Lab Assignment: Simple Linear Regression & Correlation

Paul Isaacson

Mar. 17, 2017

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

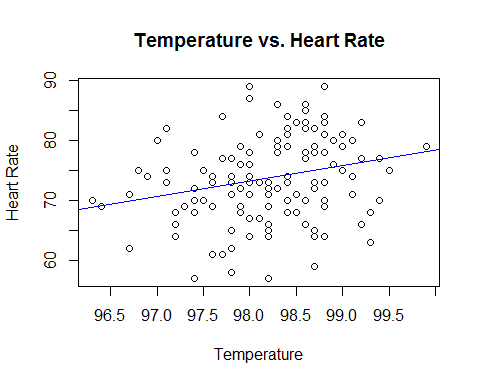
Create a scatterplot with heart rate in the vertical axis and plot the estimated linear regression line in the scatterplot. Include descriptive labels for the x and y-axes (not just the variable names as they are in the data file).

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

library(DS705data)  
data(normtemp)  
  
normtemp <- na.omit(normtemp)  
  
linear.model <- with(normtemp, lm(hr~temp))  
with(normtemp, plot(hr~temp, xlab='Temperature', ylab='Heart Rate', main='Temperature vs. Heart Rate'))  
abline(linear.model, col='blue')



It could possibly be a plausible model, however, the correlation is very low.

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

hr

We are making the following assumptions:

Errors have mean 0

Errors have the same variance for all x

Errors are independent of each other

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

linear.model

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

Replace the ## symbols with your slope and intercept.

hr = -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(linear.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

t value = 2.878 At , we reject the null hypothesis(p-value = 0.0047). There is strong evidence to support a linear relationship between body temp and heart rate.

### 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€œslopeâ€ in your interpretation).

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

confint(linear.model)

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

We are 95% confident that as body temperature raises by one degree, the heart rate will rise on average between 0.80 and 4.34.

### 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(linear.model, x, interval='confidence')

## fit lwr upr  
## 1 74.69257 73.30616 76.07897

We are 95% confident that the heart rate for an adult with temperature will be between 73.30 and 76.08.

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

x <- data.frame(temp = 98.6)  
predict(linear.model, x, interval='prediction')

## fit lwr upr  
## 1 74.69257 60.95531 88.42982

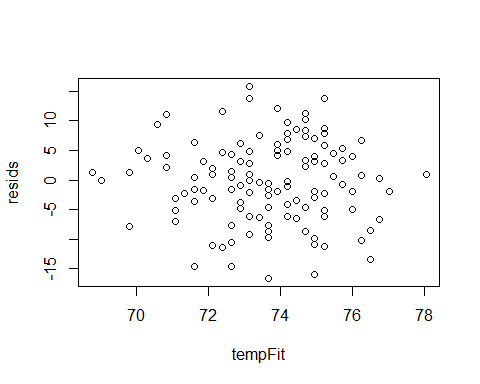
With 95% confidence, we predict the expected heart rate for a randomly selected adult with body temperature F to be between 60.96 and 88.43.

### Part 1h

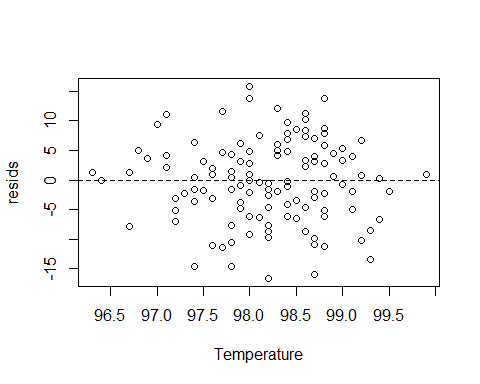
Obtain the residuals and plot them against the predicted values and also against the independent variable. Also construct a histogram, normal probability plot, and boxplot of the residuals and perform a Shapiro-Wilk test for normality. Based on your observation of the plot of residuals against the predicted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

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

resids <- linear.model$resid  
tempFit <- linear.model$fitted.values  
plot(tempFit, resids)

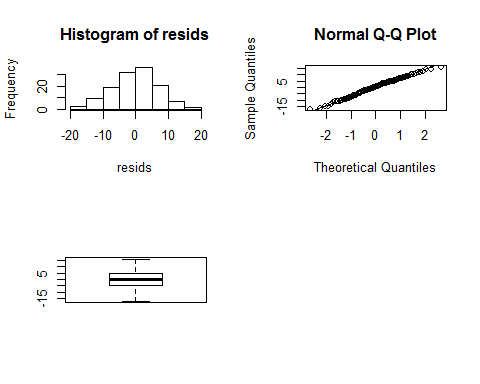


plot(normtemp$temp, resids, xlab='Temperature')  
abline(h=0, lty='dashed')



par(mfrow=c(2,2))  
hist(resids)  
qqnorm(resids)  
qqline(resids)  
boxplot(resids)  
shapiro.test(resids)

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



It does appear that the error does have the same variance for all of x (temp). Based on the boxplot and Shapiro Wilk test (p-value = 0.6027), it does appear that the errors are normally distributed.

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

The points that are furthest away from the x-axis have high leverage. They are also considered influential if the points are outliers along the y-axis. It seems in our case, it appears the values on the left are pulling the line down and the values far to the right are pulling the line up. Without these points, perhaps the line would be more horizontal.

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

linear.model.full <- with(normtemp, lm(hr ~ factor(temp)))  
anova(linear.model, linear.model.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

F = 1.4035

With , we are not able to reject the null hypothesis (p-value = 0.1103). That is to say, we cannot determine that the full model explains significantly more of the variance in the response than the linear model.

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

require(lmtest)  
bptest(linear.model)

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

BP = 0.19584

At , We are unable to reject the null hypothesis (p-value = 0.6581). That is to say, we cannot say with certainly that the variances are unequal.

### Part 1l

Calculate and interpret the Pearson correlation coefficient .

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

with(normtemp, cor.test(temp, hr, method='pearson'))

##   
## Pearson's product-moment correlation  
##   
## data: temp and hr  
## t = 2.8782, df = 126, p-value = 0.004699  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.07821862 0.40447500  
## sample estimates:  
## cor   
## 0.2483778

The Pearson correlation for these two variables is 0.25. This is very low.

### Part 1m

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

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

with(normtemp, cor.test(temp, 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 -|-|-|-|-|-|-|-|-|-|-|-

summary(linear.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 adjusted r-squared is 0.05. This suggests that only 5% of the variance in hr is explained by the linear model.

### Part 1o

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

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

The linear model does not explain the relationship between the two variables very well. This is evident by the adjusted r-squared of 0.05. As such, I would not use the model to make predictions.

### Part 1p

Calculate the Spearman correlation coefficient (just for practice).

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

with(normtemp, 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 = 254620, 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 heart rate 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(mar=c(2.5,4,.15,.15))  
xplot <- data.frame( temp = seq( 96, 100, length=128) )  
fittedC <- predict(linear.model,xplot,interval="confidence")  
fittedP <- predict(linear.model,xplot,interval="prediction")  
  
ylimits <- c(min(fittedP[,"lwr"]),max(fittedP[,"upr"]))  
par(las=1)  
with(normtemp, plot(temp, hr, ylim=ylimits, xaxt='n', yaxt='n', xlab=''))  
axis(2,mgp=c(3,.5,0))  
axis(1,mgp=c(3,.3,0))  
mtext(side=1,text='temp',line=1.5)  
abline(linear.model)  
   
lines(xplot[,1], fittedC[, "lwr"], lty = "dashed", col='darkgreen',lwd=2)  
lines(xplot[,1], fittedC[, "upr"], lty = "dashed", col='darkgreen',lwd=2)  
lines(xplot[,1], fittedP[, "lwr"], lty = "dotted", col='blue',lwd=2)  
lines(xplot[,1], fittedP[, "upr"], lty = "dotted", col='blue',lwd=2)  
  
lines(xplot[98.21563,1]\*c(1,1),fittedP[98.21563,c(2,3)],col="red",lty="dotted",lwd=2)

