Simple and Multiple Linear Regression

Matt Allen

07/04/2018

Knit a Word file from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word file via D2L Dropbox.

## 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.

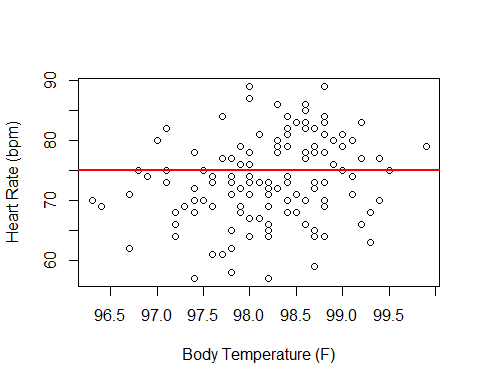
Does it appear that a linear model is at least possibly a plausible model for the relationship between heart rate and body temperature? Explain your answer.

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

# Get data  
require(DS705data)

## Loading required package: DS705data

data("normtemp")  
# Remove Missing Values  
normtemp <- na.omit(normtemp)  
loess\_line <- loess(hr~temp,normtemp)  
{  
 # Create Scatter Plot  
 plot(normtemp$hr~normtemp$temp, xlab ="Body Temperature (F)", ylab = "Heart Rate (bpm)")   
 lines(predict(loess\_line), col='red', lwd=2)  
}



From visual inspection of the scatter plot, the points appear to be scattered randomly, but it is possible that there is a positive relationship between body temperature and heart rate. Based on the loess line, however, the relationship between heart rate and body temperature appears to be random. The line has zero slope, so a linear model should be created to confirm.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

where is body temperature in degrees Fahrenheit, is heart rate in beats per minute, is the heart rate when body temperature is 0, and is the rate of increase in heart rate for an increase of 1 degree in body temperature.

Model Assumptions:

1. errors have mean 0
2. errors have the same variance for all
3. errors are independent of 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 -|-|-|-|-|-|-|-|-|-|-|-

# Obtain estimated slope and y-intercept.  
linear.heartratetemp.model<-with(normtemp,lm(hr~temp))  
linear.heartratetemp.model

##   
## Call:  
## lm(formula = hr ~ temp)  
##   
## Coefficients:  
## (Intercept) temp   
## -179.119 2.574

Based on the output from the model, the intercept and coefficient of the estimated regression equation is

= -179.119 + 2.574 \* temp

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Summary of linear model  
summary(linear.heartratetemp.model)

##   
## Call:  
## lm(formula = hr ~ temp)  
##   
## 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

The test statistic is t = 2.878. The p-value is 0.0047. At a 5% level of significance, there is evidence that a positive relationship exists between heart rate and body temperature.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# 95% confidence interval of rate of change heart rate for 1 degree change in temperature  
confint(linear.heartratetemp.model)

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

We are 95% confident that the population mean of heart rate increases 0.80 to 4.34 beats per minute for a one degree increase in 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 -|-|-|-|-|-|-|-|-|-|-|-

# Compute confidence interval  
temp\_df <- data.frame( temp = 98.6 )  
predict(linear.heartratetemp.model, temp\_df , interval="confidence")

## fit lwr upr  
## 1 74.69257 73.30616 76.07897

We are 95% confident that, for a body temperature of F, the average heart rate is between 73.3 and 76.08 beats per minute.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Compute prediction interval  
temp\_df <- data.frame( temp = 98.6 )  
predict(linear.heartratetemp.model, temp\_df, interval="prediction")

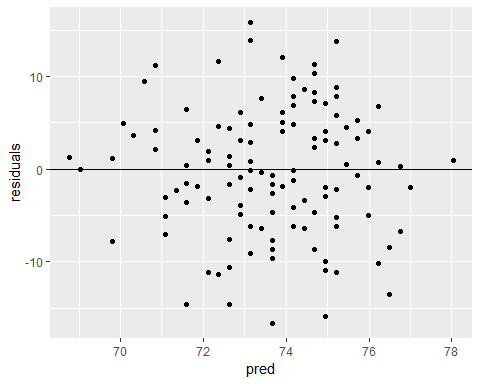
## fit lwr upr  
## 1 74.69257 60.95531 88.42982

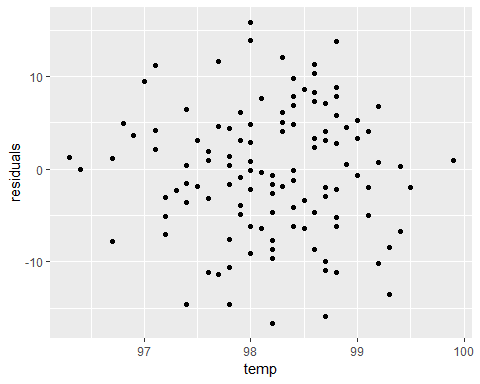
We are 95% confident that, for a body temperature of F of a randomly selected adult, their heart rate will be between 60.96 and 88.43 beats per minute.

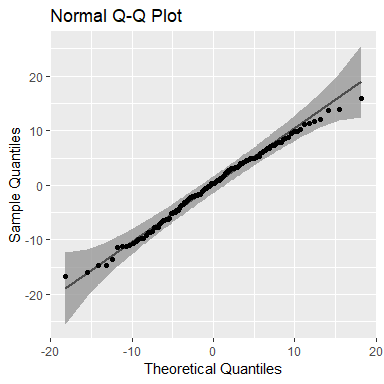
### Part 1h

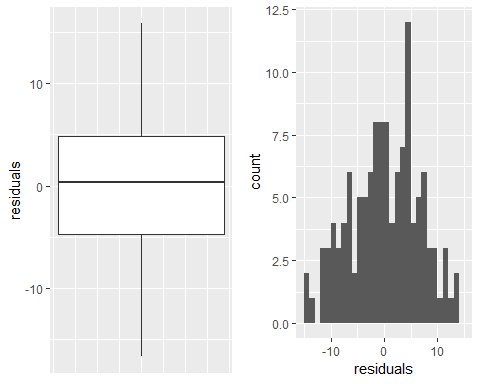
Obtain the residuals and plot them against the predicted values and also against the independent variable. Also construct a histogram, normal probability plot, and boxplot of the residuals and perform a Shapiro-Wilk test for normality. Based on your observation of the plot of residuals against the predicted 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

#Check that the mean of the residuals is 0.  
mean(resids)

## [1] -3.249557e-17

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?

Model Assumptions:

1. errors have mean 0, look at boxplot, histogram or take mean of residuals.
2. errors have the same variance for all
3. errors are independent of each other
4. errors are normally distributed.

From the plot of residuals against the predicted values, it can be seen that the residuals are scattered evenly around the zero horizontal line. This suggests that the relationship is linear, and that the errors have the same variance. The errors appear scattered, which would suggest that the errors are independent of each other. The mean of the reisuals is close to zero. Also, based on the sharpiro-wilk test, we fail to reject that the residuals are normally distributed. We can conclude that the regression line does appear to be a good fit and that the model assumptions are satisfied.

### 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 are high and low points in the middle. I think that there strong influence in high and low direction cancels out an effect on the estimated regression equation. However, due to greater variability, it may have some influence on the confidence interval.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Create full model.  
heartratetemp.model.full<-with(normtemp,lm(hr~factor(temp)))  
results <- anova( linear.heartratetemp.model, heartratetemp.model.full )  
#results$F  
#results$'Pr(>F)'  
results

## 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

$$ H\_0: \mbox{line model}, \hspace{1in} H\_a: \mbox{full model} $$

.

The F-statistic 1.403484 and the p-value is 0.1103044. At a 5% level of significance, we fail to reject the null hypothesis that the line model is a good fit for the data.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Breusch-Pagan test for the constancy of error variance  
require(lmtest) # install if needed

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.4.4

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.4.4

##   
## Attaching package: 'zoo'

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

bptest(linear.heartratetemp.model)

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

$$ H\_0: \mbox{equal variances}, \hspace{.5in} H\_1: \mbox{unequal variances}$$

The BP statistic is 0.19584 and the p-value is 0.6581. At a 5% level of significance, we fail to reject the null hypothesis that the model has equal variances.

### Part 1l

Calculate and interpret the Pearson correlation coefficient .

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

# Calculate and interpret Pearson correlation coefficient  
correlation.test <- with(normtemp, cor.test(temp, hr, method="pearson") )  
correlation.test$estimate

## cor   
## 0.2483778

The pearson correlation coefficient of .25 indicates a weak positive relationship between body temperature and heart rate.

### Part 1m

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

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

# Construct a 95% CI for the Pearson correlation coefficient.  
correlation.test$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 -|-|-|-|-|-|-|-|-|-|-|-

##   
## Call:  
## lm(formula = hr ~ temp)  
##   
## 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

The adjusted R-squared is 0.05424. That means only 5.4% of the variation is explained by the model.

### Part 1o

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

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

Based on the adjusted R squared, the model does not have signficant predictive powers, and probably should not be used to predict heart rate. However, there is some explanatory power in temperature to heart rate, and the model could be updated to include more factors like age for example in order to make better predictions.

### Part 1p

Calculate the Spearman correlation coefficient (just for practice).

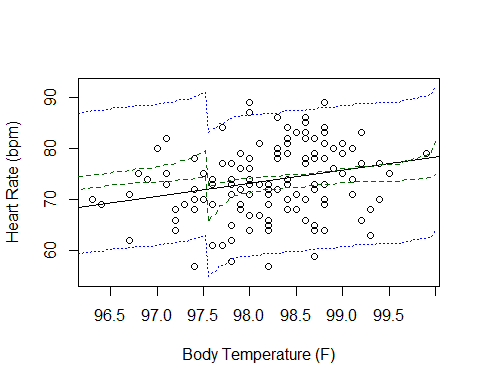
### -|-|-|-|-|-|-|-|-|-|-|- Answer 1p -|-|-|-|-|-|-|-|-|-|-|-

## Warning in cor.test.default(temp, 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.)



## 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 -|-|-|-|-|-|-|-|-|-|-|-

# Get the data  
data("JobProf")  
linear.model <- lm(y~x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3, data=JobProf)  
summary(linear.model)

##   
## 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 -|-|-|-|-|-|-|-|-|-|-|-

## Loading required package: HH

## Warning: package 'HH' was built under R version 3.4.4

## Loading required package: lattice

## Loading required package: latticeExtra

## Warning: package 'latticeExtra' was built under R version 3.4.4

## Loading required package: RColorBrewer

##   
## Attaching package: 'latticeExtra'

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

## Loading required package: multcomp

## Warning: package 'multcomp' was built under R version 3.4.4

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.4.4

## Loading required package: survival

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 3.4.4

## Loading required package: MASS

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

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

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

All of the VIFs are greater than 10. There is a high degree of collinearity.

### 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 model p-value from the F-statistic is 4.042e-10 which indicates a very good fit, but the individual p-values for each term is large indicating that none of the individual terms are signficant. The small p-value on the F statistic implies overall the predictors together are significant in explaining the variation in the predictor variable. The high p-values on individual terms are caused by the high degree of collinearity between the terms. Due to the high degree of collinearity and after accounting for the other variables, adding a new variable does not explain much more of the variation, and thus has a high p-value.

### Part 2d

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

Is the first order model significant?

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

# Estimate and summarize a first order model.  
linear.model.witoutinteractionterms <- lm(y~x1 + x2 + x3, data=JobProf)  
summary(linear.model.witoutinteractionterms)

##   
## 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 model is more significant since it has a smaller p-value in the F-statistic, and two out of the three terms are now statistically significant.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

#Compare the models  
results <- anova( linear.model, linear.model.witoutinteractionterms )  
#results$F  
#results$'Pr(>F)'  
results

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

$$ H\_0: \mbox{interaction model}, \hspace{1in} H\_a: \mbox{no interaction model} $$

.

The F-statistic 0.7444 and the p-value is 0.5395. At a 5% level of significance, we fail to reject the null hypothesis that the model with interaction terms is a good fit for the data.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Add quadratic term  
x2sq =JobProf$x2^2  
linear.model.with.xsquared <- lm(y~x1 + x2 + x3 + x2sq, data=JobProf)  
summary(linear.model.with.xsquared)

##   
## 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

The p-value for the x2 squared term is 0.353. At at 5% level of signficance, it should not be retained.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Estimate final and smaller model  
linear.model.final.first.order <- lm(y~x1 + x3, data=JobProf)  
summary(linear.model.final.first.order)

##   
## 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 -|-|-|-|-|-|-|-|-|-|-|-

# 90% confidence interval  
confint(linear.model.final.first.order, level = .9)

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

A 90% confidence interval for x3 is [1.61, 2.03]. One additional point on the third aptitude test corresponds to an increase in between 1.61 and 2.03 points in the job proficiency score.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

# Compute prediction interval  
temp\_df <- data.frame( x1=99,x2=112,x3=105 )  
predict(linear.model.final.first.order, temp\_df, interval="prediction")

## fit lwr upr  
## 1 98.33819 87.16155 109.5148

With 95% confidence, a randomly selected employee who scores x1=99, x2=112, and x3=105 on the aptitude tests will have a job proficiency score of between 87 and 110.

## Exercise 3

Consider the scenario from Exercises 12.5 and 12.7 on page 725 of Ott’s textbook. There are two categorical variables (Method and Gender) and one quantitative variable (index of English proficiency prior to the program). See the textbook for details on how the qualitative variables are coded using indicator variables.

### Part 3a

Use data in the file English.rda in the DS705data package to estimate the coefficients for the model in Exercise 12.5:

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 -|-|-|-|-|-|-|-|-|-|-|-

# Get the data and estimate coefficients.  
data("English")  
model <- lm(y~x1 + x2, data=English)  
summary(model)

##   
## 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

The intercept is 44.75. x1 is 61.4. x2 is 3.95. The estimated mean English proficiency score for method 1 is 61.4 + 44.75 = 106.15, method 2 is 3.95 + 44.75 = 48.7, and method 3 is just the value of the intercept 44.75.

### Part 3b

Fit the model for Exercise 12.7:

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

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

# Fit model  
model <- lm(y~x4 + x1 + x2 + x1:x4 + x2:x4, data=English)  
summary(model)

##   
## Call:  
## lm(formula = y ~ x4 + x1 + x2 + x1:x4 + x2:x4, data = English)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.845 -4.696 -0.110 4.178 19.470   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.2585 13.5220 2.903 0.00534 \*\*   
## x4 0.1220 0.2983 0.409 0.68411   
## x1 -20.3014 18.4731 -1.099 0.27666   
## x2 -9.4661 18.4964 -0.512 0.61089   
## x4:x1 1.7797 0.4039 4.407 5.02e-05 \*\*\*  
## x4:x2 0.3038 0.4104 0.740 0.46237   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.246 on 54 degrees of freedom  
## Multiple R-squared: 0.9463, Adjusted R-squared: 0.9413   
## F-statistic: 190.2 on 5 and 54 DF, p-value: < 2.2e-16

Model 1:

Model 2:

Model 3: