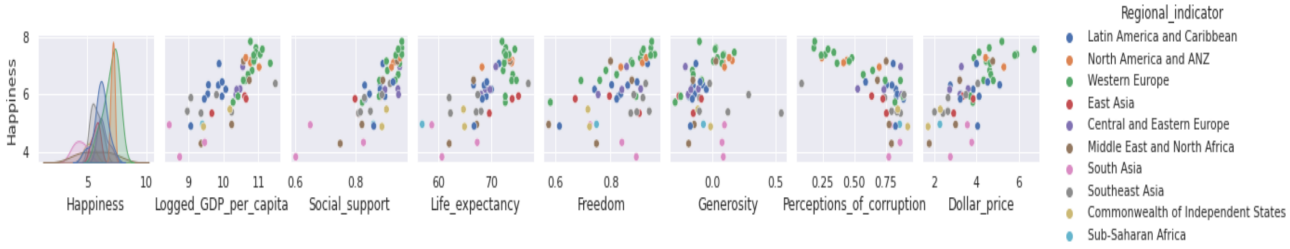
Project Part Two: Straight-line Regression

Michael Ortiz, Joshua Head, Alexander Zommer

July 11, 2021

1 Exploratory Data Analysis

Below are the scatter plots of our 7 independent variables vs our dependent variable, the perception of happiness, as well as the distribution of said dependent variable and the color coded categorical variable "Regional indicator". 3 of the graphs exhibit pretty solid linear trends, with plots 1, 3, and 7 showing the best trend. Plots 4 and 5 have both very dispersed and scattered data, and it looks like plot 6 would be better modeled by some sort of log graph. The two graphs that seem to exhibit the best linear trend are scatter plots 3 and 7, the plots of healthy life expectancy vs perceived happiness, and the happy meal price vs perceived happiness.



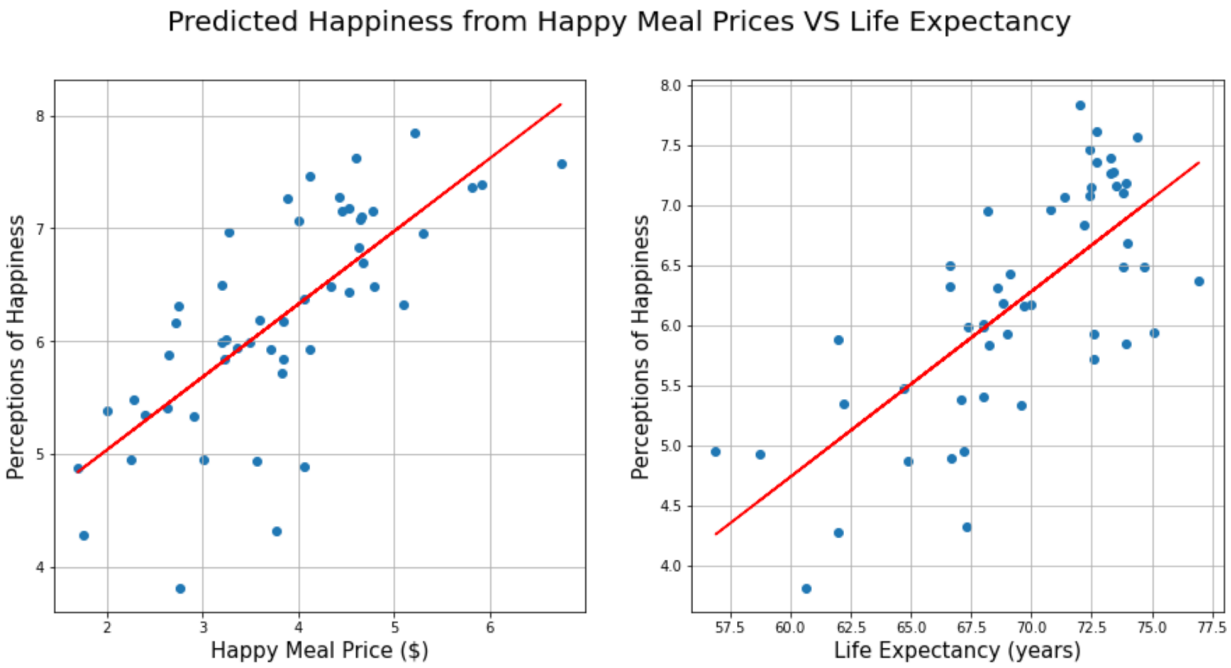
2 Straight-line Regression Models

2.1 Least-squares Models

In the happy meal price vs happiness model our y-intercept is 3.7424. For the slope estimate, we computed 0.6462. The model is then $\widehat{happiness} = 3.7424 + 0.6462(price)$.

Similarly, for the life expectancy vs happiness model, our y-intercept is -4.5287 and the slope is 0.1545. The model is then $\widehat{happiness} = -4.5287 + 0.1545(years)$. As the models show, they both fit the data fairly well and either independent variable could serve as a strong predictor of happiness. Both have nearly identical r

scores (discussed in greater detail in section **2.5**) and clearly show a linear relationship. Ultimately we chose to go with the model using life expectancy as a predictor because the data appear more evenly spread about the regression line, and it appears to lack a handful of outliers present in the other model.



2.2 Interpreting the Model

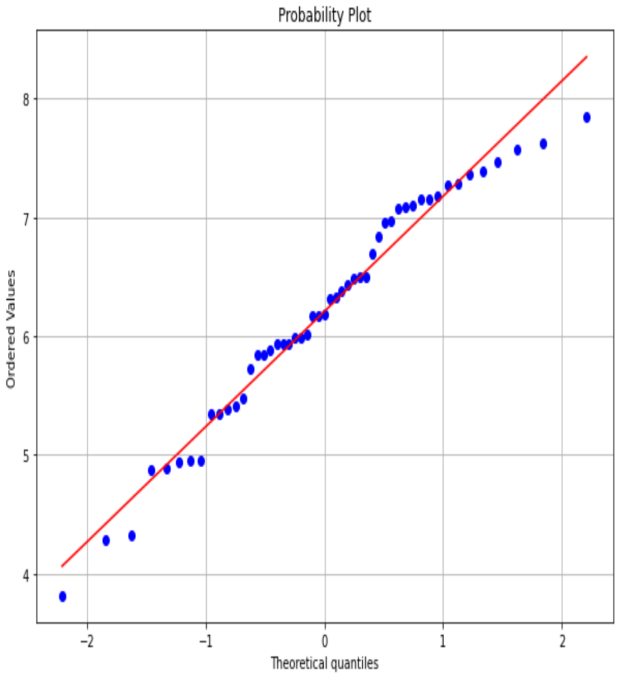
The y-intercept of the life expectancy vs perceived happiness model implies that a life expectancy of zero would expect a happiness perception of -4.53 (Standard error is 1.477) rounded to the hundredth decimal. Both of these values are nonsensical since our happiness variable is continuous ranging from 0 to 10, and (thankfully) no country on earth has a life expectancy anywhere close to 0 years. Therefore, this interpretation of the y-intercept does not apply to our situation.

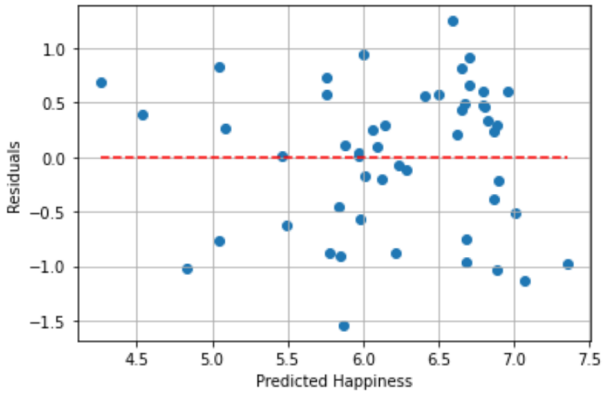
Our slope estimate is .1545 with standard error of 0.021. This implies that for ever one year increase in life expectancy, expected perception of happiness increases by .1545. This interpretation is completely valid for our modelling situation.

Checking regression assumptions

- *Existence and Independence*: All the observations in our dataset are independent of each other.
- *Homoscedasticity and Normality*: We describe these assumptions in terms of error. Below are models

showing both the Q-Q plot and distribution of our residuals.





The first graph above shows normality amongst the residuals except for the higher range, so there are possible deviations at the outliers. The second graph displays the residual plot to check homogeneity of variance. It seems that the variance throughout the model stays consistent. These plots appear convincing enough to suggest the model meets these two assumptions.

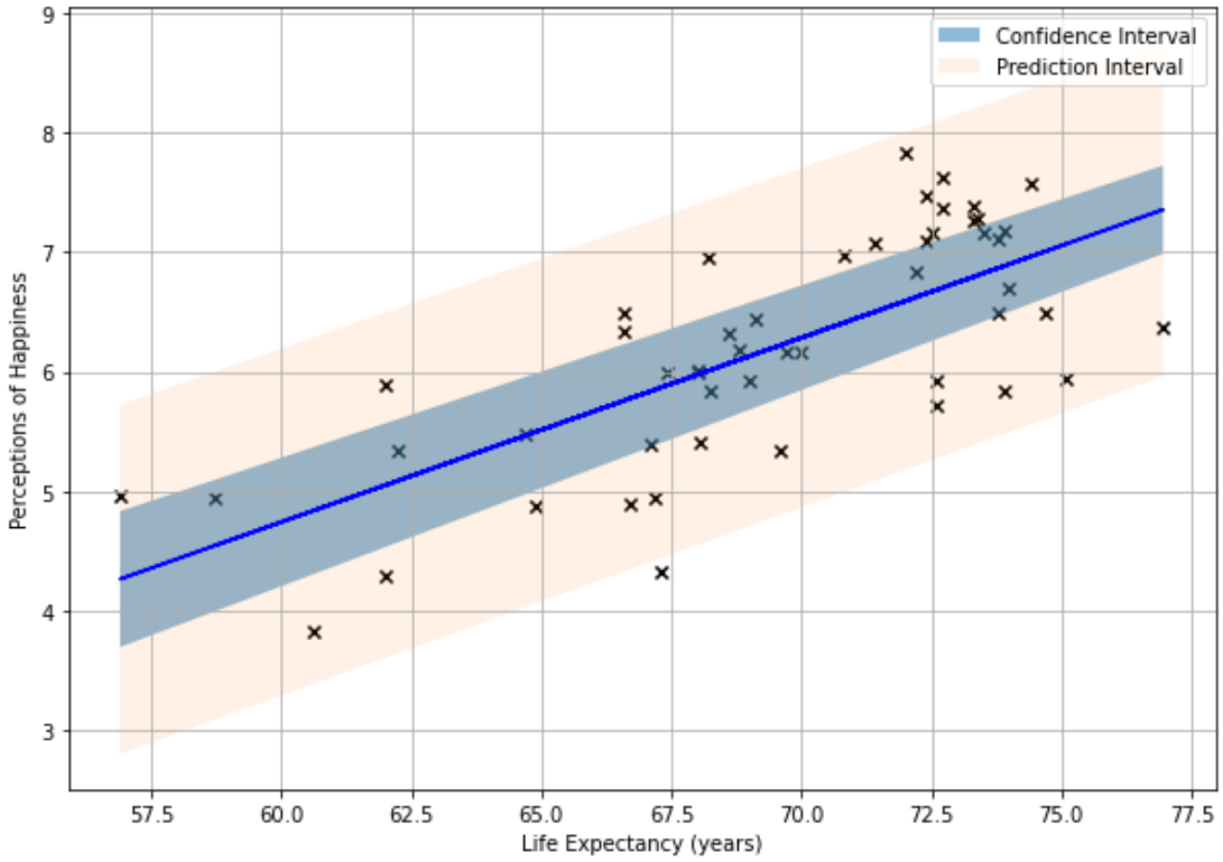
- *Linearity*: Our models in section **2.1** already provide convincing evidence that the linearity assumption is met. However, more detailed confidence interval estimates of parameters are provided in section **2.2**.

2.3 Parameter Confidence Intervals

We computed a 95% confidence interval of $(0.1118302, 0.19706996)$ for our slope estimate and a 95% confidence interval of $(-7.49693523, -1.56056379)$ for our y-intercept. The first interval supports our linearity assumption: since it does not contain zero there is enough evidence to reject the null hypothesis that the slope is zero. Therefore a linear relationship exists between happiness and life expectancy. Our interval for the intercept also does not contain zero so we can reject the null hypothesis that the intercept is zero. This is not a very meaningful result since we have already concluded that our intercept is not applicable to our situation.

2.4 Confidence and Prediction Bands

We visualize 95% confidence and prediction bands for our model below.



As an example, take the life expectancy observation from Austria where the value is 73.3. Our model predicts a mean happiness score of 6.792441. We compute a 95% confidence interval which turns out to be $(6.543936, 7.040947)$. The interpretation of this interval is that we are 95% confident that the true happiness score mean is between 6.543936 and 7.040947.

2.5 Determining r and r^2

It can be shown that our predictor has a positive correlation with a correlation coefficient r of 0.721. With an r^2 value of 0.52, our model is able to account for 52% of error, and with a sample size of 51, we can be assured that our model is adequate.