Lab Assignment: Simple Linear Regression & Correlation

Spencer Swartz

March 20, 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),].

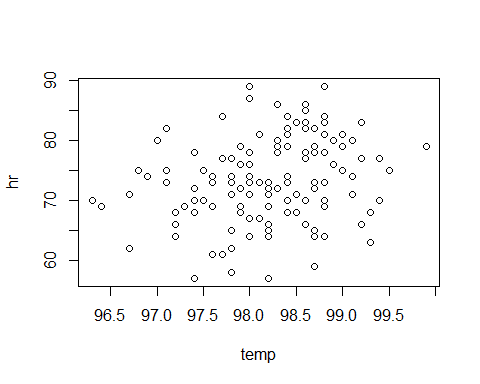
Does it appear that a linear model is at least possibly a plausible model for the relationship between hear rate and body temperature? Explain your answer.

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

require(DS705data)

## Loading required package: DS705data

data(normtemp)  
normtemp <- normtemp[c(-129,-130),]  
#normtemp  
with(normtemp,plot(temp,hr))



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

## cor   
## 0.2483778

It does not seems possible that that a linear model could work here including the correlation at about .25.

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

* Body Temp.
* Heart Rate
* predicted value of
* estimated -intercept
* estimated slope of line

### 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 <- with( normtemp, lm( hr ~ temp ) )  
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

Replace the ## symbols with your slope and intercept.

hr = -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 -|-|-|-|-|-|-|-|-|-|-|-

#summary(linear.model)  
with(normtemp, cor.test( temp, hr) )

##   
## 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

Hypotheses:

Test Statistic:

t = 2.8782

p-value:

p = 0.0047

Conclusion:

At = 0.05 we reject with a p-value of 0.0047. In conclusion there is a correlation to the data points.

### 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(linear.model)

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

We are 95% confident that the population mean heart rate increases 0.80 to 4.34 minutes for each additional degree 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( linear.model, 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, the average BPM is between 73.3 and 76.1.

### 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

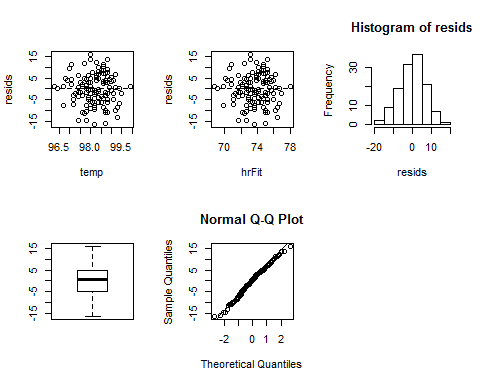
We are 95% confident that, for a body temperature of 98.6, the BPM will be between 60.96 and 88.43.

### 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$resid # extract residuals from model  
temp <- normtemp$temp  
hr <- normtemp$hr  
hrFit <- linear.model$fitted.values  
par(mfrow=c(2,3))  
plot(temp,resids); abline(h=0,lty='dashed')  
plot(hrFit,resids); abline(h=0,lty='dashed')  
hist(resids)  
boxplot(resids)  
qqnorm(resids)  
qqline(resids)  
par(mfrow=c(1,1))



shapiro.test(resids)

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

The model seems to be a good fit as we cannot claim that the sample does not come from a normal distribution based on the test and the charts above.

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

It does not look as if there are any observations that appear to be higher influence or high leverage points.

### 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 <- with( normtemp, lm( hr ~ temp ) )  
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

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

Test Statistic: F = 1.4

P-value: pvalue= .11

conclusion: the large P value confirms that the linear model fits better, or equivalently, there is a lack of fit the full 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 -|-|-|-|-|-|-|-|-|-|-|-

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

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.3.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.3.3

##   
## Attaching package: 'zoo'

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

bptest(linear.model)

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

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

test statistic: BP = 0.1958

decision rule: = 0.05

p-value: pvalue = 0.6581

conclusion: we cannot reject (pvalue = 0.6581). We cannot rule out that the data comes from a equal variance.

### Part 1l

Calculate and interpret the Pearson correlation coefficient .

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

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

## cor   
## 0.2483778

there is not a strong correlation between the 2 variables.

### Part 1m

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

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

with(normtemp, cor.test( temp, hr)$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

rsq <- linear.model$r.squared  
rsq.adj <- linear.model$adj.r.squared

the adjusted R^2 suggests that about 5% of the total variation in production times is explained by the linear relationship between body temperature an BPM. This also means that 95% of the variation is unexplained by the linear relationship.

### Part 1o

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

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

No, although the model fits well with the data, the variation in the points and R^2 show that there is a lot of noise in the data that is not explained by body temperature.

### 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 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(mar=c(2.5,4,.15,.15))  
xplot <- data.frame( temp = seq( 50, 350, length=301) )  
fittedC <- predict(linear.model,xplot,interval="confidence")  
fittedP <- predict(linear.model,xplot,interval="prediction")  
  
# scatterplot  
ylimits <- c(min(fittedP[,"lwr"]),max(fittedP[,"upr"]))  
par(las=1)  
plot(temp,hr)  
abline(linear.model)  
  
# plot the confidence and prediction bands  
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[201,1]\*c(1,1),fittedP[201,c(2,3)],col="red",lty="dotted",lwd=2)

