**Linear regression using R computer labs**

Contents

[Cars.csv data set\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2](#_Toc306139112)

[Lab 1: Simple linear regression 3](#_Toc306139120)

[**Lab 2: Diagnostics,**](#_Lab_Session_4:) **categorical variables, and ANCOVA\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 5**

[Lab 3: Multiple regression and model selection 8](#_Toc306139123)

## Cars.csv data set

The Cars.csv data set is the Cars.sav dataset with SPSS saved as a comma separated file. It contains data on specifications of 406 vehicles from 1970 to 1982. Among the variables in the data set are information on fuel consumption (mpg), horsepower, weight, acceleration, origin (Europe, Japan, U.S.), and number of cylinders.

The Cars.sav data set contains categorical variables (such as origin), numerical discrete variables (such as number of cylinders), and continuous variables (such as weight, and acceleration). Descriptive statistics for all numeric variables are shown below, along with frequencies for some variables.

## 

## Lab 1: Simple linear regression

We will be examining the linear relationship between Horsepower (Y) and Weight (X) in this problem, using the cars.csv data set. R programming statements are in italics.

1. Read in cars.csv

*cars = read.table(“cars.csv”, header=TRUE, sep=”,”)*

1. Create a scatter plot of horsepower and weight.

*with(cars, plot(horse, weight, xlab=”Horsepower”, ylab= “Weight”, pch=16))*

1. Calculate the correlation between horsepower and weight and determine the statistical significance of the correlation.

*with(cars, cor.test(horse, weight))*

1. Create a simple linear regression model object that models weight in terms of horsepower

*horse\_model=lm(weight~ horse, data=cars)*

1. Get basic results from *horse\_model*

*summary(horse\_model)*

1. Get 95% confidence intervals for the parameters in *horse\_model*

*confint(horse\_model, level=0.95)*

1. Add the regression line to the scatter plot

*abline(horse\_model, col=”red”, lwd=2)*

1. Add 95% confidence band to the scatter plot

*band = data.frame(horse=seq(min(cars$horse, na.rm=TRUE), max(cars$horse, na.rm=TRUE)))*

*y=predict.lm(horse\_model, band, interval=”confidence”, level=0.95)*

*points(band$horse, y[,2], pch=”.”,cex=2, col=”blue”)*

*points(band$horse, y[,3], pch=”.”, cex=2,col=”blue”)*

1. Add 95% prediction interval to the scatter plot

*prediction=predict.lm(horse\_model, band, interval=”prediction”, level=0.95)*

*points(band$horse, prediction[,2], pch=”.”,cex=2, col=”green”)*

*points(band$horse, prediction[,3], pch=”.”, cex=2, col=”blue”)*

1. Add text for the equation and R2 value to the scatter plot

*text(175, 2000, paste("Weight = ", round(horse\_model$coef[1]),"+",*

*round(horse\_model$coef[2]),"\* horsepower",cex=0.5))*

*text(120,1500, paste(“R^2=”, round(summary(horse\_model)$r.squared,2)))*

## Lab 2: Diagnostics, categorical variables, and ANCOVA

1. Create a basic residual plot.

*with(horse\_model, plot(fitted.values, residuals, pch=16))*

*abline(0,0)*

1. Standardize the residual plot by standardizing the fitted values and using studentized-deleted residuals. (Taken from <http://www2.kenyon.edu/Depts/Math/hartlaub/Math305%20Fall2011/R.htm>). Create a histogram and Q-Q plot of the studentized-deleted residuals.

*stud.del.resids = rstudent(horse\_model)*

*zpred = (horse\_model$fitted.values-mean(horse\_model$fitted.values))/sd(horse\_model$fitted.values)*

*with(horse\_model, plot(zpred,stud.del.resids,pch=16))*

*hist(stud.del.resids,breaks=40,prob=TRUE)*

*abline(0,0)*

*x=seq(min(stud.del.resids),max(stud.del.resids),by=0.01)*

*y=dnorm(x,mean=mean(stud.del.resids),sd=sd(stud.del.resids))*

*points(x,y,col="blue",pch=20)*

*qqnorm(stud.del.resids)*

*abline(0,1)*

1. Save Cook’s distance and DFFITS and identify potential outliers.

*horse\_model$model$cook = cooks.distance(horse\_model)*

*horse\_model$model$dffits = dffits(horse\_model)*

*n=nrow(horse\_model$model)*

*outliers=subset(horse\_model$model, cook>4/n | dffits > 2\*sqrt(2/n) )*

1. Create new dataset consisting just of cars with weight at least 1,000 pounds. Remove as well the car with 225 horsepower and weight 3,086. Create new model and residual plot.

*cars2 = subset(cars, weight >1000 & (weight >3500 | horse<200))*

*horse\_model2 = lm(weight ~ horse, data=cars2)*

*with(horse\_model2, plot(fitted.values, residuals, pch=16))*

*abline(0,0)*

1. Create a log-transformed weight horse variable and model weight in terms of this variable. Note that none of the observations have *0* weight. Examine the model and residual plot.

*cars2$log\_horse = log(cars2$horse)*

*horse\_model3 = lm(weight~log\_horse, data=cars2)*

*summary(horse\_model3)*

*with(horse\_model3,plot(fitted.values,residuals,pch=16))*

*abline(0,0)*

1. Graph a box-plot and calculate summary statistics for mpg by origin.

*cars2$origin=factor(cars2$origin, levels=c(1,2,3), labels=c(“American”,”European”,”Japanese”)*

*boxplot(mpg~origin, xlab=”Origin”, ylab=”Miles per gallon”, data=cars2*

*with(cars2, tapply(mpg, origin, summary))*

*with(cars2,tapply(mpg,origin,function(x){sd(x,na.rm=TRUE)}))*

1. Fit a regression model for mpg with origin as the predictor

*mpg\_model=lm(mpg ~ origin, data=cars2)*

*summary(mpg\_model)*

1. Change the reference category to “Japanese” and refit the model

*cars2$origin=relevel(cars2$origin,ref=3)*

*mpg\_model2=lm(mpg ~ origin, data=cars2)*

*summary(mpg\_model2)*

1. Create a scatter plot of weight and horsepower that is colored based on

origin (taken from *https://stackoverflow.com/questions/7466023/how-to-give-color-to-each-class-in-scatter-plot-in-r)*. Does the relationship between

weight and horsepower vary by origin?

*with(cars2,plot(horse,weight,col=c("red","blue","green")[origin],pch=16))*

*legend(x=”topright”, legend=levels(cars2$origin), col=c(“red”,”blue”,”green”), pch=16)*

1. Create a model for weight and horsepower that allows for different slopes

and intercepts by origin.

*horse\_model4 = lm(weight~horse + origin + horse\*origin, data =cars2)*

*summary(horse\_model4)*

## Lab 3: Multiple linear regression and model selection

## 

We will be investigating the variables that are associated with mpg in the cars dataset.

1. Create a scatterplot matrix (taken from the *car* package). Let’s restrict our attention to mpg, engine, horse, weight, and acceleration.

*library(car)*

*pairs(~mpg+engine+horse+weight+accel,data=cars2)*

1. Investigate the correlation between the variables in the dataset.

*cor(cars2,use=”complete.obs”)*

1. Create a multiple linear regression model for mpg.

*mpg\_model=lm(mpg ~ engine + horse + weight + accel, data=cars2)*

*summary(mpg\_model)*

1. Look at a residual plot by weight.

*with(mpg\_model, plot(model$weight, residuals, pch=16)*

1. Calculate variance inflation factors.

*vif(mpg\_model)*

1. Create a second mpg model using just weight.

*mpg\_model2=lm(mpg~weight, data=cars2)*

*summary(mpg\_model2)*

1. Calculate the partial correlation of mpg with horse after controlling for weight and create added variable plot.

*complete = subset(cars2, !is.na(mpg))*

*partial\_horse=lm(horse~weight, data=complete)*

*partial\_mpg=lm(mpg~weight, data=complete)*

*cor(partial\_horse$residuals, partial\_mpg$residuals)*

*plot(partial\_horse$residuals,partial\_mpg$residuals,pch=16, xlab=”Horsepower residuals”, ylab=”MPG residuals”)*

1. Add horse back into the model and compare with mpg\_model2.

*mpg\_model3=lm(mpg~weight+horse, data=cars2)*

*summary(mpg\_model3)*

*anova(mpg\_model2, mpg\_model3)*

1. Log-transform mpg and model in terms of weight and horse.

*cars2$log\_mpg = log(cars2$mpg)*

*mpg\_model4=lm(log\_mpg~weight+horse, data=cars2)*

*summary(mpg\_model4)*

*with(mpg\_model4, plot(model$weight, residuals, pch=16))*