## Making Predictions with Simple Linear Regression Model

Needed libraries

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.5.2

## -- Attaching packages ----------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.7  
## v tidyr 0.8.2 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts -------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

Read-in the data. Before doing this make sure that you have placed the CreditData.csv file (downloadable from Canvas) in your project’s working directory.

credit = read\_csv("CreditData.csv")

## Parsed with column specification:  
## cols(  
## AnnualIncome = col\_double(),  
## HouseholdSize = col\_integer(),  
## YrsEdAfterHS = col\_integer(),  
## HrWkTV = col\_integer(),  
## AnnualCharges = col\_double()  
## )

Get rid of missing data rows

credit = credit %>% drop\_na() #delete any row with an NA value

Build a regression model with AnnualIncome to predict AnnualCharges.

mod1 = lm(AnnualCharges ~ AnnualIncome, credit) #create linear regression model  
summary(mod1) #examine the model

##   
## Call:  
## lm(formula = AnnualCharges ~ AnnualIncome, data = credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12284.4 -3938.1 14.4 3947.9 13232.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3146.361 185.193 16.99 <2e-16 \*\*\*  
## AnnualIncome 121.355 2.529 47.98 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5027 on 4998 degrees of freedom  
## Multiple R-squared: 0.3153, Adjusted R-squared: 0.3152   
## F-statistic: 2302 on 1 and 4998 DF, p-value: < 2.2e-16

Prediction for AnnualIncomes of 75 and 100?

#Manually  
3146.361 + 121.355\*75

## [1] 12247.99

3146.361 + 121.355\*100

## [1] 15281.86

#Using predict function  
testdata = data.frame(AnnualIncome = c(75,100))  
predict(mod1, newdata = testdata, interval = "predict")

## fit lwr upr  
## 1 12247.98 2392.131 22103.82  
## 2 15281.85 5424.762 25138.93

#Prediction intervals  
temp\_var = predict(mod1, interval = "prediction")

## Warning in predict.lm(mod1, interval = "prediction"): predictions on current data refer to \_future\_ responses

new\_df = cbind(credit, temp\_var)  
  
ggplot(new\_df, aes(x = AnnualIncome, y = AnnualCharges)) +   
 geom\_point() +   
 geom\_smooth(method = "lm", se = FALSE, color = "red") +  
 geom\_line(aes(y=lwr), color = "red", linetype = "dashed") +  
 geom\_line(aes(y=upr), color = "red", linetype = "dashed") +  
 theme\_bw()

