Simple Linear Regression: Frequentist Inference

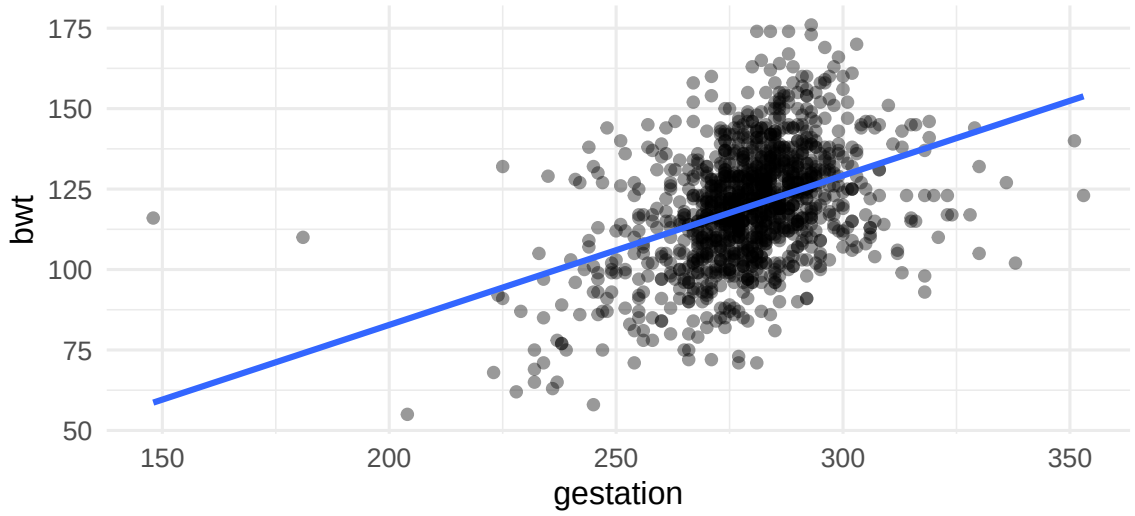
Lecture 11A handout

Name: _____

Baby weights

The `babies` data in the {openintro} package has 1,236 babies born at Kaiser Foundation hospitals in the East Bay (San Francisco area) between 1960 and 1967. Today we use `bwt` (birth weight, in ounces), `gestation` (length of the pregnancy, in days), and `smoke` (1 if the mother smoked during pregnancy, 0 if not).

```
ggplot(babies, aes(x = gestation, y = bwt)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE)
```



Response variable (y): _____ Explanatory variable (x): _____

Math class vs. statistics class. Real data do not fall exactly on a line, so the statistics equation has one more term. Name each term.

Table 1: Equation of a line in math class and in statistics class

Math class	Name	Statistics class	Name
$y = b + mx$		$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	
y	_____	y_i	_____
b	_____	β_0	_____
m	_____	β_1	_____
x	_____	x_i	_____
		ϵ_i	_____

Fitting the model

`lm()` (linear model) comes with R in the {stats} package, which loads automatically. `tidy()` is from the {broom} package: it shows the estimates as a table.

```
model_g <- lm(bwt ~ gestation, data = babies)
tidy(model_g)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -10.1      8.32       -1.21 2.27e- 1
2 gestation      0.464    0.0297     15.6  3.22e-50
```

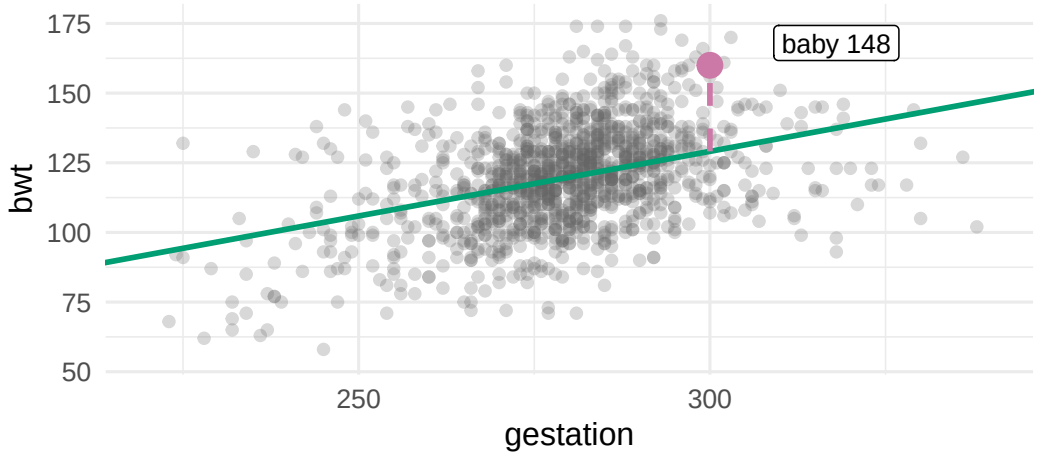
Fitted equation: $\widehat{\text{bwt}} =$ _____ + _____ $\times$ gestation

a. Expected birth weight of a baby born after 300 days of gestation: _____

b. Interpret the slope in context.

c. Interpret the intercept. Does it make sense? (We will come back to this at the end of class.)

d. Baby number 148 had a gestation of 300 days and weighed 160 ounces (the pink point below). Find the **residual**, $e_{148} = y_{148} - \hat{y}_{148}$. Is the baby above or below the line?



`lm()` finds the **least squares** line: the line with the smallest _____, $\sum e_i^2$.

Inference for the slope

If there were **no** relationship between gestation and birth weight, the population slope would be _____.

H_0: _____ H_A: _____

The `p.value` in the `gestation` row of `tidy(model_g)` is _____.

`confint()` is also from the {stats} package. It gives 95% confidence intervals.

```
confint(model_g)
```

```
                 2.5 %    97.5 %
(Intercept) -26.3915884 6.2632199
gestation     0.4059083 0.5226169
```

What do you conclude at $\alpha = 0.05$? Interpret the 95% confidence interval for the slope.

A categorical (binary) explanatory variable: smoking

```
model_s <- lm(bwt ~ smoke, data = babies)
tidy(model_s)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   123.       0.649    190.   0
2 smoke          -8.94     1.03      -8.65 1.55e-17
```

```
confint(model_s)
```

```
                2.5 %     97.5 %
(Intercept) 121.77391 124.320430
smoke       -10.96413  -6.911199
```

$\widehat{\text{bwt}} =$ _____ + _____ $\times$ smoke. `smoke` can only be 0 or 1:

- Expected birth weight, mother did not smoke (smoke = 0): _____
- Expected birth weight, mother smoked (smoke = 1): _____

So the intercept is _____ and the slope is _____.

Interpret the slope and its 95% confidence interval. Is there a difference?

Can we conclude that smoking **causes** lower birth weight? Can we generalize to babies born today?

Centering

The intercept of `model_g`, -10.1 ounces, is the expected birth weight at 0 days of gestation. That does not make sense. A common fix is to **center** the explanatory variable: subtract its mean from every value.

$$\text{gestation_c} = \text{gestation} - \overline{\text{gestation}}$$

```
mean(babies$gestation, na.rm = TRUE)
```

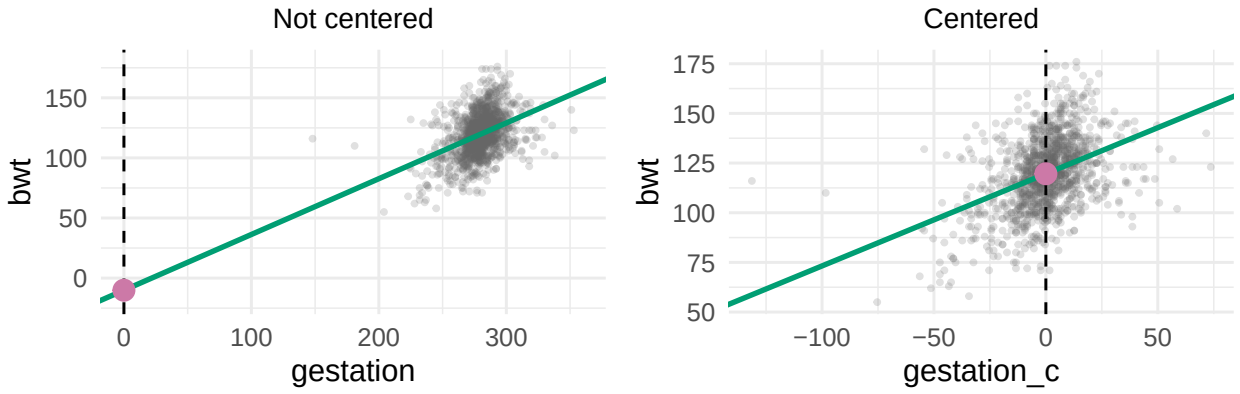
```
[1] 279.3385
```

```
babies <- babies |>
  mutate(gestation_c = gestation - mean(gestation, na.rm = TRUE))


model_c <- lm(bwt ~ gestation_c, data = babies)
tidy(model_c)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  120.       0.477      251.  0
2 gestation_c    0.464    0.0297      15.6 3.22e-50
```

a. What is `gestation_c` for a baby born after 279 days (about average)? After 300 days? After 260 days?



b. Interpret the new intercept, 120, in context. Does it make sense now?

c. Compare the slope of `model_c` with the slope of `model_g`. Why did it not change?

People often center so that the intercept describes a typical value of x. We will see this again on Wednesday!

Summary

- Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$; in R, _____ ({stats}) with `tidy()` ({broom}) and `confint()` ({stats}).
- Slope (numeric x): expected change in y for _____. Binary x: the slope is _____.
- Residual: _____. Do not _____.
- Centering x (subtracting its mean) changes the intercept to _____; the slope _____.
- $H_0 : \beta_1 = 0$. A small p-value, or a confidence interval that does not include _____, is evidence of a relationship, but not of _____.