Final report

- Group Topic: The Causal Effect of Waist Circumference on Resting Pulse Rate in Adults
- Team members: Alex Ho, Alexa Sheldon, Ena Kovac

Causal Question

Describe your causal question in a way that someone who has not taken this class would understand. Why are you interested in this question? How could answering this question allow for better decision making? Include any necessary background or context. Cite outside sources you use.
Answer We are interested in understanding whether having a high-risk waist circumference causes higher resting pulse rate in adults. Resting pulse rate is a widely used indicator of cardiovascular health, and elevated pulse has been associated with hypertension, arrhythmias, and increased risk of cardiovascular disease. Waist circumference, which reflects central (visceral) adiposity, is increasingly recognized as an important marker of cardiometabolic risk, sometimes beyond what body mass index (BMI) captures.

This question matters because most available evidence documents correlations between obesity-related measures and pulse rate, but correlation alone does not tell us whether excess central adiposity itself contributes to higher pulse. Answering this question causally could help clinicians and public health practitioners better assess cardiovascular risk and prioritize interventions that focus on reducing central obesity, rather than relying solely on BMI, when making treatment or prevention decisions.

Describe your causal question in the language of causal inference we've learned in this course: What is the treatment? What is the outcome? What are the potential outcomes? Write these out in words and in the math notation we have used in class.
Answer

- **Treatment (A):** Waist circumference category
  We define treatment as having a *low-risk waist circumference* and control as having a *high-risk waist circumference.*
- **Outcome (Y):** Resting pulse rate (beats per minute)
- **Potential outcomes:**
  For each individual i, we define:
  - $Y_i(1)$: resting pulse rate if the individual had a low-risk waist circumference
  - $Y_i(0)$: resting pulse rate if the individual had a high-risk waist circumference

The causal estimand of interest is the **average treatment effect on the treated (ATT)**:

$$\mathbb{E}[Y(1) - Y(0) \mid A = 1]$$

Causal Diagram

Draw a DAG representing your causal question that includes at least three relevant variables besides treatment and outcome that are included in your dataset. You may include more than three variables. You may include variables that are not in your dataset, but at least 3 of your variables (excluding treatment and outcome) must be included in your dataset. If you use letters to denote variables, make sure they are clearly defined
Answer

Our DAG includes the following variables:

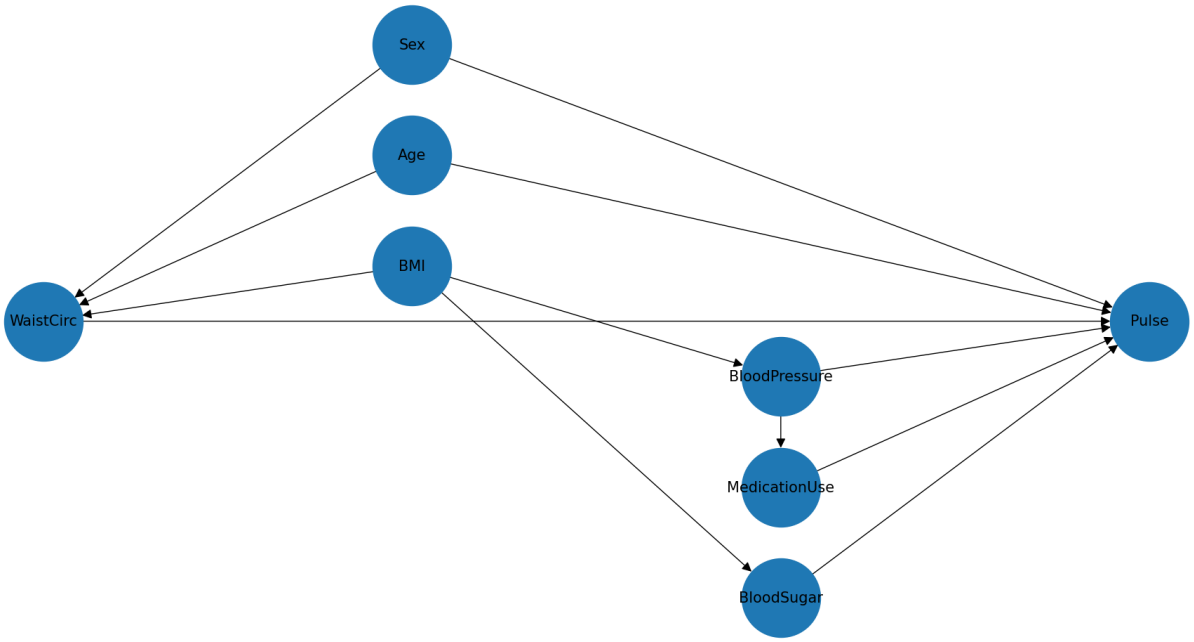

Figure 1: DAG

- **Treatment:** WaistCirc (A) – high- versus low-risk waist circumference
- **Outcome:** Pulse (Y) – resting pulse rate

Covariates from the dataset:
- Sex
- Age
- BMI (body-mass-index category)
- BloodPressure (blood pressure category)
- BloodSugar (indicator for elevated blood sugar/diabetes)
- MedicationUse (indicator for antihypertensive medication use)

The DAG has directed edges:

- Into WaistCirc from Sex, Age, and BMI
- From WaistCirc into Pulse
- Into Pulse from Sex, Age, BloodPressure, BloodSugar, and MedicationUse
- From BMI into BloodPressure and BloodSugar
- From BloodPressure into MedicationUse

Explain your DAG: tell us in words what is meant by each edge in your DAG.
Answer
- **Sex → WaistCirc, Sex → Pulse:** Biological sex affects fat distribution and cardiovascular physiology, and females tend to have higher resting pulse than males.
- **Age → WaistCirc, Age → Pulse:** Aging is associated with increased central adiposity and cardiovascular changes that affect pulse.

- **BMI → WaistCirc:** Higher overall adiposity increases the likelihood of high-risk waist circumference.
- **BMI → BloodPressure, BMI → BloodSugar:** Higher BMI worsens metabolic and cardiovascular health.
- **BloodPressure → MedicationUse:** Individuals with hypertension are more likely to take antihypertensive medications.
- **BloodPressure → Pulse:** Elevated blood pressure reflects increased cardiac workload.
- **BloodSugar → Pulse:** Diabetes and poor glycemic control affect autonomic regulation.
- **MedicationUse → Pulse:** Some antihypertensive medications directly affect heart rate.
- **WaistCirc → Pulse:** Central adiposity may increase pulse through inflammation, sympathetic activation, and cardiovascular strain.

Discuss your DAG. How realistic is it? Are there variables or edges you excluded from your DAG that someone else might argue should be included? Playing devil's advocate, how would you critique the reliability of your DAG?
Answer Our DAG is a reasonable simplification of well-established biological relationships, but it omits several plausible confounders such as physical activity, smoking, diet, stress, alcohol use, and genetics. These factors may affect both waist circumference and resting pulse. Additionally, because the data are cross-sectional, the assumed temporal ordering may not fully reflect reality. Measurement error and misclassification are also possible. Overall, the DAG is a simplified representation of our assumptions rather than a definitive causal structure.

Method and Identification

What method are you using to estimate a causal effect? What causal effect are you estimating (e.g., ATE vs LATE vs ATT)? What assumptions are required to identify the causal effect via your chosen method?
Answer

We estimate the **Average Treatment Effect on the Treated (ATT)** of having a low-risk waist circumference (treatment) versus a high-risk waist circumference (control) on resting pulse rate.

Our analysis uses a two-step approach: matching followed by a parametric g-formula. First, we perform nearest-neighbor matching with Mahalanobis distance while specifying the ATT as the estimand. Matching is done on sex, age, BMI, blood pressure category, blood sugar status, and antihypertensive medication use. This step improves covariate balance and ensures that individuals with low-risk waists are compared to similar individuals with high-risk waists.

Second, we apply a parametric g-formula using a weighted linear regression model for resting pulse as a function of waist circumference category and the same set of covariates. Using this fitted model, we predict counterfactual pulse outcomes for treated individuals under both the treated (low-risk waist) and control (high-risk waist) conditions while holding covariates fixed. The ATT is computed as the average difference between these two predicted outcomes among individuals who actually have low-risk waists.

Identification of this causal effect relies on several assumptions. We assume conditional exchangeability given the observed covariates, meaning that after adjustment there are no unmeasured confounders of waist circumference and resting pulse among the treated. We also assume positivity, so that individuals with covariate values observed among the treated could plausibly have either low- or high-risk waist circumference. Additionally, we assume consistency, meaning each individual's observed pulse corresponds to the potential outcome under their observed waist category, and that the linear outcome model used for the g-formula is reasonably well specified.

Under these assumptions, our matching and parametric g-formula approach identifies the ATT of low-risk versus high-risk waist circumference on resting pulse.

Explain what conditional exchangeability means in the context of your causal question. Is it important? Why or why not? How do sufficient adjustment sets relate to conditional exchangeability?
Answer for the observed covariates, individuals with low-risk and high-risk waist circumference are comparable with respect to factors that influence resting pulse rate. In our context, this implies that once we condition on sex, age, BMI, blood pressure, blood sugar, and antihypertensive medication use, waist circumference classification is independent of the potential resting pulse outcomes among those with low-risk waists.

This assumption is crucial because without conditional exchangeability, differences in resting pulse could be driven by confounding rather than a causal effect of waist circumference. Sufficient adjustment sets are directly related to conditional exchangeability: a sufficient adjustment set blocks all non-causal (backdoor) paths between treatment and outcome in the DAG, thereby making conditional exchangeability plausible.

Assuming your DAG is true, list out all non-causal paths between treatment and outcome and list one sufficient adjustment set to identify the causal effect of the treatment on the outcome. If a sufficient adjustment set does not exist, add additional variables to your DAG so that one does exist.
Answer Assuming the DAG is correct, the non-causal (backdoor) paths between waist circumference and resting pulse include paths through common causes of both variables. These include pathways through sex and age, as well as metabolic pathways through BMI that operate via blood pressure, blood sugar, and antihypertensive medication use. For example, waist circumference and pulse are associated through the paths WaistCirc $\leftarrow$ Sex $\rightarrow$ Pulse, WaistCirc $\leftarrow$ Age $\rightarrow$ Pulse, WaistCirc $\leftarrow$ BMI $\rightarrow$ BloodPressure $\rightarrow$ Pulse, WaistCirc $\leftarrow$ BMI $\rightarrow$ BloodSugar $\rightarrow$ Pulse, and WaistCirc $\leftarrow$ BMI $\rightarrow$ BloodPressure $\rightarrow$ MedicationUse $\rightarrow$ Pulse.

A sufficient adjustment set that blocks all of these non-causal paths is sex, age, BMI, blood pressure, blood sugar, and antihypertensive medication use.

Discuss the plausibility of conditional exchangeability in your setting. If your sufficient adjustment set contains variables that are not in your dataset, discuss the implications.
Answer Conditional exchangeability is plausible but imperfect in our setting. While we adjust for several important demographic and metabolic confounders available in the Add Health dataset, there are likely unmeasured factors such as physical activity, diet, smoking, stress, and socioeconomic or neighborhood characteristics that influence both waist circumference and resting pulse rate. These variables are not included in our dataset, which means that some residual confounding may remain. As a result, our causal estimates should be interpreted cautiously, recognizing the possibility of bias due to unmeasured confounders.

Discuss any other identification assumptions for your method here, such as positivity and consistency. What do they mean in the context of your causal question and are they plausible?
Answer In addition to conditional exchangeability, our analysis relies on several other identification assumptions. The positivity assumption requires that, for individuals with covariate values observed among the treated group, there is a non-zero probability of having either low-risk or high-risk waist circumference. This appears reasonable in our data, though extreme combinations of covariates may be rare. The consistency assumption requires that each individual's observed resting pulse corresponds to the potential outcome under their observed waist circumference category, which is plausible given the clear definition of waist risk categories. We also assume no interference between individuals and that the parametric outcome model used in the g-formula is reasonably well specified.

Together, these assumptions allow us to interpret our estimated ATT as a causal effect under the stated design and modeling choices.

Discussion: Analysis and Results

Give some context for your dataset. Who is included in your dataset? How was the data collected? When was the data collected? Make sure to cite the dataset.
Answer The dataset we were working with was Wave V from Add Health. The National Longitudinal Study of Adolescent to Adult Health (Add Health) follows a sample of about 20,000 adolescents who were in grades 7-12 during the 1994-95 school year over many years, with Wave V being in 2016-2018. We used data from Demographics, Cardiovascular Measures, and Anthropometrics.

Wave V was a mixed-mode survey with some respondents having in-home interviews. This included repeat anthropometric, cardiovascular, metabolic, and inflammatory measures indicative of change in various health situations such as diabetes. https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/ZYRZ5J

Discuss any choices you made regarding data cleaning and processing: Did your data have missing values or outliers? How did you handle them? Were there any variables you dichotomized (i.e. made binary), or variables that you changed the format (e.g. yes/no to 1/0)?
Answer
Any individuals that had invalid or missing data in any of the variables were excluded from the analysis (including waist size, heart rate, body-mass-index, blood pressure, blood sugar level, anti-hypertensive medication use, however sex and age was present for each individual). Originally there were around 1,839 individuals, but throughout the cleaning process, we cut it down to 1,706 individuals who had valid data in the variables we were looking at. Outliers were not discounted since many of the variables that were analyzed were classifications (including waist, body-mass-index, blood-sugar level, and a flag for anti-hypertensive medication use).

Waist size was reported as a classification between high risk and low risk (was dependent on sex) in the Add Health data. This means it was already dichotomized, but it was classified with 1 being low risk and 2 being high risk. I changed this to be 0 being high risk and 1 being low risk.

Discuss the impact of any choices you made regarding your dataset, such as choices you made in data cleaning or processing.
Answer
Using corresponding variable names for cardiovascular statistics (H5PR for pulse rate) and anthropometrics statistics (H5WSTCLS for waist size), we processed the data to exclude non-answers in waist circumference and pulse PR (ensuring that only relevant entries were considered). This may have reduced the total number of individuals for our analysis, but was necessary for the purpose of assessing our causal question. Without insights into these measures, we could not use every individual evaluated in Wave V, and we lose some potential confounder about whether or not they could be evaluated on that metric. For example, we do not know if low risk or high risk waist sized people would be more or less likely to have a valid pulse rate measurement.

Once we merged all of the data to just contain individuals with valid data for all of the relevant variables, we found that we had more individuals who were at high risk for waist (our control) than low risk waist (our treatment). This helps us to find more similar counterparts for each treated person if you have more control to select from. However, since the treatment group is smaller, it can be more impacted by randomness or outliers.

Explain how you estimated a causal effect.

- If you used matching, explain and discuss your choices. What formula did you use and why? What matching strategy did you use and why? Are there any advantages or drawbacks to the strategy you chose? How many units did your matching drop? How was the covariate balance in your matched sample? Discuss the implications of any choices you made and the quality of your matching.

- If you didn't use matching, explain any choices you made related to the method you used and discuss their implications. Think about advantages or drawbacks to any choices you made, possible bias-variance trade-offs, and assessing how well your method did.

Answer
We utilized matching to understand the effect of those who had a low risk (or smaller) waist size on their pulse rate (so we did ATT with low risk as treatment). Since we had various levels of classification (and differing ranges), we utilized Mahalanobis distance to account for the different scales of sex, age, bmi, blood pressure, blood sugar, anti-hypertensive medication use (as well as correlation between our different variables).

Once matched, we had 60 more units from the control group, so we dropped those 60 that were not nearest to any of the treatment individuals. We used 'nearest' to determine matching since we only had 823 matched units and had already reduced our number of individuals by requiring that they had valid entries for all relevant variables. There is a downside to this since all treated individuals were matched and there could have been an outlier in that set that had to get matched with a control that was not similar to it (but the control was the most similar to it).

In terms of the quality of our matching, the distribution of our matched set of units was quite similar to our data before matching. It has the same min and max, while the mean is only less by .24 and the median was less by .25. Since the distributions were similar, we know that our data is likely quite representative of our individuals with all the relevant valid data.

Report your causal effect estimate and interpret it in the context of your causal question.
Answer From our results (using parametric g-formula), we see that for the variable "waist_class_bin" (the binary value assigned to the waist classes, for lower or higher circumferences), the calculated coefficient causal effect value is "-2.934693" (when rounding, approximately -2.93). In the context of our analysis, this estimated result for the causal effect means that moving from the high-risk to low-risk waist circumference causally lowers resting heart rate by around 2.93 bpm for people who actually have low-risk waists (the Average Treatment Effect on the Treated, ATT).

Discuss the limitations of your analysis: what are the limitations of your dataset? Is there other data you would have wanted to have to bolster your analysis? Playing devil's advocate, how would you critique the reliability of you causal estimate?
Answer Our analysis has several important limitations related to both the dataset and the methodological choices we made. First, the Add Health Wave V data are cross-sectional for the purposes of our analysis, which limits our ability to establish temporal ordering. Although our DAG encodes a plausible causal direction from waist circumference to resting pulse, reverse causality or simultaneous determination cannot be fully ruled out.

Second, while matching improves balance between treated and control units, our use of nearest-neighbor matching may pair some treated individuals with controls that are only weakly similar, especially for treated units with uncommon covariate profiles. This issue is compounded by the fact that the treated group (low-risk waist circumference) is smaller than the control group, making the ATT estimate more sensitive to randomness and potential outliers among treated units.

Third, several important confounders are not available in the dataset, including physical activity, diet, smoking status, stress, fitness level, and socioeconomic or neighborhood-level factors. These omitted variables may affect both waist circumference and resting pulse, raising concerns about residual confounding and weakening the plausibility of conditional exchangeability.

Additionally, many of the variables used in our analysis are categorical classifications rather than continuous measures, such as BMI category, blood pressure stage, and waist circumference risk classification. This coarsening may obscure within-category variation and reduce precision. In particular, the estimated effects of BMI and medication use are imprecise, with confidence intervals that overlap zero, suggesting substantial uncertainty about their true effects.

Finally, the parametric g-formula relies on correct specification of the outcome model. If the linear model does not adequately capture the true relationship between covariates and resting pulse, our causal estimates may be biased even after matching. Taken together, these limitations suggest that while our estimated ATT is consistent with prior literature and biologically plausible, it should be interpreted as suggestive rather than definitive evidence of a causal effect.

Code:

```
# This is where your code goes for the data cleaning/processing and analysis.
# Make your code clean and easy-to-follow. Add short comments to explain what you are doing.
# If a classmate who wasn't as familiar with R were to read through this section,
# would they be able to follow along?

library(haven)
# reading in xpt files from ADD Health
demo <- read_xpt("pdemo5.xpt")
cardio <- read_xpt("pcardio5.xpt")
anthro <- read_xpt("panthro5.xpt")

## ===== Isolating Variables to be Used =====

# For our Treatment variable, we use anthro$H5WSTCLS
# with waist circumference classifications (1 = low risk, 2 = high risk)

# For our Outcome variable, we use cardio$H5PR, which measures heart rate (per minute)
# numerically


## ===== Cleaning the data =====

library(dplyr)
library(tibble)
# pruning for non-answers in waist circumference (94, 96, 97, 99 are invalid)
waist <- enframe(anthro$H5WSTCLS, name = "individual", "waist_class")
waist_fil <- waist %>% filter(waist_class <=2)
waist_fil <- waist_fil %>% mutate(waist_class_bin = waist_class %% 2)
# pruning for non-answers in pulse PR (9996 9997 9999 are invalid)
hr <- enframe(cardio$H5PR, name = "individual", "hr")
hr_fil <- hr %>% filter(hr < 9996)
# joining on individuals remaining
waist_hr <- merge(waist_fil, hr_fil, by = "individual")
#print(waist_hr)



## ===== Other variables =====

# No invalid entries for sex and age
# Biological sex
sex <- enframe(demo$H5Q011, name = "individual", "sex")
# Age by year
age <- enframe(demo$H5AGE, name = "individual", "age")
# BMI (classification), valid bmi class is <= 6
bmi <- enframe(anthro$H5BMICLS, name = "individual", "bmi")
bmi_fil <- bmi %>% filter(bmi <= 6)
# blood pressure (classification), valid bp class is <= 5
bp <- enframe(cardio$H5BPCLS5, name = "individual", "bp") #
bp_fil <- bp %>% filter(bp <= 5)
# blood sugar level, valid bs class is 0 or 1
```

```r
bs <- enframe(cardio$H5Q045B, name = "individual", "bs")
bs_fil <- bs %>% filter(bs == 0 | bs == 1)
# flag for anti-hypertensive medication use
# valid aht med flag is 0 or 1
med_use <- enframe(cardio$H5AHT, name = "individual", "med_use")
med_use_fil <- med_use %>% filter(med_use == 0 | med_use == 1)


## ===== Merging all variables together =====

output_tb <- waist_hr %>%
  merge(sex, by = "individual") %>%
  merge(age, by = "individual") %>%
  merge(bmi_fil, by = "individual") %>%
  merge(bp_fil, by = "individual") %>%
  merge(bs_fil, by = "individual") %>%
  merge(med_use_fil, by = "individual")
#print(output_tb)


## ===== Matching =====

library("tidyverse")
library("MatchIt")
library("marginaleffects")
# Perform matching using nearest neighbors with mahalanobis distance
matching_all <- matchit(waist_class_bin ~ sex + age + bmi + bp + bs + med_use,
                 data = output_tb,
                 method = "nearest",
                 estimand = "ATT",
                 distance = "mahalanobis")
summary(matching_all)$nn
```

```
##               Control Treated
## All (ESS)         883     823
## All               883     823
## Matched (ESS)     823     823
## Matched           823     823
## Unmatched          60       0
## Discarded           0       0
```

```r
summary(select(output_tb, hr))
```

```
##        hr
##  Min.   : 42.00
##  1st Qu.: 66.50
##  Median : 74.00
##  Mean   : 74.89
##  3rd Qu.: 82.00
##  Max.   :123.00
```

```r
summary(select(match.data(matching_all), hr))
```

```
##        hr
##  Min.   : 42.00
##  1st Qu.: 66.50
##  Median : 73.75
##  Mean   : 74.65
##  3rd Qu.: 81.50
##  Max.   :123.00
```

```r
## ===== Estimation of Causal Effect =====

# utilize parametric g-formula (implementation from lab 8 "matching_lab.Rmd")
# Use matched data & weights from MatchIt
matched_data <- match.data(matching_all)

outcome_model <- lm(
  hr ~ waist_class_bin + sex + age + bmi + bp + bs + med_use,
  data   = matched_data,
  weights = weights
)

summary(outcome_model)
```

```
##
## Call:
## lm(formula = hr ~ waist_class_bin + sex + age + bmi + bp + bs +
##     med_use, data = matched_data, weights = weights)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.918  -7.658  -0.863   6.989  48.757
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      69.8306     5.8032  12.033  < 2e-16 ***
## waist_class_bin  -2.9347     0.7826  -3.750 0.000183 ***
## sex               4.5981     0.6220   7.393 2.28e-13 ***
## age              -0.1666     0.1426  -1.168 0.242848
## bmi              -0.1006     0.3134  -0.321 0.748330
## bp                2.3451     0.2694   8.706  < 2e-16 ***
## bs                3.7461     1.4573   2.571 0.010239 *
## med_use           0.4941     0.9448   0.523 0.601025
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.18 on 1638 degrees of freedom
## Multiple R-squared:  0.1048, Adjusted R-squared:  0.1009
## F-statistic: 27.38 on 7 and 1638 DF,  p-value: < 2.2e-16
```

```r
# Calculate ATT via parametric g-formula
# Predict Y under treat=1 and treat=0 for the actually treated units
```

```r
treated_rows <- subset(matched_data, waist_class_bin == 1)

treated_as_treated <- treated_rows
treated_as_control <- treated_rows
treated_as_control$waist_class_bin <- 0

# Y^{a=1}
y1_hat <- predict(outcome_model, newdata = treated_as_treated)
# Y^{a=0}
y0_hat <- predict(outcome_model, newdata = treated_as_control)

# ATT (avg causal effect of small vs. large waist on heart rate)
att_gformula <- mean(y1_hat) - mean(y0_hat)
att_gformula
```

```
## [1] -2.934693
```

```r
# This outputs the calculated causal effect result: "-2.934693"


## ===== Marginal effects (additional) from lab =====

# Additional information (other than ATT of possible interest for future exploration)
avg_comparisons(
  outcome_model,
  variables = "waist_class_bin",
  vcov = ~subclass,
  newdata = subset(matched_data, waist_class_bin == 1)
)
```

```
##
##  Estimate Std. Error     z Pr(>|z|)    S 2.5 % 97.5 %
##     -2.93      0.771 -3.81   <0.001 12.8 -4.45  -1.42
##
## Term: waist_class_bin
## Type: response
## Comparison: 1 - 0
```