Biostatistics week 9: Simple linear regression

Schedule for today

- Chi-squared and t-test review activity
- Simple linear regression workshop
- Challenge time

Outline

These slides cover the basics of *simple linear regression*:

- Linear models
- Simple linear regression
- Predictor significance
- Model significance
- Model fit
- Assumption checking
- Alternate models

What makes people happy?

- Most countries measure success by the gross national product (GNP), or how economically productive the country is
- Bhutan has long measured gross national happiness (GNH) instead, focusing on how happy its citizens are instead of how much they produce
- The Global Happiness Council (GHC) has recently started to follow the lead of Bhutan and develop a World Happiness Report
- The 2018 World Happiness report (http://worldhappiness.report/ed/2018/) measures mean national happiness and characteristics of a country that influence happiness
- The data for country happiness and several related characteristics is linked as an Excel file on the World Happiness Report website

Bring in the world happiness data

- The data are in raw form here: https://s3.amazonaws.com/happiness-report/2018/WHR2018Chapter2OnlineData.xls
- The codebooks are linked here: https://s3.amazonaws.com/happiness-report/2018/Appendix1ofChapter2.pdf

```
# DATA IMPORT
# Bring in the raw data from online using the readxl package to read Excel
# and the httr package to read and store the URL
# the readxl package can read directly from on your computer, but needs
# intermediate help to read directly from the internet
library(readxl)
library(httr)
# get URL where data are
happyURL <- "https://s3.amazonaws.com/happiness-report/2018/WHR2018Chapter2OnlineData.xls"
# put data in a temp file on your computer
GET(happyURL, write_disk(tf <- tempfile(fileext = ".xls")))
# load the fifth spreadsheet from the temp file for characteristics
happyWorld <- read_xls(tf, sheet = 5)
```


Select relevant variables and rename

```
# DATA MANAGEMENT
# remove unneeded columns
happyWorld <- happyWorld[ , c(-2, -4, -5, -9,
                              -11, -13, -15)]
# add better variable names to the columns
names(happyWorld) <- c("country", "happiness", "gdpPerPerson",
                       "lifeExpect", "socialSupport",
                       "freedom", "generosity", "corruption")
```


Happiness variables

In importing the data, I selected the following variables from the available data and renamed them for easier use:

- **country:** name of the country
- **happiness:** mean national score for where you are between best possible life for you (10) and the bottom of the ladder represents the worst possible life for you (0)
- **gdpPerPerson:** gross domestic product per person
- **lifeExpect:** when born, years of expected healthy life ahead
- **socialSupport:** having someone to count on in times of trouble
- **freedom:** satisfied with freedom to choose what to do with your life
- **generosity:** have you donated money to a charity in the past month
- **corruption:** is corruption widespread in business and government

Linear models in general

Linear models use the basic idea of a line to explain and predict things about relationships among variables. To build and use linear models, it is useful to remember the equation for a line:

$y = mx + b$

Where:

- m is the slope of the line
- b is the y-intercept of the line, or the value of y when x = 0
- x and y are the coordinates of each point along the line

Linear model in statistics

In statistics, the same formula is written many ways, two of the most common are:

y = b_0 + $b_1 x$

y = c + $b_1 x$

- b_1 is the slope
- b_0 or c is the y-intercept
- x and y are the coordinates of each point along the line

Linear model vocabulary

y = b_0 + $b_1 x$

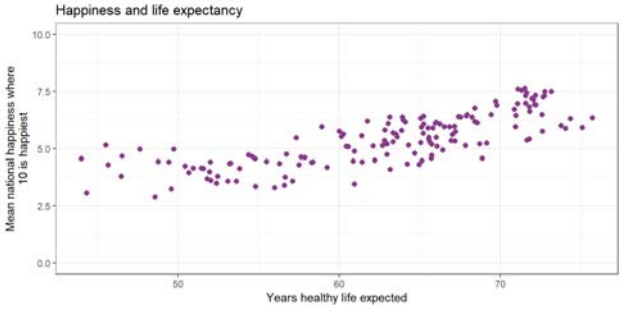
- When using the linear model in practice, the y variable is called the *dependent* or *outcome* variable.
- The x variable(s) is/are called the *independent* or *predictor* variable(s).
- Occasionally you might even see a β rather than a b, however, typically Greek letters are only used for *population* values so in work with samples the lower-case b is more appropriate.

Using the linear model

For example, we could use the linear model to see how life expectancy is related to happiness in countries around the world.

$$happiness = b_0 + b_1 life.expectancy$$

To get an idea of what you might find, start with a plot:

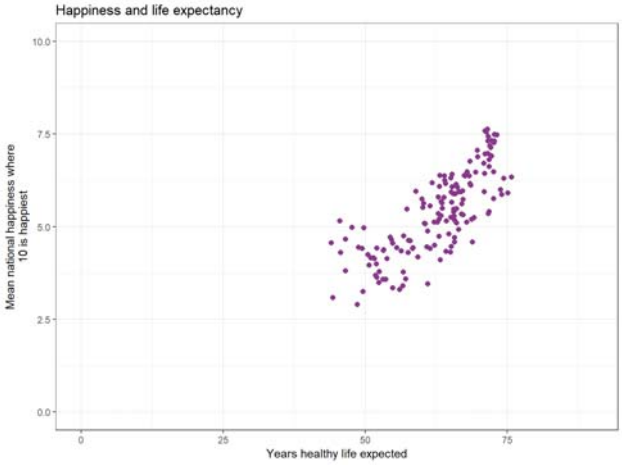


Interpreting the plot

- As years of healthy life expected increase, happiness increases
  - *this suggests a positive slope*
- The y-intercept is hard to guess because the x-axis does not go to zero, so it is hard to anticipate where a line through the data would cross if extended to 0
- Try a new graph, changing the limits of x and y so they go to 0:

```
# plot of life expectancy and happiness
# see https://ggplot2.tidyverse.org/ for info on how to use ggplot2
# and many options for plotting
library(ggplot2)
ggplot(happyWorld, aes(y = happiness, x = lifeExpect)) +
  geom_point(size = 2, colour = "#88398a") +
  labs(y = "Mean national happiness where\n10 is happiest",
       x = "Years healthy life expected") + theme_bw() +
  ylim(0,10) + xlim(0,90)+
  ggtitle("Happiness and life expectancy")
```

Graph with axes going to 0



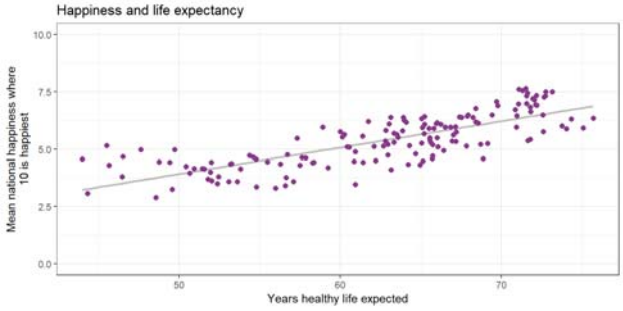
- It looks like a line through the data would cross the y-axis below 0
  - *this suggests a negative y-intercept*

How is the regression line determined?

- The formulas used aim to get the line as close as possible to all of the points
  - *The formula for the slope is:*

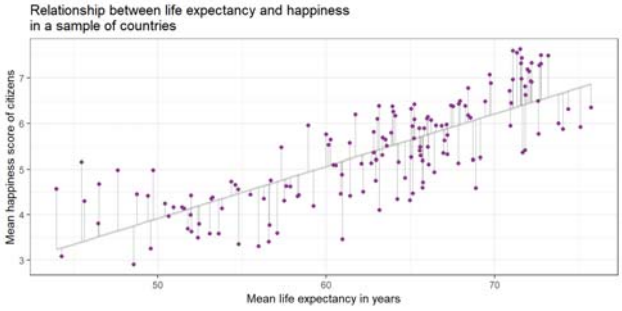
$$b_1 = \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}$$

- The intercept is then computed by entering the slope, the mean of x, and the mean of y into the formula for a line and solving for b_0
- Statistically, this approach minimizes the distances between each point and the regression line



Visualizing residuals

- The line represents the predicted values of happiness for each value of life expectancy
- The line is not perfect, it gets close to the points but does not reach them, leaving *residual* or unexplained information
- Residuals are the difference between what we observed in each country and what the regression line predicted (similar to chi-squared idea of observed and expected!), like this:



Finding the actual slope and intercept

To find the slope and the intercept of the regression line, use the *linear model* or *lm* command in R:

```
# simple linear regression to find slope
# and intercept
happiByLifeExpect <- lm(happiness ~ lifeExpect, data = happyWorld)
summary(happiByLifeExpect)
```

```
##
## Call:
## lm(formula = happiness ~ lifeExpect, data = happyWorld)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71250 -0.51951  0.07019  0.47598  1.75022
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.80228    0.46222  -3.899 0.000145 ***
## lifeExpect   0.11446    0.00732  15.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7003 on 151 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.6182, Adjusted R-squared:  0.6157
## F-statistic: 244.5 on 1 and 151 DF,  p-value: < 2.2e-16
```

Interpreting the output

The output shows a y-intercept of -1.80 and a slope of .11. Substitute these into the model:

$$happiness = -1.80228 + .11446 lifeExpect$$

Use this to predict mean happiness for countries whose people have a 60 year life expectancy:

$$happiness = -1.80228 + .11449 * 60 = 5.06$$

Based on the linear regression model, countries with a 60 year life expectancy have a mean happiness of 5.06 on a scale of 0 to 10.

Cool but where are the p-values?

In addition to knowing the model and predicting values, there are three other things to determine from a regression model:

- Is the slope statistically significantly different from 0?
- Is the model statistically significantly better than the mean?
- How well does the model fit the data?

Is the slope statistically significantly different from 0?

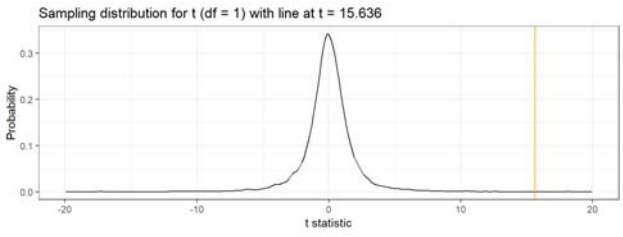
- Write the null and alternate hypotheses:
  - *H0: There is no difference between the slope of the line and 0 (slope = 0).*
  - *HA: There is a difference between the slope of the line and 0 (slope does not equal 0).*
- Calculate the t-statistic for a one-sample t-test comparing the slope to 0

$$t = \frac{b_1 - 0}{se}$$

- Reject or retain the null
- Make a conclusion

Calculate t and determine probability

$$t = \frac{b_1 - 0}{se} = \frac{0.11446 - 0}{0.00732} = 15.636$$



- The area under the curve of the sampling distribution contains all possible values of the t-statistic for samples with 1 d.f. that *came from a population where slope = 0*
- The probability of getting a t-statistic of 15.636 (or larger) if the sample came from *a population where slope = 0* is the area under the curve to the right of the orange line
- This is a very tiny probability of getting a sample where the t-statistic is this big (or bigger) *if slope = 0 in the population*
- So, *slope is probably not 0 in the population that this sample came from*

Make a conclusion

- There is sufficient evidence to reject the null hypothesis
- The slope is *statistically significantly* different from 0, that is, it is statistically unlikely that it came from a population where slope = 0.
- Interpret the findings:

Life expectancy is a statistically significant predictor of happiness (b_1 = .11; t = 15.64; p < .05). In the sample, for every one year life expectancy goes up in a country, happiness in that country goes up .11 on a scale from 0 to 10.

The conclusion is missing something

- Typically we would like to be able to say something about what is going on in the population that our sample comes from
- Confidence intervals are useful for this
- Confidence intervals can be computed for the intercept and slope using the confint command:

```
# confidence intervals for the slope and intercept
confint(happiByLifeExpect)
```

```
##                   2.5 %     97.5 %
## (Intercept) -2.71554025 -0.8890128
## lifeExpect   0.09999428  0.1289204
```

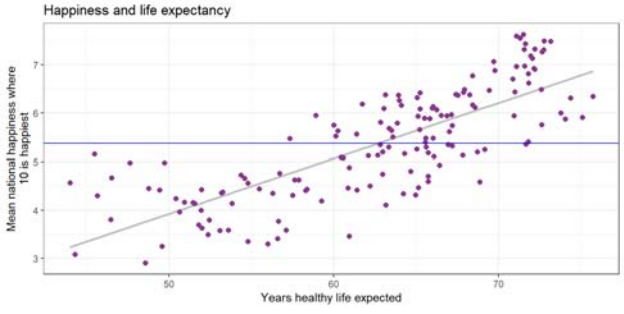
- So, the slope in the sample is .11, but in the population the likely value of the slope is between .10 and .13.

Interpret the results

Life expectancy is a statistically significant predictor of happiness (b_1 = .11; t = 15.64; p < .05). In the sample, for every one year life expectancy goes up in a country, happiness in the country goes up .11. The true slope for all countries is between .10 and .13, so the sample of countries comes from a population of countries where there is likely a .10 to .13 increase in happiness with every one-year increase in life expectancy of its citizens (95% CI: .10 - .13).

But that's not all!

- We are not only interested in the slope, but also the whole regression equation
- That is, does the regression line gets us any closer to understanding the outcome compared the mean value of the outcome would.
  - *Essentially, is the gray line better than the blue line at getting close to the data points?*
- For a regression equation with only one variable in it (i.e., simple linear regression), this is going to seem redundant to examining the slope



Write a null and alternate hypothesis

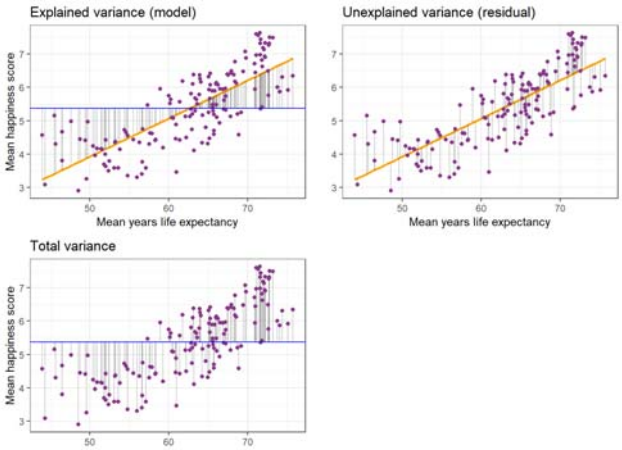
- H0: There is no difference between the regression model and the mean in explaining happiness.
- HA: The regression model is better than the mean in explaining happiness.

Calculate the test statistic

- The test statistic for the null hypothesis in linear regression is the F-statistic
- The F-statistic is a ratio of explained information to unexplained information

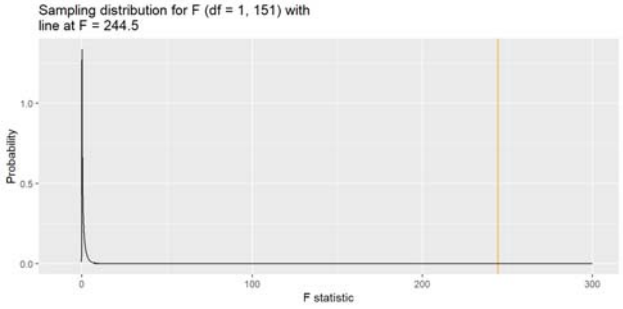
$$F = \frac{MS_M}{MS_R}$$

- MS_M is the mean squared difference between model and mean (MS_M is mean squares MODEL) (left)
- MS_R is the mean squared difference between model and observation, also known as the residual (MS_R is mean squares RESIDUAL) (right)



The probability of F

- The area under the curve of the sampling distribution contains all possible values of the F-statistic for samples with 1 and 151 d.f. that *came from a population where the null was true*
- The probability of getting an F-statistic of 244.5 (or larger) for a sample from *a population where the null was true* is the area under the curve to the right of the orange line
- This is a very tiny probability of getting a sample where the F-statistic is this big (or bigger) *if the null was true in the population*
- So, *the null is not likely true*



Write a conclusion/interpretation

- There is sufficient evidence to reject the null hypothesis (F(1, 151) = 244.5; p < .05)
- A linear model with life expectancy as the independent variable was significantly better than the mean at explaining or predicting happiness
- The countries likely came from a population of countries where happiness can be explained or predicted by life expectancy

Altogether:

A linear model predicting mean happiness in countries by life expectancy in years was statistically significantly better than the mean (F(1, 151) = 244.5; p < .05). Life expectancy is a statistically significant predictor of happiness (b_1 = .11; t = 15.64; p < .05). In the sample, for every one year life expectancy goes up in a country, happiness in the country goes up .11. The true slope for all countries is between .10 and .13, so the sample of countries comes from a population of countries where there is a .10 to .13 increase in happiness with every one-year increase in life expectancy of its citizens (95% CI: .10 - .13).

How well does the model fit?

- So far we know:
  - *The value and meaning of the slope and intercept*
  - *The significance of the slope*
  - *The significance of the overall model*
- We can also predict values of happiness by using the model
- The final measure to report with a linear model is the model fit
- The model fit is calculated by:
  - *using the model to predict the outcome for each observation*
  - *finding the correlation between the observed values of the outcome and the predicted values*
  - *squaring the correlation to get the percent of variance*

Getting model fit

- To get the amount of variance in happiness the model with life expectancy explains:
  - *Use the model to predict the happiness score based on life expectancy for each country* $happiness = -1.80228 + .11446 lifeExpect$
  - *Conduct a correlation between these predicted values of happiness and the observed happiness in each country*
  - *Square the correlation to get an* R^2 *or the percent of variation in happiness explained by the model*
- Or, get it from the output: R^2 = 0.62
- 61.82% of the variance in happiness is accounted for by the model

Reporting regression results

Once you know all of these things, you can report your results. Regression result reporting includes:

- an interpretation of the value of the slope in the sample and the likely value of the slope in the population (b_1 and its 95% CI)
- the significance of the slope (t and p)
- the significance of the model (F and p)
- the fit of the model (R-squared)

Interpretation: An linear regression analysis of the relationship between life expectancy in a country and happiness found a statistically significant (t = 15.64; p < .05) slope of .11. For every one year increase in life expectancy, there is a .11 increase in happiness (b_1 = .11; 95% CI: .10 - .13). The regression model was better than the mean value of happiness at explaining happiness [F(1, 151) = 244.5; p < .05] and the model explained 61.8% of the variation in the outcome (R^2 = .618).

Assumption checking for linear regression

There are four primary assumptions for simple linear regression (LINE acronym):

- Linearity
- Independence of residuals
- Normality of residuals
- Equal variance (homoscedasticity)

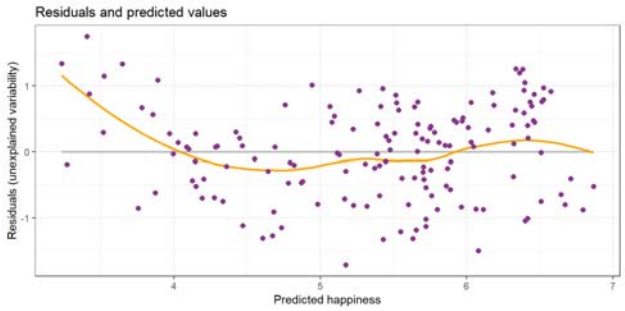
To test assumptions, you will need predicted values and residuals added to the data frame. To do this you will first have to remove missing values:

```
# make smaller data frame with complete cases and
# variables of interest
happyWorldSmall <- na.omit(happyWorld[ , c("happiness","lifeExpect")])
# add predicted values and residuals to the data
happyWorldSmall$predicted <- predict(lm(happiness ~ lifeExpect, data = happyWorldSmall))
happyWorldSmall$residuals <- residuals(lm(happiness ~ lifeExpect, data = happyWorldSmall))
```


Check assumption 1: Linearity

Because linear regression can have more than one predictor, instead of examining scatterplots of the outcome with each predictor, this assumption is most commonly tested by plotting one of two options:

- observed values and predicted values
- residuals and predicted values (preferred)



- In this case, an orange Loess curve shows some major deviation from linear at the lower happiness scores.
- Bigger residuals mean the model is WORSE at predicting lower happiness scores
- This assumption is NOT MET

Check assumption 2: Independence of residuals

- If residuals are related (e.g., the residual of one observation is dependent on the residual of another), this suggests the observations are related
- Check using the Durbin-Watson test
  - *Durbin-Watson tests the null hypothesis that the residuals are independent*

```
# test the residuals for independence
library(lmtest)
dwtest(happiByLifeExpect)
```

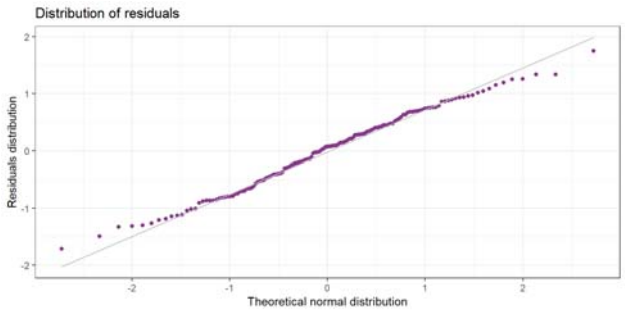
```
##
##  Durbin-Watson test
##
## data:  happiByLifeExpect
## DW = 2.1541, p-value = 0.8305
## alternative hypothesis: true autocorrelation is greater than 0
```

- p-value is quite high, we retain the null
- The assumption is MET

Check assumption 3: Normality of residuals

- QQ-plot compares the distribution of variable to a normal distribution
- If the variable is normally distributed it will fall along the line

```
# plot outcome
ggplot(happyWorldSmall, aes(sample = residuals)) +
  geom_qq(col = "#88398a") +
  stat_qq_line(col = "gray") +
  labs(x = "Theoretical normal distribution",
       y = "Residuals distribution") + theme_bw() +
  ggtitle("Distribution of residuals")
```



Looks pretty normal, check statistically

The Shapiro-Wilk test tests the null hypothesis that the residuals are normal.

```
# check normality statistically with the Shapiro-Wilk test
shapiro.test(happyWorldSmall$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  happyWorldSmall$residuals
## W = 0.99045, p-value = 0.3909
```

We failed to reject the null hypothesis that the residuals are normal. Assumption is MET.

Check Assumption 4: Equal variance (homoscedasticity)

Use the Breusch-Pagan test to test the null that the variance is constant.

```
# testing for equal variance
library(lmtest)
testVar <- bptest(happyWorldSmall$happiness ~ happyWorldSmall$lifeExpect)
testVar
```

```
##
##  studentized Breusch-Pagan test
##
## data:  happyWorldSmall$happiness ~ happyWorldSmall$lifeExpect
## BP = 0.2833, df = 1, p-value = 0.5945
```

- The p-value is high, we retain the null that the variance is constant. Assumption is MET.

Assumption checking results

- Met three of the four assumptions
- Failed linearity

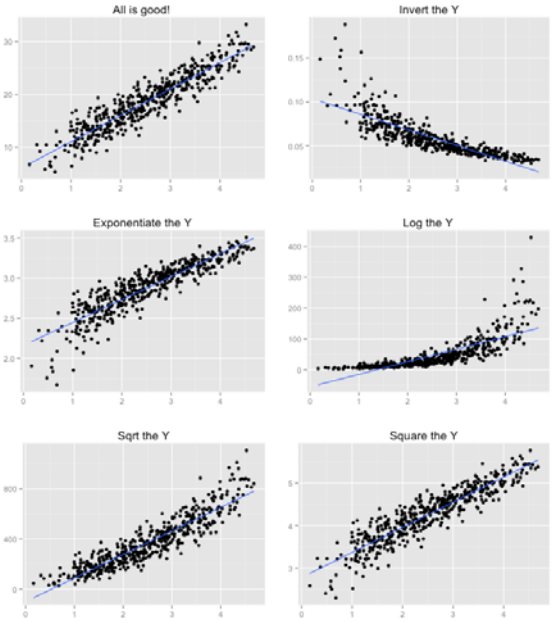
Alternatives

There is no one specific test that is the alternative to simple linear regression. Some of the options for dealing with failed assumptions are:

- Report the results only using descriptive statistics (no generalizing!)
- Include other variables that may explain the outcome
- Recode the dependent or independent variable(s) into categories and analyze with the appropriate test
- Transform the outcome or predictor(s)
  - *Only an option when the non-linear relationship is monotonic (curves up or curves down, not both)*
- Use a spline when it looks like there may be two or more distinct relationships
  - *models parts of the data separately*
- Give up and have a snack

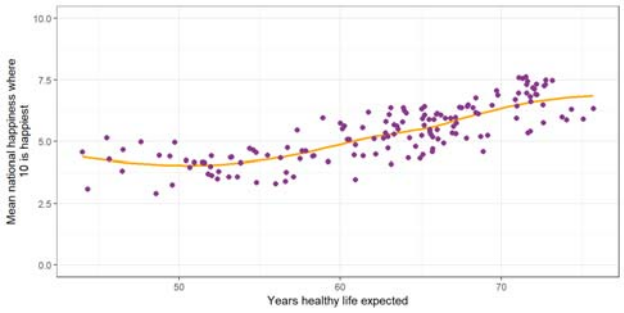
Transforming for failing linearity

- Only useful when the relationship is *monotonic*
- If your plot to check linearity looks like any one of these, use the transformation indicated:



Is relationship monotonic?

- Nope. Cannot use transformation.



- If you could, here are some examples of how to create the new variables:

```
# square the outcome
# make a new variable of the squared outcome
# use the new variable as the outcome and try again
happyWorldSmall$squareHappiness <- happyWorldSmall$happiness^2
# square root of the outcome
happyWorldSmall$sqRootHappiness <- sqrt(happyWorldSmall$happiness)
```


Recode the variables into categories and use ANOVA or Kruskal-Wallis

- Example of recoding life expectancy into a 3-category variable
- Use ANOVA or the non-parametric alternative

```
# example of recoding life expectancy
# recode 0 to 55 years old as low
# 55.1
library(car)
happyWorldSmall$lifeExpCat <- recode(happyWorldSmall$lifeExpect,
                                     "lo:55 = 'low';
                                     55.1:70 = 'moderate';
                                     70.1:hi = 'high'")
table(happyWorldSmall$lifeExpCat)
```

```
##
##     high      low moderate
##       30       32       91
```


Try a spline (this is fancy!)

- Sometimes it might appear that there are two or more different relationships going on between the independent and dependent variables
- A spline anlaysis estimates the slope of the regression line on either side of a *knot* or point where the relationship appears to change

```
##
## Call:
## lm(formula = happiness ~ bs(lifeExpect, degree = 1, knots = 55),
##     data = happyWorldSmall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.56588 -0.51236  0.09558  0.47668  1.29645
##
## Coefficients:
##                                    Estimate Std. Error t value
## (Intercept)                          4.2088     0.2502   16.82
## bs(lifeExpect, degree = 1, knots = 55)1  -0.1384     0.3078   -0.45
## bs(lifeExpect, degree = 1, knots = 55)2   2.9709     0.2593   11.46
##                                    Pr(>|t|)
## (Intercept)                          <2e-16 ***
## bs(lifeExpect, degree = 1, knots = 55)1    0.654
## bs(lifeExpect, degree = 1, knots = 55)2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6563 on 150 degrees of freedom
## Multiple R-squared:  0.6669, Adjusted R-squared:  0.6625
## F-statistic: 150.2 on 2 and 150 DF,  p-value: < 2.2e-16
```

Interpreting spline output (visualize)

- Intercept is the value of the outcome at the knot (lifeExpect = 55)
- First estimate can be converted to a slope by dividing the estimate for each range over the range of values it covers
- slope below 55: $b_1 = -.1384/(55 - 43.99)$ = -0.01
- slope from 55 to maximum:
  $b_2 = (2.97 - (-.1384))/(75.72 - 55)$ = 0.15

The End

- Challenge is on GitHub