Biostatistics I: Linear Regression

Effect Plots

Nicole S. Erler

Department of Biostatistics, Erasmus Medical Center

✉ n.erler@erasmusmc.nl

@N_Erler

Erasmus MC
University Medical Center Rotterdam

The Multiple Linear Regression Model

$$y_i = f(\mathbf{x}_i)^\top \boldsymbol{\beta} + \varepsilon_i \qquad \mathrm{E}(\varepsilon_i) = \mathbf{0}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

Interpretation:
β_j is the expected change in y that is associated with an increase in x_j (or $f(x_j)$) of 1 unit **while all other covariates are kept constant**.

⇨ The interpretation of the coefficients of each of the covariate terms is **independent of all other covariate terms**.

But:
When the model includes **interaction terms** or **non-linear effects** that involve multiple coefficients for the same variable, the corresponding regression coefficients **cannot be interpreted independently**.

Example: Child Growth

```r
library("splines")
mod1 <- lm(weight ~ ns(age, df = 3) * sex + height + race + kcal_sd, data = child)
```

Example: Child Growth

```r
library("splines")
mod1 <- lm(weight ~ ns(age, df = 3) * sex + height + race + kcal_sd, data = child)
```

```r
summary(mod1)
```

```
## [...]
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8.8194     5.0408    1.75    0.081 .
## ns(age, df = 3)1          32.0575     4.1071    7.81  5.6e-14 ***
## ns(age, df = 3)2          50.0534     7.2659    6.89  2.3e-11 ***
## ns(age, df = 3)3          51.2129     4.2022   12.19  < 2e-16 ***
## sexgirl                   -2.3740     2.1136   -1.12    0.262
## height                     0.0862     0.0539    1.60    0.111
## raceasian                  0.5653     0.7585    0.75    0.457
## raceother                 -0.7135     0.7314   -0.98    0.330
## kcal_sd                    0.2140     0.3046    0.70    0.483
## ns(age, df = 3)1:sexgirl   3.2505     2.5435    1.28    0.202
## ns(age, df = 3)2:sexgirl  -6.3503     5.3730   -1.18    0.238
## ns(age, df = 3)3:sexgirl -15.5742     2.3841   -6.53  2.0e-10 ***
## [...]
```

Fitted Values & Confidence Intervals

We obtain the **fitted values** from a linear regression model by multiplying a **design matrix** $\mathbf{X}$ with the vector of **parameter estimates** for the regression coefficients $\hat{\boldsymbol{\beta}}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Fitted Values & Confidence Intervals

We obtain the **fitted values** from a linear regression model by multiplying a **design matrix** $\mathbf{X}$ with the vector of **parameter estimates** for the regression coefficients $\hat{\boldsymbol{\beta}}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

A $(1-\alpha)\times$ 100% **confidence interval** for $\hat{\mathbf{y}}$ can be obtained as

$$\hat{\mathbf{y}} \pm z_{1-\alpha/2}\mathrm{se}(\hat{\mathbf{y}}),$$

where

$$\mathrm{se}(\hat{\mathbf{y}}) = \sqrt{\mathrm{var}(\hat{\mathbf{y}})} = \sqrt{\mathrm{diag}\left(X\mathrm{var}(\hat{\boldsymbol{\beta}})X^{\top}\right)}.$$

Example: Effects of Race

Create data containing the (hypothetical) observations:

```
effectDF <- data.frame(
  race = levels(child$race),
  age = 14,
  height = 160,
  sex = "boy",
  kcal_sd = 0
)
```

```
##        race age height sex kcal_sd
## 1 caucasian  14    160 boy       0
## 2     asian  14    160 boy       0
## 3     other  14    160 boy       0
```

Example: Effects of Race

Create data containing the (hypothetical) observations:

```
effectDF <- data.frame(
  race = levels(child$race),
  age = 14,
  height = 160,
  sex = "boy",
  kcal_sd = 0
)
```

```
##        race age height sex kcal_sd
## 1 caucasian  14    160 boy       0
## 2     asian  14    160 boy       0
## 3     other  14    160 boy       0
```

Create the design matrix:

```
X <- model.matrix(~ ns(age, df = 3) * sex + height + race + kcal_sd, data = effectDF)
```

Problem: Knots are positioned based on the new data.
⇨ The coefficients have a different interpretation!

Example: Effects of Race

Create the **design matrix** via the `terms`-object:

```
Terms <- terms(mod1)
X <- model.matrix(delete.response(Terms),
                  data = effectDF,
                  xlev = mod1$xlevels)
```

Example: Effects of Race

Create the **design matrix** via the `terms`-object:

```
Terms <- terms(mod1)
X <- model.matrix(delete.response(Terms),
                  data = effectDF,
                  xlev = mod1$xlevels)
```

Extract the **parameter estimates**:

```
betas <- coef(mod1)
```

Extract the **variance-covariance matrix** of the regression coefficients:

```
V <- vcov(mod1)
```

Example: Effects of Race

Calculate the fitted values, standard errors, and confidence intervals:

```
fit <- X %*% betas
se <- sqrt(diag(X %*% V %*% t(X)))

lwr <- fit - qnorm(0.975) * se
upr <- fit + qnorm(0.975) * se
```

Example: Effects of Race

Calculate the fitted values, standard errors, and confidence intervals:

```r
fit <- X %*% betas
se <- sqrt(diag(X %*% V %*% t(X)))

lwr <- fit - qnorm(0.975) * se
upr <- fit + qnorm(0.975) * se
```

Combine the input data with the fitted values, etc.:

```r
effectDF <- data.frame(effectDF, fit = fit, se = se, lwr = lwr, upr = upr)
```

Example: Effects of Race

Calculate the fitted values, standard errors, and confidence intervals:

```
fit <- X %*% betas
se <- sqrt(diag(X %*% V %*% t(X)))

lwr <- fit - qnorm(0.975) * se
upr <- fit + qnorm(0.975) * se
```

Combine the input data with the fitted values, etc.:

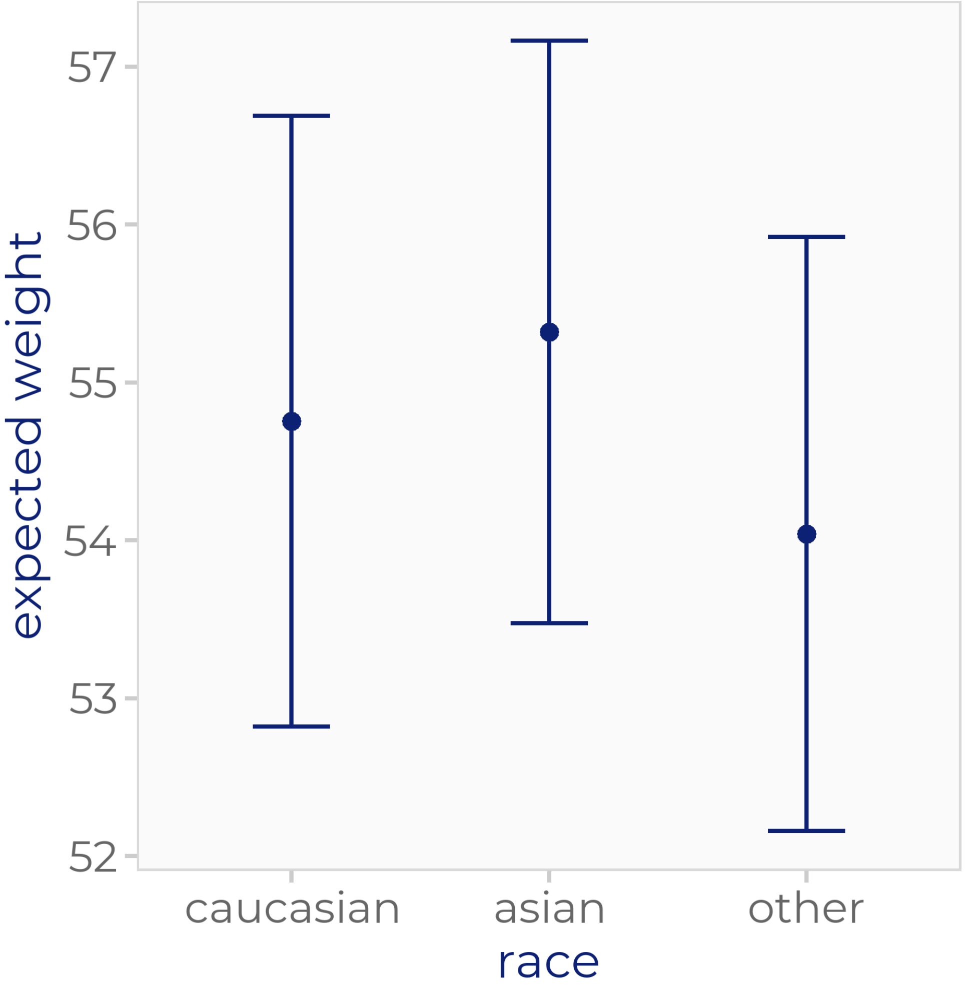
```
effectDF <- data.frame(effectDF, fit = fit, se = se, lwr = lwr, upr = upr)
```

Alternatively:

```
pred <- predict(mod1, newdata = effectDF, interval = "confidence")
effectDF <- cbind(effectDF, pred)
```

Effect Plot for Race

```
library("ggplot2")

ggplot(effectDF,
       aes(x = race, y = fit)) +
  geom_point() +
  geom_errorbar(aes(ymin = lwr, ymax = upr)) +
  ylab("expected weight")
```

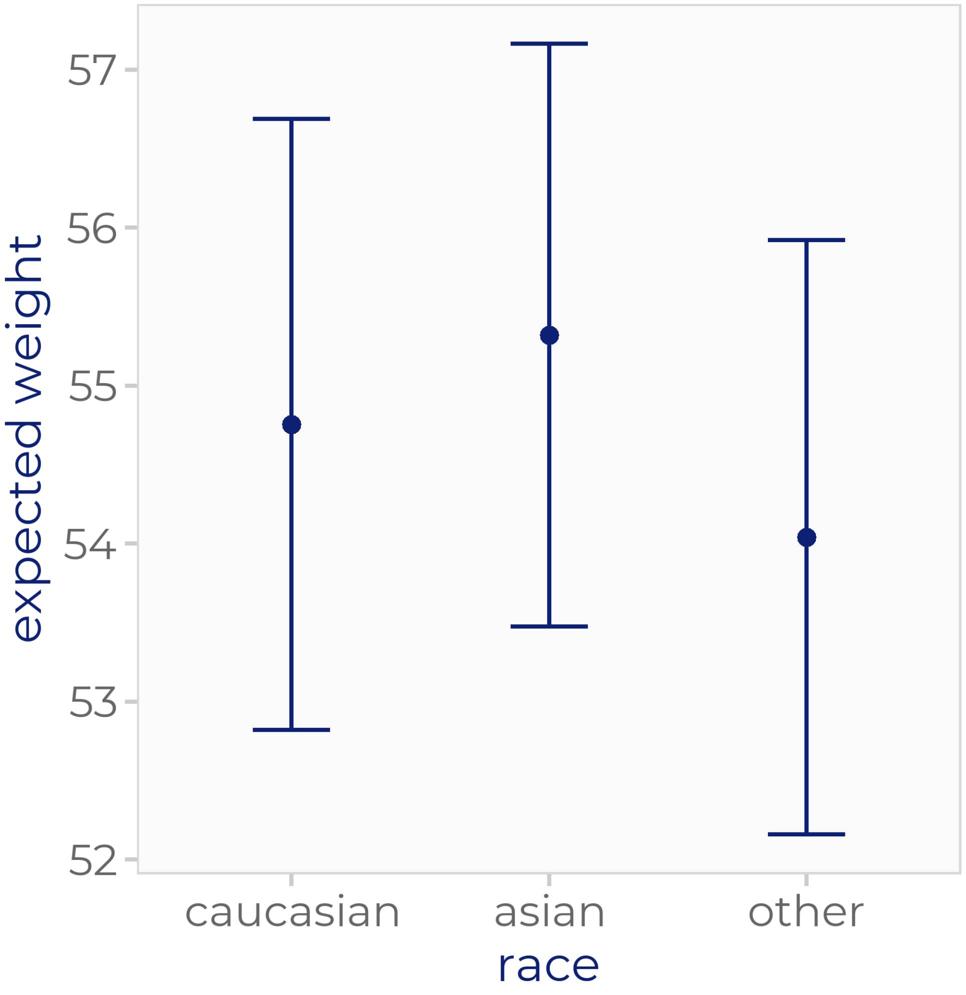


Effect Plot for Race

```
library("ggplot2")

ggplot(effectDF,
       aes(x = race, y = fit)) +
  geom_point() +
  geom_errorbar(aes(ymin = lwr, ymax = upr)) +
  ylab("expected weight")
```

Expected weight for

- boys, who are
- 14 years of age and
- 160 cm tall,
- with a standardized kcal intake of 0.



Example: Race and Sex

Specify the new (hypothetical) data:

```
effectDF2 <- expand.grid(
  race = levels(child$race),
  age = 14,
  height = 160,
  sex = levels(child$sex),
  kcal_sd = 0
)
```

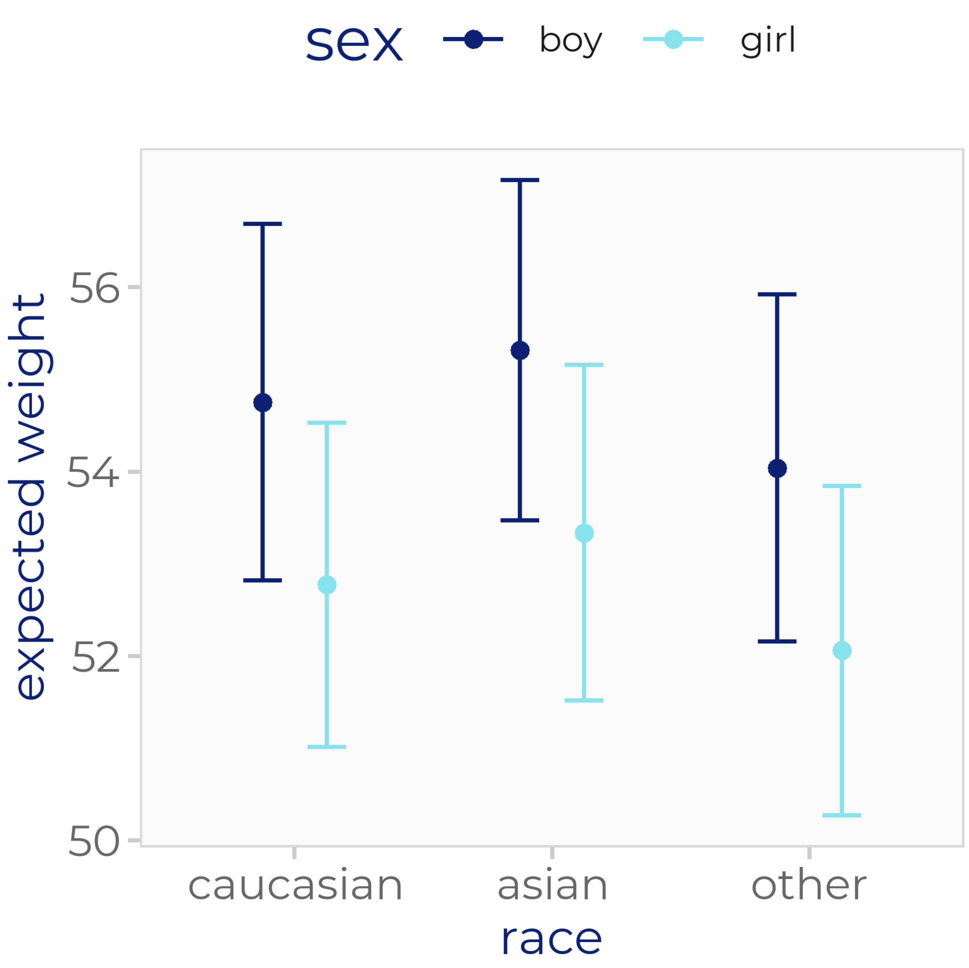
```
##        race age height  sex kcal_sd
## 1 caucasian  14    160  boy       0
## 2     asian  14    160  boy       0
## 3     other  14    160  boy       0
## 4 caucasian  14    160 girl       0
## 5     asian  14    160 girl       0
## 6     other  14    160 girl       0
```

Obtain fitted values and confidence intervals:

```
pred2 <- predict(mod1, newdata = effectDF2,
                 interval = "confidence")
effectDF2 <- cbind(effectDF2, pred2)
```

Example: Race and Sex

```
ggplot(effectDF2,
       aes(x = race, y = fit, color = sex)) +

  geom_point(
    position = position_dodge(width = 0.5)) +

  geom_errorbar(
    aes(ymin = lwr, ymax = upr),
    position = position_dodge(width = 0.5)) +

  ylab("expected weight") +

  theme(legend.position = "top")
```



Choosing Reference Values

Any variable not of interest is set to a "reference" value, for example,

- the median,
- the reference category, or
- the largest category.

The "reference" values should be representative / typical values.

Choosing Reference Values

Any variable not of interest is set to a "reference" value, for example,

- the median,
- the reference category, or
- the largest category.

The "reference" values should be representative / typical values.

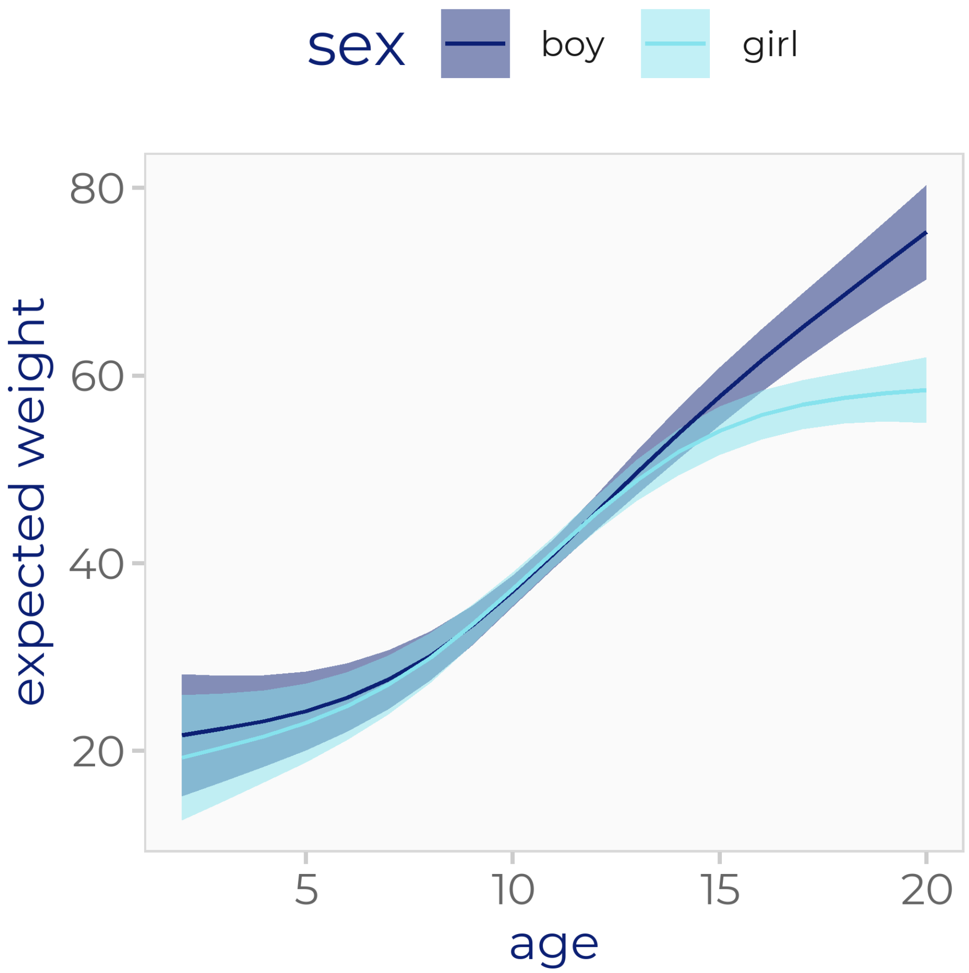
Example: Age and Sex

```
effectDF3 <- expand.grid(
  age = 2:20,
  sex = levels(child$sex),
  height = median(child$height),
  race = "asian",
  kcal_sd = 0
)
```

```
##     age  sex height  race kcal_sd
## 1     2  boy  143.1 asian       0
## 2     3  boy  143.1 asian       0
## 3     4  boy  143.1 asian       0
## ...
## 19   20  boy  143.1 asian       0
## 20    2 girl  143.1 asian       0
## 21    3 girl  143.1 asian       0
## ...
```

Effect Plot for Sex and Age

```
ggplot(effectDF3,
       aes(x = age, y = fit, color = sex)) +

  geom_ribbon(
    aes(ymin = lwr, ymax = upr, fill = sex),
    alpha = 0.5, color = NA) +

  geom_line() +

  ylab("expected weight") +

  theme(legend.position = "top")
```

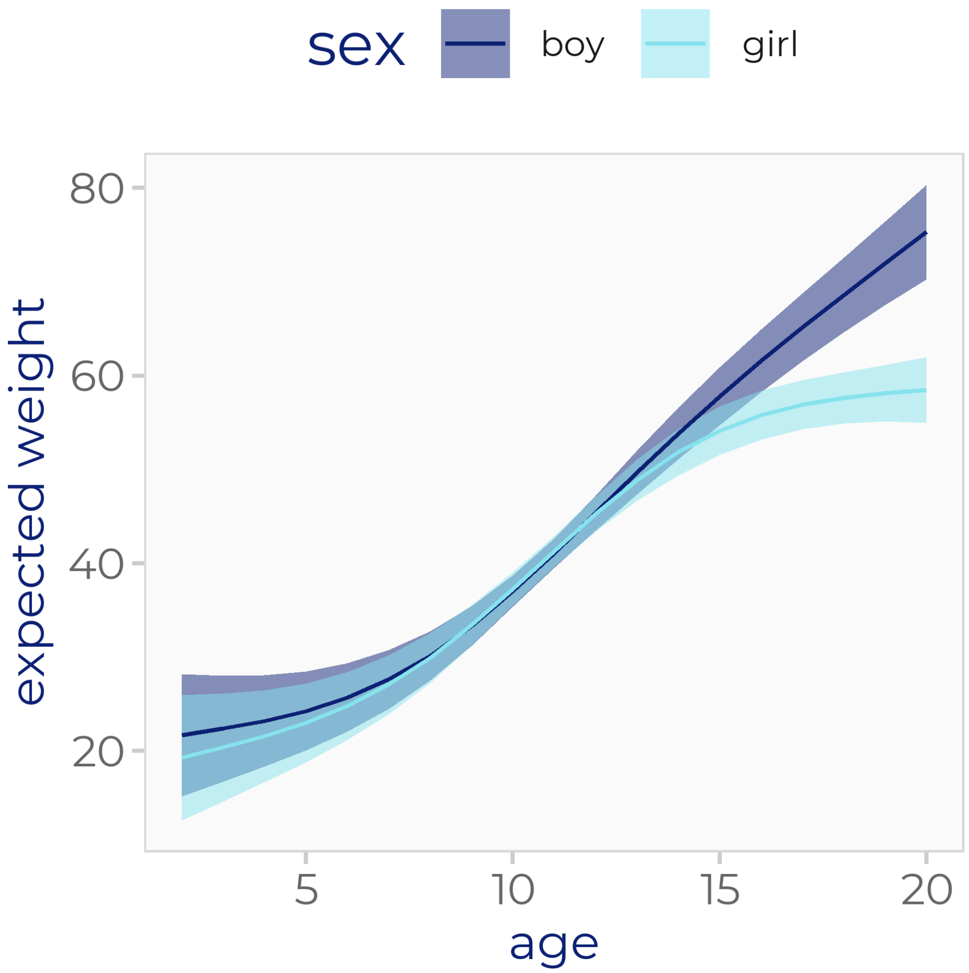


Effect Plot for Sex and Age

```
ggplot(effectDF3,
       aes(x = age, y = fit, color = sex)) +

  geom_ribbon(
    aes(ymin = lwr, ymax = upr, fill = sex),
    alpha = 0.5, color = NA) +

  geom_line() +

  ylab("expected weight") +

  theme(legend.position = "top")
```

Expected weight for boys and girls, all

- 143.1 cm tall,
- asian,
- with a standardized kcal intake of 0.



Effect Plot for Sex and Age

Note:
Due to the strong association between age and height, some combinations of covariate values are **not realistic**.

```
##    age  sex height  race kcal_sd  fit  lwr  upr
## 1    2  boy    143 asian       0 21.7 15.2 28.2
## 19  20  boy    143 asian       0 75.3 70.3 80.4
## 20   2 girl    143 asian       0 19.3 12.6 26.0
## 38  20 girl    143 asian       0 58.5 55.0 62.0
```

Effect Plot for Sex and Age

Note:
Due to the strong association between age and height, some combinations of covariate values are **not realistic**.

```
##     age  sex height  race kcal_sd  fit  lwr  upr
## 1     2  boy    143 asian       0 21.7 15.2 28.2
## 19   20  boy    143 asian       0 75.3 70.3 80.4
## 20    2 girl    143 asian       0 19.3 12.6 26.0
## 38   20 girl    143 asian       0 58.5 55.0 62.0
```

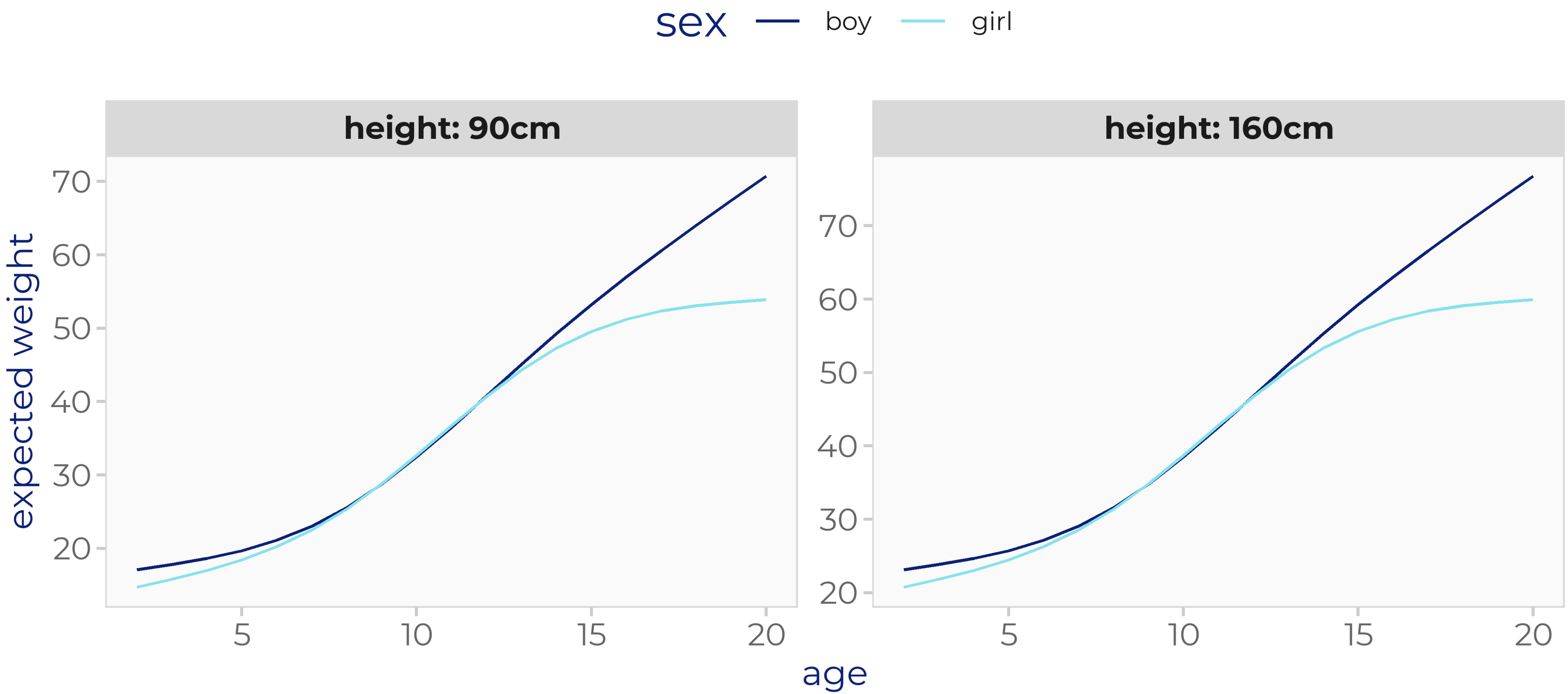
Because the model **does not include an interaction** of height different choices for height only cause a **shift on the y-axis**.

- The **shape** of the age-sex-weight relationship remains the same, irrespective of height.
- The **width of the confidence band** may change with height.

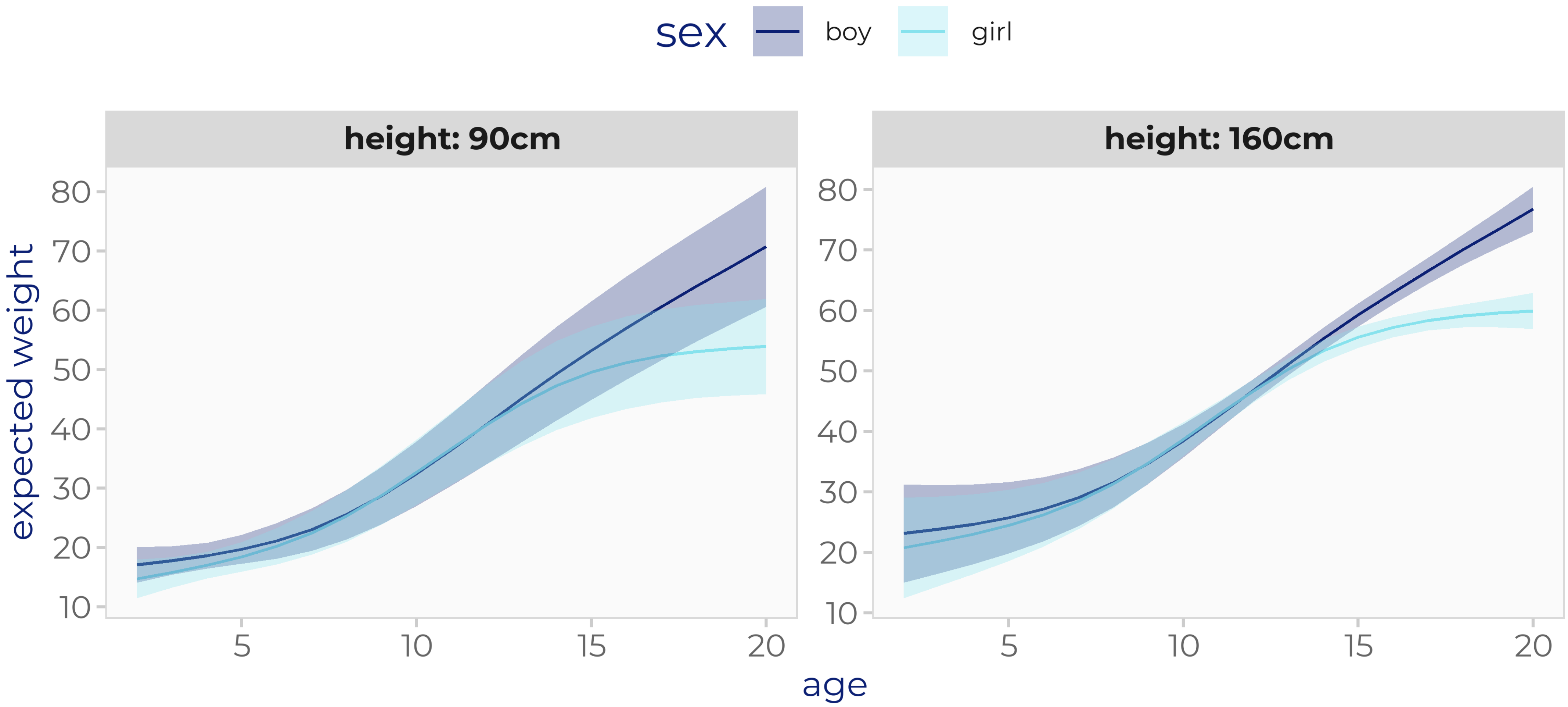
⇨ **Interpretation of the shape or differences** in expected weight between boys and girls is possible.

⇨ Interpretation of the **absolute values** of expected weight may not be meaningful.

Effect Plot for Sex and Age



Effect Plot for Sex and Age



Uncertainty is Important!

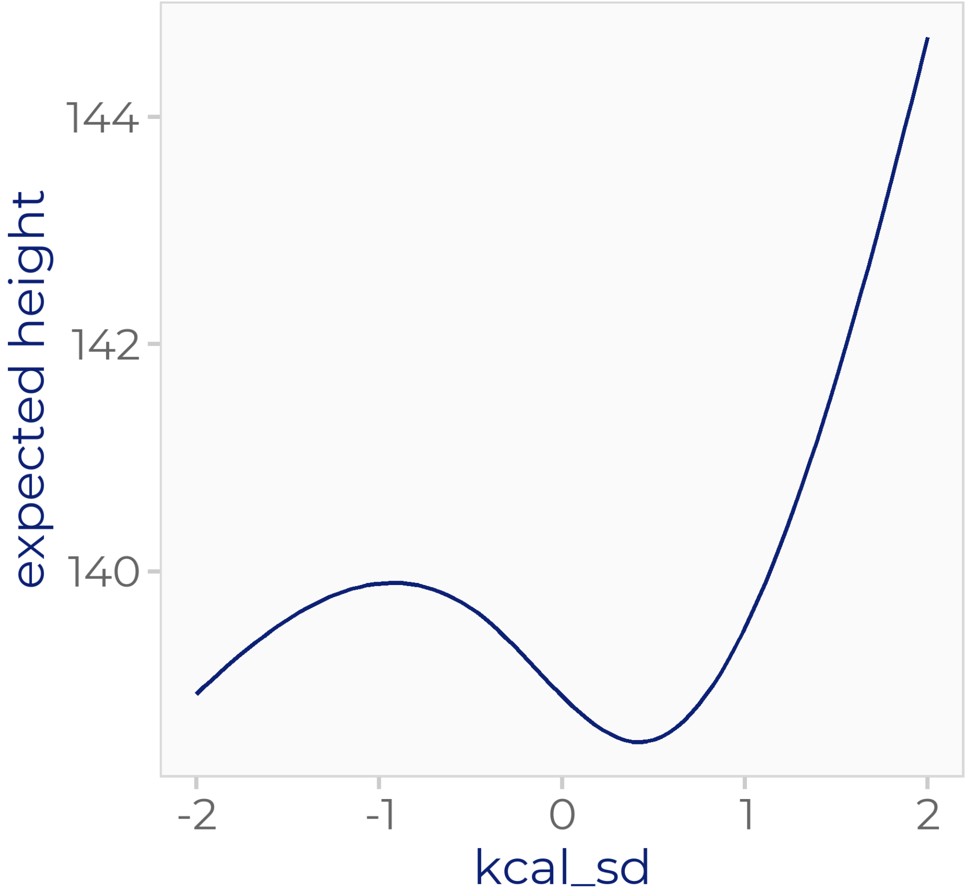
Why are confidence intervals/bands so important for interpretation of effect plots?

```r
mod_kcal <- lm(height ~ ns(kcal_sd, df = 3),
               data = child)

predDF <- data.frame(
  kcal_sd = seq(-2, 2, length = 100)
)

predDF <- cbind(
  predDF,
  predict(mod_kcal, newdata = predDF,
          interval = "confidence")
)

p_naive <- ggplot(predDF,
                  aes(x = kcal_sd, y = fit)) +
  geom_line() +
  ylab("expected height")

p_naive
```

Uncertainty is Important!

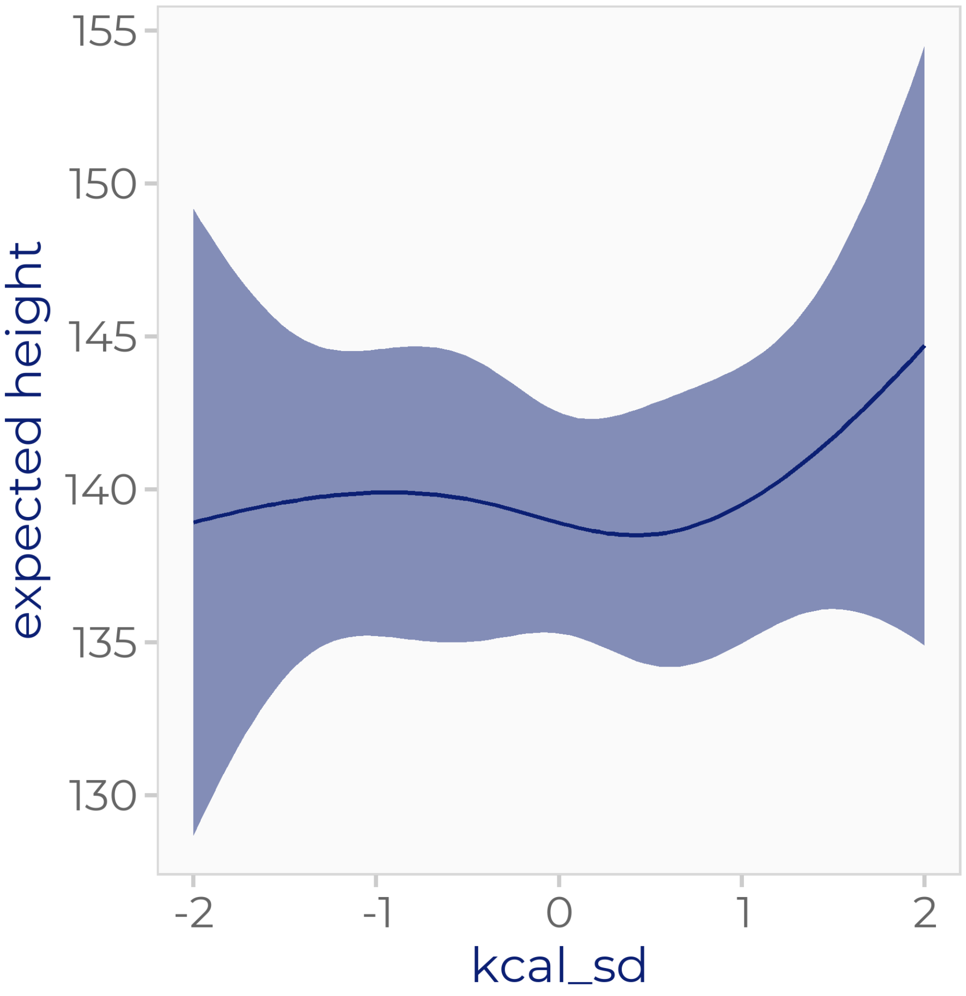
Why are confidence intervals/bands so important for interpretation of effect plots?

```
mod_kcal <- lm(height ~ ns(kcal_sd, df = 3),
               data = child)

predDF <- data.frame(
  kcal_sd = seq(-2, 2, length = 100)
)

predDF <- cbind(
  predDF,
  predict(mod_kcal, newdata = predDF,
          interval = "confidence")
)

p_naive <- ggplot(predDF,
                  aes(x = kcal_sd, y = fit)) +
  geom_line() +
  ylab("expected height")

p_naive
```



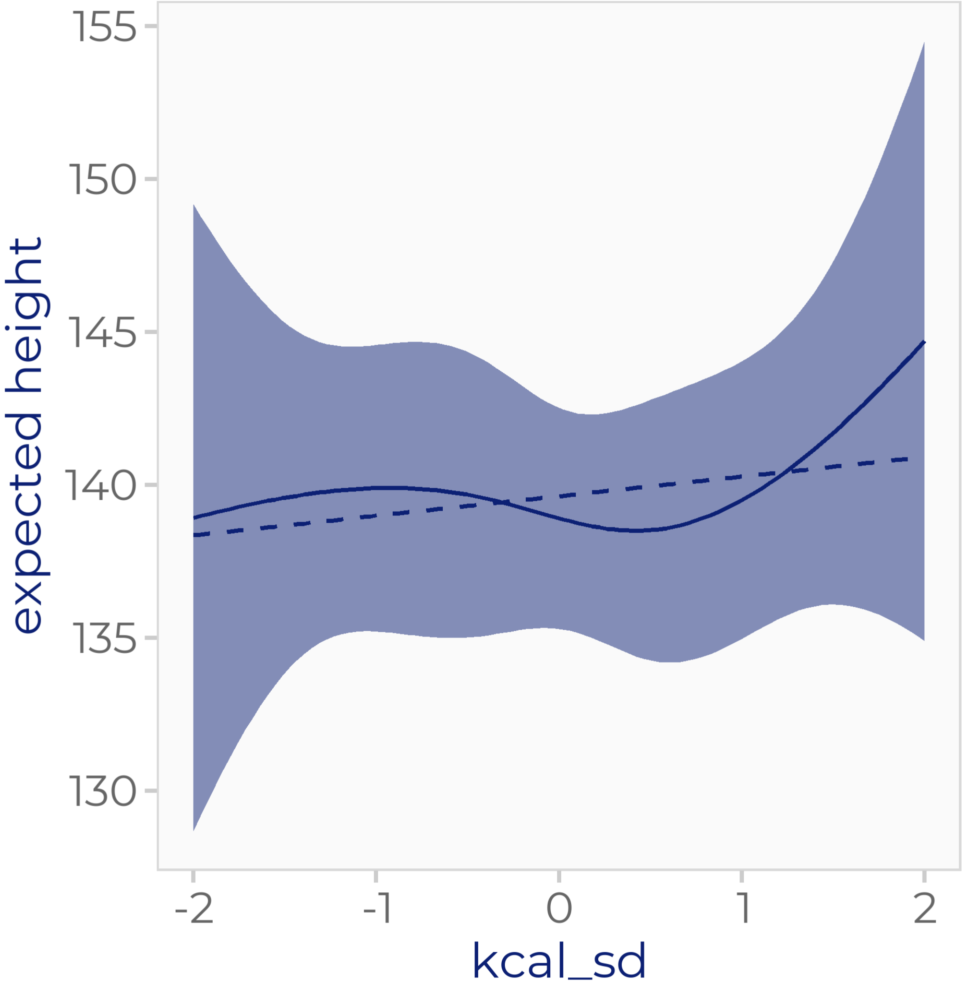
Uncertainty is Important!

```
p_naive +
  geom_ribbon(aes(ymin = lwr, ymax = upr),
              alpha = 0.5)
```



Uncertainty is Important!

```r
mod_kcal_lin <- lm(height ~ kcal_sd,
                   data = child)

predDF <- cbind(
  predDF,
  fit_lin = predict(mod_kcal_lin,
                    newdata = predDF)
)


p_naive %+% predDF +
  geom_ribbon(aes(ymin = lwr, ymax = upr),
              alpha = 0.5) +
  geom_line(aes(y = fit_lin), lty = 2)
```



Effect Plots

- Are often necessary to be able to interpret **complex associations**.
- Can help to evaluate the **clinical relevance** of effects.
- Should use **meaningful** (combinations of) **"reference" values** for the covariates.

Effect Plots

- Are often necessary to be able to interpret **complex associations**.
- Can help to evaluate the **clinical relevance** of effects.
- Should use **meaningful** (combinations of) **"reference" values** for the covariates.

Interpretation:

Expected response for a particular (hypothetical) observation that has a particular combination of covariate values.

- No categorization of continuous variables.
- No stratification.

Effect plots show a **visualization of the model**, not of the original data.

Effect Plots

- Are often necessary to be able to interpret **complex associations**.
- Can help to evaluate the **clinical relevance** of effects.
- Should use **meaningful** (combinations of) **"reference" values** for the covariates.

Interpretation:
Expected response for a particular (hypothetical) observation that has a particular combination of covariate values.

- No categorization of continuous variables.
- No stratification.

Effect plots show a **visualization of the model**, not of the original data.

There are **R packages** to facilitate creation of effect plots, for example, "visreg", "effects", "ggeffects", ...