Lecture 7: Numeric Prediction

Outline

- Numeric Prediction
- Regression Methods
  - Simple Linear Regression
  - Multiple Linear Regression
  - Regression Trees and Model Trees

Numeric Prediction

The construction and evaluation of models used to generate **predictions of numeric values** of a **target variable**.

- Complement of classification which generates predictions of categorical values.
- Predictors, features or input variables (or attributes) may be either numeric or categorical.

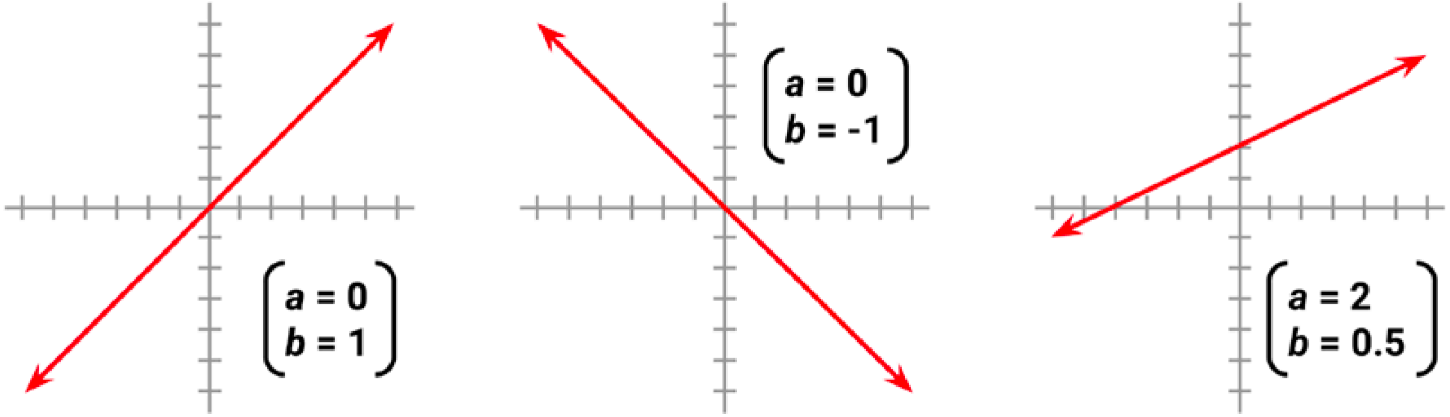
Regression Methods

- Regression is concerned with specifying relationships between
  - A single numeric **dependent variable** (the target variable)
  - One or more **independent variables** (the predictors)
  - The dependent variable depends upon the value of the independent variable or variables.

$$y = a + bx$$

Regression Methods

- The simplest forms of regression assume that the relationship between the independent and dependent variables follows a straight line.



The machine's job is to identify values of *a* and *b* so that the specified line is best able to relate the supplied *x* values to the values of *y*.

- Lines can be defined in a slope-intercept form: $y = a + bx$
  - The letter y indicates the **dependent variable** and x indicates the **independent variable**
  - The **slope term** *b* specifies how much the line rises for each increase in x
  - Positive values defines lines that slope upward while negative values define lines that slope downward.
  - The term *a* is known as the **intercept** because it specifies the point where the line crosses, or intercepts the vertical y axis.

Regression Methods

- Regression analysis is an umbrella for a large number of methods that can be adapted to many data mining tasks.
  - The most basic **linear regression** models – those that use straight lines
    - When there is only a single independent variable it is known as **simple linear regression**
    - In the case of two or more independent variables, it is known as **multiple linear regression** or **multiple regression**
  - Regression can be also used for classification task
    - Logistic regression is used to model a binary categorical outcome.

Simple Linear Regression

Predicting Medical Expenses – Example

- Predicting medical expenses using multiple linear regression
  - In order for a health insurance company to make money, it needs to collect more in yearly premium than it spends on medical care to its beneficiaries.
  - Insurance companies develop models that can accurately forecast medical expenses for its clients.
  - Goal: use patient data to estimate the medical care expenses given the health condition of a patient.
  - The estimate can be used to set the price of yearly premiums.

Simple Linear Regression

- Insurance data

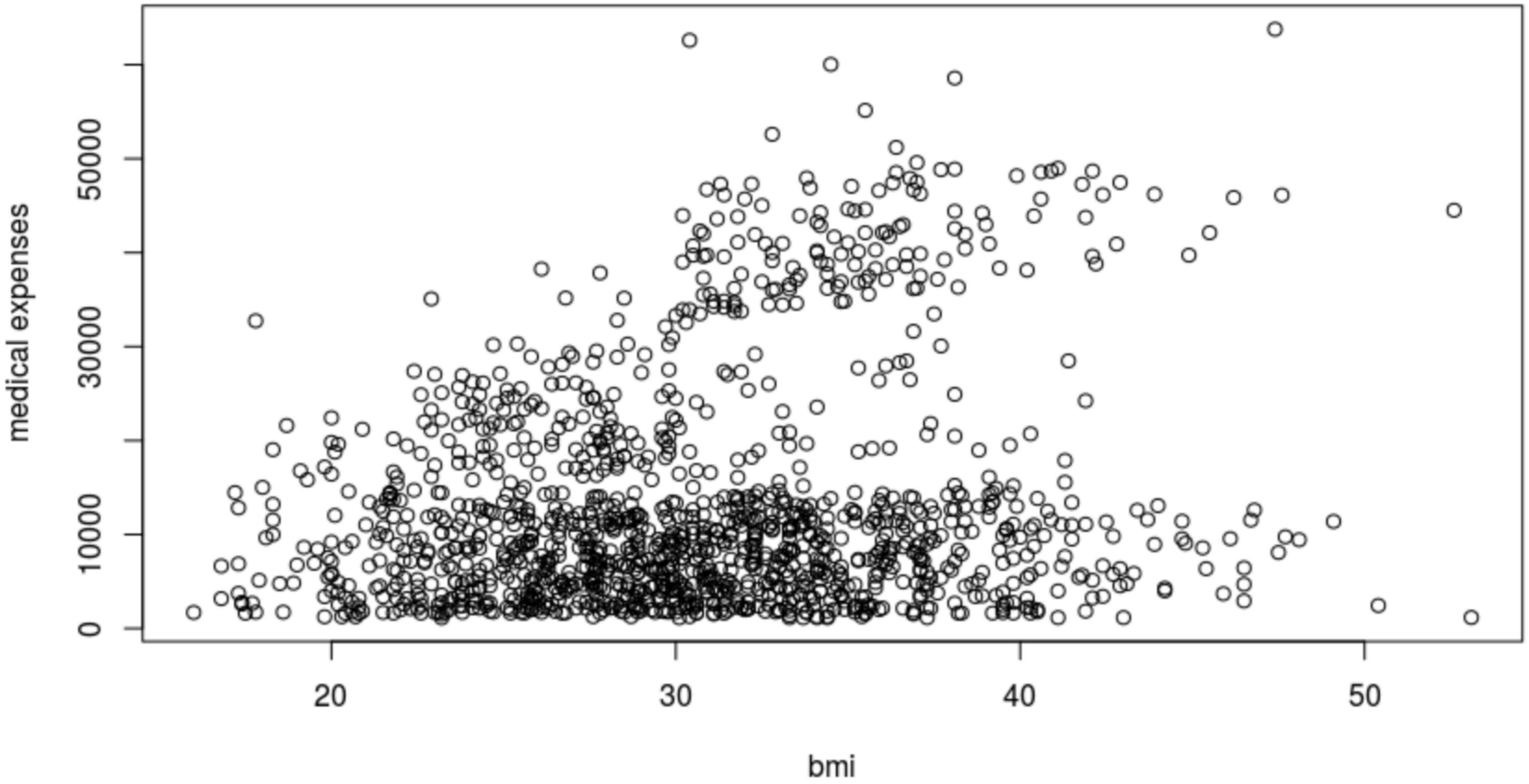
age	sex	bmi	children	smoker	region	expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41
60	female	25.8	0	no	northwest	28923.14
25	male	26.2	0	no	northeast	2721.32

Simple Linear Regression

- Simple Linear Regression
  - Dependent variable: expenses
  - Independent variable: bmi

Simple Linear Regression

- Scatterplot of bmi and expenses



Simple Linear Regression

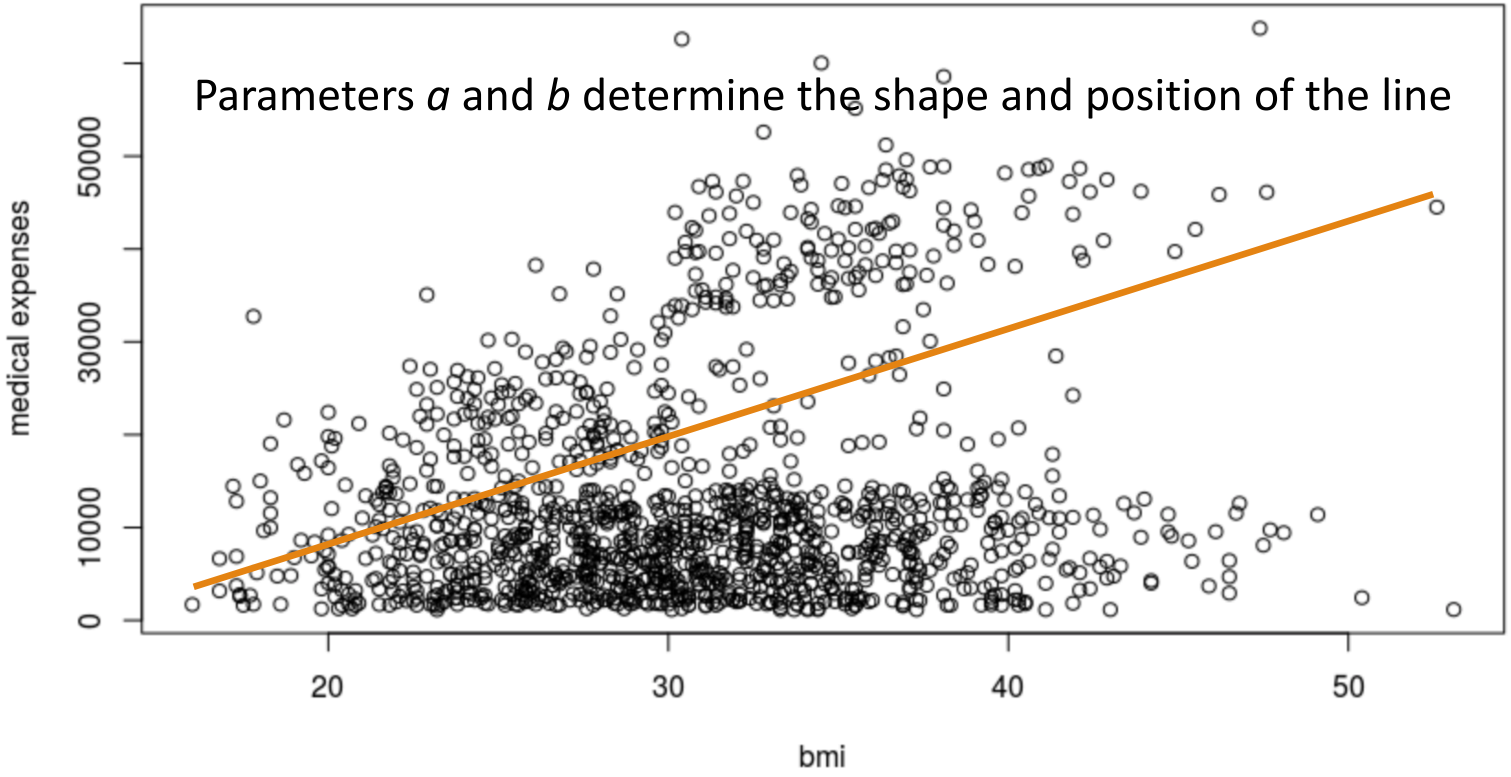
- To model the relationship between bmi and expenses, we can turn to simple linear regression.
- A simple linear regression model defines the relationship between a dependent variable and a single independent predictor variable using a line defined by an equation in the following form

$$y = a + bx$$

Simple Linear Regression

- Suppose we know that the estimated regression parameters in the equation for the shuttle launch data are: a = 785.25 and b = 409.90.

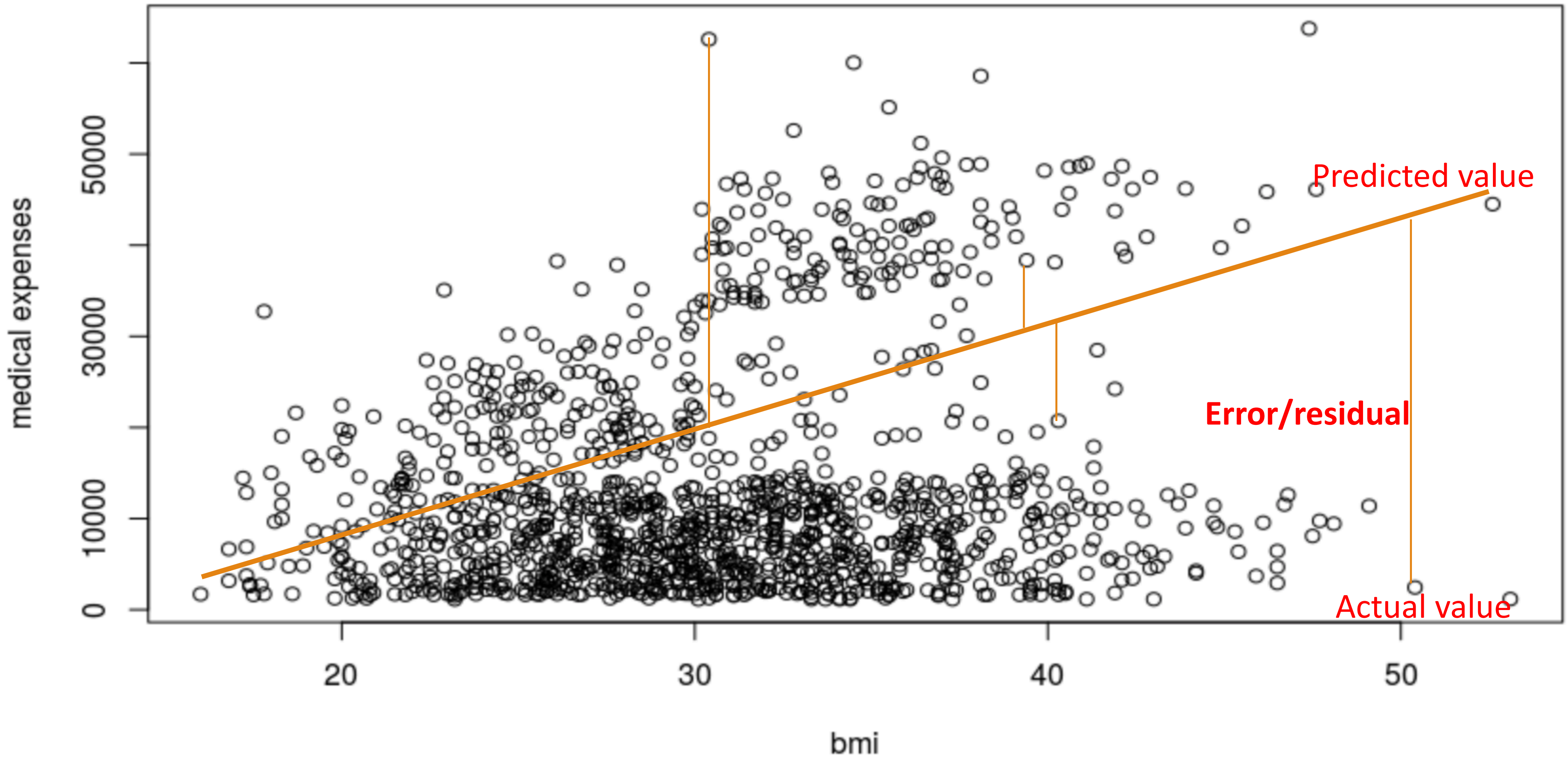
The line doesn't pass through each data point exactly. Instead, it cuts through the data somewhat evenly, with some predictions lower or higher than the line.



Ordinary Least Squares (OLS) Estimation

- In order to determine the optimal estimates of *a* and *b*, an estimation method known as **Ordinary Least Squares** (**OLS**) was used.
- In OLS regression, the slope and intercept are chosen so that they minimize the sum of the squared errors, that is, the vertical distance between the predicted *y* value and the actual *y* value. These errors are known as **residuals**.

Ordinary Least Squares (OLS) Estimation



Ordinary Least Squares (OLS) Estimation

- These errors are known as $\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$
  - This equation defines *e* (the error) as the difference between the actual *y* value and the predicted *y* value. The error values are squared and summed across all the points in the data.
  - The goal of OLS regression: minimize the squared errors
    - $a = \bar{y} - b\bar{x}$
    - $b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
    - R provides functions to perform the estimation automatically.

Multiple Linear Regression

Multiple Linear Regression

- Most real-world analyses have more than one independent variable: using multiple linear regression for most numeric prediction tasks.

Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features, so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

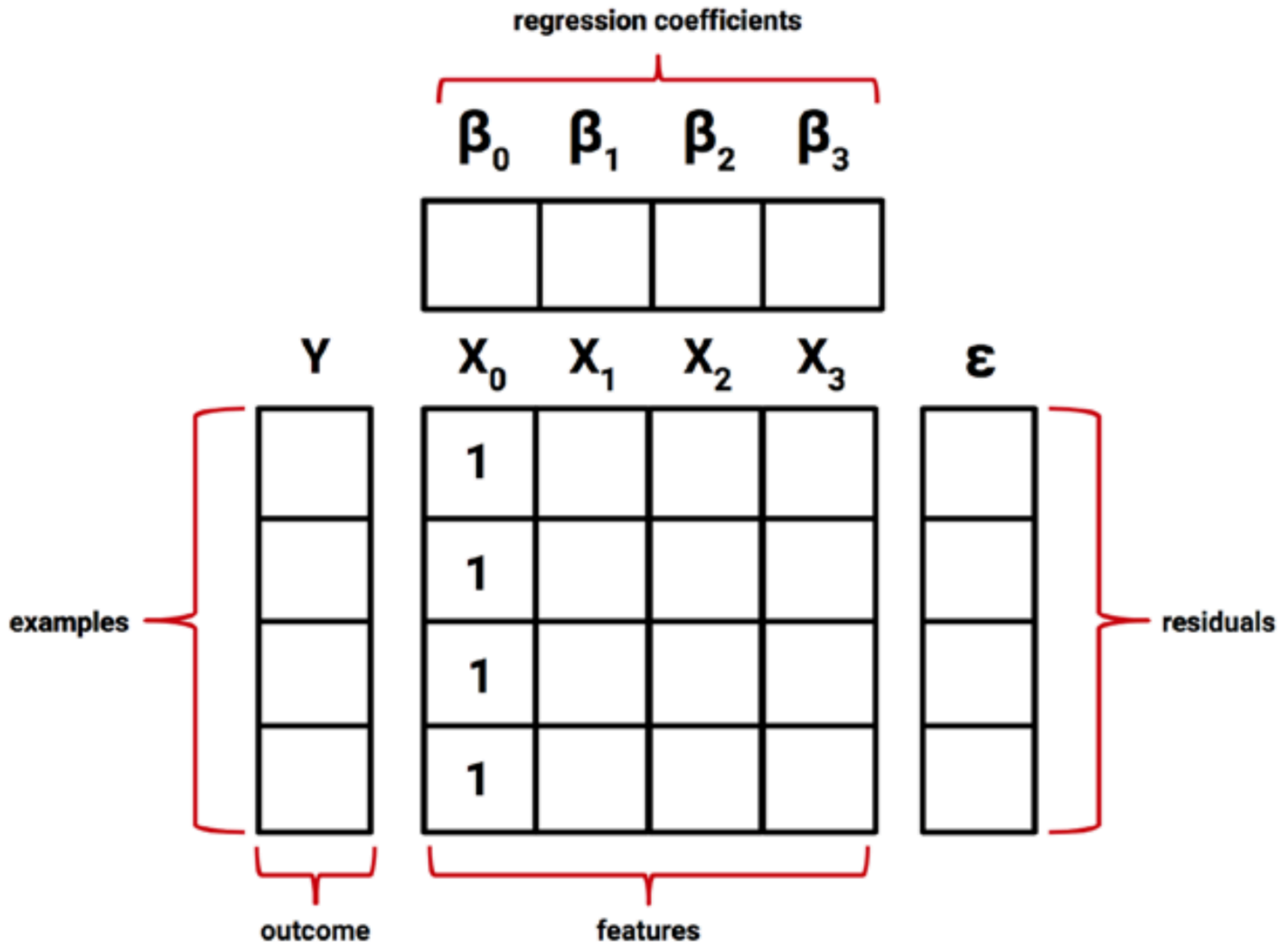
Multiple Linear Regression

- Multiple linear regression is an extension of simple linear regression.
  - It has additional terms for additional independent variables
  - The goal: find values of beta coefficients that minimize the prediction error of a linear equation.
- Multiple linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

- The dependent variable y is specified as the sum of an intercept term β_0 plus the product of the estimated β value and the x values for each of the i features.
- An error term (epsilon) has been added as a reminder that the predictions are not perfect. The error term represents the residual term noted previously.

Multiple Linear Regression



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$

Multiple Linear Regression

- Insurance data

age	sex	bmi	children	smoker	region	expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41
60	female	25.8	0	no	northwest	28923.14
25	male	26.2	0	no	northeast	2721.32

Multiple Linear Regression

```
> summary(insurance)
      age            sex           bmi           children      smoker          region       expenses    
 Min.   :18.00   female:662   Min.   :16.00   Min.   :0.000   no :1064   northeast:324   Min.   : 1122  
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325   1st Qu.: 4740  
 Median :39.00                Median :30.40   Median :1.000              southeast:364   Median : 9382  
 Mean   :39.21                Mean   :30.67   Mean   :1.095              southwest:325   Mean   :13270  
 3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000                              3rd Qu.:16640  
 Max.   :64.00                Max.   :53.10   Max.   :5.000                              Max.   :63770  
```

Multiple Linear Regression

- The intercept is the predicted value of expenses when the independent variables are equal to zero.
  - Intercept is of little value alone because it is impossible to have values of zero for all features
- 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.

```
Call:
lm(formula = expenses ~ ., data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11575.9  -2809.9   -832.8   1524.5  29716.8 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12305.15    1147.21 -10.726  < 2e-16 ***
age                247.66      14.01  17.674  < 2e-16 ***
sexmale            228.55     392.78   0.582  0.56078    
bmi                346.93      33.97  10.213  < 2e-16 ***
children           467.67     164.30   2.846  0.00452 ** 
smokeryes        23919.45     483.93  49.427  < 2e-16 ***
regionnorthwest     12.00     566.76   0.021  0.98311    
regionsoutheast   -658.35     563.93  -1.167  0.24333    
regionsouthwest   -553.26     560.17  -0.988  0.32358    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5978 on 929 degrees of freedom
Multiple R-squared:  0.7605,	Adjusted R-squared:  0.7585 
F-statistic: 368.8 on 8 and 929 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

- lm() function automatically applied a technique known as **dummy coding** to each of the factor-type variables we included in the model.
- The results of the linear regression model make logical sense: 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.

```
Call:
lm(formula = expenses ~ ., data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11575.9  -2809.9   -832.8   1524.5  29716.8 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12305.15    1147.21 -10.726  < 2e-16 ***
age                247.66      14.01  17.674  < 2e-16 ***
sexmale            228.55     392.78   0.582  0.56078    
bmi                346.93      33.97  10.213  < 2e-16 ***
children           467.67     164.30   2.846  0.00452 ** 
smokeryes        23919.45     483.93  49.427  < 2e-16 ***
regionnorthwest     12.00     566.76   0.021  0.98311    
regionsoutheast   -658.35     563.93  -1.167  0.24333    
regionsouthwest   -553.26     560.17  -0.988  0.32358    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5978 on 929 degrees of freedom
Multiple R-squared:  0.7605,	Adjusted R-squared:  0.7585 
F-statistic: 368.8 on 8 and 929 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

1. The **residuals** section provides summary statistics for the errors in our predictions, some of which are apparently quite substantial.
   - 50 percent of errors fall within the 1Q and 3Q values (the first and third quartile), so the majority of predictions were between $2,809.90 over the true value and $1,523.50 under the true value.

```
Call:
lm(formula = expenses ~ ., data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11575.9  -2809.9   -832.8   1524.5  29716.8 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12305.15    1147.21 -10.726  < 2e-16 ***
age                247.66      14.01  17.674  < 2e-16 ***
sexmale            228.55     392.78   0.582  0.56078    
bmi                346.93      33.97  10.213  < 2e-16 ***
children           467.67     164.30   2.846  0.00452 ** 
smokeryes        23919.45     483.93  49.427  < 2e-16 ***
regionnorthwest     12.00     566.76   0.021  0.98311    
regionsoutheast   -658.35     563.93  -1.167  0.24333    
regionsouthwest   -553.26     560.17  -0.988  0.32358    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5978 on 929 degrees of freedom
Multiple R-squared:  0.7605,	Adjusted R-squared:  0.7585 
F-statistic: 368.8 on 8 and 929 DF,  p-value: < 2.2e-16
```

1

2

3

Multiple Linear Regression

2. For each estimated regression coefficient, the **p-value**, denoted Pr(>|t|), provides an estimate of the probability that the true coefficient is zero given the value of the estimate.
   - 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.
   - p-values less than the significance level are considered **statistically significant**.

```
Call:
lm(formula = expenses ~ ., data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-11575.9  -2809.9   -832.8   1524.5  29716.8   1

Coefficients:                                          2
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -12305.15    1147.21 -10.726  < 2e-16 ***
age                247.66      14.01  17.674  < 2e-16 ***
sexmale            228.55     392.78   0.582  0.56078
bmi                346.93      33.97  10.213  < 2e-16 ***
children           467.67     164.30   2.846  0.00452 **
smokeryes        23919.45     483.93  49.427  < 2e-16 ***
regionnorthwest     12.00     566.76   0.021  0.98311
regionsoutheast   -658.35     563.93  -1.167  0.24333
regionsouthwest   -553.26     560.17  -0.988  0.32358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
                                                        3
Residual standard error: 5978 on 929 degrees of freedom
Multiple R-squared:  0.7605,    Adjusted R-squared:  0.7585
F-statistic: 368.8 on 8 and 929 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

3. The **multiple R-squared value** provides a measure of how well our model as a whole explains the values of the dependent variable.
   - The model explains nearly 76 percent of the variation in the dependent variable.
   - High R-squared on training data indicates overfitting.

```
Call:
lm(formula = expenses ~ ., data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max
-11575.9  -2809.9   -832.8   1524.5  29716.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -12305.15    1147.21 -10.726  < 2e-16 ***
age                247.66      14.01  17.674  < 2e-16 ***
sexmale            228.55     392.78   0.582  0.56078
bmi                346.93      33.97  10.213  < 2e-16 ***
children           467.67     164.30   2.846  0.00452 **
smokeryes        23919.45     483.93  49.427  < 2e-16 ***
regionnorthwest     12.00     566.76   0.021  0.98311
regionsoutheast   -658.35     563.93  -1.167  0.24333
regionsouthwest   -553.26     560.17  -0.988  0.32358
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5978 on 929 degrees of freedom
Multiple R-squared:  0.7605,	Adjusted R-squared:  0.7585
F-statistic: 368.8 on 8 and 929 DF,  p-value: < 2.2e-16
```

1

2

3

R squared and adjusted R squared

- R-squared = Explained variation / Total variation
- Adjusted R-squared is the variation of R-squared that adjusts to the number of predictors in a model.
- Adjusted R square calculates the proportion of the variation in the dependent variable accounted by the explanatory variables.

R squared and Adjusted R Squared

- The total sum of squares:

$$SS_{tot} = \sum_i (y_i - \bar{y_i})^2$$

- The sum of squares of residuals, also called the residual sum of squares

$$SS_{res} = \sum_i (y_i - \hat{y_i})^2$$

- The most general definition of the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- The adjusted R²

$$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-p-1} = R^2 - (1 - R^2)\frac{p}{n-p-1}$$

p is the total number of independent variables in the model (not including the constant term), and *n* is the sample size.

Multiple Linear Regression Evaluation

- Splitting method for evaluation

age	sex	bmi	children	smoker	region	expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41
60	female	25.8	0	no	northwest	28923.14
25	male	26.2	0	no	northeast	2721.32

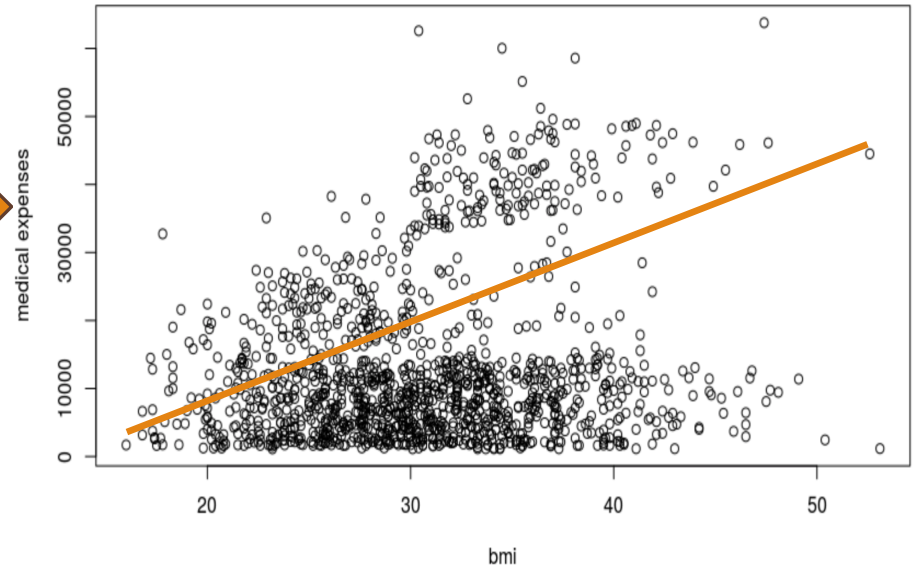
70% training data

30% testing data

Multiple Linear Regression Evaluation

- Train Multiple Linear Regression model on **training data (70%)**

age	sex	bmi	children	smoker	region	expenses
19	female	27.9	0	yes	southwest	16884.92
18	male	33.8	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.7	0	no	northwest	21984.47
32	male	28.9	0	no	northwest	3866.86
31	female	25.7	0	no	southeast	3756.62
46	female	33.4	1	no	southeast	8240.59
37	female	27.7	3	no	northwest	7281.51
37	male	29.8	2	no	northeast	6406.41
60	female	25.8	0	no	northwest	28923.14
25	male	26.2	0	no	northeast	2721.32



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$$

Calculate the coefficients

Multiple Linear Regression Evaluation

- Make predictions on **testing data (30%)** and **training data (70%)**

age	sex	bmi	children	smoker	region	expenses	Predicted expenses	
19	female	27.9	0	yes	southwest	16884.92	15445.54	70% training data
18	male	33.8	1	no	southeast	1725.55	1455.21	
28	male	33	3	no	southeast	4449.46	6799.98	
33	male	22.7	0	no	northwest	21984.47	12351.13	
32	male	28.9	0	no	northwest	3866.86	2434.12	
31	female	25.7	0	no	southeast	3756.62	1235.12	
46	female	33.4	1	no	southeast	8240.59	6394.12	
37	female	27.7	3	no	northwest	7281.51	5450.74	
37	male	29.8	2	no	northeast	6406.41	2509.23	30% testing data
60	female	25.8	0	no	northwest	28923.14	39455.26	
25	male	26.2	0	no	northeast	2721.32	2311.65	

Evaluation Metrics for Numeric Prediction

- **Mean Absolute Error (MAE)**
  - The **mean absolute error (MAE)** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.
  - This is known as a scale-dependent accuracy measure and therefore cannot be used to make comparisons between series using different scales

$$\text{MAE} = \frac{1}{\text{n}}\sum_{t=1}^{n} |y_t - \hat{y}_t|$$

- **Root Mean Squared Error (RMSE)**
  - RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n}}$$

- **Mean absolute percentage error (MAPE)**
  - The **mean absolute percentage error** (**MAPE**) measures this accuracy as a percentage.
  - It cannot be used if there are zero values because there would be a division by zero.

$$\text{MAPE} = \frac{1}{\text{n}}\sum_{t=1}^{n} \left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

Not depend on the scale

- **Relative absolute error (RAE)**
  - The **relative absolute error** takes the total absolute error and normalizes it by dividing by the total absolute error of the mean estimator.

$$\text{RAE} = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\sum_{t=1}^{n}|y_t - \bar{y}_t|}$$

Evaluation Metrics for Numeric Prediction

```
> mmetric(datTrain$expenses,prediction_on_train2,metric=c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE 
4130.17999 5949.05058   42.70693   44.99144 
> mmetric(datTest$expenses,prediction_on_test2,metric=c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE 
4319.79104 6275.06831   44.63077   48.62537 
```

Explanatory vs Predictive Power

- **Explanatory power**
  - Of explaining which and how predictors affect the dependent variable most significantly.
  - Model elements and metrics to be used: correlation coefficients, beta coefficients, p-values, R-squared
- **Predictive power**
  - Evaluate prediction accuracy and generalizability.
  - Performance metrics to be used: MAE, RMSE, MAPE, and RAE

Improving Model Performance: Adding non-linear relationships

- In linear regression, the relationship between independent variable and the dependent variable is assumed to be **linear**. Yet this may not necessarily be true.
  - The effect of age on medical cost may not be constant throughout all the age values
  - To account for a non-linear relationship, we can add a higher order term to the regression model. We will be modeling relationship like this:

$$y = \alpha + \beta_1 x + \beta_2 x^2$$

Improving Model Performance: Adding non-linear relationships

```
Call:
lm(formula = expenses ~ ., data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-11072.2  -2769.9   -946.4   1266.9  31341.4 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -5586.259   2002.048  -2.790 0.005374 ** 
age              -121.903     94.806  -1.286 0.198832    
sexmale           -43.789    389.571  -0.112 0.910528    
bmi               345.390     34.058  10.141  < 2e-16 ***
children          621.231    165.583   3.752 0.000186 ***
smokeryes       23847.010    481.129  49.565  < 2e-16 ***
regionnorthwest  -823.050    551.922  -1.491 0.136238    
regionsoutheast -1725.564    563.951  -3.060 0.002279 ** 
regionsouthwest  -949.741    562.308  -1.689 0.091555 .  
age2                4.813      1.182   4.071 5.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5921 on 928 degrees of freedom
Multiple R-squared:  0.7612,	Adjusted R-squared:  0.7589 
F-statistic: 328.6 on 9 and 928 DF,  p-value: < 2.2e-16
```

Increase age from 20 to 30

- (900 – 400) * 4.813 + (30-20)* (– 121.903)

Increase age from 40 to 50

- (2500 – 1600) * 4.813 + (50-40)* (– 121.903)

Improving Model Performance: Adding non-linear relationships

```
> mmetric(datTrain$expenses,prediction_on_train3,metric=c("MAE","RMSE","MAPE","RAE"))
      MAE      RMSE      MAPE       RAE
4046.8831 5889.5345   40.9591   44.8450
> mmetric(datTest$expenses,prediction_on_test3,metric=c("MAE","RMSE","MAPE","RAE"))
       MAE        RMSE        MAPE         RAE
4341.51061 6307.86394   44.62597   46.93653
```

Improving Model Performance: Converting a numeric variable into a binary indicator

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

Improving Model Performance:
Converting a numeric variable into a binary indicator

- To create the feature, we can use the `ifelse()` function
- For BMI greater than or equal to 30, we will return 1, otherwise 0:

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

- Include the `bmi30` variable in our improved model, either replacing the original `bmi` variable or in addition
  - Depending on whether or not we think the effect of obesity occurs in addition to a separate linear BMI effect.

Improving Model Performance: 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.
  - 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.

Improving Model Performance

```
Call:
lm(formula = expenses ~ . + bmi30 * smoker, data = datTrain)

Residuals:
     Min       1Q   Median       3Q      Max 
-16954.0  -1659.8  -1195.4   -509.8  23935.7 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -670.4243  1602.0246  -0.418  0.67569    
age               -17.4262    70.3592  -0.248  0.80444    
sexmale          -255.8048   288.6159  -0.886  0.37568    
bmi               133.0085    40.6899   3.269  0.00112 ** 
children          656.4572   127.4982   5.149  3.2e-07 ***
smokeryes       13467.7242   521.2449  25.838  < 2e-16 ***
regionnorthwest    50.1895   415.8849   0.121  0.90397    
regionsoutheast  -653.1946   414.0931  -1.577  0.11504    
regionsouthwest -1300.5816   411.9940  -3.157  0.00165 ** 
age2                3.4873     0.8752   3.984  7.3e-05 ***
bmi30           -1029.9937   501.4441  -2.054  0.04025 *  
smokeryes:bmi30 19477.2235   710.6592  27.407  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4385 on 926 degrees of freedom
Multiple R-squared:  0.8716,    Adjusted R-squared:  0.8701 
F-statistic: 571.3 on 11 and 926 DF,  p-value: < 2.2e-16
```

For non-obese patients:

- Smoker increases the expenses by 13467.7242

For obese patients:

- Smoker increases the expenses by 13467.7242 + 19477.2235

Improving Model Performance

```
> mmetric(datTrain$expenses,prediction_on_train4,metric=c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE 
2306.75318 4356.49881   26.94607   25.12824 
> mmetric(datTest$expenses,prediction_on_test4,metric=c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE 
2485.97932 4601.72277   28.31190   27.98322 
```

An Improved Regression Model

- Based on domain 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

Regression Trees and Model Trees

Trees for Numeric Prediction

- A decision tree can also be used for numeric prediction by making only small adjustments to the tree-growing algorithm.

- Trees for numeric prediction fall into two categories:
  - Regression trees
  - Model trees

Trees for Numeric Predictions

- Regression tree
  - Similar in construction to classification tree
  - Predicted value of each leaf node is typically reported as mean value of output variable for all training observations belonging to node.
  - SDR (Standard deviation reduction) or squared error reduction is used to choose a predictor to split a node, similar to the way classification error is used in decision tree.
- Model tree
  - A Model Tree is grown in a similar way to Regression Tree
  - In each of the leaf, an MLR model is built using data in the leaf

Regression Tree

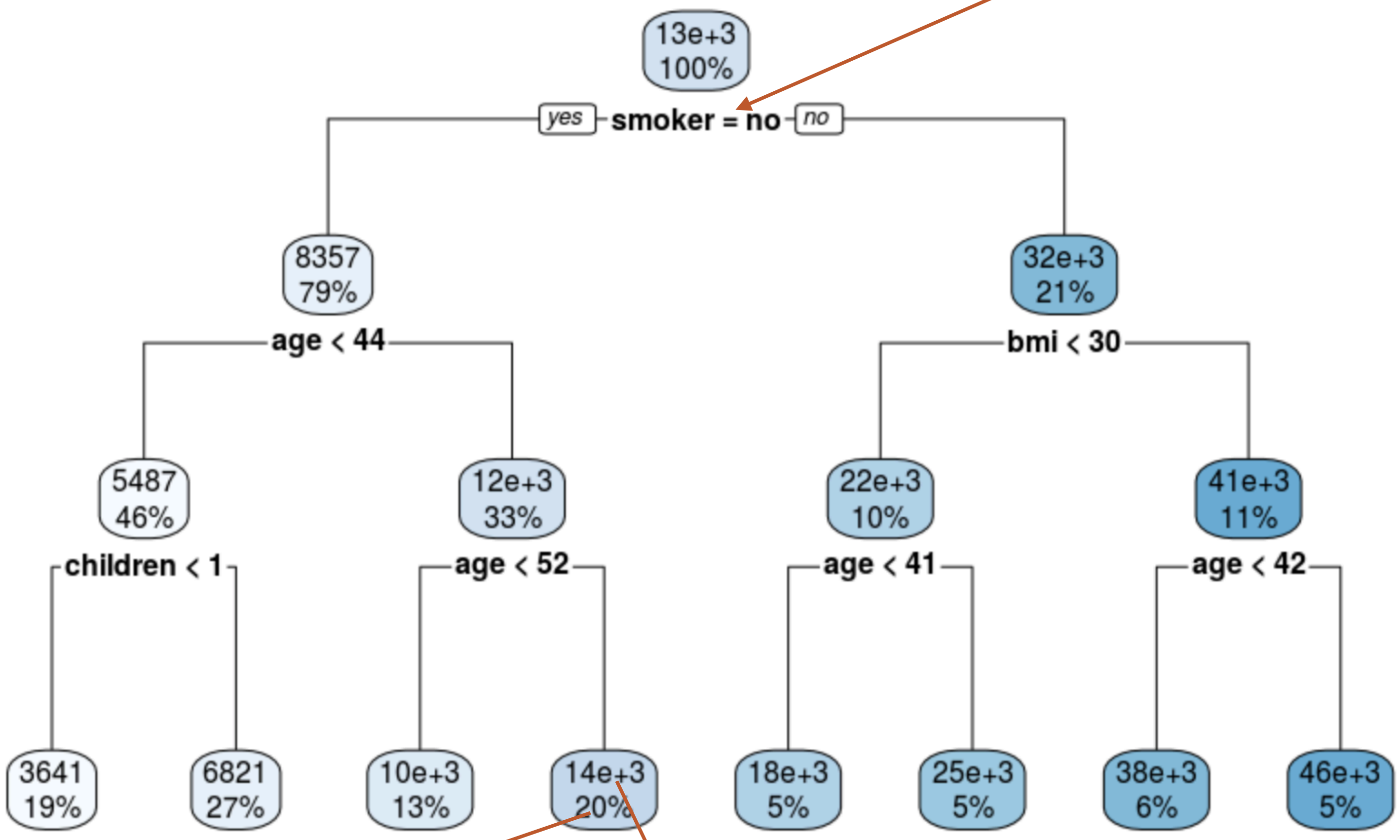
- Decision trees offer distinct advantages over traditional regression models.
  - Decision trees may be better suited for tasks with many complex, non-linear relationships among features and outcome.

Regression Tree

- Trees for numeric prediction are built in much the same way as they are for classification.
  - Beginning at the root node, the data is partitioned according to the feature that will result in the greatest increase in homogeneity in the outcome.
  - For numeric decision trees, homogeneity is measured by statistics such as variance, standard deviation, or absolute deviation from the mean.

Regression Tree

Reduce the squared error most



Percentage of data

Predicted y value

Squared error

Predicted y value

```
node), split, n, deviance, yval
      * denotes terminal node

 1) root 938 138628400000 13329.830
   2) smoker=no 742  25763850000  8356.999
     4) age< 43.5 429  10087290000  5486.668
       8) children< 0.5 180   2649069000  3640.711 *
       9) children>=0.5 249   6381467000  6821.094 *
     5) age>=43.5 313   7297775000 12291.090
      10) age< 51.5 124   2595795000 10240.390 *
      11) age>=51.5 189   3838386000 13636.530 *
   3) smoker=yes 196  25051740000 32155.540
     6) bmi< 30.1 91   2544511000 21541.950
      12) age< 41 48    747311800 18459.850 *
      13) age>=41 43    832241900 24982.430 *
     7) bmi>=30.1 105   3372033000 41353.980
      14) age< 41.5 59    667122000 37857.140 *
      15) age>=41.5 46   1058133000 45839.060 *
```

Trees for Numeric Prediction

```
> mmetric(datTrain$expenses,prediction_on_train_rpart,metric = c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE
2704.66368 4473.27130   35.77833   29.46281
> mmetric(datTest$expenses,prediction_on_test_rpart,metric = c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE
2914.88757 4900.81937   37.83747   32.81119
```

Model Trees

- A model tree improves on regression trees by replacing the leaf nodes with regression models.
  - This often results in more accurate results than regression trees, which use only a single value for prediction at the leaf nodes.

Model Trees

```
M5 pruned model tree:
(using smoothed linear models)

smoker=yes <= 0.5 : LM1 (742/37.018%)
smoker=yes >  0.5 :
|   bmi <= 30.1 : LM2 (91/29.743%)
|   bmi >  30.1 : LM3 (105/26.163%)
```

Root relative squared error

```
LM num: 1
expenses =
	254.6446 * age
	+ 4.9247 * sex=male
	+ 564.9711 * children
	+ 9.3549 * smoker=yes
	+ 1966.8791 * region=northeast,northwest,southeast
	- 698.2696 * region=southeast
	- 3595.6991

LM num: 2
expenses =
	230.3609 * age
	+ 17.6683 * sex=male
	+ 566.2788 * bmi
	+ 23.7739 * children
	+ 33.5622 * smoker=yes
	+ 1699.4966 * region=northeast,northwest,southeast
	- 3688.6548

LM num: 3
expenses =
	259.8892 * age
	+ 17.6683 * sex=male
	+ 559.4363 * bmi
	+ 23.7739 * children
	+ 33.5622 * smoker=yes
	+ 1699.4966 * region=northeast,northwest,southeast
	+ 7354.11
```

Model Trees

```
> mmetric(datTrain$expenses,prediction_on_train_M5P,metric = c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE
2391.30212 4411.51175   26.32135   26.04926
> mmetric(datTest$expenses,prediction_on_test_M5P,metric = c("MAE","RMSE","MAPE","RAE"))
       MAE       RMSE       MAPE        RAE
2659.35371 4819.90417   27.37270   29.93479
```