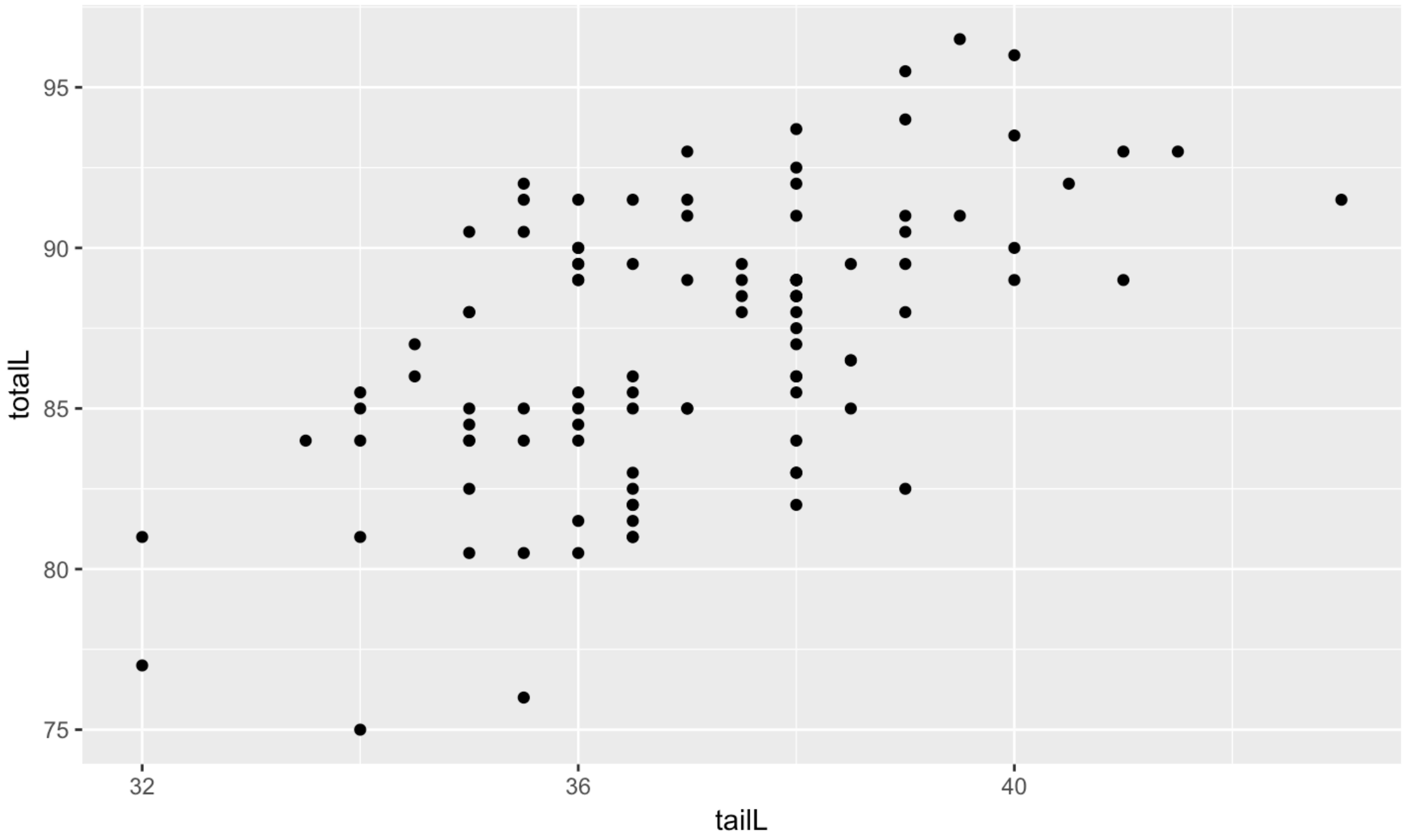
CORRELATION AND REGRESSION

Visualization of Linear Models

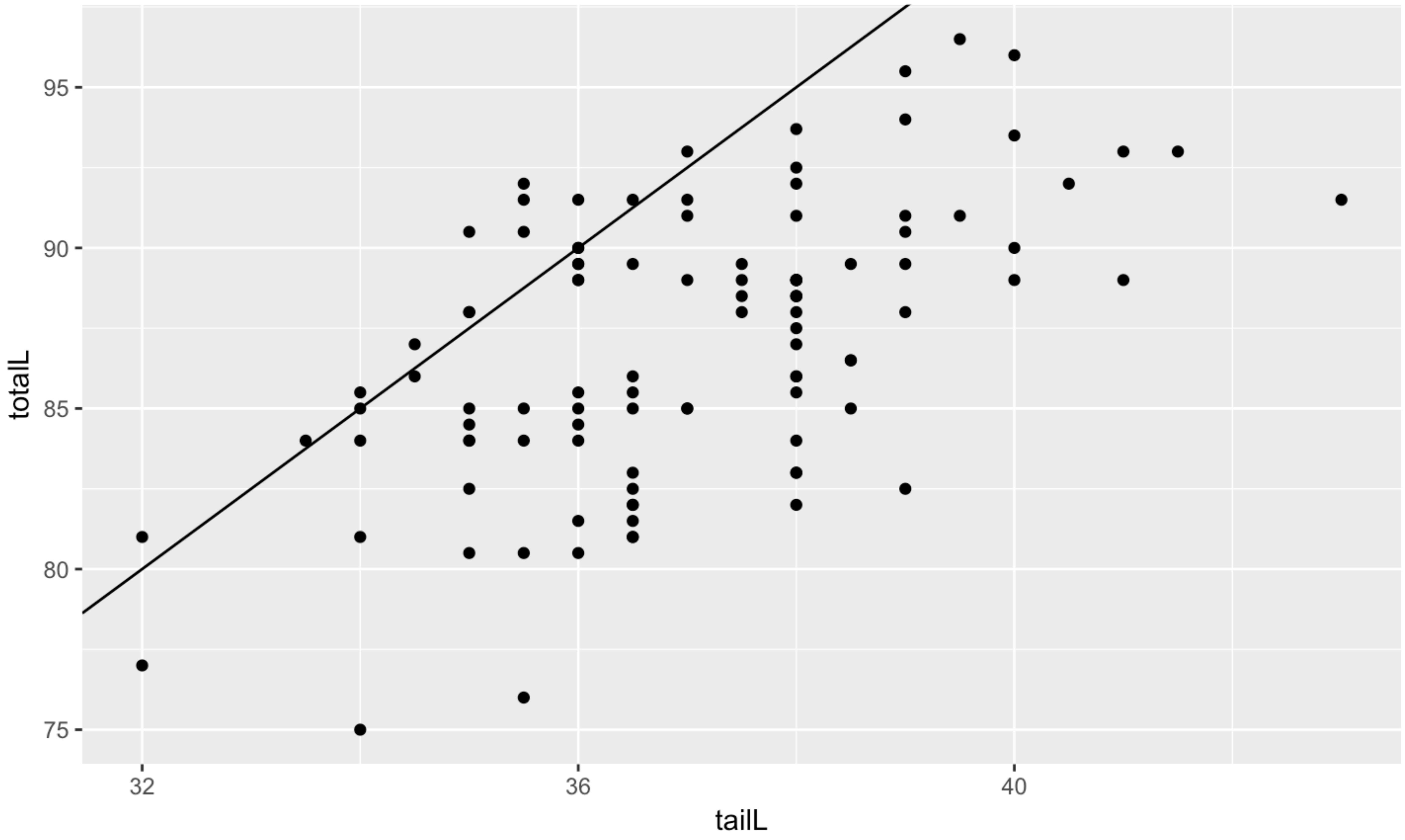
Possums

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point()
```



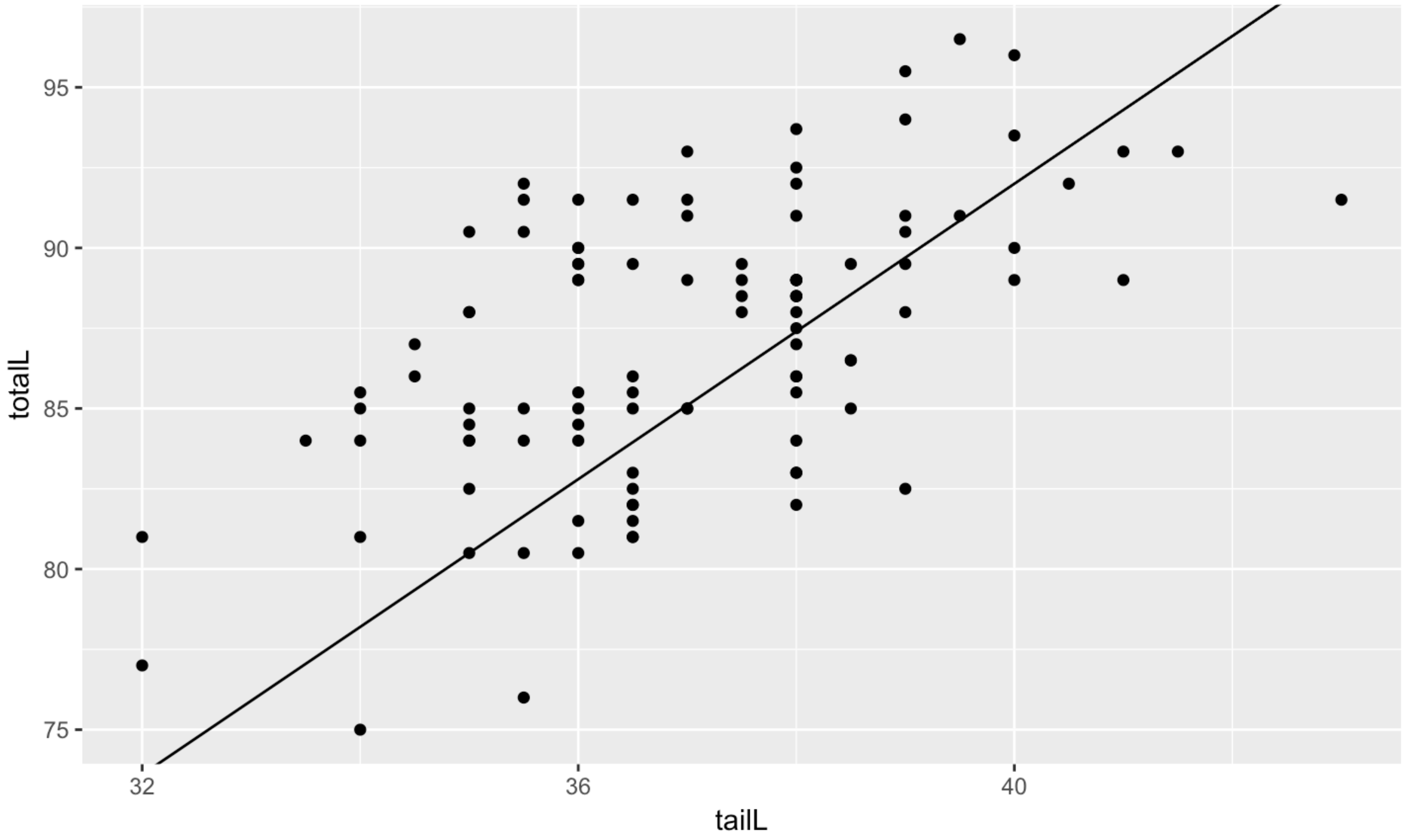
Through the origin

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_abline(intercept = 0, slope = 2.5)
```



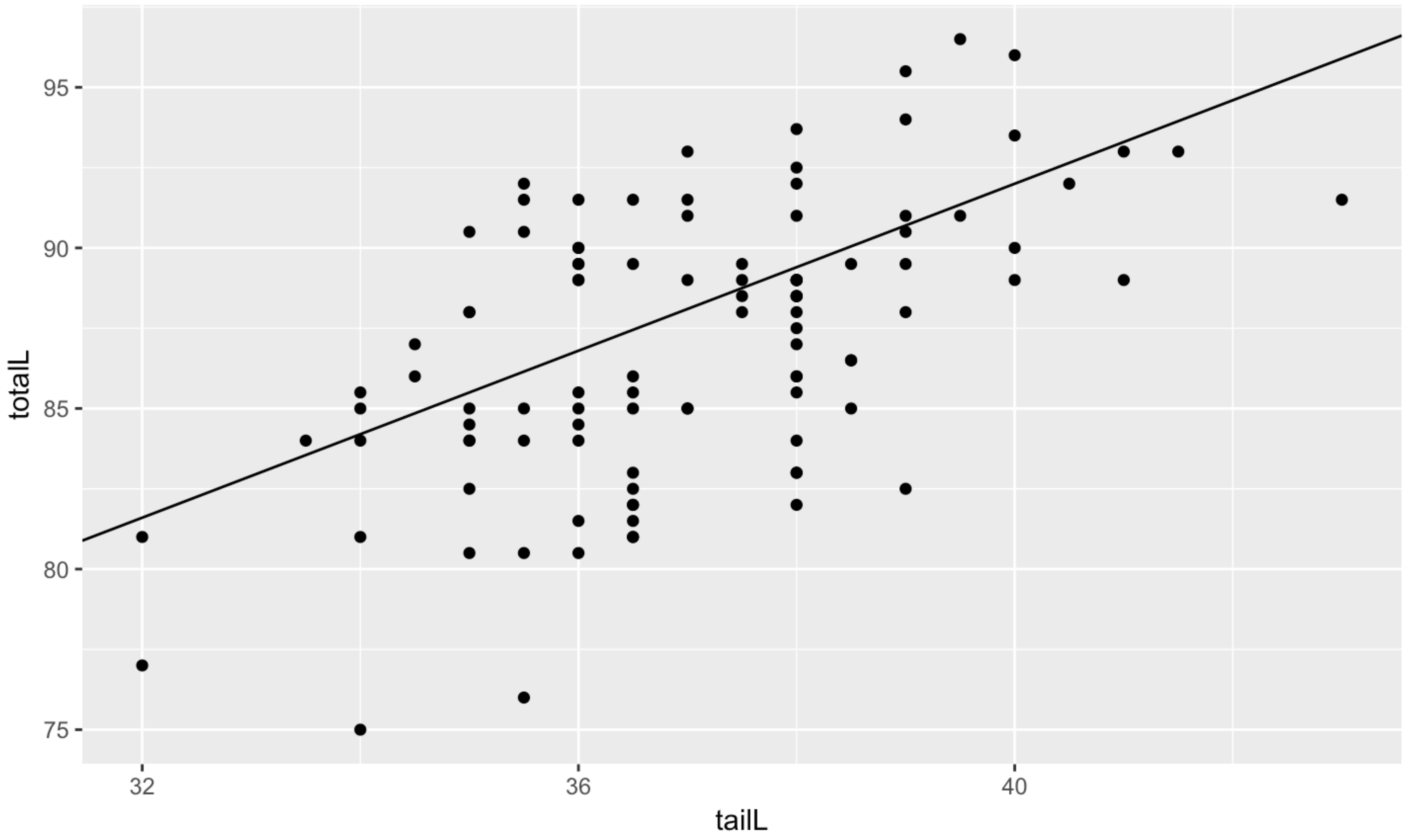
Through the origin, better fit

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_abline(intercept = 0, slope = 1.7)
```



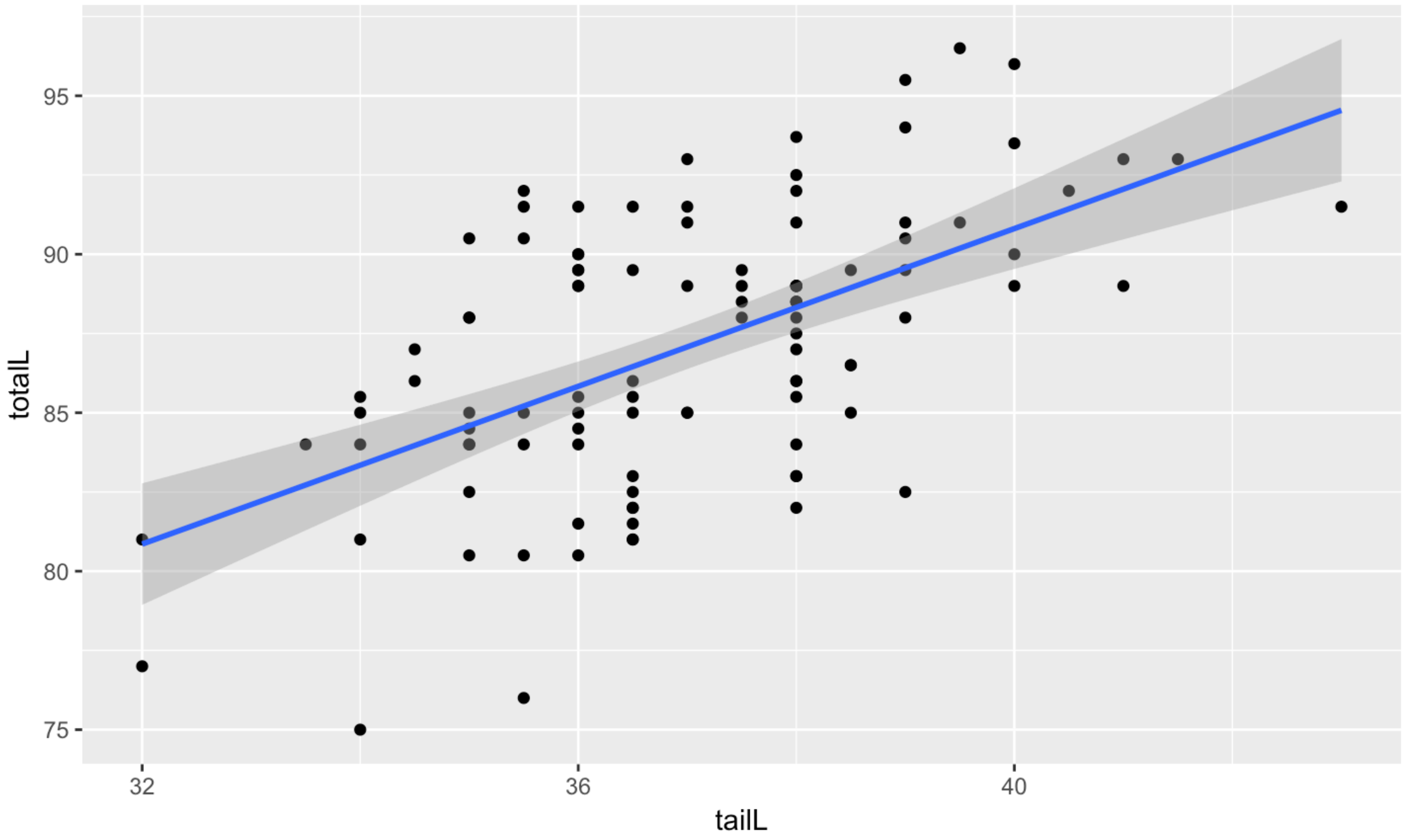
Not through the origin

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_abline(intercept = 40, slope = 1.3)
```



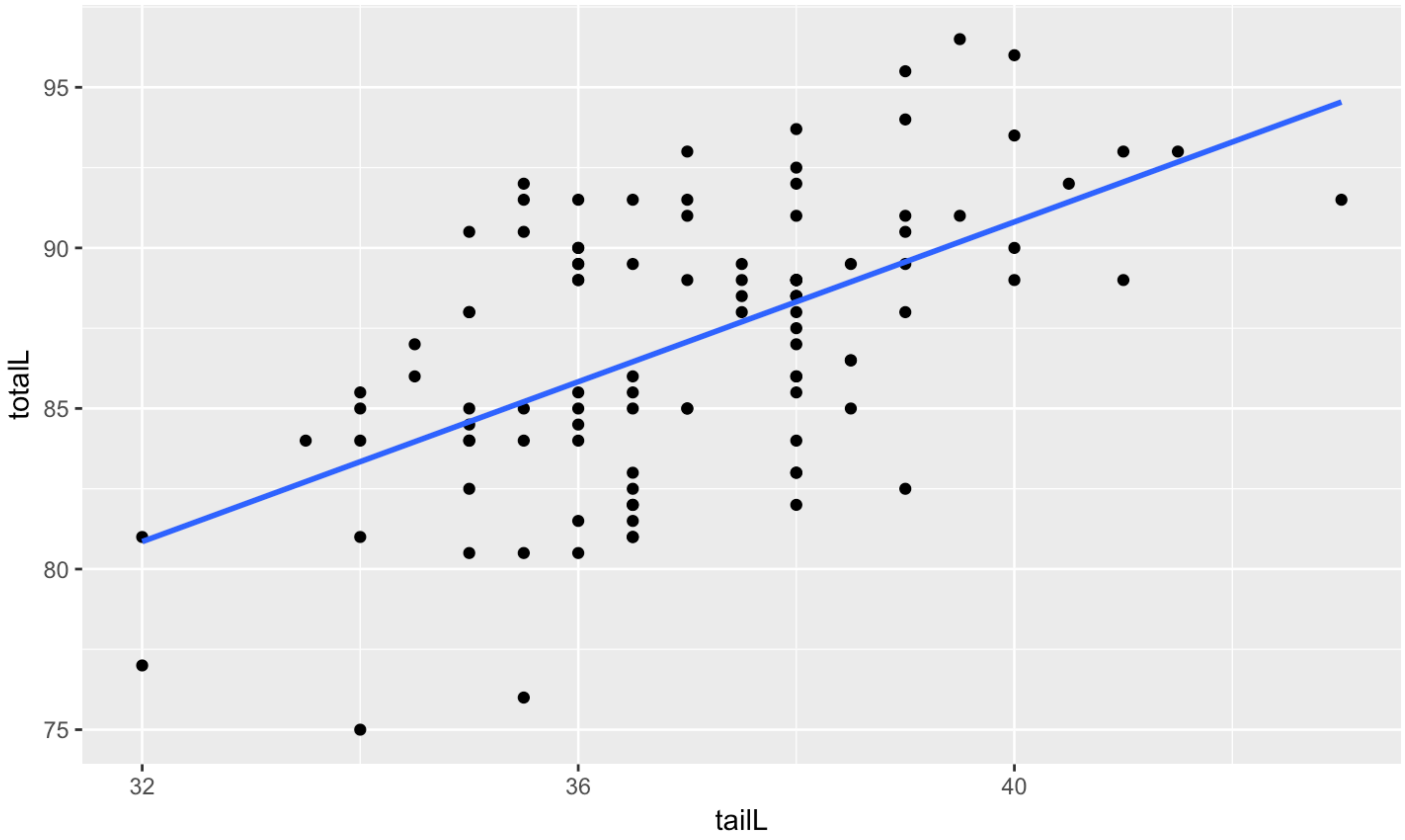
The "best" fit line

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_smooth(method = "lm")
```



Ignore standard errors

```
> ggplot(data = possum, aes(y = totalL, x = tailL)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```



Let's practice!

Understanding the linear model

Generic statistical model

$$response = f(explanatory) + noise$$

Generic linear model

*response = intercept + (slope * explanatory) + noise*

Regression model

$$Y = \beta_0 + \beta_1 \cdot X + \epsilon\,, \qquad \epsilon \sim N(0, \sigma_\epsilon)$$

Fitted values

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

Residuals

$$e = Y - \hat{Y}$$

Fitting procedure

- Given n observations of pairs (x_i, y_i)...
- Find $\hat{\beta}_0, \hat{\beta}_1$ that minimize $\sum_{i=1}^{n} e_i^2$

Least squares

- Easy, deterministic, unique solution
- Residuals sum to zero
- Line must pass through $(\bar{x}, \bar{y})$
- Other criteria exist—just not in this course

Key concepts

- Y-hat is expected value given corresponding X
- Beta-hats are estimates of true, unknown betas
- Residuals (e's) are estimates of true, unknown epsilons
- "Error" may be misleading term—better: noise

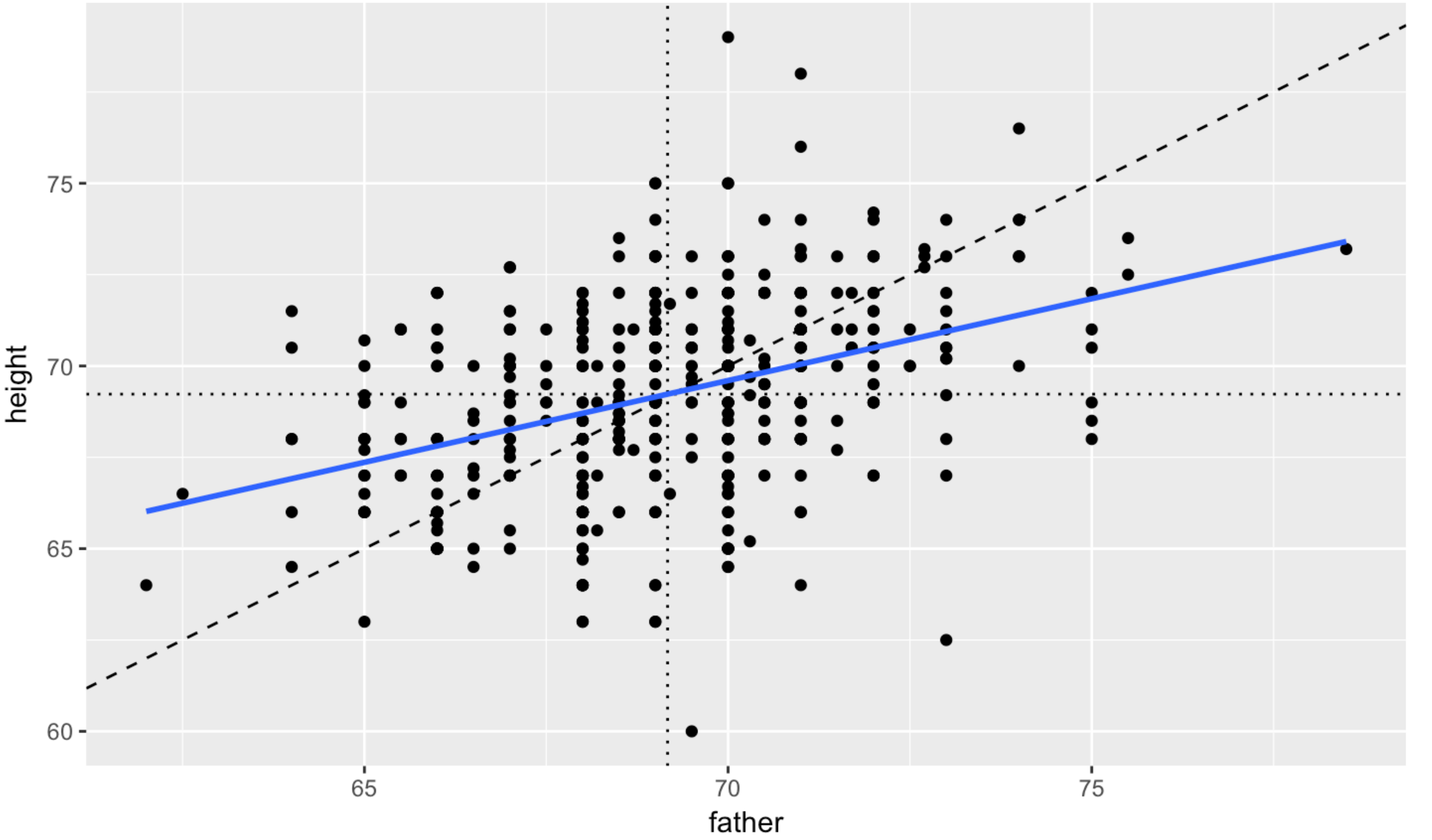
Let's practice!

Regression vs. regression to the mean

Heredity

- Galton's "regression to the mean"
- Thought experiment: consider the heights of the children of NBA players

Galton's data



Regression modeling

- "Regression": techniques for modeling a quantitative response
- Types of regression models:
  - Least squares
  - Weighted
  - Generalized
  - Nonparametric
  - Ridge
  - Bayesian
  - ...

Let's practice!