SOC6302 Statistics for Sociologists

Monica Alexander

Week 10: Multiple Linear Regression

Recap

- Y_i is the dependent variable or response variable
- X_{i1} and X_{i2} are the independent variables, explanatory variables or predictors

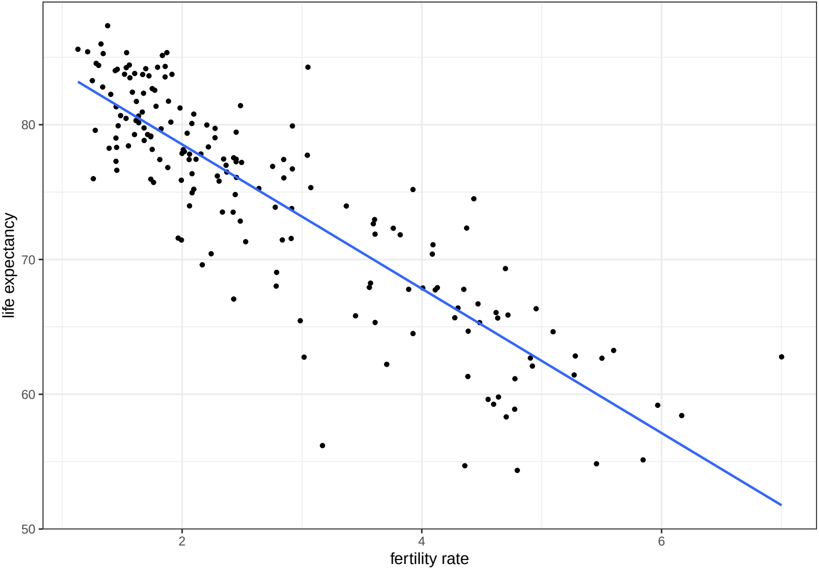
Example:

- $\{Y_1, Y_2, \ldots, Y_{176}\}$ is life expectancy by country in 2017
- $\{X_{1,1}, X_{2,1}, \ldots, X_{176,1}\}$ is TFR by country in 2017
- $\{X_{1,2}, X_{2,2}, \ldots, X_{176,2}\}$ is child mortality by country in 2017

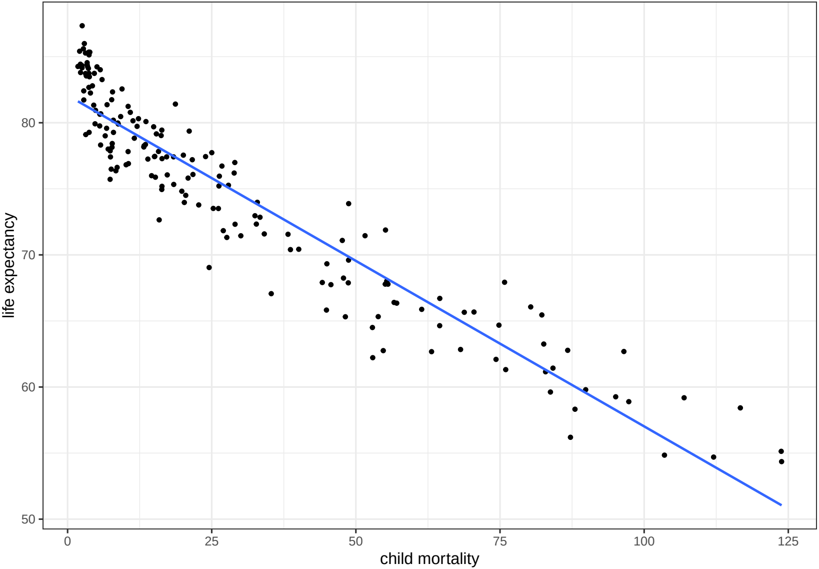
Research question:

- How does life expectancy differ across different levels of fertility and child mortality
- In other words, is life expectancy associated with fertility and child mortality, and if so, how?

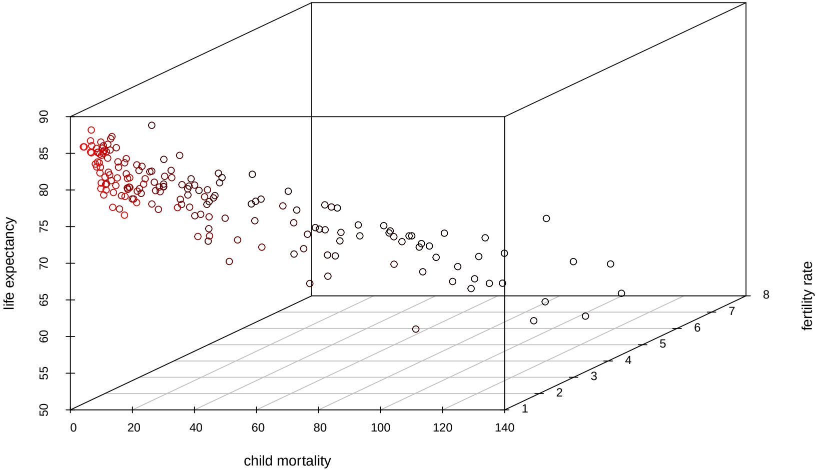
Scatter plot of fertility and life expectancy



Scatter plot of child mortality and life expectancy



Scatter plot of all variables



MLR model

With two covariates, the MLR model is

$$\begin{aligned} Y_i &= E\left(Y_i \mid X_{i1}, X_{i2}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

Specifically, the most basic MLR model is a simple linear function of X_{i1} and X_{i2}, and three parameters, β_0, β_1 and β_2.

MLR in R

- Can estimate MLR exactly the same way as SLR, just add additional variables with a + in the formula in `lm`
- Residuals, fitted values etc extracted in the same way

```r
mod <- lm(life_expectancy~tfr+child_mort, data = country_ind_2017)
summary(mod)
```

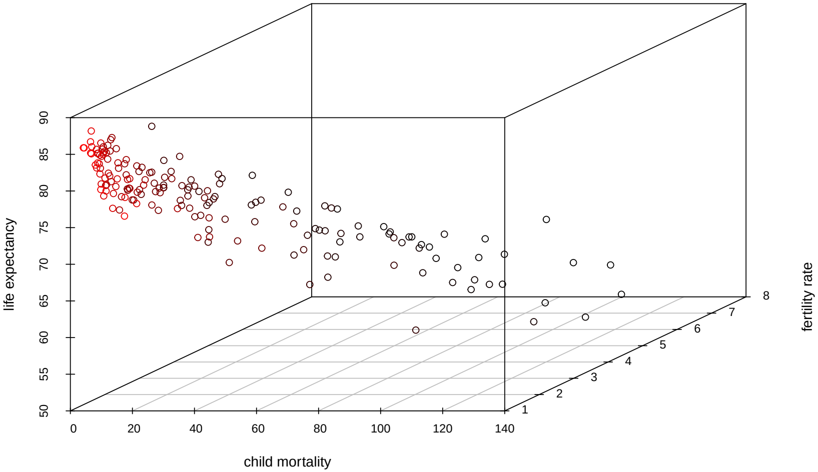
```
##
## Call:
## lm(formula = life_expectancy ~ tfr + child_mort, data = country_ind_2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7103 -1.6787 -0.1197  1.7379  5.7605
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.80622    0.56028 149.578  < 2e-16 ***
## tfr         -1.07102    0.30293  -3.536 0.000522 ***
## child_mort  -0.21031    0.01301 -16.171  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.527 on 173 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.9004
## F-statistic: 792.1 on 2 and 173 DF,  p-value: < 2.2e-16
```

OLS Estimation

- $E\left(Y_i \mid X_{i1}, X_{i2}\right)$, and by extension, β_0, β_1 and β_2 are unknown population quantities, so we need a way of estimating the MLR from sample data
- Similar to the SLR case, we will use ordinary least squares (OLS) to choose estimators for $\{\beta_0, \beta_1, \beta_2\}$, denoted $\left\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right\}$, that minimize the sum of squared residuals. This can be written as

$$
\begin{aligned}
\sum_i \hat{\varepsilon}_i^2 &= \Sigma_i \left(Y_i - \hat{E}\left(Y_i \mid X_{i1}, X_{i2}\right)\right)^2 \\
&= \sum_i \left(Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2}\right)\right)^2
\end{aligned}
$$

OLS Estimation: minimizing square residuals



OLS Estimation

The OLS estimators for the MLR model parameters are:

$$\hat{\beta}_1 = \frac{\sum_i \left(\tilde{Y}_i \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \Sigma_i \left(\tilde{Y}_i \tilde{X}_{i2}\right) \Sigma_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}{\sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) \Sigma_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right) \Sigma_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}$$

$$\hat{\beta}_2 = \frac{\sum_i \left(\tilde{Y}_i \tilde{X}_{i2}\right) \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) - \Sigma_i \left(\tilde{Y}_i \tilde{X}_{i1}\right) \Sigma_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}{\sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) \Sigma_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right) \Sigma_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}$$

$$\hat{\beta}_0 = \frac{1}{n} \sum_i Y_i - \hat{\beta}_1 \left(\frac{1}{n} \sum_i X_{i1}\right) - \hat{\beta}_2 \left(\frac{1}{n} \sum_i X_{i2}\right) = \bar{Y}_i - \hat{\beta}_1 \bar{X}_{i1} - \hat{\beta}_2 \bar{X}_{i2}$$

where $\tilde{Y}_i = Y_i - \bar{Y}_i$, $\tilde{X}_{i1} = X_{i1} - \bar{X}_{i1}$, and $\tilde{X}_{i2} = X_{i2} - \bar{X}_{i2}$.

OLS Estimation

$$\hat{\beta}_1 = \frac{\sum_i \left(\tilde{Y}_i \tilde{X}_{i1}\right) \sum_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \Sigma_i \left(\tilde{Y}_i \tilde{X}_{i2}\right) \Sigma_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}{\sum_i \left(\tilde{X}_{i1} \tilde{X}_{i1}\right) \Sigma_i \left(\tilde{X}_{i2} \tilde{X}_{i2}\right) - \sum_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right) \Sigma_i \left(\tilde{X}_{i1} \tilde{X}_{i2}\right)}$$

Covariation between Y_i and X_{i1} that is independent of X_{i2} divided by variation in X_{i1} that is independent of X_{i2}.

(Similarly for $\hat{\beta}_2$, but it is the covariation between Y_i and X_{i2} that is independent of X_{i1} divided by variation in X_{i2} that is independent of X_{i1}.)

OLS Estimation

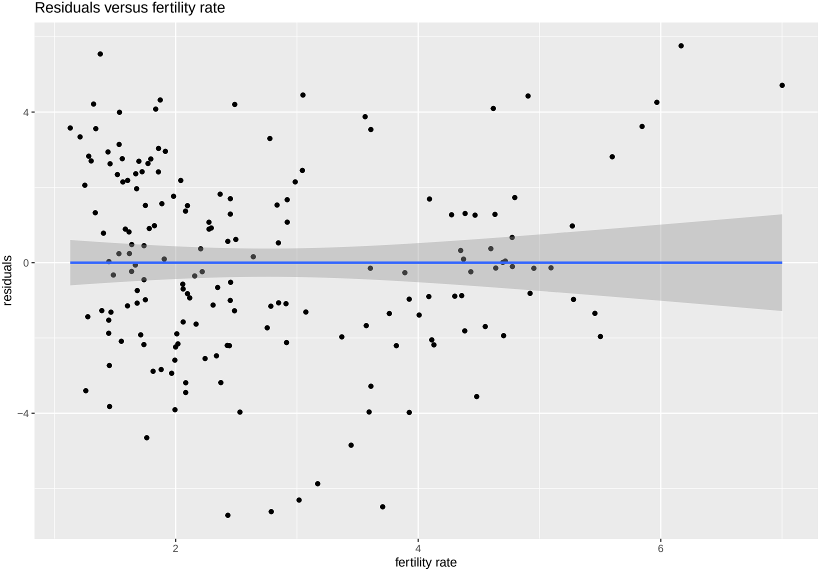
The ‘partial effect’ estimators can also be expressed as:

$$\hat{\beta}_1 = \frac{\sum_i \left(Y_i - \frac{1}{n}\Sigma_i Y_i\right)\left(X_{i1}^r - \frac{1}{n}\Sigma_i X_{i1}^r\right)}{\Sigma_i \left(X_{i1}^r - \frac{1}{n}\Sigma_i X_{i1}^r\right)^2}$$

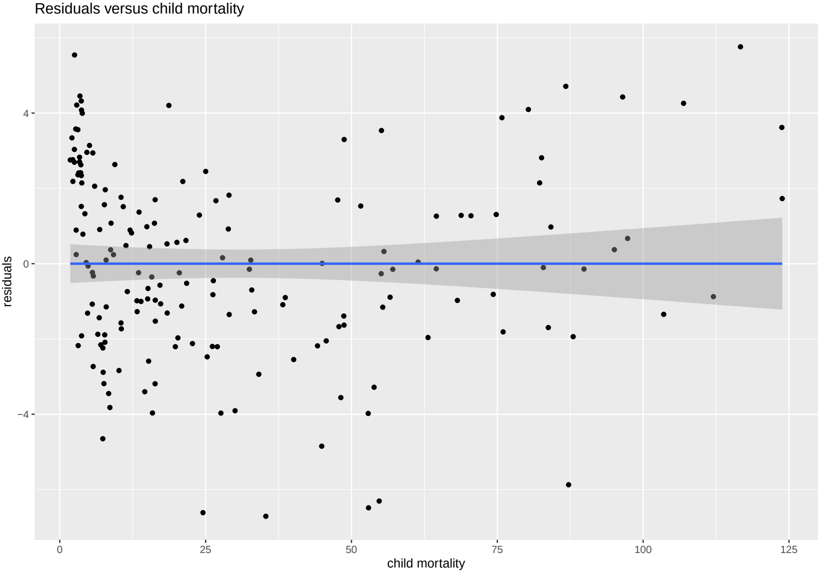
where $X_{i1}^r = X_{i1} - \hat{E}\left(X_{i1} \mid X_{i2}\right)$ are the residuals from an SLR of X_{i1} on X_{i2}.

In a similar way, $\hat{\beta}_2$ can be expressed in terms of the residuals from an SLR of X_{i2} on X_{i1}.

Residuals


Residuals versus fertility rate
residuals
4
0
−4
2
4
6
fertility rate

Residuals


Residuals versus child mortality
residuals
4
0
−4
0
25
50
75
100
125
child mortality

Residuals

The residuals $\hat{\varepsilon}_i$ have two important properties

1. The sum to zero
2. The are uncorrelated with X_{i1} and X_{i2}

```r
mod <- lm(life_expectancy~tfr+child_mort, data = country_ind_2017)
sum(resid(mod))
```

```
## [1] -1.096345e-15
```

Variance decomposition

Recall that the variance of Y_i can be decomposed into two components: a component 'explained by X_{i1} and X_{i2}' and a component 'unexplained by X_{i1} and X_{i2}'.

$$\text{total sum of squares} = \text{model sum of squares} + \text{reisdual sum of squares}$$

$$SST = SSM + SSR$$

$$\sum_i \left(Y_i - \bar{Y}_i\right)^2 = \sum_i \left(\widehat{Y}_i - \bar{Y}_i\right)^2 + \quad \sum_i \left(Y_i - \widehat{Y}_i\right)^2$$

Variance decomposition

Recall from SLR that we can use this to assess model fit, through the R^2:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$$

Adjusted R^2

- The addition of more explanatory variables with MLR will **always** increase the value of R^2
- Because of this, researchers occasionally use a goodness of fit measure called the ‘adjusted R^2’ which includes a small ‘penalty’ for the number of explanatory variables in the model

$$R^2_{adj} = 1 - \frac{SSR/n - k - 1}{SST/n - 1}$$

where k is the number of explanatory variables.

```r
mod <- lm(life_expectancy~tfr+child_mort, data = country_ind_2017)
summary(mod)
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + child_mort, data = country_ind_2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7103 -1.6787 -0.1197  1.7379  5.7605
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.80622    0.56028 149.578  < 2e-16 ***
## tfr         -1.07102    0.30293  -3.536 0.000522 ***
## child_mort  -0.21031    0.01301 -16.171  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.527 on 173 degrees of freedom
## Multiple R-squared:  0.9015, Adjusted R-squared:  0.9004
## F-statistic: 792.1 on 2 and 173 DF,  p-value: < 2.2e-16
```

Outlook for the rest of the content

- More than two variables
- Assumptions
- More on categorical variables
- Polynomial regression
- Interactions

Interpretation of MLR with $k = 3$

```
summary(lm(life_expectancy~tfr+child_mort+maternal_mort, data = country_ind))
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + child_mort + maternal_mort,
##     data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1077  -1.8229  -0.0268   1.9726   7.9812
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.6668192  0.2067649 399.811  < 2e-16 ***
## tfr           -0.7195526  0.1073323  -6.704 2.81e-11 ***
## child_mort    -0.2068586  0.0058846 -35.153  < 2e-16 ***
## maternal_mort -0.0003844  0.0006366  -0.604    0.546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.837 on 1580 degrees of freedom
## Multiple R-squared:  0.8954, Adjusted R-squared:  0.8952
## F-statistic:  4510 on 3 and 1580 DF,  p-value: < 2.2e-16
```

Interpretation

The MLR assumptions

The five SLR assumptions we discussed are also important in the MLR context.

1. no model misspecification
2. there is independent variation in all of the explanatory variables
   - In other words, none of the explanatory variables are constants, and there are no perfect linear relationships among the explanatory variables
   - e.g. can't have $X_{i1} = X_{i2} + X_{i3}$
3. All variables are from a simple random sample
   - This assumption implies that all members of a population have an equal probability of selection, that all possible samples of size n have an equal probability of selection, and that each observation is independent of all the others

The MLR assumptions

4. The variance of $\varepsilon_i = Y_i - E\left(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right)$ is the same across all values of the explanatory variables i.e. $\mathrm{Var}\left(\varepsilon_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right) = \sigma^2$
   - This is called homoskedasticity
5. The normality assumption $\varepsilon_i = Y_i - E\left(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right)$ is normally distributed

Implications for inference

Under the five assumption discussed, the SE-standardized $\hat{\beta}_k$

$$T_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \beta_k}{se\left(\widehat{\beta}_k\right)}$$

follows a t-distribution with $n - (k + 1)$ degrees of freedom.

- Hypothesis testing is thus similar to SLR, through the use of t-tests.
- In regression, we are interested in testing the null hypothesis that $\beta_k = 0$.

Example R output

```
summary(lm(life_expectancy~tfr+child_mort+maternal_mort, data = country_ind))
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + child_mort + maternal_mort,
##     data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1077  -1.8229  -0.0268   1.9726   7.9812
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.6668192  0.2067649 399.811  < 2e-16 ***
## tfr           -0.7195526  0.1073323  -6.704 2.81e-11 ***
## child_mort    -0.2068586  0.0058846 -35.153  < 2e-16 ***
## maternal_mort -0.0003844  0.0006366  -0.604    0.546
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.837 on 1580 degrees of freedom
## Multiple R-squared:  0.8954, Adjusted R-squared:  0.8952
## F-statistic:  4510 on 3 and 1580 DF,  p-value: < 2.2e-16
```

Interpretation

More on categorical variables: changing the reference category

More than one category

- Let's model the association of life expectancy and TFR, and region of the world
- Region is a category
- In R, we can directly put a categorical variable into MLR and it gets converted to a series of indicator variables

More than one category

```
summary(lm(life_expectancy~tfr+region, data = country_ind))
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + region, data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8281  -2.1921   0.4836   2.4323   9.6510
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        81.9036     0.5389 151.994  < 2e-16 ***
## tfr                                -3.0314     0.1203 -25.205  < 2e-16 ***
## regionDeveloped regions             4.6512     0.4933   9.429  < 2e-16 ***
## regionEastern Asia                  2.0510     0.8493   2.415 0.015854 *
## regionLatin America and Caribbean   1.8448     0.4967   3.714 0.000211 ***
## regionNorthern Africa               2.5312     0.7649   3.309 0.000957 ***
## regionOceania                       0.3735     0.6632   0.563 0.573389
## regionSouth-eastern Asia            0.3469     0.5928   0.585 0.558527
## regionSouthern Asia                -1.8089     0.6069  -2.980 0.002923 **
## regionSub-Saharan Africa           -5.9669     0.5377 -11.097  < 2e-16 ***
## regionWestern Asia                  2.8872     0.5701   5.064 4.59e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.746 on 1573 degrees of freedom
## Multiple R-squared:  0.8185, Adjusted R-squared:  0.8174
## F-statistic: 709.6 on 10 and 1573 DF,  p-value: < 2.2e-16
```

Interpretation

The above model is

$$
\begin{aligned}
E\left(Y_i \mid X_{i1}, X_{i2}\right) = {} & \beta_0 + \beta_1 X_{i1} + \beta_2 \boldsymbol{I}\left(X_{i2} = \text{“Developed Regions”}\right) \\
& + \beta_3 \boldsymbol{I}\left(X_{i2} = \text{“Eastern Asia”}\right) \\
& + \beta_4 \boldsymbol{I}\left(X_{i2} = \text{“Latin America and Caribbean”}\right) \\
& + \dots \\
& + \beta_{10} \boldsymbol{I}\left(X_{i2} = \text{“Western Asia”}\right)
\end{aligned}
$$

Now the reference category is the Caucasus and Central Asia region.

- The intercept is the expected value of Y_i when TFR is zero AND at the reference level of X_{i2}
- Each of the β_2 to β_{10} gives the difference in the expected value of Y_i between the reference level of X_{i2} and when X_{i2} equals that particular category.

Interpretation

$$\begin{aligned}
E\left(Y_i \mid X_{i1}, X_{i2}\right) = \beta_0 &+ \beta_1 X_{i1} + \beta_2 I\left(X_{i2} = \text{“Developed Regions”}\right) \\
&+ \beta_3 I\left(X_{i2} = \text{“Eastern Asia”}\right) \\
&+ \beta_4 I\left(X_{i2} = \text{“Latin America and Caribbean”}\right) \\
&+ \dots \\
&+ \beta_{10} I\left(X_{i2} = \text{“Western Asia”}\right)
\end{aligned}$$

- $\hat{\beta}_2 = 4.65$ holding TFR constant, the expected life expectancy in Developed Regions is 3 years higher than Caucasus and Central Asia
- What is $\hat{\beta}_4$?
- How do we get the expected value of life expectancy at TFR $= 0$ for Southern Asia?

Changing the reference category

- Note that in the last example the reference category was chosen by R by default (first alphabetically)
- We can change this, by doing the following:
  - Converting region to a factor
  - Changing the reference level of the factor
- Factors in R are characters that have a specified order

Changing the reference category

```r
country_ind <- country_ind %>%
   # change the original variable to a factor
  mutate(region = factor(region)) %>%
  # relevel the region factor make dev regions the reference
  mutate(region_2 = fct_relevel(region, "Developed regions", after = 0))

unique(country_ind$region)
```

```
##  [1] Southern Asia               Developed regions
##  [3] Northern Africa             Sub-Saharan Africa
##  [5] Latin America and Caribbean Caucasus and Central Asia
##  [7] Western Asia                South-eastern Asia
##  [9] Eastern Asia                Oceania
## 10 Levels: Caucasus and Central Asia Developed regions ... Western Asia
```

```r
unique(country_ind$region_2)
```

```
##  [1] Southern Asia               Developed regions
##  [3] Northern Africa             Sub-Saharan Africa
##  [5] Latin America and Caribbean Caucasus and Central Asia
##  [7] Western Asia                South-eastern Asia
##  [9] Eastern Asia                Oceania
## 10 Levels: Developed regions Caucasus and Central Asia ... Western Asia
```

Changing the reference category

```
summary(lm(life_expectancy~tfr+region_2, data = country_ind))
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr + region_2, data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8281  -2.1921   0.4836   2.4323   9.6510
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         86.5548     0.2709 319.493  < 2e-16 ***
## tfr                                 -3.0314     0.1203 -25.205  < 2e-16 ***
## region_2Caucasus and Central Asia   -4.6512     0.4933  -9.429  < 2e-16 ***
## region_2Eastern Asia                -2.6001     0.7457  -3.487 0.000502 ***
## region_2Latin America and Caribbean  -2.8063     0.3022  -9.287  < 2e-16 ***
## region_2Northern Africa             -2.1200     0.6664  -3.181 0.001495 **
## region_2Oceania                     -4.2776     0.5724  -7.473 1.29e-13 ***
## region_2South-eastern Asia          -4.3043     0.4463  -9.644  < 2e-16 ***
## region_2Southern Asia               -6.4601     0.4754 -13.588  < 2e-16 ***
## region_2Sub-Saharan Africa         -10.6180     0.4467 -23.770  < 2e-16 ***
## region_2Western Asia                -1.7639     0.4273  -4.128 3.84e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.746 on 1573 degrees of freedom
## Multiple R-squared:  0.8185, Adjusted R-squared:  0.8174
## F-statistic: 709.6 on 10 and 1573 DF,  p-value: < 2.2e-16
```

Changing the reference category

- Compare the results in the previous slide to the results where we just used region
- The coefficients have changes because the reference category has changed
- The significance has also changed because the reference category has a larger sample size (more countries in Developed Regions)

Extensions I: polynomial regression

Motivation

- One of the most common types of model misspecification involves nonlinearity
- We've already talked about one solution - logs!
- A regression model can also accommodate certain types of nonlinearity pretty well through the use of polynomial terms
- A polynomial regression function of degree m can be expressed as:

$$E\left(Y_i \mid X_{i1}, \ldots, X_{ik}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \beta_{k+1} X_{ik}^2 + \cdots + \beta_{k+m} X_{ik}^m$$

- Adding polynomial terms to a MLR model allows the CEF to change nonlinearly with an explanatory variable

Polynomial regression with $m = 2$

- Consider the following MLR model with a linear and quadratic term for X_{i3}

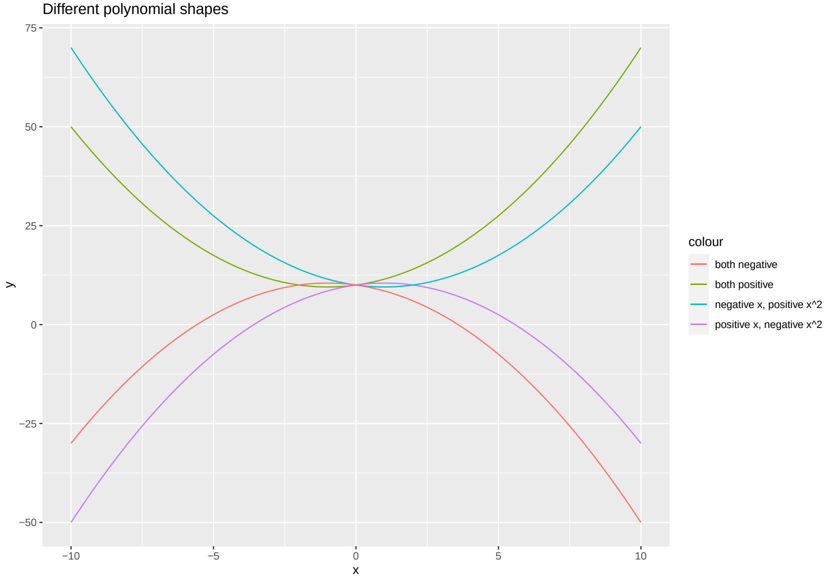
$$E\left(Y_i \mid X_{i1}, X_{i2}, X_{i3}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i3}^2$$

- The partial effect of X_{i3} in this model is

$$\beta_3 + 2\beta_4 X_{i3}$$

  which indicates that the change in the conditional expectation associated with a unit increase in X_{i3} depends on the reference value of X_{i3}

Different shapes



Example

Income versus age in the American Community Survey

```
summary(lm(incwage~age, data = acs))
```

```
##
## Call:
## lm(formula = incwage ~ age, data = acs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75341 -35352 -15431  12414 657601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31725.81    2639.32  12.020   <2e-16 ***
## age           564.75      56.73   9.956   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 68050 on 6944 degrees of freedom
## Multiple R-squared:  0.01407,    Adjusted R-squared:  0.01393
## F-statistic: 99.11 on 1 and 6944 DF,  p-value: < 2.2e-16
```

Example: income versus age

```r
acs <- acs %>%
  mutate(age_sq = age^2)
summary(lm(incwage~age+ age_sq, data = acs))
```

```
##
## Call:
## lm(formula = incwage ~ age + age_sq, data = acs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -72818 -33425 -12407  11612 644040
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -96803.133   7470.885  -12.96   <2e-16 ***
## age           6921.217    351.182   19.71   <2e-16 ***
## age_sq         -70.544      3.849  -18.33   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 66460 on 6943 degrees of freedom
## Multiple R-squared:  0.05958,    Adjusted R-squared:  0.05931
## F-statistic: 219.9 on 2 and 6943 DF,  p-value: < 2.2e-16
```

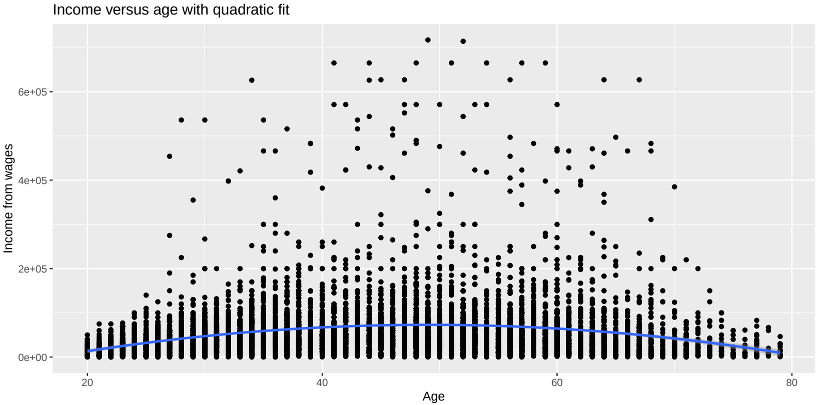
Example: income versus age

$$Y_i = -96803 + 6921X - 70X^2$$

- How should we interpret the partial effect of age?
- When age $=$ 25, effect $=$ 3396.217
- When age $=$ 30, effect $=$ 2691.217
- When age $=$ 70, effect $=$ -2948.783

Example: income versus age

- This figure shows a plot of income against age that additionally displays the quadratic fitted line
- The quadratic fit indicates that mean incomes increase faster at younger ages, slower during middle age, and eventually decline among the oldest workers, holding other factors constant



Polynomial regression: summary

- Incorporating polynomial terms into a regression model is a flexible way to approximate nonlinear relationships
- The interpretation of parameter estimates is more difficult, but estimation and inferential procedures we covered previously can all be implemented exactly as before

Extensions II: Interaction terms

Motivation

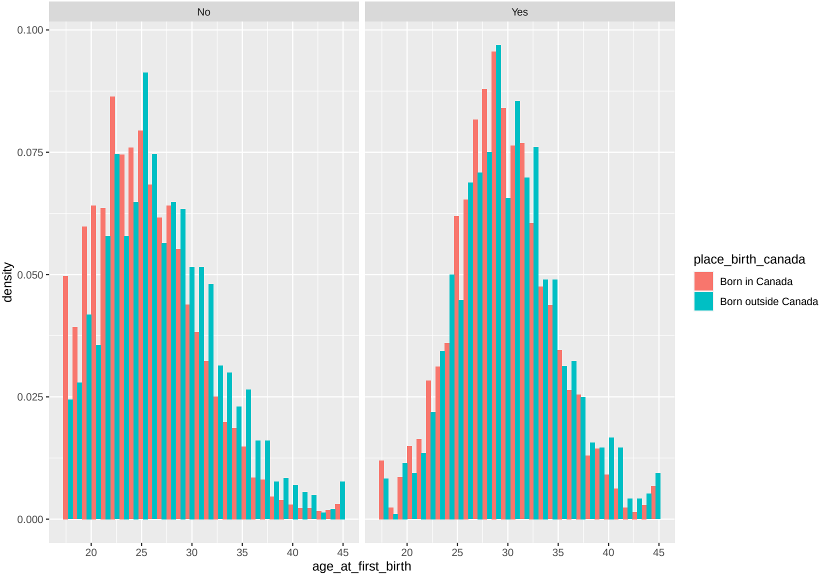
Example from last week, using GSS data

- Question: are people born outside of Canada more likely to start having children later compared to those born in Canada?
- Does the answer to this question persist after we take into account education?

Variables:

- Age at first birth
- Place of birth (Canada, outside Canada)
- Bachelor or higher (yes/no)

Looking at distributions



Effect moderation

- Effect moderation refers to the situation where the partial effect of one explanatory variable differs or changes across levels of another explanatory variable
  - e.g. the association between income and age may vary by education level
- All of the models we have considered thus far constrain the partial effects of the explanatory variables to be invariant, but this may not be appropriate

We can accommodate effect moderation through the use of **interaction terms**

Interaction terms

Example of an MLR model with an interaction term:

$$\begin{aligned} Y_i &= E\left(Y_i \mid X_{i1}, X_{i2}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} \end{aligned}$$

- How should we interpret the parameters in an MLR model with interaction terms?
- First, let's take a look at how $E\left(Y_i \mid X_{i1}, X_{i2}\right)$ changes with a unit increase in X_{i1}

Interaction terms

$$E\left(Y_i \mid X_{i1}, X_{i2}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

In this model, the change in the expected value of Y_i associated with a unit increase in X_{i1} is given by

$$E\left(Y_i \mid X_{i1} = x_1 + 1, X_{i2} = x_2\right) - E\left(Y_i \mid X_{i1} = x_1, X_{i2} = x_2\right) = \beta_1 + \beta_3 x_2$$

- The partial effect of X_{i1} now depends on the value to which we set the other explanatory variable, X_{i2}
- Note that when $X_{i2} = 0$, this expression simplifies to β_1, or in other words, β_1 is the change in the expected value of Y_i associated with a unit increase in X_{i1} specifically when $X_{i2} = 0$

Interaction terms

$$E\left(Y_i \mid X_{i1}, X_{i2}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}$$

Now let's look at the other variable, X_{i2}. In this model, the change in the expected value of Y_i associated with a unit increase in X_{i2} is given by

$$E\left(Y_i \mid X_{i1} = x_1, X_{i2} = x_2 + 1\right) - E\left(Y_i \mid X_{i1} = x_1, X_{i2} = x_2\right) = \beta_2 + \beta_3 x_2$$

- The partial effect of X_{i2} now depends on the value to which we set the other explanatory variable, X_{i1}
- Note that when $X_{i1} = 0$, this expression simplifies to β_2, or in other words, β_2 is the change in the expected value of Y_i associated with a unit increase in X_{i2} specifically when $X_{i1} = 0$

Interaction terms

- The previous two slides may take a little getting used to
- In reality, one of our explanatory variables (say X_{i2}) is a binary variable (so either 0 or 1)
- This simplifies the interpretation of the interaction term

Example

- What is the association between TFR, life expectancy and region?
- Does the association between TFR and life expectancy differ based on whether country is in Developed Regions or not?

Example in R

```r
country_ind_2017 <- country_ind %>%
  filter(year==2017) %>%
  mutate(dev_region = ifelse(region=="Developed regions", "yes", "no"))

summary(lm(tfr ~ life_expectancy + dev_region + life_expectancy*dev_region, data = country_ind_2017))
```

```
##
## Call:
## lm(formula = tfr ~ life_expectancy + dev_region + life_expectancy *
##     dev_region, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23326 -0.29618 -0.02426  0.28744  2.54832
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    13.52646    0.52158  25.933  < 2e-16 ***
## life_expectancy                -0.14454    0.00722 -20.019  < 2e-16 ***
## dev_regionyes                 -12.95159    2.91594  -4.442 1.59e-05 ***
## life_expectancy:dev_regionyes   0.15711    0.03557   4.417 1.76e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6164 on 172 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7745
## F-statistic: 201.4 on 3 and 172 DF,  p-value: < 2.2e-16
```

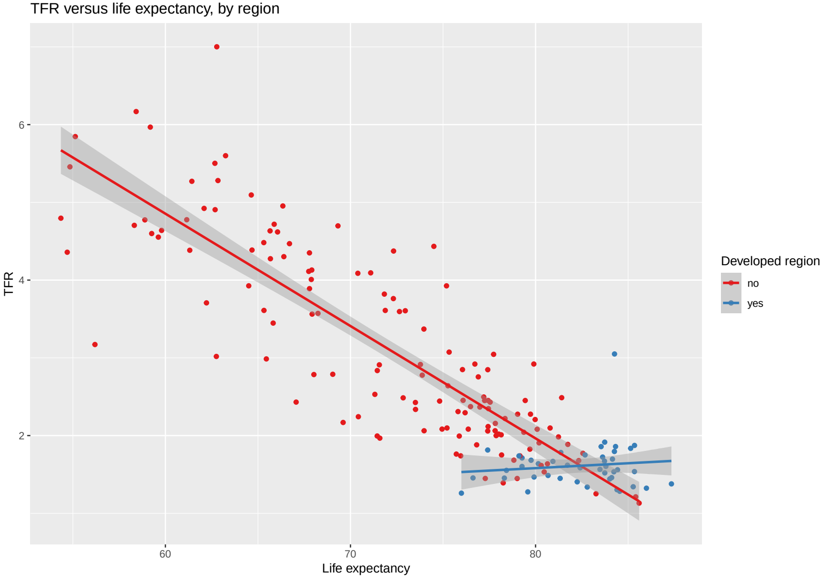
Example

$$Y_i = 13.5 - 0.14X_1 - 13.0X_2 + 0.16X_1X_2$$

Some interpretations

- for non-developed regions, 1 year increase in life expectancy associated with 0.14 decrease in TFR
- for developed regions, a 1 year increase in life expectancy associated with a 0.02 increase in TFR

Visualizing interactions



GSS example

```
##
## Call:
## lm(formula = age_at_first_birth ~ place_birth_canada + has_bachelor_or_higher +
##     place_birth_canada * has_bachelor_or_higher, data = filter(gss,
##     place_birth_canada != "Don't know"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1140  -3.7254  -0.5726   3.1041  19.4041
##
## Coefficients:
##                                                                 Estimate
## (Intercept)                                                     25.59587
## place_birth_canadaBorn outside Canada                            1.52951
## has_bachelor_or_higherYes                                        3.97672
## place_birth_canadaBorn outside Canada:has_bachelor_or_higherYes -0.98812
##                                                                 Std. Error
## (Intercept)                                                        0.05837
## place_birth_canadaBorn outside Canada                              0.14271
## has_bachelor_or_higherYes                                          0.12286
## place_birth_canadaBorn outside Canada:has_bachelor_or_higherYes    0.23958
##                                                                 t value
## (Intercept)                                                     438.485
## place_birth_canadaBorn outside Canada                            10.718
## has_bachelor_or_higherYes                                        32.368
## place_birth_canadaBorn outside Canada:has_bachelor_or_higherYes  -4.124
##                                                                 Pr(>|t|)
## (Intercept)                                                      < 2e-16 ***
## place_birth_canadaBorn outside Canada                            < 2e-16 ***
## has_bachelor_or_higherYes                                        < 2e-16 ***
## place_birth_canadaBorn outside Canada:has_bachelor_or_higherYes 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.112 on 12473 degrees of freedom
```

GSS example: interpretation

Summary

- The linear regression model is more flexible than it may appear at first
- There are a variety of extensions and adaptions to conventional regression models that can mitigate the problems associated with misspecifcation
- All models are wrong, but some models are less wrong than others, and some are useful