How to run Regression Analyses using R

Presenters: Tom Coyne and Jon Wayland

Spring 2013

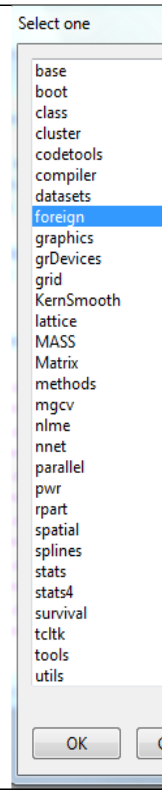
R is a free software downloadable at http://www.r-project.org/

Notes:	**Code and Output:**
1. R Console Setup: > prompts you for formula or function. The result appears on the next line(s).	
2. Comments begin with # Anything in the line following a # is a comment.	# This is a comment!

3. Loading a Package

Many functions and data sets are available in packages that be downloaded from a CRAN site.
We generally use PA 1 (Carnegie Mellon)
We will be loading a package called "foreign".

1) Select "**Packages**" at the top of the screen.
2) Select **"Load package**…"
3) Select the package "**foreign**" and press "**ok**".
4) Activate the package using the library command.




```
# local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})

> library(foreign)
```

4. Importing a Dataset

We will look at the predictive ability of:

- Odometer reading (thousands of Miles),
- Age (years)
- Type(Coupe vs. Sedan)

on Honda prices(dollars).

```
> honda = read.spss(file.choose(), to.data.frame = TRUE)
> attach(honda)
> head(honda) # this shows the first 6 cases of your data frame.  Odometer is measured in thousands of miles, age in years, and price in dollars.
```

	ID	ODOMETER	AGE	TRIM	TYPE	PRICE
1	45	102	2	EX	COUPE	7555
2	68	20	0	LX	SEDAN	15995
3	91	38	0	LX	SEDAN	14995
4	112	48	0	LX	SEDAN	14995
5	133	22	0	EX	COUPE	15995
6	154	26	0	EX	SEDAN	15995

5. Regression

Fitting a model to data and using it to predict dependent, or outcome, variables

i) Simple Regression: Predicting an outcome variable from one independent, or predictor, variable.

ii) Multiple Regression: Predicting an outcome variable from two or more predictor variables.

6. Simple Regression

We will run a simple linear regression analysis to predict the price of Hondas using the number of miles on the odometer as the predictor.

Predictor variable = ODOMETER

Outcome variable = PRICE

```
> lmSLR = lm(PRICE~ODOMETER)
> summary(lmSLR)

Residuals:
    Min      1Q  Median      3Q     Max
-5194.3 -1074.7     5.9  1241.8  5290.8

Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)  15856.741    750.186    21.14   <2e-16 ***
ODOMETER       -92.831      8.343   -11.13   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2035 on 63 degrees of freedom
Multiple R-squared: 0.6628,    Adjusted R-squared: 0.6574
F-statistic: 123.8 on 1 and 63 DF,  p-value: < 2.2e-16
```

Multiple R-squared = .663. Mileage on the odometer accounts for 66.3% of the variation in Honda prices.

F-statistic = 123.8 with , p-value < 2.2e-16.

Overall, this means the regression model predicts Honda prices well.

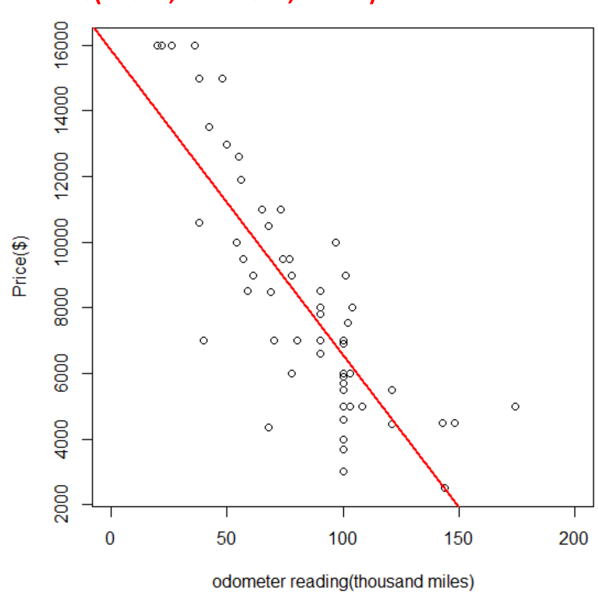
8. Confidence Intervals for the Parameters

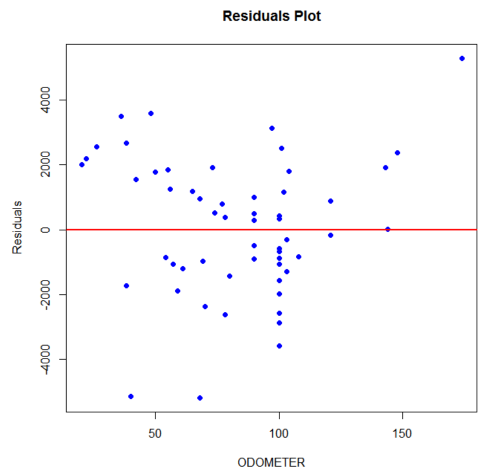
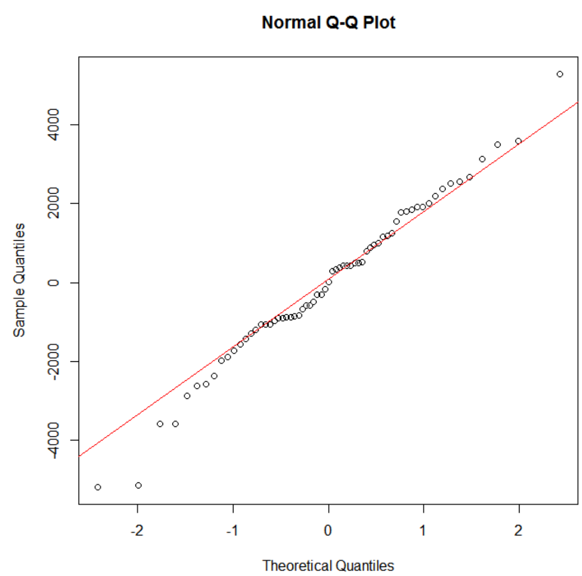
```
>round(confint(lmSLR),1)

                2.5 %    97.5 %
(Intercept)  14357.6   17355.9
ODOMETER      -109.5     -76.2
```

8. Plotting the Data

```
> plot(PRICE~ODOMETER, xlab="odometer reading(Thousand miles)", ylab="Price($)", xlim=c(0,200))
> abline(lmSLR, col="red", lwd=2)
```



8. Plotting the Residuals To check for linearity, plot the residuals vs. the odometer reading	**> resid = residuals(lmSLR)** **> plot(resid~ODOMETER, main = "Residuals Plot", ylab = "Residuals", xlab = "ODOMETER", col = "blue", pch=16)** **> abline(h=0, col="red", lwd=2)** 
9. Plotting the Residuals To check for normality, generate the QQ plot of the residuals:	> qqnorm(resid) > qqline(resid, col="red") 
## 10. Multiple Linear Regression Involves the same concept as simple linear regression, but includes two or more predictor variables. Predictor variables must be continuous or dichotomous categorical variables.	
11. Assumptions Multicollinearity: The predictor variables should not correlate too highly with each other (r > .80)	
12. Checking Assumptions: Multicollinearity To check the assumption of multicollinearity, we will run a correlation matrix for the predictor variables in R. Rule of Thumb: |r| < .8	**> cor(AGE, ODOMETER, use = "pairwise.complete.obs", method = "pearson")** [1] 0.6182078

13. Multiple Linear Regression – Continuous Variables Only

Predictor variables: AGE, ODOMETER
Outcome variable: PRICE

```
> lmMLR = lm(PRICE~ODOMETER + AGE)
> summary(lmMLR)
```

14. Interpretation

Multiple R-squared = .884. 88.% of the variation in Honda prices can be attributed to the mileage on the odometer and the age of the car.
F-statistic = 245.7, p-value < 2.2e-16.

b-values found under Estimate column. For every 1000-mile increase in the odometer reading, we would expect the price of the car to decrease by $50.27. For every one-year increase in age, we would expect the price of the car to decrease by $579.41.

t-values are significant at the 0.001 level, indicating that both odometer and age are significant predictors of car prices.

Beta values are then obtained to more accurately determine the importance of each predictor in the model.

```
Call:
lm(formula = PRICE ~ ODOMETER + AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-2409.15 -825.74  -93.09  749.81 2480.65

Coefficients:
            Estimate  Std. Error  t value   Pr(>|t|)
(Intercept) 15328.018   438.414   34.962   < 2e-16 ***
ODOMETER      -50.268     6.167   -8.152  2.16e-11 ***
AGE          -579.405    51.898  -11.164   < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1182 on 62 degrees of freedom
Multiple R-squared: 0.888,     Adjusted R-squared: 0.8844
F-statistic: 245.7 on 2 and 62 DF,  p-value: < 2.2e-16
```

15. Obtaining Standardized Beta Estimates

We must install and load the "QuantPsyc" package

We then create an object that only includes the predictor variables.

Age of car is a stronger predictor of car price than odometer value. As age increases by 1 standard deviation, the price decreases by 0.604 standard deviations. As the odometer reading increases by 1 standard deviation, the price decreases by 0.441 standard deviations.

```
>install.packages("QuantPsyc")
# to load spss file, select "packages" in the top menu and choose "Load Packages"
# from the package list, choose "QuantPsyc"
>library(QuantPsyc)

>lm.beta( lmMLR )
 ODOMETER       AGE
-0.4408387 -0.6037712
```

16. Multiple Linear Regression – Adding a Dichotomous Categorical Variable

New variable: TYPE

```
> table(TYPE)
TYPE
COUPE SEDAN
   23    42
```

Predictor Variables: AGE, ODOMETER, TYPE
Outcome Variable: PRICE

```
> lmMLR2 = lm(PRICE~ODOMETER + AGE + TYPE)

> summary(lmMLR2)
```

17. Interpretation

Multiple R-squared = .888. 88.8% of the variation in Honda prices can be attributed to the mileage on the odometer, the age of the car, and the type of the car.

F-statistic = 161.6, p-value < 2.2e-16.

```
Call:
lm(formula = PRICE ~ ODOMETER + AGE + TYPE)

Residuals:
    Min      1Q  Median      3Q     Max
-2341.10 -869.73  -29.39  711.20 2561.36
```

b-values found under Estimate column. For every 1000-mile increase in the odometer reading, we would expect the price of the car to decrease by $50.95. For every one-year increase in age, we would expect the price of the car to decrease by $580.93. We would expect a sedan to cost $114 more than a coupe. t-values are significant at the 0.001 level, indicating that both odometer and age are significant predictors of car prices.	Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 15235.884 508.875 29.940 < 2e-16 *** ODOMETER -49.951 6.271 -7.965 5.05e-11 *** AGE -580.930 52.432 -11.080 3.05e-16 *** TYPESEDAN 113.557 311.855 0.364 0.717 --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 1191 on 61 degrees of freedom Multiple R-squared: 0.8882, Adjusted R-squared: 0.8827 F-statistic: 161.6 on 3 and 61 DF, p-value: < 2.2e-16
18. Obtaining Standardized Beta Estimates Create data frame that includes three predictor variables. Age of car is a stronger predictor of car price than odometer value. As age increases by 1 standard deviation, the price decreases by 0.605 standard deviations. As the odometer reading increases by 1 standard deviation, the price decreases by 0.438 standard deviations.	**>lm.beta(lmMLR2)** ODOMETER AGE TYPESEDAN -0.4380576 -0.6053604 NA