Lab Assignment: Simple Linear Regression & Correlation

Christian Acosta

Jul. 25, 2016

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, 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. Find these rows and delete them from the data frame before proceeding. To delete rows 10 and 42 you could do something like this: df <- df[c(-10,-42),].

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

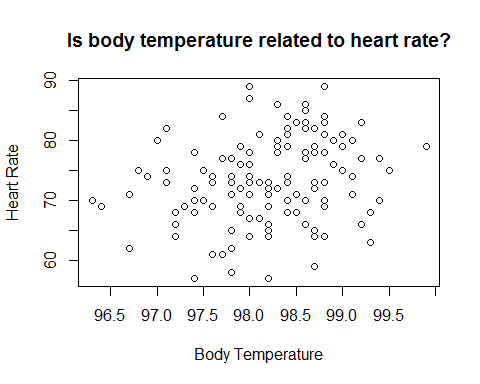
require(DS705data)

## Loading required package: DS705data

data("normtemp")  
head(normtemp); #normtemp #temp, hr

## temp hr  
## 1 96.3 70  
## 2 96.7 71  
## 3 96.9 74  
## 4 97.0 80  
## 5 97.1 73  
## 6 97.1 75

cleanedTemp <- normtemp[c(-129, -130), ]  
#cleanedTemp  
attach(cleanedTemp)  
plot(cleanedTemp, main = "Is body temperature related to heart rate?", xlab = "Body Temperature", ylab = "Heart Rate")



#with(cleanedTemp, plot(temp, hr))

## While this data is probably not going to be very accurate at predicting values based on body temperature, the scatterplot shows that there is at least a potential relationship between body temperature and heart rate.

### 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.

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

y = B0 + B1x

y = predicted heart rate based on body temperature. B0 = y-intercept B1 = slope of line (the rate at which body temperature affects heart rate). x = specific body temperature being used to predict heart rate.

The linear model assumes equal spread of residuals and data normality (which would be affected based on pre-existing conditions of subjects).

## From a practical standpoint, the model assumes that body temperature falls in the 'normal' range for humans provided by the data (96.3 to 99.9). Body temperatures used outside this range may tend to be less accurate - especially given that temperatures outside of this range may show that the subject could be ill.

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

tempHRModel <- lm(hr ~ temp)  
summary(tempHRModel)

##   
## 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

hr = -179.12 + 2.57\*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(tempHRModel)

##   
## 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

Null Hypothesis: There is no relationship (slope = 0) between body temperature and heart rate. Alternative Hypothesis: There is a relationship between body temperature and heart rate.

The test statistic is 8.284.

Conclusion: Reject the null hypothesis at alpha = 0.05.

## There is enough evidence to suggest that there is a positive relationship between body temeperature and heart rate in the overall population (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 âslopeâ in your interpretation).

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

confint(tempHRModel)

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

## We are 95% confident that the population mean heart rate (in beats per minute) increases 0.804 to 4.344 for each one degree increase 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 -|-|-|-|-|-|-|-|-|-|-|-

x <- data.frame(temp=98.6)  
predict(tempHRModel, x, interval="confidence")

## fit lwr upr  
## 1 74.69257 73.30616 76.07897

## We are 95% confident that, for a body temperature of 98.6 degrees, the average heart rate is between 73.31 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 -|-|-|-|-|-|-|-|-|-|-|-

predict(tempHRModel, x, interval="prediction")

## fit lwr upr  
## 1 74.69257 60.95531 88.42982

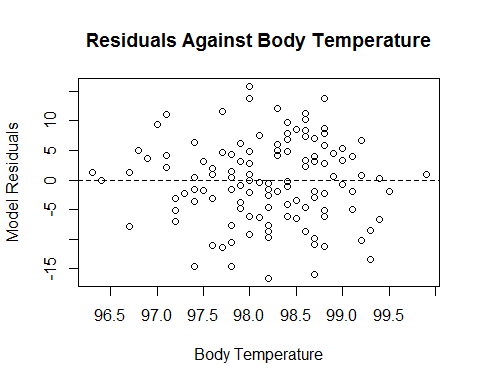
## We are 95% confident that, for a body temperature of 98.6 degrees, the heart rate will be between 60.96 and 88.43 beats per minute.

### Part 1h

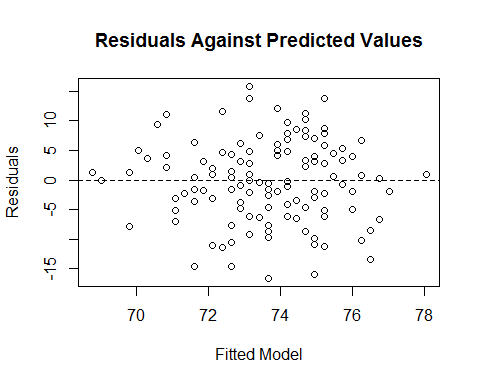
Obtain the residuals and plot them against the predicted values and also against the independent variable. Also construct a histogram, normal probability plot, and boxplot of the residuals and perform a Shapiro-Wilk test for normality. Based on your observation of the plot of residuals against the predicted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

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

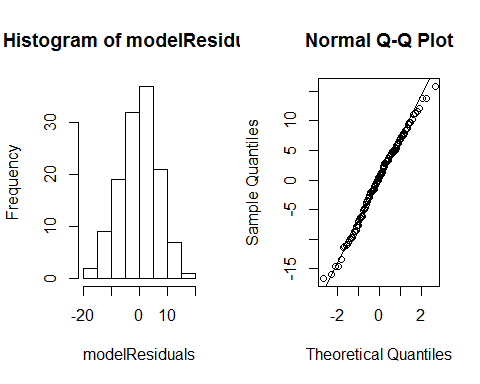
#Obtain residuals, plot against predicted values and against the independant variable.  
modelResiduals <- tempHRModel$residuals  
modelFit <- tempHRModel$fitted.values  
plot(temp, modelResiduals, main = "Residuals Against Body Temperature", xlab = "Body Temperature", ylab = "Model Residuals"); abline(h=0, lty="dashed") #Plot against independant variable temperature.



#modelResiduals  
#temp  
  
plot(modelFit, modelResiduals, main = "Residuals Against Predicted Values", xlab="Fitted Model", ylab="Residuals"); abline(h=0, lty="dashed") #Plot against predicted values



#Construct a histogram, normal probability plot, boxplot of the residuals.  
par(mfrow=c(1,2)); hist(modelResiduals); qqnorm(modelResiduals); qqline(modelResiduals)



#Perform a Shapiro-Wilk test for normality.   
shapiro.test(modelResiduals) #p-value of .6027

##   
## Shapiro-Wilk normality test  
##   
## data: modelResiduals  
## W = 0.9912, p-value = 0.6027

## The regression line appears to be a good fit for the model. Both the independent variable and predicted values vs. residuals appear fairly normally distributed (one many argue that it is a 'diamond' shape). In addition, the histogram of residuals shows a normal bell curve while the Q-Q plot shows relatively close clustering along the Q-Q line. The Shapiro test for normality was inconclusive in that we can be confident that the residuals are 'normal.'

### 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 a few data points that stand out from others. For example, there is one point where the body temperature is 99.9 degrees. While the heart rate for this individual is reasonable at 79 BPM, the linear model 'predicts' that as body temperature rises, heart rate rises. Because this individual data point BPM doesn't represent a high heart rate value it has a 'dampening' effect on the overall model. If this point was removed, I'd expect the slope equation to be higher.

## Another data point that is influential would be the point that lies at approximately (98.9, 89). This point represents the highest observed heart rate with a higher-than-average body temperature. While the point isn't 'extreme', it does increase the slope of the model equation.

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

#REMINDER: Lack of Fit F-test requires x values to have multiple observed y-values (which we do).   
tempHRModel.full <- with(cleanedTemp, lm(hr ~ factor(temp)))  
anova(tempHRModel, tempHRModel.full)

## 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

Null Hypothesis: Full model does NOT explain significantly more of the variance in the response variable (heart rate) than the linear model. Alternative Hypothesis: Full model explains significantly more of the variance in the response variable (heart rate) than the linear model.

Test statistic: 1.4035.

Conclustion: Fail to reject the null hypothesis at alpha = 0.05.

There is not enough evidence to suggest that the full model explains significantly more of the variance in the response variable (heart rate) compared to the linear model (p = 0.1103).

### Part 1k

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

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

#install.packages('lmtest')  
require(lmtest)

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.1.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.1.3

##   
## Attaching package: 'zoo'

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

bptest(tempHRModel)

##   
## studentized Breusch-Pagan test  
##   
## data: tempHRModel  
## BP = 0.1958, df = 1, p-value = 0.6581

Null Hypothesis: Equal Variance in residual plot. Alternative Hypothesis: Unequal Variance in residual plot.

Test statistic: 0.1958.

Conclusion: There is not enough evidence to reject the null hypothesis at alpha = 0.05.

There is insufficient evidence to claim that there is unequal variance in the residual plot in the body temperature/heart rate model.

### Part 1l

Calculate and interpret the Pearson correlation coefficient .

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

with(cleanedTemp, cor.test(temp, hr)$estimate)

## cor   
## 0.2483778

#.248

## The correlation coefficient of 0.248 suggests that there is a weak positive linear relationship between body temperature and heart rate.

### Part 1m

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

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

with(cleanedTemp, cor.test(temp, hr)$conf.int)

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

## We are 95% confident that the correlation coefficient for the linear model rests between 0.078 and 0.404.

### Part 1n

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

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

summary(tempHRModel)$r.squared

## [1] 0.06169154

## Only 6.2% of the variation in heart rate is explained by the linear relationship between body temperature and heart rate. This leaves 93.8% left unexplained by the linear relationship. This is not a particularly strong model.

### Part 1o

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

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

## No. While the model was useful in showing that there was at least SOME relationship between body temperature and heart rate, a poor pearson coefficient value (0.248) and R^2 (6.2% explained variation) suggests that this model is insufficient for predicting heart rate. More investigation is needed into other causal mechanisms that drive increased heart rate.

### Part 1p

Calculate the Spearman correlation coefficient (just for practice).

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

cor.test(temp, hr, method="spearman")

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

##   
## Spearman's rank correlation rho  
##   
## data: temp and hr  
## S = 254618.4, p-value = 0.001936  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## 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.)

par(mfrow=c(1,1))  
confPlot <- data.frame(temp = seq(95, 100, length = 5))  
fittedC <- predict(tempHRModel, confPlot, interval="confidence")  
fittedP <- predict(tempHRModel, confPlot, interval="prediction")  
#?seq  
#Create the scatterplot  
ylims <- c(min(fittedP[,"lwr"]), max(fittedP[,"upr"]))  
plot(temp, hr, ylim = ylims, main = "Does heart rate help predict body temperature?", xlab = "Body Temperature", ylab = "Heart Rate")  
abline(tempHRModel)  
  
#Plotting the confidence and prediction bands.  
newX <- seq(95, 100, length = 5)  
lines(newX, fittedC[,"lwr"], lty = "dashed", col="darkgreen")  
lines(newX, fittedC[,"upr"], lty = "dashed", col="darkgreen")  
lines(newX, fittedP[,"lwr"], lty = "dashed", col="blue")  
lines(newX, fittedP[,"upr"], lty = "dashed", col="blue")

