Block 9
Complex Regressors

Complex Regressors

Factors

One-Factor Models

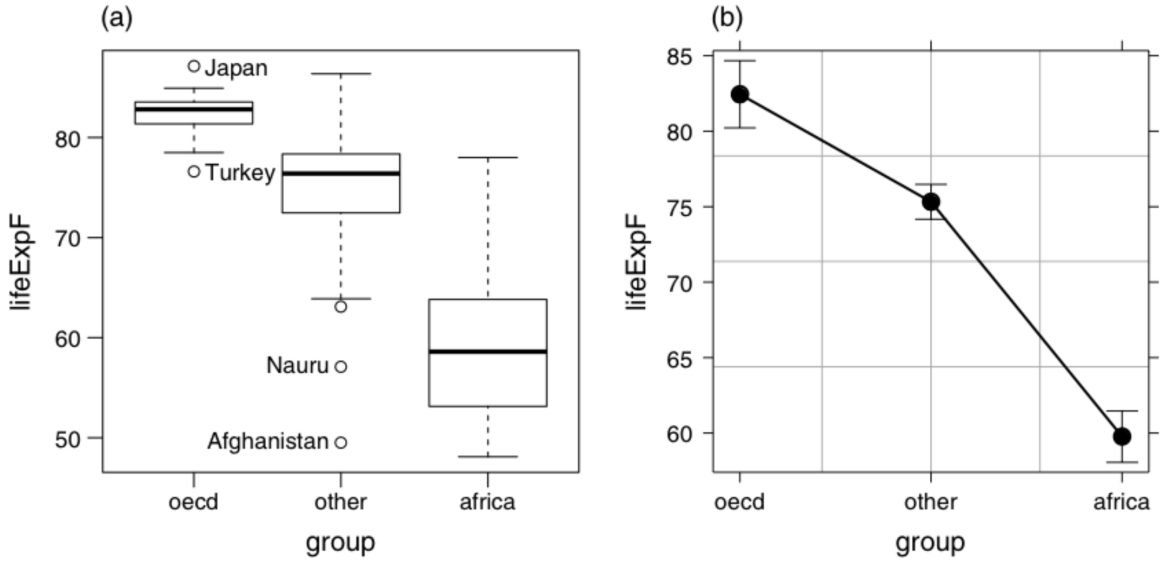


Figure 5.1 UN data. (a) Boxplot of `lifeExpF` separately for each `group` in the UN data. (b) Effect plot for `group` for the one-way model.

- when no other predictor in model other than group.

We model factors using dummy variables.

Factor with k levels, has $(k-1)$ dummy variables.

Example: For j^{th} dummy variable $u_{ij}, j = 1, \ldots, d$, has i^{th} value u_{ij} for $i = 1, \ldots, n$:

$$u_{ij} = \begin{cases} 1 & \text{, if group } i = j\text{th category} \\ 0 & \text{, otherwise} \end{cases}$$

The values of the dummy variables for the first 10 cases in the example are as follows:

	Group	U_1	U_2	U_3
Afghanistan	other	0	1	0
Albania	other	0	1	0
Algeria	africa	0	0	1
Angola	africa	0	0	1
Anguilla	other	0	1	0
Argentina	other	0	1	0
Armenia	other	0	1	0
Aruba	other	0	1	0
Australia	oecd	1	0	0
Austria	oecd	1	0	0

If we add an intercept, $\Rightarrow$ overparameterized model
because $u_1 + u_2 + u_3 = 1$ and column of 1's for intercept.
Solve this by dropping one of dummy variables.

$$E(\text{lifeExpF}|\text{group}) = \beta_0 + \beta_2 U_2 + \beta_3 U_3$$

Since the first level of group will be implied when $U_2 = U_3 = 0$,

$$E(\text{lifeExpF}|\text{group} = \text{oecd}) = \beta_0 + \beta_2 0 + \beta_3 0 = \beta_0$$

and so β_0 is the mean for the first level of *group*.
For the second level, $U_2 = 1$ and $U_3 = 0$,

$$E(\text{lifeExpF}|\text{group} = \text{other}) = \beta_0 + \beta_2 1 + \beta_3 0 = \beta_0 + \beta_2$$

and $\beta_0 + \beta_2$ is the mean for the second level of *group*.

Similarly, for the third level, $U_2 = 0$ and $U_3 = 1$,

$$E(\text{lifeExpF}|\text{group} = \text{africa}) = \beta_0 + \beta_2 0 + \beta_3 1 = \beta_0 + \beta_3$$

Table5.1 Regression Summary for Model (5.4)

	Estimate	**Std. Error**	**t-Value**	**Pr**$(>\|t\|)$
(Intercept), $\hat{\beta}_0$	82.4465	1.1279	73.09	0.0000
other, $\hat{\beta}_2$	-7.1197	1.2709	-5.60	0.0000
africa, $\hat{\beta}_3$	-22.6742	1.4200	-15.97	0.0000

$\hat{\sigma} = 6.2801$ with 196 df, $R^2 = 0.6191$

Means for three groups are:

$$\hat{E}(\text{lifeExpF}|\text{group=oecd}) = \hat{\beta}_0 + \hat{\beta}_2 \cdot 0 + \hat{\beta}_3 \cdot 0 = 82.45$$

$$\hat{E}(\text{lifeExpF}|\text{group=other}) = \hat{\beta}_0 + \hat{\beta}_2 \cdot 1 + \hat{\beta}_3 \cdot 0 = 82.45 - 7.12$$

$$\hat{E}(\text{lifeExpF}|\text{group=africa}) = \hat{\beta}_0 + \hat{\beta}_2 \cdot 0 + \hat{\beta}_3 \cdot 1 = 82.45 - 22.67$$

Comparison of Level Means

With analysis of factors, we often care about comparison of means (adjusted for other factors and regressors).
Pairwise comparisons of means:
- requires SE of the differences between each pair of means.
Example: Comparing other to africa

$$\hat{\beta}_2 - \hat{\beta}_3 = -7.12 - (-22.67) = 15.55$$

SEs for this difference can be calculated.

Let $a = (0, 1, -1, 0)'$

$$\Rightarrow \ell = a'\beta = \beta_2 - \beta_3 \quad \text{(contrast)}$$

$$SE(\hat{\ell}|X) = \hat{\sigma}\sqrt{a'(X'X)^{-1}a}$$

$$= \hat{\sigma} \quad \underbrace{\sqrt{C_{22} + C_{33} - 2C_{23}}}_{C_{ij} \text{ is the } (i,j)^{th} \text{ element of } (X'X)^{-1}}$$

Then use t-value (t-test) to generate p-value:

$$\frac{\hat{\ell}}{SE(\hat{\ell})} \overset{H_0}{\sim} t_{n-(p+1)}$$

Table5.2 Pairwise Comparisons of Level Means for Group

Compariso	Estimate	SE	t-Value	p-Value
oecd - other	7.12	1.27	5.60	0.000
oecd - africa	22.67	1.42	15.97	0.000
other - africa	15.55	1.04	14.92	0.000

Health Equity Measurement: Considerations for Selecting a Benchmark

Prepared by SHADAC

STATE Health & Value STRATEGIES | *Driving Innovation Across States*

A grantee of the Robert Wood Johnson Foundation

September 2023

Approaches to Benchmarking: Reference Groups

As previously noted, a health equity measurement benchmark is essentially a starting point against which performance of different subgroups are measured. One commonly used approach in measuring health equity, therefore, is to compare the performance of subgroups against each other, with the understanding that any gaps in performance between groups represent a disparity.[1] For instance, one might compare health insurance coverage rates among people with higher incomes (who typically have higher rates of coverage) against rates of people with moderate incomes and lower incomes—effectively treating the health insurance coverage rate for people with higher incomes as the measurement benchmark against which the other two groups' rates are compared.

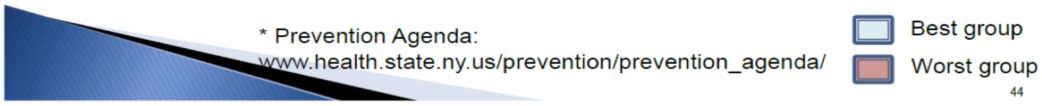
Best-Performing Group

One way to measure health equity is to compare the performance of each population subgroup against the group with the "best" performance on a given metric (i.e., having the highest rate where higher is considered better or the lowest rate where lower is considered better). We use the term "best-performing" group for the sake of clarity and to maintain consistency with literature written on the topic of benchmarking for health equity measurement. However, it is important to note that subgroups' performance on health measures is typically heavily influenced by health systems and other social factors; in reality, stratifying measures by demographic categories provides data on which groups are being well-served by health systems and social structures, rather than which groups are performing well themselves.

For example, New York used the best-performing approach in its report examining health disparities in the state (Figure 1).[4] For each individual measure, the state highlights the best-performing group in blue. The best-performing group varies by measure; in many cases—but not all—White non-Hispanic (NH) groups performed the best. For one measure—early stage cervical cancer—Hispanic individuals performed the best.

Figure 1. Example: Measures of Access to Quality Healthcare in New York Displayed Using a Best-Performing Benchmark

Indicators	NYS	White NH	Black NH	Asian NH	Hispanic	Index of Disparity
Percent of adults with health care coverage	88.0%	92.0%	86.5%	NA	75.1%	7%
Percent of adults with regular health care provider	86.5%	90.0%	89.0%	NA	76.1%	6%
Percent of adults who have seen a dentist in the past year	74.2%	76.4%	68.8%	NA	73.3%	4%
Early Stage Breast Cancer	63.5%	65.5%	55.6%	63.6%	58.0%	6%
Early Stage Cervical Cancer	46.1%	45.9%	39.2%	47.2%	54.7%	9%
Early Stage Colorectal Cancer	45.0%	45.8%	43.5%	42.6%	40.0%	5%



Source: Tobias L. *New York State - Managing High Need Medicaid Patients*. NY: New York State Department of Health; 2011. https://www.health.ny.gov/statistics/community/minority/docs/2011-08-09_health_disparities_work_grp_present.pdf.

Best-Performing Approach Advantages

- Using the best-performing group as the benchmark implicitly establishes an expectation that the "best" level of performance is achievable. The assumption is that if a high level of performance can be achieved for one subgroup, then the performance of other subgroups could be raised to that level as well.
- By setting a relatively high level of performance as the benchmark, this approach sets an expectation of overall performance improvement for all but the best-performing group. Consequently, setting the benchmark to the best-performing group encourages improvement efforts focused on subgroups experiencing inequities, rather than encouraging further improvement on the best-performing group.

Best-Performing Approach Disadvantages

- As described earlier, in cases where a metric is tracked over time or where multiple metrics are being monitored, the best-performing group may differ over time or across measures. Both of those challenges could pose difficulties for producing the measures and for interpreting the measures, as the benchmarks themselves would be subject to change.
- Additionally, because any subgroup could potentially have the best performance on a given measure, there is the potential for volatility in the level of performance for the benchmark, as shown in Figure 1. The resulting volatility within the benchmark could cause confusion among people interpreting the data if it is not clearly displayed and explained, and it could, in turn, cause volatility in measures of health equity.

- While not necessarily unique to this benchmarking approach, the measurement of health inequities without context can result in misunderstandings about the causes of inequities. By using the best-performing group as the benchmark against which other subgroups are measured, this approach may reinforce the idea that groups experiencing inequities are somehow deficient rather than focusing on addressing underlying factors, such as systemic racism, that cause or contribute to the cause of the inequities themselves.
- This approach also has the potential to reinforce problematic narratives. For instance, when presenting data by race and ethnicity, measures showing Asian people as having the best performance could reinforce the "**model minority**" myth. For example, in 2021, population data showed that Asian Americans had a lower burden of COVID-19 mortality than the overall population.[5] However, more disaggregated data found that Chinese patients had the highest mortality rate of any racial or ethnic group. Researchers noted that "racism underlying the 'Model Minority' myth harms Asian Americans by perpetuating the perception that they do not have disparities and therefore are unworthy of resources, which leads to lack of data or inaccurate data for this population that then reinforces the misperception that Asian Americans do not have disparities."[6]
- Because the "best-performing" group approach is likely to result in White people being selected as a reference group, this approach reinforces the idea that the experience of White people is the norm and is an example of "White framing." [7,8] Using care received by White patients as the benchmark would not necessarily encourage the highest quality of care possible for people of color.

Most Socially Advantaged Group

Another approach to selecting a reference group benchmark for health equity measurement is to identify the subgroup experiencing the highest level of social advantage, which would be considered the group at the top of the social hierarchy with the most wealth, income, opportunities, and power—and which is least likely to experience racism and other forms of social oppression.[9] The rationale for this approach is that the most socially advantaged group would not be subject to the same disadvantages that cause health inequities for other groups, so it can serve as a benchmark for assessing health inequities rooted in discrimination and systemic biases.

Michigan, for example, uses the White population as its reference population in its annual Medicaid Health Equity Report (Figure 2). The state indicates that "the White population served as the reference population for all pairwise comparisons because, the White population is not exposed to racial/ethnic discrimination, any disparities from this population rate can be an indicator of the health effects of discrimination and racism."[10]

Figure 2. Example: Michigan Health Equity Measures Displayed Using a Socially Advantaged (White) Benchmark

Health Equity Summary

Michigan Medicaid Managed Care - All Plans

Please note that some of the tables in this report utilize color coding, in addition to labeling. The word "below" is in red and the word "above" is in green. Where applicable, a legend is provided below the table to provide further clarification.

Table 3a: *Summary Table - Difference from Reference (White)*

Race/Ethnicity	Breast Cancer Screening	Cervical Cancer Screening	Chlamydia Screening in Women - Total	Post-partum Care	Childhood Immunizations - Combination 3	Immunizations for Adolescents - Combination 1	Lead Screening in Children
Asian American/ Native Hawaiian/ Pacific Islander	NS	NS	Above	NS	Above	NS	NS
African American	Below	Above	Above	Below	Below	Below	Below
White	Reference	Reference	Reference	Reference	Reference	Reference	Reference
Hispanic	Above	Above	Above	Below	Above	Above	Above
American Indian/ Alaska Native	Below	NS	Above	NS	NS	NS	NS
All Plans	NS	NS	Above	Below	Below	NS	NS

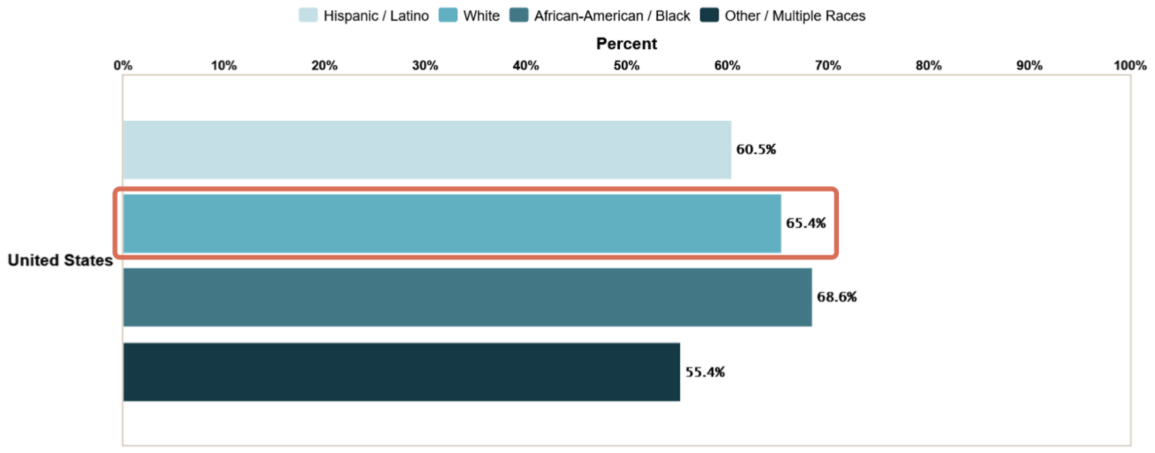
Rate is significantly higher than the Reference

Rate is significantly lower than the Reference

NS = *Not significantly different from the Reference*

Source: Michigan Department of Health & Human Services. *Medicaid Health Equity Project Year 9 Report (HEDIS 2019)*. MI: MDHHS; 2020. https://www.michigan.gov/-/media/Project/Websites/mdhhs/Folder50/Folder5/2019_Health_Equity_All-Plan_Report_Final_Digital_-_Accessible.pdf?rev=53548a8d823c42208ccfcd5d11e0d808

Figure 3. Example: Percent of U.S. Adults Who Have Received Recommended Cancer Screenings Displayed Using a White Benchmark (circled in orange)



Source: SHADAC. SHADAC analysis of the Behavioral Risk Factor Surveillance System public use files, State Health Compare. University of Minnesota. https://statehealthcompare.shadac.org/table/284/percent-of-adults-who-have-received-recommended-cancer-screenings-by-race-ethnicity#1/39,40,41,43/32/333,moe.

Socially Advantaged Approach Advantages

- In contrast with the best-performing group approach to benchmarking, the most socially advantaged group approach carries the benefit of typically employing a singular, consistent group across measures and over time. This can make the most socially advantaged group approach simpler to operationalize and interpret, because the benchmark group will be consistent when health equity is being measured across multiple metrics or over time.
- While not necessarily an inherent advantage or basis for using the most socially advantaged group approach, if White people are selected as the most socially advantaged group, it could limit volatility in benchmark estimates. That is because White people comprise a majority or plurality of the population in most parts of the United States, limiting the potential for small sample sizes to result in volatile benchmark estimates.

Socially Advantaged Approach Disadvantages

- Because the most socially advantaged group approach is likely to result in White people being selected as a reference group, this approach reinforces the idea that the experience of White people is the norm and is an example of "White framing."
- Although the White population will often be selected as the most socially advantaged group, the group is not a monolith, and White individuals with intersecting identities (such as poverty level, education, geography, or disability status) may experience different outcomes that might not be apparent. For example, MN Community Measurement's 2020 Health Care Disparities Report finds that even though White Minnesotans overall have the highest rate of colorectal cancer screening, White patients born outside the United States have significantly lower rates of colorectal cancer screening compared to White patients born inside the United States.[12]
- Although the most socially advantaged group may *often* have the best performance on health metrics, it may not *always* have the best performance. This may result in some situations in which people interpreting the data may be confused and wonder why sub-optimal performance is being used as the benchmark (see Figure 3).

Approaches to Benchmarking: Reference Points

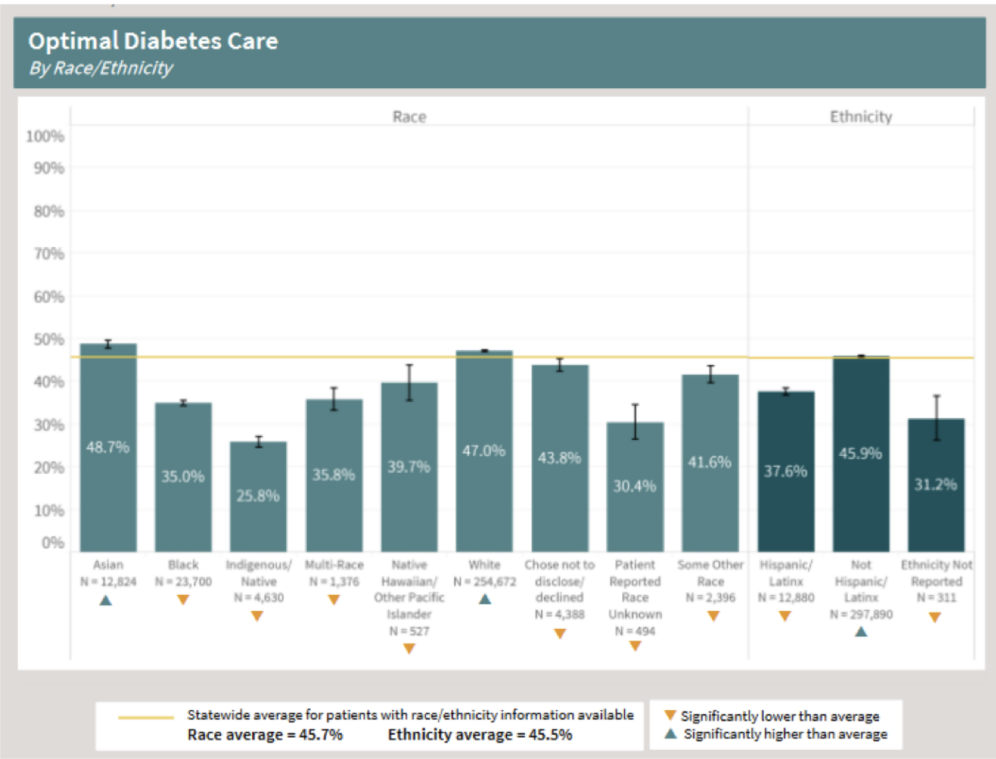
An alternative to using an existing reference group as the benchmark against which other groups are measured is to use a reference point that does not represent any particular group, but rather can be a point calculated on a broader scale, irrespective of any single group's measured performance. Two common options for this approach are to use either a **total population average**, such as the average for the entire U.S. or state population, or to use an entirely separate **target- or goal-setting approach** that may not depend on the existing performance of the overall population or any particular subgroups.

Population Average

Rather than using the measured performance of a particular population subgroup as a reference group, another approach is to use the average performance across the entire population as a benchmark. This straightforward approach entails calculating the population average for a given metric and level of analysis (e.g., state), then using that estimate to measure how much population subgroups differ from the population average.

For example, MN Community Measurement, a non-profit that serves as a contractor to the state of Minnesota for collecting and reporting quality data, uses the statewide average as the benchmark for comparison (while also comparing demographic subgroups to one another) in publications such as the 2020 Minnesota Health Care Disparities report (Figure 4).[12]

Figure 4. Example: Optimal Diabetes Care Measures in Minnesota Displayed Using a Population (State) Average Benchmark



Source: Donovan J, Nelson G. *Minnesota Health Care Disparities by Race, Hispanic Ethnicity, Language, and Country of Origin*. MN: MN Community Measurement; 2020. https://mncmsecure.org/website/Reports/Community%20Reports/Disparities%20by%20RELC/2020%20Disparities%20by%20RELC%20Chartbook%20-%20FINAL.pdf.

Population Average Approach Advantages

- Average population performance is already commonly reported in many health measures, so its use in health equity measurement is likely to be familiar to people interpreting the measurement results.
- The population average rate for any health measure will be less volatile than any single subgroup due to its size, so this approach will limit swings in the size of disparities that occur simply due to changes within any subgroup that would be used as the benchmark reference group (as would happen when using either the best-performing or most socially advantaged group).
- Using the total population average avoids any confusion with the reference group changing across measures and over time, as could happen with the best-performing group approach. It also avoids the disadvantages associated with selecting the most socially advantaged group (e.g., White framing).

Population Average Approach Disadvantages

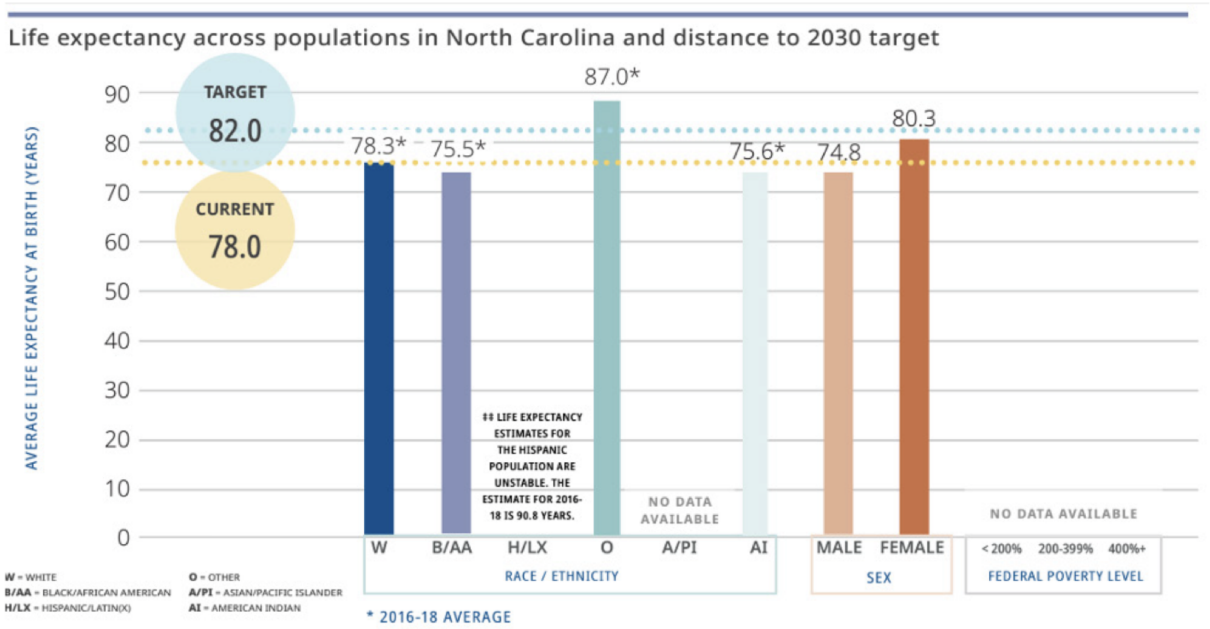
- Unlike with the best-performing group approach, using the population average as a reference point will always result in some groups performing better and some groups performing worse than the benchmark; consequently, this approach doesn't necessarily set expectations for raising overall performance.
- Interpreting results using the population average benchmark may not be as intuitive as using the best-performing group benchmarking approach, since some subgroups will have a better performance than average while others will be worse (i.e., positive and negative disparities). Additionally, because this approach compares each subgroup to the average, it does not explicitly report the full extent of disparities (i.e., the difference between the top-performing subgroup and bottom-performing subgroup).
- A potential concern in choosing any benchmarking approach, but particularly when selecting the population average, is that any approach displaying measures by average and by subgroup will result in comparison of subgroups, which without additional narrative context, could be interpreted as judgment on the worthiness of a group (e.g., groups with performance closer to or clustered around the reference point [the average] may be overlooked in favor of focusing on groups with the "highest" negative disparities) or can ignore the consequences of systemic racism on performance.
- Another consideration is the fact that since the population average performance would be influenced both by the demographic makeup of the population and the performance rates for those groups, the average performance is influenced by changes in population makeup as well as changes in performance. While this is unlikely to pose a serious risk when considering an entire state population, it could pose complications when considering smaller geographies within a state, which may see their demographics change more quickly.

Target or Goal Setting

One way to circumvent many of the potential disadvantages of the previously described options is to identify a "target" or "goal" benchmark that is potentially independent of the current performance of the total population or any specific subgroups. Such a target could be identified through various means, such as a review of research that may identify an optimal level of performance, borrowing targets used by other consensus decisions, selecting the best performance from a peer entity (e.g., another state or country), or taking the population average performance and adding an amount of expected improvement on top of it.

North Carolina is another example of a state that uses a target goal to benchmark its Healthy North Carolina 2030 heath indicators (Figure 5).[14] Here the state used historical data to forecast a target value for life expectancy in 2030.

Figure 5. Example: Life Expectancy in North Carolina Displayed Using a Target Goal Benchmark



Source: North Carolina Institute of Medicine. *A Path Toward Health – Chapter 7 – Health Outcomes*. NC: NCIOM;2020. https://nciom.org/wp-content/uploads/2020/04/Ch-7_HNC-2030_Health-Outcomes2.pdf. Accessed September 8, 2023.

Target or Goal Setting Approach Advantages

- By untying the measurement benchmark from performance for any population group (whether total population or by specific subgroup), the target-setting approach allows for simultaneous encouragement toward an overall improvement in performance as well as reduction in disparities.
- Using an approach that sets a fixed target or goal mitigates a commonly cited concern that measurement of health equity could be gamed, perversely, through a decline in performance of best-performing subgroups, rather than the preferred result of improving performance of under-performing subgroups.
- As with using population averages and the best-performing group, targets or goals are intuitive measurement concepts that are relatively straightforward to interpret. And, similar to the best-performing group approach to benchmarking, target or goal setting sets an aspirational performance goal while avoiding the pitfalls of changing reference groups between measures and over time.

Target or Goal Setting Approach Disadvantages

- Depending on the metric, it may take more work to identify and set a target than it would to simply select an alternative benchmark that already exists (e.g., the best-performing group or the most socially advantaged group) or is easy to calculate (e.g., the population average).
- Because the target approach might simultaneously set expectations for overall improved performance as well as improving health equity, it runs the risk of diluting the focus on health equity—both in the effort that is put into performance improvement and in how the intent of the benchmark is viewed.

Adding a Continuous Predictor to Model (ANCOVA)

- Suppose we add log(ppgdp) to the model.

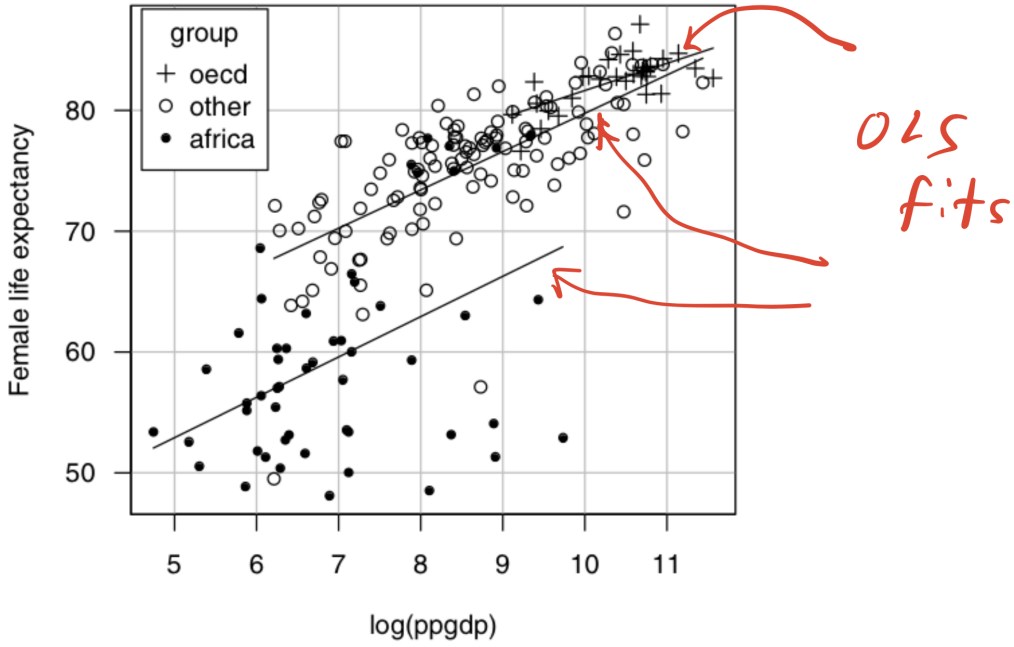


Figure 5.2 Plot of `lifeExpF` versus log(`ppgdp`) for the UN data. Points are marked with different symbols for the 3 levels of `group`, and OLS lines are shown for each of the 3 levels of `group`.

Can write this out as for $group = j$:

$$E(\text{lifeExp} \mid \log(\text{ppgdp}) = x,\ \text{group} = j) = \zeta_{0j} + \zeta_{1j}x \quad (j = 1, \ldots, d)$$

$$\Rightarrow 2d = 6 \text{ parameters} \quad \text{(separate slopes and intercepts)}$$

Can parametrize differently as:

$$E(\text{lifeExp} \mid \log(\text{ppgdp}) = x, \text{group}) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x + \beta_{12}U_2 x + \beta_{13}U_3 x$$

So,

$$\zeta_{01} = \beta_0 \qquad \zeta_{11} = \beta_1 \quad \text{(baseline)}$$

$$\zeta_{02} = \beta_0 + \beta_{02} \qquad \zeta_{12} = \beta_1 + \beta_{12}$$

$$\zeta_{03} = \beta_0 + \beta_{03} \qquad \zeta_{13} = \beta_1 + \beta_{13}$$

Table 5.3 Regression Summary for Model (5.7)

	Estimate	Std. Error	*t*-Value	Pr(>|t|)
(Intercept), $\hat{\beta}_0$	59.2137	15.2203	3.89	0.0001
`other`, $\hat{\beta}_{02}$	–11.1731	15.5948	–0.72	0.4746
`africa`, $\hat{\beta}_{03}$	–22.9848	15.7838	–1.46	0.1470
log(`ppgdp`), $\hat{\beta}_1$	1.5544	1.0165	1.53	0.1278
`other`: log(`ppgdp`), $\hat{\beta}_{12}$	0.6442	1.0520	0.61	0.5410
`africa`: log(`ppgdp`), $\hat{\beta}_{13}$	0.7590	1.0941	0.69	0.4887

$\hat{\sigma} = 5.1293$ with 193 *df*, $R^2 = 0.7498$.

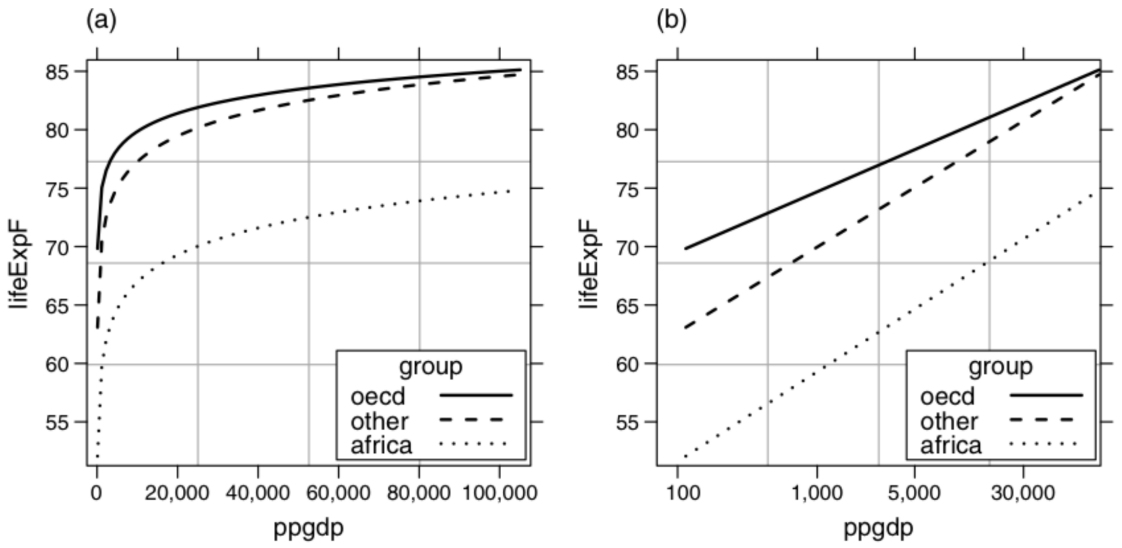


Figure 5.3 Effects plot for the interaction model (5.7) for the UN data. (a) `ppgdp` on the horizontal axis. (b) `ppgdp` in log-scale.

Case Study: Lead Exposure and Neuro-Psychological Function

Task: examine association between lead exposure and developmental features in children.

Study: Done in El Paso, TX

Control group ($n = 78$)

- blood-lead levels < 40 mg/100ml in both 1972 and 1973

Exposed group ($n = 46$)

- blood-lead levels ≥ 40 mg/100ml in either 1972 or 1973

Response: # finger wrist taps / 10 seconds in dominant hand

Dummy Variable for Two-Level Categorical Predictors

- Categories of predictor: A, B (for example)
- First category = reference cell, gets a zero
- Second category gets a 1.0
- Formal definition of dummy variable: $x = I[\text{category} = B]$,
  $I[w] = 1$ if w is true, 0 otherwise
- $\alpha + \beta x = \alpha + \beta I[\text{category} = B] =$
  α for category A subjects
  $\alpha + \beta$ for category B subjects
  β = mean difference $(B - A)$

Two-Sample *t*-test vs. Simple Linear Regression

- They are equivalent in every sense:
  - P-value
  - Estimates and C.L.s after rephrasing the model
  - Assumptions (equal variance assumption of two groups in t-test is the same as constant variance of $Y|X$ for every X)

$$\hat{\alpha} = \bar{y}_A$$

$$\hat{\beta} = \bar{y}_B - \bar{y}_A$$

$$SE(\hat{\beta}) = SE\left(\bar{Y}_B - \bar{Y}_A\right)$$

Analysis of Covariance

- Multiple regression can extend the t-test
  - More than 2 groups (multiple dummy variables can do multiple-group ANOVA)
  - Allow for categorical or continuous adjustment variables (covariates, co-variables)
- Model: $MAXFWT = \alpha + \beta_1 \text{age} + \beta_2 \text{sex} + e$
- Rosner coded $sex = 1, 2$ for male, female.
  - Does not affect interpretation of β_2 but makes interpretation of α more tricky (mean $MAXFWT$ when $age = 0$ and $sex = 0$ which is impossible by this coding).
- Better coding would have been $sex = 0, 1$ for male, female
  - $\alpha =$ mean $MAXFWT$ for a zero year-old male
  - $\beta_1 =$ increase in mean $MAXFWT$ per 1-year increase in age
  - $\beta_2 =$ mean $MAXFWT$ for females minus mean $MAXFWT$ for males, holding age constant
- Create derived variable that indicates exposure in either year.
  - Call the variable *exposure* and use the following formula for its derivation:
    `group != 'blood lead < 40mg/100ml in 1972 & 1973'`
- Model: $MAXFWT = \alpha + \beta_1 \text{exposure} + \beta_2 \text{age} + \beta_3 \text{sex} + e$
  exposure = **TRUE** (1) for exposed, **FALSE** (0) for unexposed
- $\beta_1 =$ mean $MAXFWT$ for exposed minus mean for unexposed, holding age and sex constant

Table 11.12 Simple linear-regression model comparing exposed and control children for MAXFWT (n = 95)

The REG Procedure

Model: MODEL1

Dependent Variable: maxfwt

Number of Observations Read	120
Number of Observations Used	95
Number of Observations with Missing Values	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	940.63327	940.63327	9.02	0.0034
Error	93	9697.30357	104.27208		
Corrected Total	94	10638			

Root MSE	10.21137	R-Square	0.0884
Dependent Mean	52.85263	Adj R-Sq	0.0786
Coeff Var	19.32046		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	55.09524	1.28651	42.83	<.0001
cscn2	1	−6.65774	2.21667	−3.00	0.0034

Table 11.13 Multiple-regression model of MAXFWT on age and sex (n = 95)

The REG Procedure

Model: MODEL1

Dependent Variable: maxfwt

Number of Observations Read	120
Number of Observations Used	95
Number of Observations with Missing Values	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5438.14592	2719.07296	48.11	<.0001
Error	92	5199.79092	56.51947		
Corrected Total	94	10638			

Root MSE	7.51794	R-Square	0.5112
Dependent Mean	52.85263	Adj R-Sq	0.5006
Coeff Var	14.22435		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	31.59139	3.16011	10.00	<.0001
ageyr	1	2.52068	0.25706	9.81	<.0001
sex	1	−2.36574	1.58722	−1.49	0.1395

Table 11.14 Multiple-regression model comparing mean MAXFWT between exposed and control children after controlling for age and sex ($n = 95$)

The REG Procedure

Model: MODEL1

Dependent Variable: maxfwt

Number of Observations Read	120
Number of Observations Used	95
Number of Observations with Missing Values	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5994.81260	1998.27087	39.16	<.0001
Error	91	4643.12424	51.02334		
Corrected Total	94	10638			

Root MSE	7.14306	R-Square	0.5635
Dependent Mean	52.85263	Adj R-Sq	0.5491
Coeff Var	13.51506		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	34.12129	3.09869	11.01	<.0001
cscn2	1	−5.14717	1.55831	−3.30	0.0014
ageyr	1	2.44202	0.24540	9.95	<.0001
sex	1	−2.38521	1.50808	−1.58	0.1172

From Rosner 7th ed.

The Main Effects Model

Fig 5.3 suggests intercepts might differ, but slopes may be equal.

$$\Rightarrow E(\text{lifeExp} \mid \log(\text{ppgdp}) = x, \text{ group})$$

$$= \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x$$

Main effects model [no interactions]

main effects model

Table 5.4 Regression Summary for Model (5.8)

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept), $\hat{\beta}_0$	49.5292	3.3996	14.57	0.0000
`other`, $\hat{\beta}_{02}$	−1.5347	1.1737	−1.31	0.1926
`africa`, $\hat{\beta}_{03}$	−12.1704	1.5574	−7.81	0.0000
log(`ppgdp`), $\hat{\beta}_1$	2.2024	0.2190	10.06	0.0000

$\hat{\sigma} = 5.1798$ with 195 df, $R^2 = 0.7422$.

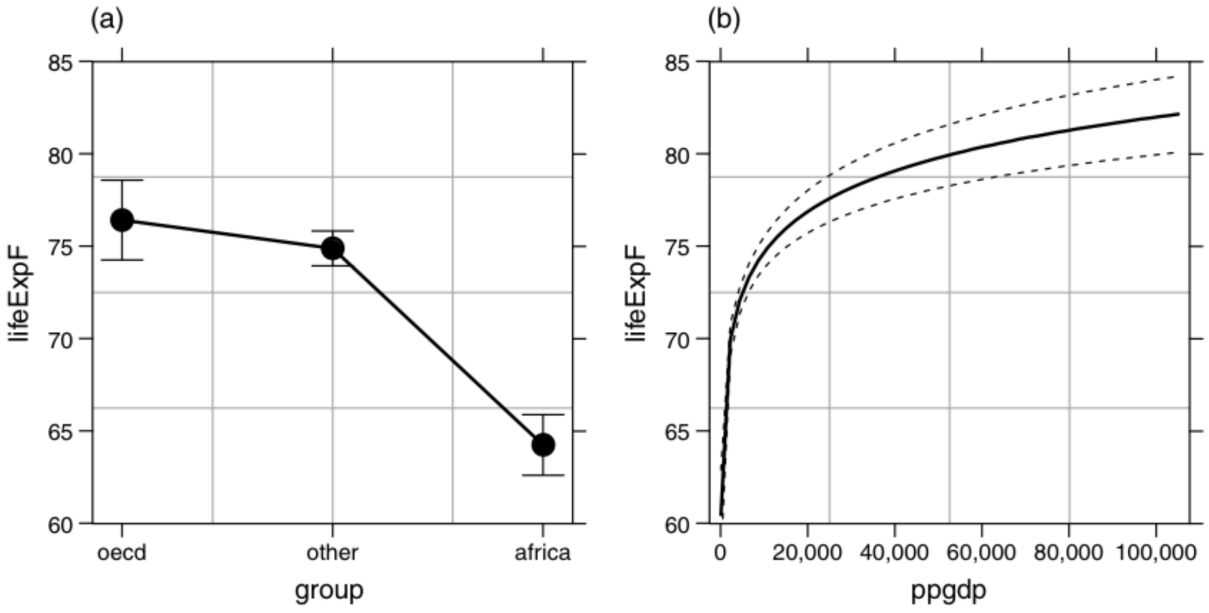


Figure 5.4 Effects plots for the main effects model (5.8) for the UN data: (a) group, (b) `ppgdp`. Dotted lines are drawn at plus and minus 1 standard error.

Many Factors

Let's say we have 3 factors, each with 3 possible levels.

$$\Rightarrow 3^3 = 27 \text{ possible combinations of the three factors}$$

Main effects means model

- Intercept and 2 dummy variables per factor

$\Rightarrow$ 7 total parameters

Second-order means model

- Adds in all 2-factor interactions

$\Rightarrow$ Total # parameters $7 + (3 \times 4) = 19$

Third-order means model

- Includes all 3-way interactions between factors

$\Rightarrow$ Total # parameters $19 + 8 = 27$

These kinds of means models are called ANOVA.

[will discuss in more detail shortly]

Polynomial Regression

Consider model:

$$E(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

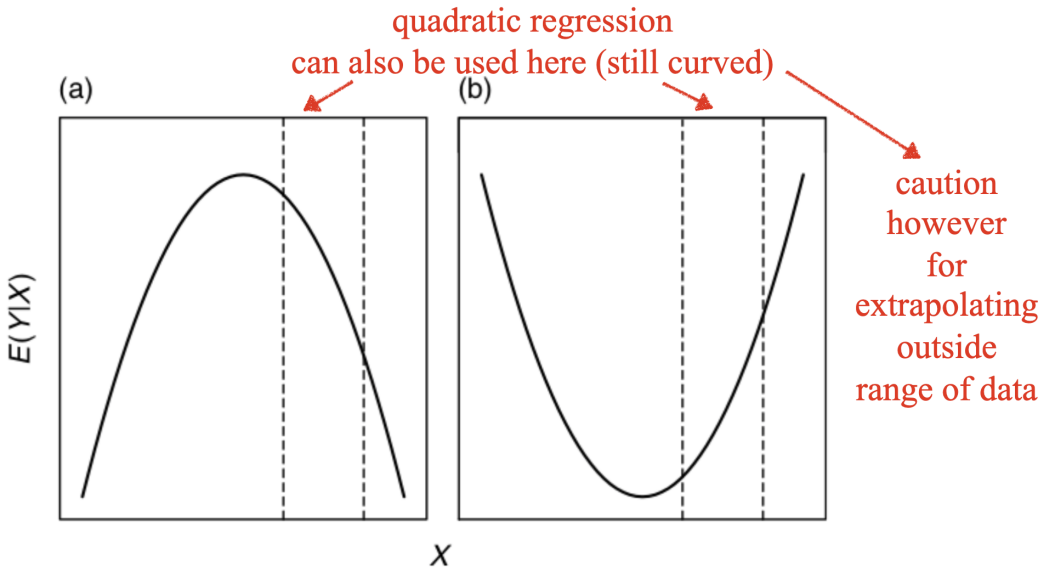


Figure 5.5 Generic quadratic curves. A quadratic is the simplest curve that can approximate a mean function with a minimum or maximum within the range of possible values of the predictor. It can also be used to approximate some nonlinear functions without a minimum or maximum in the range of interest, possibly using the part of the curve between the dashed lines.

Min or Max will occur at value of X where

$$\frac{dE(Y \mid X = x)}{dx} = 0 \quad \Rightarrow x_\mu = -\frac{\beta_1}{2\beta_2}$$

Polynomial regression model

$$E(Y \mid X = x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d$$

$$d = 2 \Rightarrow \text{quadratic} \ , \ d = 3 \Rightarrow \text{cubic}$$

[Can approx. any smooth curve with high enough order polynomial]

Polynomials With Several Predictors

Consider x_1, x_2 as predictors

$$E(Y \mid X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

$x_1 x_2$ is multiplicative interaction

With k regressors, have intercept, k main effects, $k(k-1)/2$ (two-way) interactions.

$\therefore$ with $k = 5 \Rightarrow 21$ regressors, $k = 10 \Rightarrow 66$ regressors.

An illustrative example:

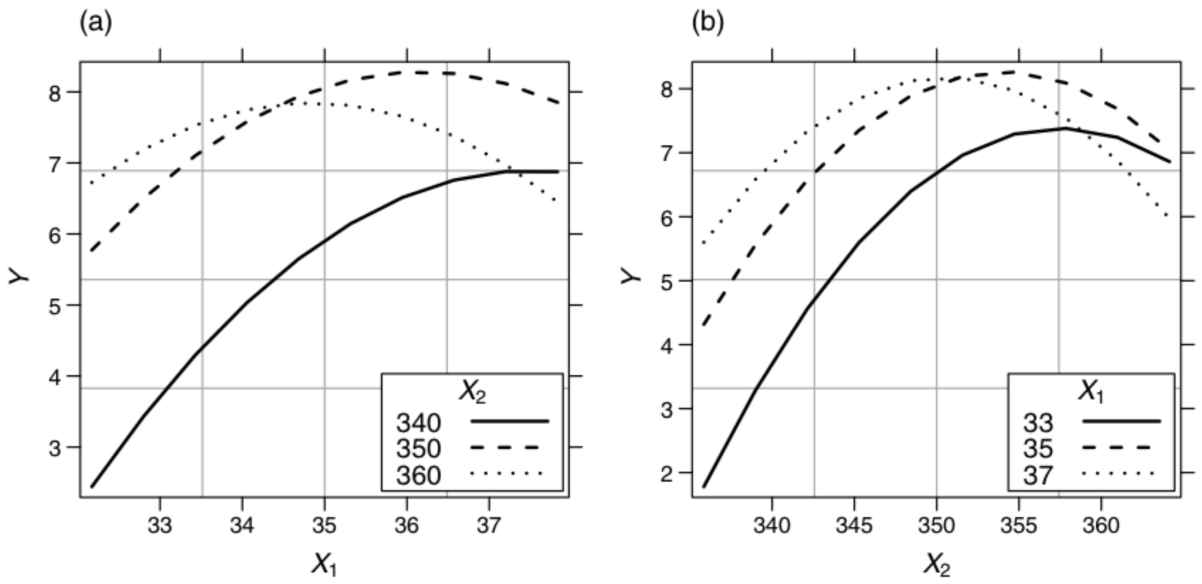


Figure 5.6 Effects plots for the cakes data, based on (5.13). Both plots show the same effects, in (a) with X_1 on the x-axis and levels of X_2 indicated by separate curves, and in (b) with X_2 on the x-axis and levels of X_1 indicated by separate curves.

Splines

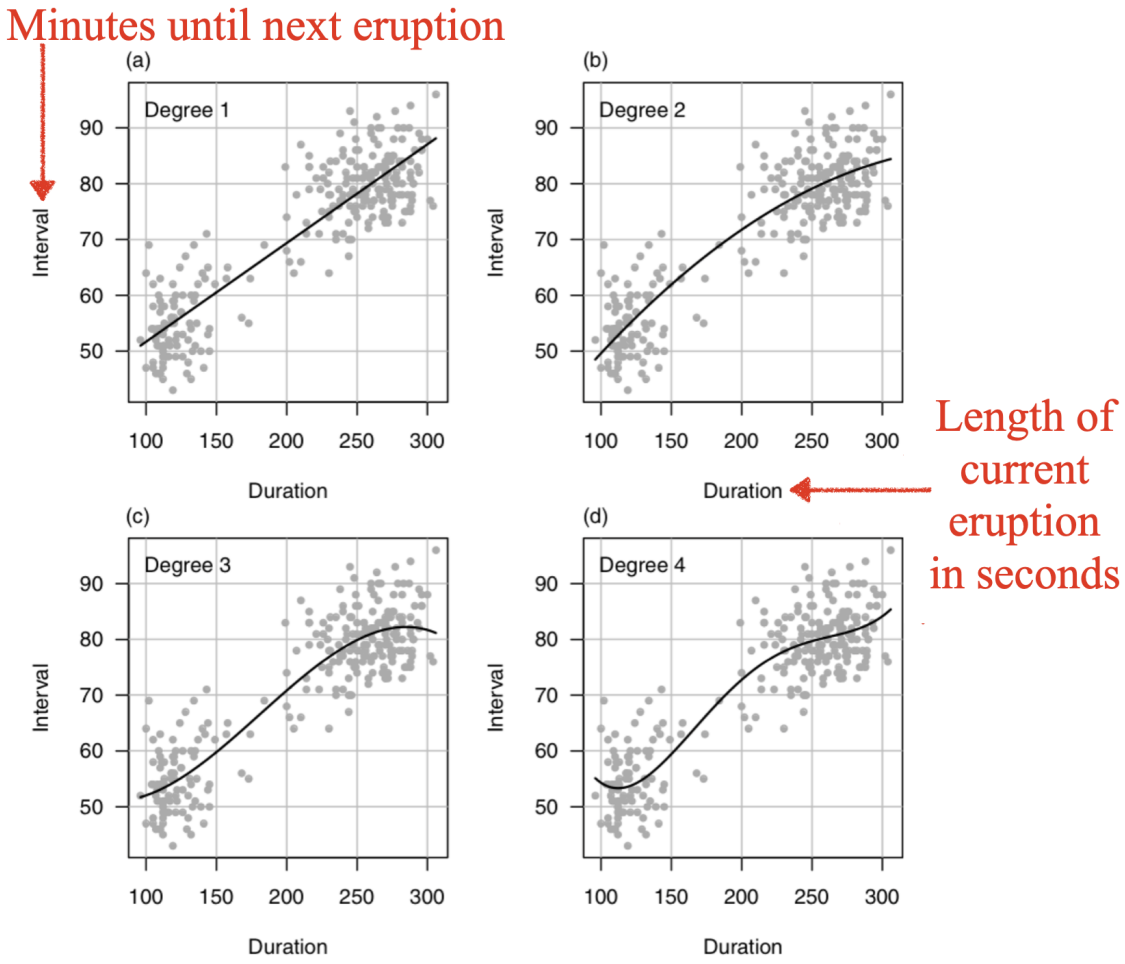


Figure 5.7 Polynomial fits for the Old Faithful Geyser data.

Increasing degree of polynomial can improve some aspects of fit but make others worse [global vs local].

Polynomial fit is weighted sum of basis functions

$$E(Y \mid X = x) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j$$

Basis functions are monomials $\{x^1, x^2, \ldots, x^d\}$

Weights are β's

Useful for global fitting because defined for all possible values of X.[not local]

Splines

- different set of basis functions
- defined locally

⇒ changing weight of one basis function will mostly affect fitted curve only for a limited range of X.

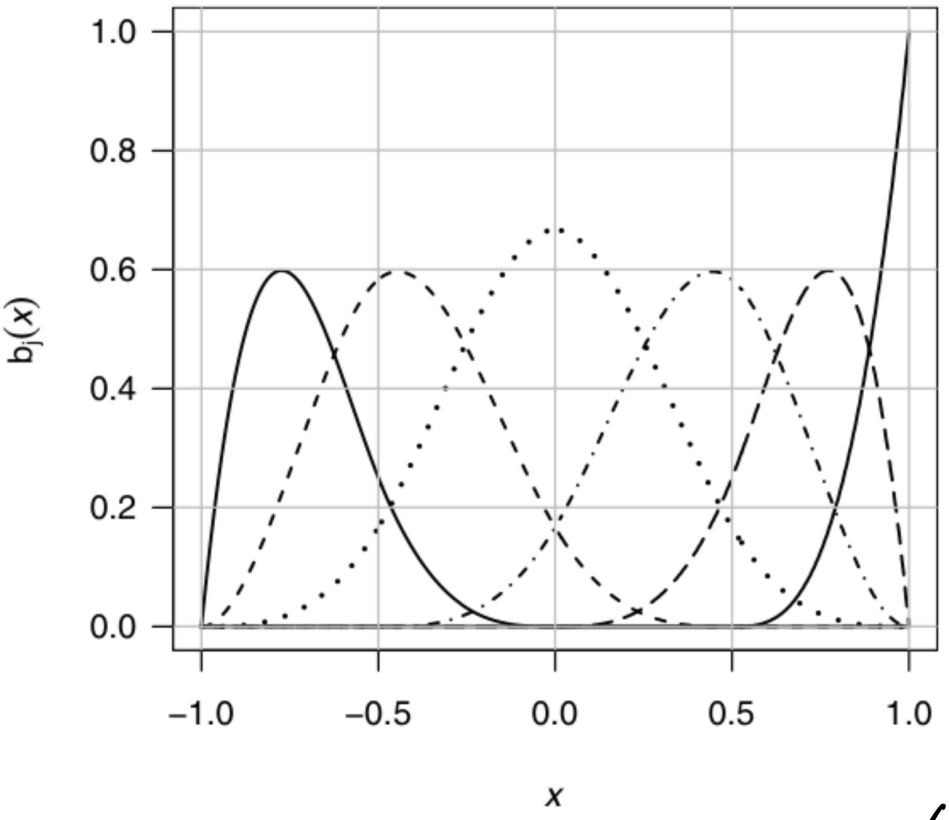


Figure 5.8 The B-spline basis with $d = 6$ basis functions. $(b_j(x))$

General Model

$$E(Y \mid X = x) = \beta_0 + \sum_{j=1}^{d} \beta_j b_j(x)$$

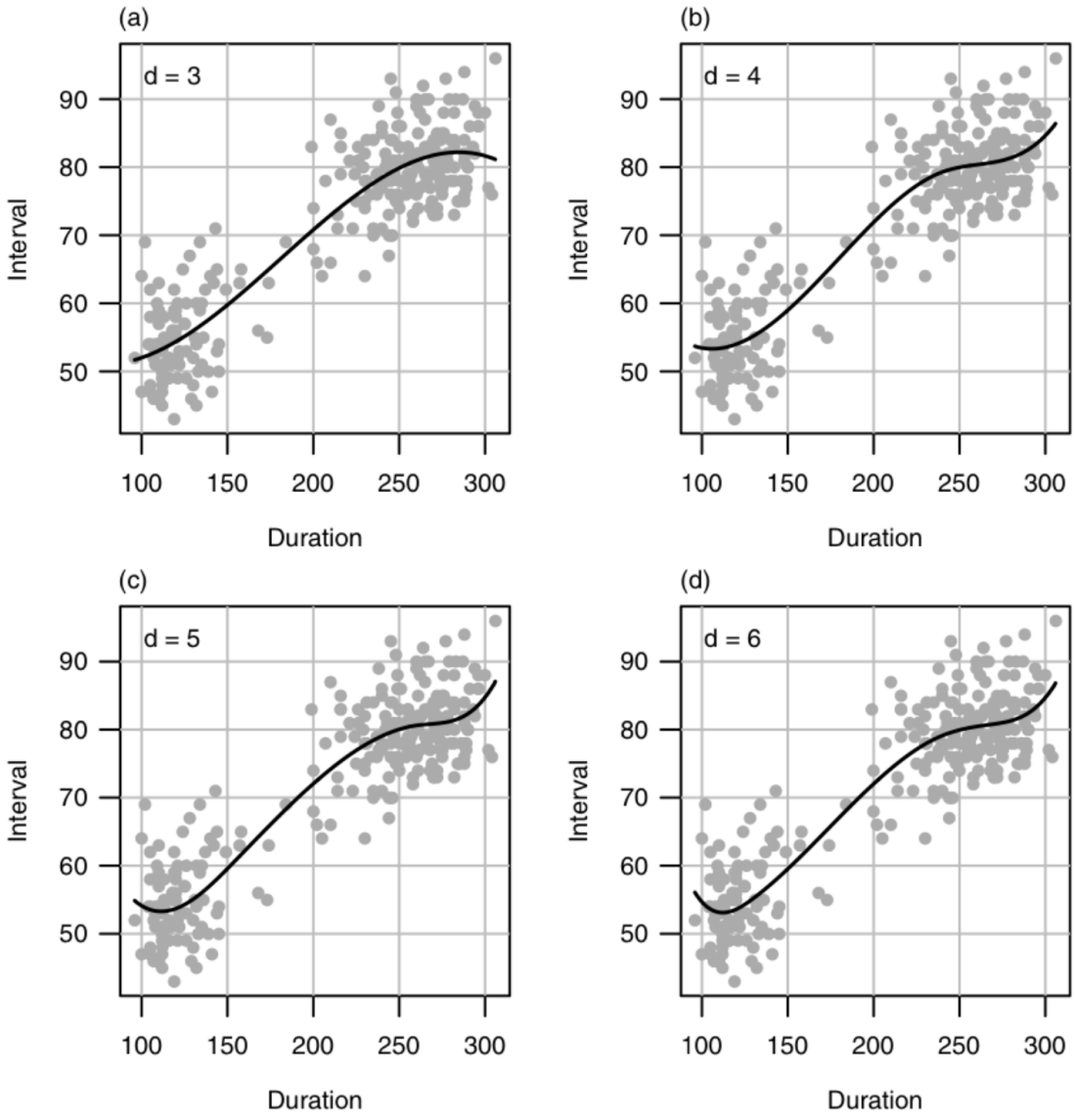


Figure 5.9 Spline fits for the Old Faithful geyser data.

How to estimate weights (coefficients)?

- use OLS

- but degree of polynomial d is a tuning parameter controlling amount of smoothness.

[May discuss more later]