coefficients:  
## (Intercept) age sexmale bmi   
## -11938.5 256.9 -131.3 339.2   
## children smokeryes regionnorthwest regionsoutheast   
## 475.5 23848.5 -353.0 -1035.0   
## regionsouthwest   
## -960.1

The intercept is the predicted value of expenses when the independent variables are equal to zero. As is the case here, quite often the intercept is of little value alone because it is impossible to have values of zero for all features. For example, since no person exists with age zero and BMI zero, the intercept has no real-world interpretation.

The beta coefficients indicate the estimated increase in expenses for an increase of one in each of the features, assuming all other values are held constant.

Although we only specified six features in our model formula, there are eight coefficients reported in addition to the intercept. This happened because the lm() function automatically applied a technique known as dummy coding to each of the factor-type variables we included in the model.

**14. What increase, or decrease will affect the medical expenses for each additional child? or additional year of age? or unit increase of BMI?**

For each additional year of age, we would expect $256.80 higher medical expenses on average.

Each additional child results in an average of $475.70 in additional medical expenses each year, and each unit increase in BMI is associated with an average increase of $339.30 in yearly medical expenses.

**15. Explain the effect of medical expenses in numbers on each one of the 3 factor features.**

Males have $131.40 less medical expenses each year relative to females and smokers cost an average of $23,847.50 more than non-smokers per year.

The coefficient for each of the three regions in the model is negative, which implies that the reference group, the northeast region, tends to have the highest average expenses.

We can summarize the final result (feature associations) we got after training the model as following:

The results of the linear regression model is: old age, smoking, and obesity tend to be linked to additional health issues, while additional family member dependents may result in an increase in physician visits and preventive care such as vaccinations and yearly physical exams.

Evaluating model performance

The parameter estimates we obtained previously tell us about how the independent variables are related to the dependent variable, but they tell us nothing about how well the model fits our data.

**16. To evaluate the model performance, use the summary () command on the stored model.**

# see more detail about the estimated beta coefficients  
summary(ins\_model)

##   
## Call:  
## lm(formula = expenses ~ ., data = insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11304.9 -2848.1 -982.1 1393.9 29992.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11938.5 987.8 -12.086 < 2e-16 \*\*\*  
## age 256.9 11.9 21.587 < 2e-16 \*\*\*  
## sexmale -131.3 332.9 -0.394 0.693348   
## bmi 339.2 28.6 11.860 < 2e-16 \*\*\*  
## children 475.5 137.8 3.451 0.000577 \*\*\*  
## smokeryes 23848.5 413.1 57.723 < 2e-16 \*\*\*  
## regionnorthwest -353.0 476.3 -0.741 0.458769   
## regionsoutheast -1035.0 478.7 -2.162 0.030782 \*   
## regionsouthwest -960.0 477.9 -2.009 0.044765 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6062 on 1329 degrees of freedom  
## Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494   
## F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16

The summary () output may seem confusing at first, but the basics are easy to pick up. The output provides three key ways to evaluate the performance, or fit, of our model:

* First, the **residuals** section provides summary statistics for the errors in our predictions. A residual is equal to the true value of the dependent variable minus the predicted value of it using the regression model.
* Second, the **p-value**, denoted Pr(>|t|), for each estimated regression coefficient. Small p-values suggest that the true coefficient is very unlikely to be zero, which means that the feature is extremely unlikely to have no relationship with the dependent variable.
* Third, the multiple **R-squared** value called the coefficient of determination, provides a measure of how well our model explains the values of the dependent variable. It is similar to the correlation coefficient, in that the closer the value is to 1.0, the better the model perfectly explains the data. The **adjusted R-squared** value corrects R-squared by penalizing models with a large number of independent variables. It is useful for comparing the performance of models with different numbers of explanatory variables.

**17. Given the preceding three performance indicators, How the model is performing? explain.**

1. Since a residual is equal to the true value minus the predicted value, the maximum error of 29981.7 suggests that the model under-predicted expenses by nearly $30,000 for at least one observation. On the other hand, 50 percent of errors fall within the 1Q and 3Q values, so the majority of predictions were between $2,850.90 over the true value and $1,383.90 under the true value.
2. Small p-value indicates features are statistically significant in predicting the dependent outcome. Our model has several highly significant variables, and they seem to be related to the outcome in logical ways.
3. Since the adjusted R-squared value is 0.7494, we know that the model explains nearly 75 percent of the variation in the dependent variable.

Improving model performance

A key difference between the regression modeling and other machine learning approaches is that regression typically leaves feature selection and model specification to the user.

Consequently, if we have subject matter knowledge about how a feature is related to the outcome, we can use this information to inform the model specification and potentially improve the model’s performance.

Model specification - adding non-linear relationships

In linear regression, the relationship between an independent variable and the dependent variable is assumed to be linear, yet this may not necessarily be true.

For example, the effect of age on medical expenditure may not be constant throughout all the age values; the treatment may become disproportionately expensive for oldest populations.

To account for a non-linear relationship, we can add a higher order term to the regression model, treating the model as a polynomial.

The difference between these two models is that an additional beta will be estimated, which is intended to capture the effect of the x-squared term. This allows the impact of age to be measured as a function of age squared. To add the non-linear age to the model, we simply need to create a new variable named **age2** to the data as follow:

#insurance$age2 <- insurance$age^2

Transformation - converting a numeric variable to a binary indicator

Suppose we have a hunch that the effect of a feature is not cumulative, rather it has an effect only after a specific threshold has been reached. For instance, BMI may have zero impact on medical expenditures for individuals in the normal weight range, but it may be strongly related to higher costs for the obese (that is, BMI of 30 or above).

We can model this relationship by creating a binary obesity indicator variable that is 1 if the BMI is at least 30, and 0 if less. The estimated beta for this binary feature would then indicate the average net impact on medical expenses for individuals with BMI of 30 or above, relative to those with BMI less than 30.

To create the feature, we can use the **ifelse()** function, which for each element in a vector tests a specified condition and returns a value depending on whether the condition is true or false.

Let’s create a new faeture named bmi30, for BMI greater than or equal to 30, where it will return 1 if BMI greater than 30 and otherwise 0:

#insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)

Model specification - adding interaction effects

So far, we have only considered each feature’s individual contribution to the outcome. What if certain features have a combined impact on the dependent variable?

For instance, smoking and obesity may have harmful effects separately, but it is reasonable to assume that their combined effect may be worse than the sum of each one alone.

When two features have a combined effect, this is known as an **interaction**. If we suspect that two variables interact, we can test this hypothesis by adding their interaction to the model.

Interaction effects are specified using the R formula syntax. To have the obesity indicator (bmi30) and the smoking indicator (smoker) interact, we would write a formula in the form

#expenses ~ bmi30\*smoker.

The multiplication operator is shorthand that instructs R to model

#expenses ~ bmi30 + smokeryes + bmi30:smokeryes.

The: (colon) operator in the expanded form indicates that bmi30:smokeryes is the interaction between the two variables.

Note that the expanded form also automatically included the bmi30 and smoker variables as well as the interaction.

Putting it all together - an improved regression model

Based on a bit of subject matter knowledge of how medical costs may be related to patient characteristics, we developed what we think is a more accurately specified regression formula.

To summarize the improvements, we:

* Added a non-linear term for age
* Created an indicator for obesity
* Specified an interaction between obesity and smoking

**18. Repeat the model fitting and evaluation by adding age2, bmi30, and bmi30.smoker. Any difference? Explain.**

# add a higher-order "age" term  
insurance$age2 <- insurance$age^2  
  
# add an indicator for BMI >= 30  
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)  
  
# create final model  
ins\_model2 <- lm(expenses ~ age + age2 + children + bmi + sex +  
 bmi30\*smoker + region, data = insurance)  
  
summary(ins\_model2)

##   
## Call:  
## lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 \*   
## smoker + region, data = insurance)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17296.4 -1656.0 -1263.3 -722.1 24160.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 134.2509 1362.7511 0.099 0.921539   
## age -32.6851 59.8242 -0.546 0.584915   
## age2 3.7316 0.7463 5.000 6.50e-07 \*\*\*  
## children 678.5612 105.8831 6.409 2.04e-10 \*\*\*  
## bmi 120.0196 34.2660 3.503 0.000476 \*\*\*  
## sexmale -496.8245 244.3659 -2.033 0.042240 \*   
## bmi30 -1000.1403 422.8402 -2.365 0.018159 \*   
## smokeryes 13404.6866 439.9491 30.469 < 2e-16 \*\*\*  
## regionnorthwest -279.2038 349.2746 -0.799 0.424212   
## regionsoutheast -828.5467 351.6352 -2.356 0.018604 \*   
## regionsouthwest -1222.6437 350.5285 -3.488 0.000503 \*\*\*  
## bmi30:smokeryes 19810.7533 604.6567 32.764 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4445 on 1326 degrees of freedom  
## Multiple R-squared: 0.8664, Adjusted R-squared: 0.8653   
## F-statistic: 781.7 on 11 and 1326 DF, p-value: < 2.2e-16

The model fit statistics help to determine whether our changes improved the performance of the regression model.

Relative to our first model, the R-squared value has improved from 0.75 to about 0.87. Similarly, the adjusted R-squared value, which takes into account the fact that the model grew in complexity, also improved from 0.75 to 0.87.

Our model is now explaining 87 percent of the variation in medical treatment costs.

Additionally, our theories about the model’s functional form seem to be validated. The higher-order age2 term is statistically significant, as is the obesity indicator, bmi30. The interaction between obesity and smoking suggests a massive effect; in addition to the increased costs of over $13,404 for smoking alone, obese smokers spend another $19,810 per year. This may suggest that smoking exacerbates diseases associated with obesity.