Simple and Multiple Linear Regression

Josh Jarvey

06/12/2020

Knit a Word file from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word file.

## Exercise 1

The data for this problem comes from a dataset presented in Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), “A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich,” Journal of the American Medical Association, 268, 1578-1580. Body temperature (in degrees Fahrenheit) and heart rate (in beats per minute) were two variables that were measured for a random sample of 130 adults. A simple linear regression was used to see if body temperature had an effect on heart rate.

The data are in the file normtemp.rda in the DS705data package, this data is included in the DS705data package so you can access it by loading the package and typing data(normtemp).

### Part 1a

Create a scatterplot with heart rate in the vertical axis and plot the estimated linear regression line in the scatterplot. Include descriptive labels for the x and y-axes (not just the variable names as they are in the data file).

Note: this data set needs a little cleaning first. The heart rates are missing for two of the rows. You can delete these rows from the data frame using the R function na.omit(). Just put the name of the data frame in the parenthesis.

Does it appear that a linear model is at least possibly a plausible model for the relationship between hear rate and body temperature? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

library(DS705data)  
library(ggformula)

## Loading required package: ggplot2

## Loading required package: ggstance

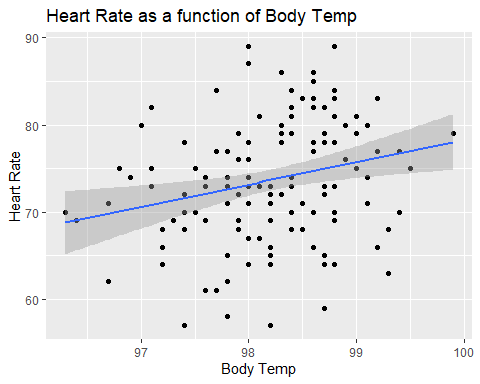
##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

data("normtemp")  
normtemp = na.omit(normtemp)  
  
gf\_point(hr~temp,data = normtemp,title = "Heart Rate as a function of Body Temp",xlab = "Body Temp", ylab = "Heart Rate") + geom\_smooth(method = "lm")

## `geom\_smooth()` using formula 'y ~ x'



From a visual perspective, there does appear to be a weak positive linear pattern between heart rate and body temp. It does seem pretty weak however.

### Part 1b

Write the statistical model for estimating heart rate from body temperature, define each term in the model in the context of this application, and write the model assumptions. (Note: the statistical model is the underlying true, but unknown, model for the population that includes the error or noise term. The model obtained in 1c, is our estimate, obtained using least-squares regression, of the the deterministic (non-random) part of the true model.)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

y\_hat = beta0 + beta1 \* x + epsilon

y\_hat: The predicted heart rate beta0: The intercept heart rate (when there is no body temp) beta1: The coefficient of body temp x: The body temp provided epsilon: The unknown error

Assumptions to be made: 1. Errors have a mean of 0 2. Errors have the same variance for all the x’s 3. Errors are independent from each other 4. Errors are normally distributed

### Part 1c

Obtain the estimated slope and y-intercept for the estimated regression equation and write the equation in the form hr (only with and replaced with the numerical estimates from your R output).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

#create a linear regression between heart rate vs. temp.   
model = lm(hr~temp, data = normtemp)  
 #grab the coefficients.  
model$coefficients

## (Intercept) temp   
## -179.119290 2.574157

Replace the ## symbols with your slope and intercept.

= -179.12 + 2.57temp

### Part 1d

Test whether or not a positive linear relationship exists between heart rate and body temperature using a 5% level of significance. State the null and alternative hypotheses, test statistic, the p-value, and conclusion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

#check summary of the model.  
summary(model)

##   
## Call:  
## lm(formula = hr ~ temp, data = normtemp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.6629 -4.7421 0.3816 4.8519 15.8519   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -179.1193 87.8417 -2.039 0.0435 \*   
## temp 2.5742 0.8944 2.878 0.0047 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.906 on 126 degrees of freedom  
## Multiple R-squared: 0.06169, Adjusted R-squared: 0.05424   
## F-statistic: 8.284 on 1 and 126 DF, p-value: 0.004699

H0: There is no linear relationship between heart rate and body temperature (beta1 = 0) Ha: There is a linear relationship between heart rate and body temperature (beta1 != 0)

t-statistic = 2.878 p-value = 0.0047

## At a 95% level of significance, there is enough evidence to claim that there is a linear relationship between the heart rate for all adults and body temperature (p=0.0047).

### Part 1e

Provide a 95% confidence interval to estimate the slope of the regression equation and interpret the interval in the context of the application (do not us the word “slope” in your interpretation).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

#generate a 95% confidence interval for the variables in my model.  
confint(model)

## 2.5 % 97.5 %  
## (Intercept) -352.9554566 -5.283124  
## temp 0.8042554 4.344058

At a 95% level of confidence, the population mean heart rate for all adults increases from 0.80 to 4.34 bpm for each additional degree in body temperature.

### Part 1f

Provide a 95% confidence interval to estimate the mean heart rate for all adults with body temperature F. Interpret the interval in the context of the problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

#using the model, predict a temp of 98.6 degrees and provide a confidence interval.  
predict(model, data.frame(temp=98.6), interval = "confidence")

## fit lwr upr  
## 1 74.69257 73.30616 76.07897

At a 95% level of confidence, the population mean heart rate for all adults with a body temperature of 98.6 degrees is between 73.30 and 76.08 bpm.

### Part 1g

Provide a 95% prediction interval to estimate the expected heart rate for a randomly selected adult with body temperature F. Interpret the interval in the context of the problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g -|-|-|-|-|-|-|-|-|-|-|-

#using the model, predict a temp of 98.6 degrees and provide a prediction interval.  
predict(model, data.frame(temp=98.6), interval = "prediction")

## fit lwr upr  
## 1 74.69257 60.95531 88.42982

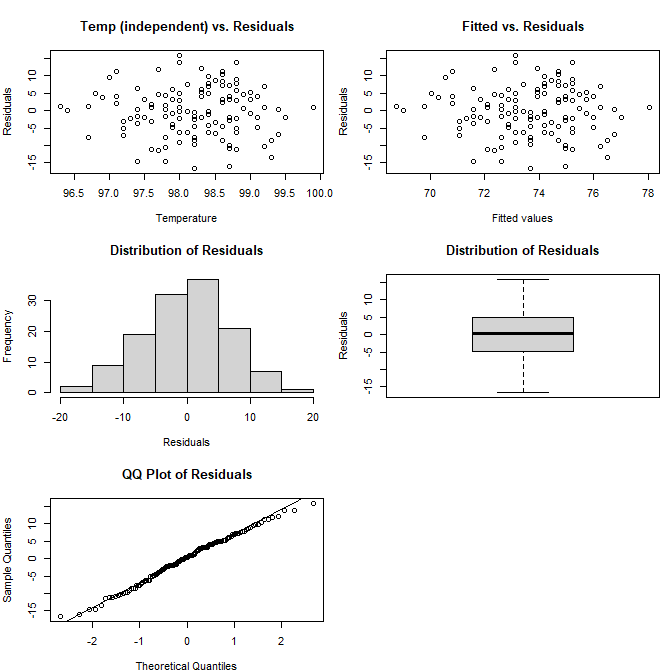
At a 95% level of confidence, the predicted heart rate for all adults with a body temperature of 98.6 degrees is between 60.96 and 88.43 bpm.

### Part 1h

Obtain the residuals and plot them against the predicted values and also against the independent variable. Also construct a histogram, normal probability plot, and boxplot of the residuals and perform a Shapiro-Wilk test for normality. Based on your observation of the plot of residuals against the predicted values, does the regression line appear to be a good fit? Do the model assumptions appear to be satisfied? Comment.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1h -|-|-|-|-|-|-|-|-|-|-|-

##   
## Shapiro-Wilk normality test  
##   
## data: resids  
## W = 0.99124, p-value = 0.6027



The linear model does appear to be a good fit since the residuals are have equal variance given both the independent variable and the fitted values. Additionally, the residuals appear to be normally distributed by both inspecting visually (histogram, boxplot, QQ Plot) and with a Shapiro test (p=0.6027).

### Part 1i

Examine the original scatterplot and the residual plot. Do any observations appear to be influential or be high leverage points? If so, describe them and what effect they appear to be having on the estimated regression equation.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1i -|-|-|-|-|-|-|-|-|-|-|-

## There doesn’t appear to be any influential or high leverage data points when reviewing the original scatterplot or the residual plot.

### Part 1j

Perform the F test to determine whether there is lack of fit in the linear regression function for predicting heart rate from body temperature. Use . State the null and alternative hypotheses, test statistic, the p-value, and the conclusion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1j -|-|-|-|-|-|-|-|-|-|-|-

#creating a "factorized" model using the independent variable  
model\_factor = lm(hr~factor(temp), data = normtemp)  
 #conducting an anova test on the base linear model and the factorized model. The base model is "nested" within the factorized model, so thats why this will work. This is an F-Teest for lack of fit. We hope to see a higher pvalue so we can keep with the basic linear model and keep it simple.   
anova(model, model\_factor)

## Analysis of Variance Table  
##   
## Model 1: hr ~ temp  
## Model 2: hr ~ factor(temp)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 126 6009.6   
## 2 96 4177.4 30 1832.2 1.4035 0.1103

H0: The simple linear model does a good job describing the variance in the response variable Ha: The simple linear model does not do a good job describing the variance in the response variable

F-statistic = 1.4035 p-value = 0.1103

At a 95% level of significance, there is not enough evidence to claim that the linear model dosen’t do a good job describing the variance in the response variable (p=0.1103).

### Part 1k

Conduct the Breusch-Pagan test for the constancy of error variance. Use α = 0.05. State the null and alternative hypotheses, test statistic, the P-value, and the conclusion.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1k -|-|-|-|-|-|-|-|-|-|-|-

#import the package to access bptest  
library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

#conduct the bptest() on the model to understand if the variances are equal vs. not equal. We want variances to be equal (and not reject the null).  
bptest(model)

##   
## studentized Breusch-Pagan test  
##   
## data: model  
## BP = 0.19584, df = 1, p-value = 0.6581

H0: The linear model has equal variance. Ha: The linear model has unequal variance.

BP-statistic = 0.19584 p-value = 0.6581

At a 95% level of significance, there is not enough evidence to claim that the linear model has unequal variance (p=0.6581).

### Part 1l

Calculate and interpret the Pearson correlation coefficient .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1l -|-|-|-|-|-|-|-|-|-|-|-

#completing a correlation test for the temp and heart rate. Sample correlation is listed below  
cor.test(normtemp$temp,normtemp$hr, method = "pearson")$estimate

## cor   
## 0.2483778

The sample correlation between body temperature and heart rate is ~+0.2484. This indicates a weak positive linear relationship between these two variables.

### Part 1m

Construct a 95% confidence interval for the Pearson correlation coefficient .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1m -|-|-|-|-|-|-|-|-|-|-|-

#pull the 95% confidence interval for the correlation.   
cor.test(normtemp$temp,normtemp$hr, method = "pearson")$conf.int

## [1] 0.07821862 0.40447500  
## attr(,"conf.level")  
## [1] 0.95

### Part 1n

Calculate and interpret the coefficient of determination (same as ).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1n -|-|-|-|-|-|-|-|-|-|-|-

(cor.test(normtemp$temp,normtemp$hr, method = "pearson")$estimate)^2

## cor   
## 0.06169154

The coefficient of determination between body temperature and heart rate, R^2, is only ~0.0617. This means that body temperature only explains about 6.17% of the variation in heart rate. This is pretty low, so there must be other variables that account for heart rate.

### Part 1o

Should the regression equation obtained for heart rate and temperature be used for making predictions? Explain your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1o -|-|-|-|-|-|-|-|-|-|-|-

No, I would not use this model for making predictions of heart rate. With body temp only accounting for 6.17% of the heart rate data, this doesn’t seem like it is enough to make solid predictions, and there are probably other variables that should be considered.

### Part 1p

Calculate the Spearman correlation coefficient (just for practice).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1p -|-|-|-|-|-|-|-|-|-|-|-

#calculating the spearman "Rank" correlation test.   
cor.test(normtemp$temp,normtemp$hr, method = "spearman")$estimate

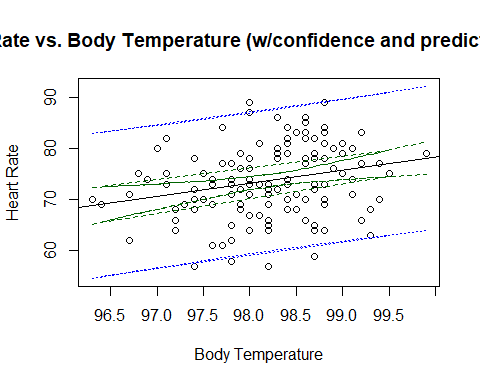
## Warning in cor.test.default(normtemp$temp, normtemp$hr, method = "spearman"):  
## Cannot compute exact p-value with ties

## rho   
## 0.2714866

### Part 1q

Create 95% prediction and confidence limits for the predicted mean heartrate for each temperature given in the sample data and plot them along with a scatterplot of the data. (Look for the slides titled “Confidence Bands” in the presentation.)

#put body temps into their own dataframe for prediction.  
xplot = data.frame(temp)  
 #create 2 datasets - one predicting mean heart rate (confidence band), one predicting point estimate heart rate (prediction band)  
fittedC = predict(model,xplot,interval = "confidence")  
fittedP = predict(model,xplot,interval = "prediction")  
 #set the y limits of the plot to be the max of the upper value in the prediction band, and the minimum of the lower value in the prediction band.   
ylimits = c(min(fittedP[,"lwr"]),max(fittedP[,"upr"]))  
  
 #plot the dataset and add a the model's linear line.  
plot(normtemp$temp,normtemp$hr,ylim = ylimits, main = "Heart Rate vs. Body Temperature (w/confidence and prediction bands)", ylab = "Heart Rate", xlab = "Body Temperature")  
abline(model)  
  
 #now add the "bands" using the upper values as the top and lower values as the bottom for each the confidence and prediction bands.  
lines(xplot$temp, fittedC[,"lwr"], lty="dashed",col="darkgreen")  
lines(xplot$temp, fittedC[,"upr"], lty="dashed",col="darkgreen")  
lines(xplot$temp, fittedP[,"lwr"], lty="dotted",col="blue")  
lines(xplot$temp, fittedP[,"upr"], lty="dotted",col="blue")



## Exercise 2

A personnel officer in a governmental agency administered three newly developed aptitude tests to a random sample of 25 applicants for entry-level positions in the agency. For the purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job.

The scores on the three tests (x1, x2, x3) and the job proficiency score (y) for the 25 employees are in the file JobProf.rda.

(Based on an exercise from Applied Linear Statistical Models, 5th ed. by Kutner, Nachtsheim, Neter, & Li)

### Part 2a

We’d like to explore using interaction terms in a statistical model including the three first-order terms and the three cross-product interaction terms:

Use R to find the corresponding estimated model and also obtain the summary().

## -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

#load the data  
library(DS705data)  
data("JobProf")  
  
 #create a model with cross-products; show summary.  
model2 = lm(y~x1+x2+x3+x1\*x2+x1\*x3+x2\*x3, data = JobProf)  
summary(model2)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x1 \* x2 + x1 \* x3 + x2 \* x3,   
## data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.513 -3.408 -1.082 2.548 11.593   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -48.965067 142.039396 -0.345 0.734  
## x1 -0.580916 0.820429 -0.708 0.488  
## x2 -0.174913 0.905654 -0.193 0.849  
## x3 1.443371 1.495901 0.965 0.347  
## x1:x2 0.004012 0.004341 0.924 0.368  
## x1:x3 0.004959 0.008893 0.558 0.584  
## x2:x3 -0.002015 0.008399 -0.240 0.813  
##   
## Residual standard error: 5.431 on 18 degrees of freedom  
## Multiple R-squared: 0.9414, Adjusted R-squared: 0.9218   
## F-statistic: 48.17 on 6 and 18 DF, p-value: 4.042e-10

### Part 2b

Use R to compute the VIF for each term in the model. Are any of the VIFs over 10? (We need to add this into Lesson 6, but it’s covered in the Lesson 8 Swirl - I’ve put an example in the chunk below. Replace the chunk with code to find the VIF’s for this model.)

## -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

#load the vif() function from the HH library. Use it to check for collinearity among variables.   
library(HH)

## Loading required package: lattice

## Loading required package: grid

## Loading required package: latticeExtra

##   
## Attaching package: 'latticeExtra'

## The following object is masked from 'package:ggplot2':  
##   
## layer

## Loading required package: multcomp

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

## Loading required package: gridExtra

vif(model2)

## x1 x2 x3 x1:x2 x1:x3 x2:x3   
## 225.6691 199.6007 142.7966 138.0512 368.6751 308.2454

Every variable within this model has a VIF score greater than 10. This indicates strong collinearity among the variables (as they must be correlated with each other (which would make sense because they are from the same person).

### Part 2c

The model from 2a is suffering from the effects of collinearity (which you should see in 2b), which inflates the standard errors of the estimated coefficients.

Using the model summary from 2a what do you notice about the overall model p-value (from the F-statistic) and the individual p-values for each term in the model? Does it make sense that the overall model shows statistical significance but no individual term does?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

The overall F-test for the model comes back as significant (p=4.042e-10) indicating that this model does have value in explaining the dependent (job performance) variable. Although the independent variables (or combinations of such) do not appear significant at the individual, together they do have value in explaining the dependent (as mentioned in the previous sentence). It is ok that none of the independents show significance but the model does - this just means the model cannot discern which independent variable is specifically adding the value.

### Part 2d

Use R to estimate and summary() the first order model corresponding to

Is the first order model significant?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

#create a 3rd model with just the first order terms.   
model3 = lm(y~x1+x2+x3, data = JobProf)  
summary(model3)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.7517 -3.0371 -0.4618 1.8358 11.7315   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -127.77378 12.88053 -9.920 2.23e-09 \*\*\*  
## x1 0.34813 0.05451 6.387 2.48e-06 \*\*\*  
## x2 0.04353 0.07362 0.591 0.561   
## x3 1.77921 0.14541 12.236 5.08e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.331 on 21 degrees of freedom  
## Multiple R-squared: 0.9341, Adjusted R-squared: 0.9247   
## F-statistic: 99.21 on 3 and 21 DF, p-value: 1.457e-12

This first-order model (without any interaction effects) is significant with the representative dataset (p=1.457e-12)

### Part 2e

Do the interaction terms in 2a really add anything significant beyond the first order model in 2d? Now we’ll compare the models with and without interaction terms to see if the interaction terms make a statistically significant improvement to the fit of our models.

Test the significance of all three coefficients for the interaction terms as a subset by using anova() to compare the model from Part 2a to the first order model from Part 2d. Use a 5% level of significance. State and and provide the R output as well as a written conclusion which includes the P-value. Should we keep the interaction terms?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e -|-|-|-|-|-|-|-|-|-|-|-

#use anova to test model3 (simpler model) vs. model 2 (all inclusive, more complex model).  
anova(model3,model2)

## Analysis of Variance Table  
##   
## Model 1: y ~ x1 + x2 + x3  
## Model 2: y ~ x1 + x2 + x3 + x1 \* x2 + x1 \* x3 + x2 \* x3  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 21 596.72   
## 2 18 530.86 3 65.861 0.7444 0.5395

H0: The simpler model is sufficient. Ha: The more complex model is required.

## At a 95% level of significance, there is not enough evidence to suggest that the more complex model is required for this dataset (p=0.5395).

### Part 2f

There are more methodical approaches to exploring different models that we’ll learn about in a later lesson, but we’ll try one more model here to get a bit more experience. In this case we’ll add a quadratic term . To do this you’ll want to create a new variable x2sq = x2^2 and include it in your model. Use R to estimate and summarize() the model corresponding to:

Examine the p-value corresponding to the quadratic term. If the quadratic term is significant at significance level , then according to the hierarchical approach we should retain it and the term. If it isn’t significant, then we won’t retain it but we’ll have to evaluate the significance of the term separately.

Should the quadratic term be retained in the model at a 5% level of significance?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2f -|-|-|-|-|-|-|-|-|-|-|-

#create a quadratic version of the x2 variable  
x2sq = JobProf$x2^2  
 #fit a new model using the x2sq  
model4 = lm(y~x1+x2+x3+x2sq, data = JobProf)  
summary(model4)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x2sq, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.852 -2.724 -0.918 1.956 10.071   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -84.293186 47.470221 -1.776 0.091 .   
## x1 0.340887 0.055159 6.180 4.89e-06 \*\*\*  
## x2 -0.891696 0.985329 -0.905 0.376   
## x3 1.820764 0.152130 11.968 1.42e-10 \*\*\*  
## x2sq 0.004530 0.004759 0.952 0.353   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.343 on 20 degrees of freedom  
## Multiple R-squared: 0.9369, Adjusted R-squared: 0.9243   
## F-statistic: 74.3 on 4 and 20 DF, p-value: 1.03e-11

We should not keep the quadratic term because it does not appear to be significant (p=0.3530)

### Part 2g

If you’ve been successful so far, then you should realize that the none of interaction terms nor the quadratic term have been significant (if you concluded otherwise, then review your work). This brings us back to the first order model in Part 2d. Look at that model summary again. There should be one term that is insignificant so omit it and use R to estimate our final and smaller first order model.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g -|-|-|-|-|-|-|-|-|-|-|-

#create a 5th model without the use of x2 (because it was non-significant)  
model5 = lm(y~x1+x3, data = JobProf)  
summary(model5)

##   
## Call:  
## lm(formula = y ~ x1 + x3, data = JobProf)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3489 -2.8086 -0.4546 2.8981 12.6469   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -127.59569 12.68526 -10.06 1.09e-09 \*\*\*  
## x1 0.34846 0.05369 6.49 1.58e-06 \*\*\*  
## x3 1.82321 0.12307 14.81 6.31e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.251 on 22 degrees of freedom  
## Multiple R-squared: 0.933, Adjusted R-squared: 0.9269   
## F-statistic: 153.2 on 2 and 22 DF, p-value: 1.222e-13

### Part 2h

From the final first order model in 2g, obtain a 90% confidence interval for the coefficient of and interpret it in the context of this problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g -|-|-|-|-|-|-|-|-|-|-|-

#calculate confidence intervals for the model terms.  
confint(model5, level = 0.90)

## 5 % 95 %  
## (Intercept) -149.3781166 -105.8132586  
## x1 0.2562572 0.4406579  
## x3 1.6118704 2.0345407

At a 90% level of confidence, the true coefficient for the x3 explanatory variable is between 1.612 and 2.035.

### Part 2i

Using the final first order model from 2g, construct a 95% prediction interval for a randomly selected employee with aptitude scores of and to forecast their proficiency rating at the end of the probationary period. Write an interpretation for the interval in the context of this problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2i -|-|-|-|-|-|-|-|-|-|-|-

#create a dataframe of scores 99,112,105. Predict proficiency rating using these scores using model5, and give a prediction interval along with it.  
predict(model5, data.frame(x1=99,x2=112,x3=105),interval = "prediction")

## fit lwr upr  
## 1 98.33819 87.16155 109.5148

At a 95% level of confidence, we expect an employee with a score of 99 on their first aptitude exam, 112 on their second aptitude exam, and 105 on their third aptitude exam, to have a proficiency rating between 87.16 and 109.51 at the end of their probationary period.

## Exercise 3

A research professor in a leading department of education is studying three different methods of teaching English as a second language. After three months in the program the participants take an exam and let be the score on the exam. The following model was used to assess the efficiencies of the three methods

where

and

### Part 3a

Use data in the file English.rda in the DS705data package to estimate the coefficients for the model:

Obtain the estimated intercept and coefficients and state the estimated mean English proficiency scores for each of the 3 methods of teaching English as a second language.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

data("English")  
model6 = lm(y~x1+x2,data = English)  
summary(model6)

##   
## Call:  
## lm(formula = y ~ x1 + x2, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.150 -5.713 -0.225 4.850 34.850   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.750 2.202 20.325 <2e-16 \*\*\*  
## x1 61.400 3.114 19.719 <2e-16 \*\*\*  
## x2 3.950 3.114 1.269 0.21   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.847 on 57 degrees of freedom  
## Multiple R-squared: 0.8953, Adjusted R-squared: 0.8916   
## F-statistic: 243.6 on 2 and 57 DF, p-value: < 2.2e-16

Method 1 = Mean English proficiency score = 44.75. (This is just the intercept by itself; x1=0 and x2=0) Method 2 = Mean English proficiency score = 106.15. (This is the intercept + x1; x2=0) Method 3 = Mean English proficiency score = 48.70. (This is the intercept + x2; x1=0; althought it doesn’t appear significant and therefore we couldnt say for sure it would actually add anything). —

### Part 3b

The researcher has given each participant a test prior to the beginning of the study and obtained an index, , of the participant’s English proficiency.

Using the estimated coefficients, write three separate estimated models, one for each method, relating the scores after 3 months in the program () to the index score prior to starting the program ().

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

#create models of methods 1,2,3, with independent variables removes as necessary since it's categorical.   
method1 = lm(y~x4, data = English)  
method2 = lm(y~x4+x1+x1\*x4, data = English)  
method3 = lm(y~x4+x2+x2\*x4, data = English)  
summary(method1)

##   
## Call:  
## lm(formula = y ~ x4, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -36.88 -23.08 -12.65 32.98 49.74   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.6160 29.7988 0.256 0.799   
## x4 1.3069 0.6557 1.993 0.051 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.18 on 58 degrees of freedom  
## Multiple R-squared: 0.0641, Adjusted R-squared: 0.04797   
## F-statistic: 3.973 on 1 and 58 DF, p-value: 0.05095

summary(method2)

##   
## Call:  
## lm(formula = y ~ x4 + x1 + x1 \* x4, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.6498 -5.3968 -0.1468 4.0783 18.8245   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.9763 9.3615 3.736 0.00044 \*\*\*  
## x4 0.2628 0.2078 1.265 0.21117   
## x1 -16.0193 15.8448 -1.011 0.31636   
## x4:x1 1.6389 0.3459 4.738 1.52e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.36 on 56 degrees of freedom  
## Multiple R-squared: 0.9425, Adjusted R-squared: 0.9394   
## F-statistic: 306 on 3 and 56 DF, p-value: < 2.2e-16

summary(method3)

##   
## Call:  
## lm(formula = y ~ x4 + x2 + x2 \* x4, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -46.712 -24.792 1.267 25.109 39.155   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.9148 33.9544 0.233 0.8165   
## x4 1.4867 0.7416 2.005 0.0498 \*  
## x2 21.8776 57.7188 0.379 0.7061   
## x4:x2 -1.0609 1.2794 -0.829 0.4105   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.8 on 56 degrees of freedom  
## Multiple R-squared: 0.2378, Adjusted R-squared: 0.1969   
## F-statistic: 5.823 on 3 and 56 DF, p-value: 0.00155

Method 1: English Proficiency Score (y) ~= 7.6160 + 1.3069x4 + error Method 2: English Proficiency Score (y) ~= 34.9763 + 0.2628x4 - 16.0193x1 + 1.6389x4x1 + error Method 3: English Proficiency Score (y) ~= 7.9148 + 1.4867x4 + 21.8776x2 - 1.0609x4x2 + error

Ultimately I would select model 2 here (that reflects method 2) since not only does it appear to have overall significance, but it also has a high Adjusted R^2 value, explaining that it accounts for most of the variability within this dataset. Model 3 is significant, but its R^2 is pretty low.