Sums of Squares

- SS(C. Total) = 123379.94
- SS(Year) = 113745.91
  - Year explains 92.2%
- SS(Year²|Year) = 9496.26
  - Year² adds 7.7%

Sums of Squares

- SS(C. Total) = 123379.94
- SS(Year²) = 114663.06
  - Year² explains 92.9%
- SS(Year|Year²) = 8579.11
  - Year adds 7.0%
Sums of Squares

- SS(C. Total) = 123379.94
- SS(Year|Year2) = 8579.11 (7.0%)
- SS(Year2) = 114663.06 (92.9%)
- SS(Year2|Year) = 9496.26
- SS(shared) = 105166.80 (85.2%)

Sums of Squares

- SS(C. Total) = 123379.94
- SS(YearCtr) = 113745.91
- YearCtr explains 92.2%
- SS(YearCtr²|YearCtr) = 9496.26
- YearCtr² adds 7.7%
Sums of Squares

SS(C. Total) = 123379.94

SS(YearCtr) = 113745.91
  92.2%

SS(YearCtr²|YearCtr) = 9496.26
  7.7%

SS(YearCtr²) = 9496.26
  YearCtr² explains 7.7%

SS(YearCtr|YearCtr²) = 113745.91
  YearCtr adds 92.2%
Sums of Squares

- $SS(\text{C. Total}) = 123379.94$
- $SS(\text{shared}) = 0.00$
- $SS(\text{YearCtr}|\text{YearCtr}^2) = 113745.91$
- $SS(\text{YearCtr}^2|\text{YearCtr}) = 9496.26$

- 92.2%
- 7.7%

Effects of Centering

- Year$^2$ shares over 85% of the explained variation with Year.
- YearCtr$^2$ shares none of the explained variation with YearCtr.

Why does this happen?

- The correlation between Year$^2$ and Year is statistically significant, multicollinearity.
- The correlation between YearCtr$^2$ and YearCtr is zero, no linear relationship.
What about 1940 & 1950?
- The predictions for 1940 and 1950 are much higher than the actual population values.
- Why?
- Can we add a term to the model that could account for this?

Dummy Variable
- A dummy of indicator variable can be used to identify individual or sets of values.
- \( X = 1 \) if Year is 1940 or 1950
- \( X = 0 \) otherwise

Quadratic with Dummy
- Predicted Population = 62.890 + 1.227*(Year – 1890) + 0.00646*(Year – 1890)^2 - 8.352*X
- Note that the other estimated slope coefficients are very close to those in the quadratic model.
Quadratic with Dummy

For 1940 and 1950, the prediction is lowered by 8.352 million.

Quadratic

1940
- Actual = 132.165
- Predicted = 138.951
- Residual = –6.786

1950
- Actual =151.326
- Predicted =158.261
- Residual = –6.936

Quadratic with Dummy

1940
- Actual = 132.165
- Predicted = 132.038
- Residual = 0.127

1950
- Actual =151.326
- Predicted =151.414
- Residual = –0.088
Change in R²

- Quadratic: $R^2 = 0.9989$
  - 99.89% explained variation
- Quadratic + Dummy: $R^2 = 0.9998$
  - 99.98% explained variation
- Only a small increase.

Significant Improvement?

- Dummy variable, $X$ added to the quadratic model.
  - $t = -9.22$, P-value < 0.001
  - Because the P-value is small, the dummy variable, $X$, adds significantly to the quadratic model.

Change in RMSE

- Quadratic:
  - RMSE = 2.767
- Quadratic + Dummy:
  - RMSE = 1.162
  - RMSE reduced quite a bit.
Plot of Residuals

- One might detect a up-down-up-down wave.
- Worst predictions are still within 2.5 million of the actual population.
- Probably can’t do any better.