STA221

Neil Montgomery

Last edited: 2017-07-14 10:48

R^2

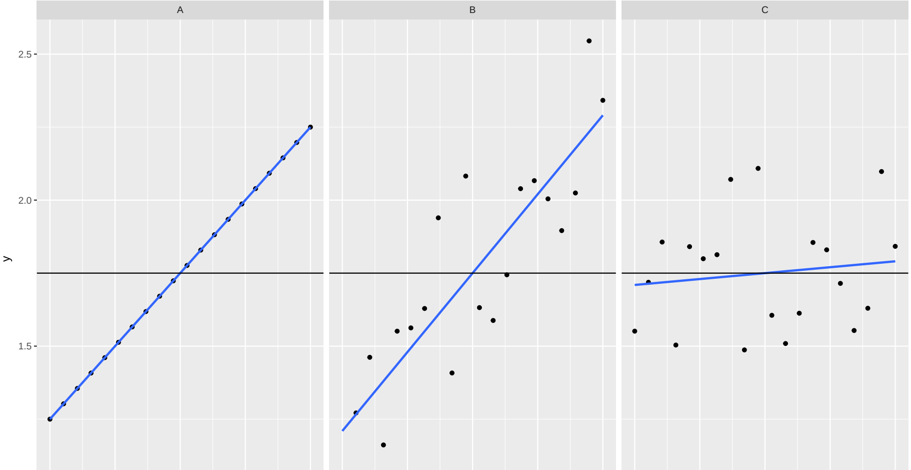
R^2

The y values are random. They aren't all the same. What "explains" the differences in the y values?

R^2

The y values are random. They aren't all the same. What "explains" the differences in the y values?

A = all "model" | B = "typical" | C = all "error":



R^2 conceptual building blocks; a "sum-of-squares" decomposition

$$\text{variation in the } y \;=\; \text{variation due to the model} \;+\; \text{variation due to error}$$

$$\sum (y_i - \overline{y})^2 = \qquad\qquad + $$

R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y $=$ variation due to the model $+$ variation due to error

$$\sum(y_i - \overline{y})^2 = \sum(\hat{y}_i - \overline{y})^2 +$$

R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y $=$ variation due to the model $+$ variation due to error

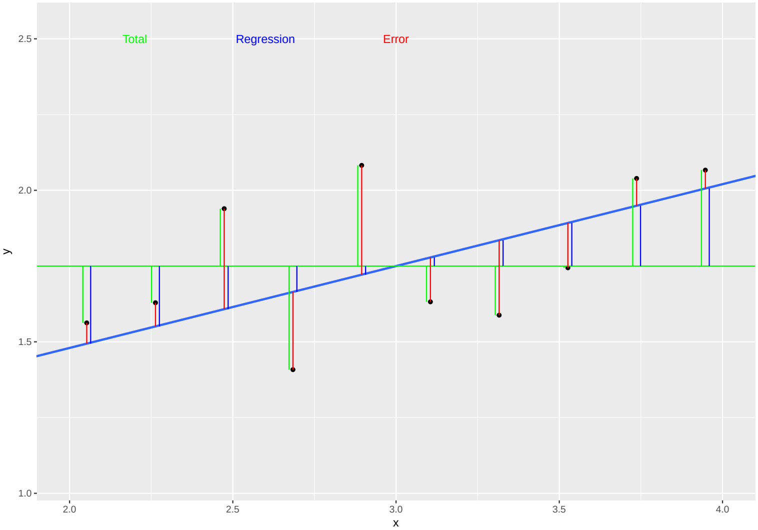
$$\sum(y_i - \overline{y})^2 = \sum(\hat{y}_i - \overline{y})^2 + \sum(y_i - \hat{y}_i)^2$$

R^2 conceptual building blocks; a "sum-of-squares" decomposition

variation in the y $=$ variation due to the model $+$ variation due to error

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum (y_i - \hat{y}_i)^2$$
$$SS_{Total} = SS_{Regression} + SS_{Error}$$

sum of squares decomposition, graphically



R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

"The proportion of variation explained by the (regression) model."

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

"The proportion of variation explained by the (regression) model."

People prone to excessive drama in their lives might call this "THE COEFFICIENT OF DETERMINATION!!!"

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

"The proportion of variation explained by the (regression) model."

People prone to excessive drama in their lives might call this "THE COEFFICIENT OF DETERMINATION!!!"

Although it is not a coefficient, and it does not really determine anything. It's just a mildly useful number.

R^2 definition

$$R^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

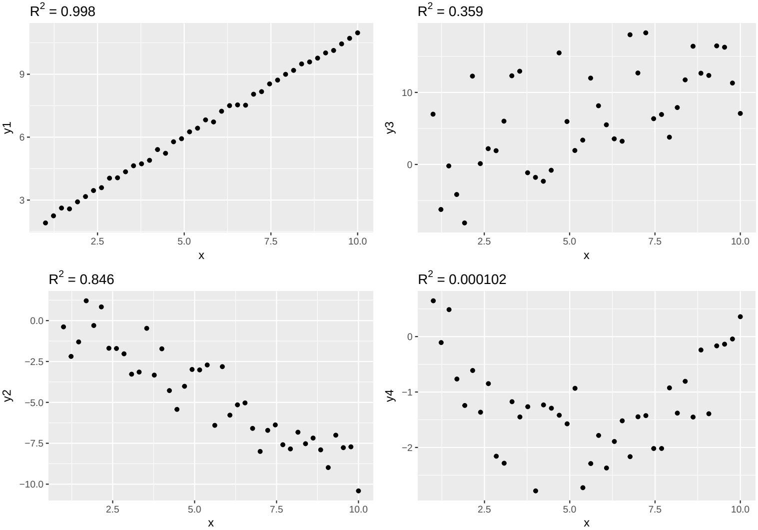
"The proportion of variation explained by the (regression) model."

People prone to excessive drama in their lives might call this "THE COEFFICIENT OF DETERMINATION!!!"

Although it is not a coefficient, and it does not really determine anything. It's just a mildly useful number.

Keep in mind it is *one number* that is being used to summarize an entire empirical bivariate relationship. And it isn't even the *best* number.

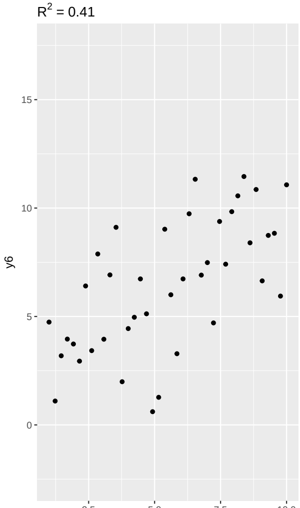
Some R^2 values

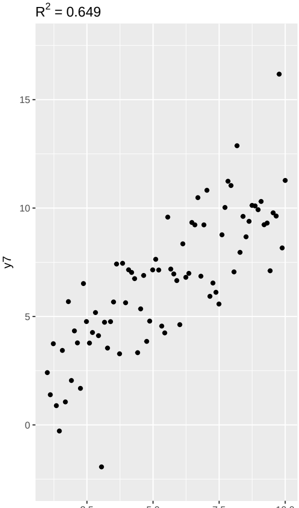


Another limitation: sample size effect

Both simulated datasets are from the ***same underlying model***

(happens to be $y = 1 + 1 \cdot x + \varepsilon$ with $\varepsilon \sim N(0, 2)$)





regression model assumption (etc.) verification

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

Also, the observations should be independent, but this is hard to verify (a plot of values versus time/order could be appropriate.)

recap model and calculation requirements

The model is:

$$y = \beta_0 + \beta_1 x + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma)$$

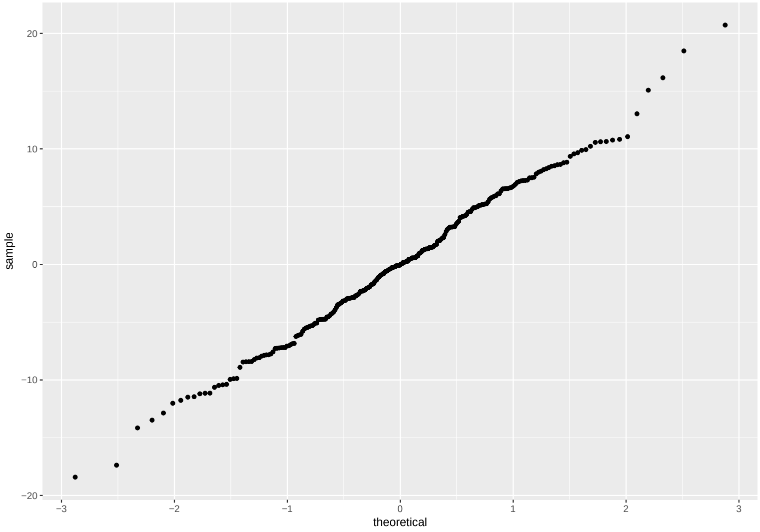
The requirements that should always be checked are:

- Linear relationship between y and x.
- Variation plus/minus the line is of constant magnitude.
- Error has a normal distribution.

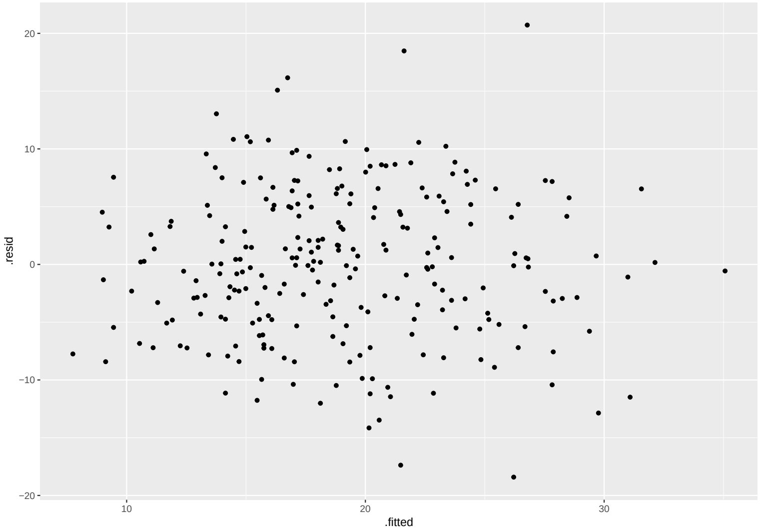
Also, the observations should be independent, but this is hard to verify (a plot of values versus time/order could be appropriate.)

We will verify graphically, using various plots of the *residuals* $\hat{\varepsilon}_i = y_i - \hat{y}_i$

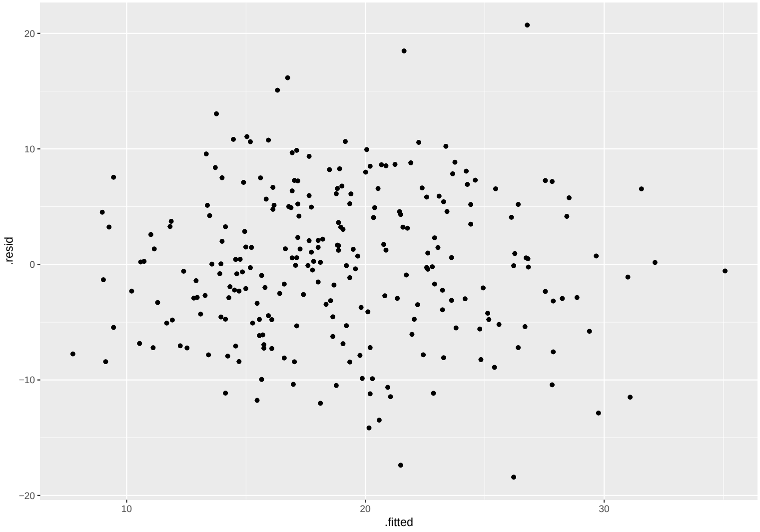
verify normality with normal quantile (or normal probability) plot of $\hat{\varepsilon}_i$



verify linearity with plot of $\hat{\varepsilon}_i$ versus $\hat{y}_i$



verify equal variance with (same!) plot of $\hat{\varepsilon}_i$ versus $\hat{y}_i$



estimation and prediction with regression models

estimate the mean response at a new x value

Suppose you want to estimate the mean “response” at some new x_ν (may or may not be one of the original x's.)

estimate the mean response at a new x value

Suppose you want to estimate the mean "response" at some new x_ν (may or may not be one of the original x's.)

The *true value* for the mean response is:

$$\mu_\nu = \beta_0 + \beta_1 x_\nu$$

estimate the mean response at a new x value

Suppose you want to estimate the mean "response" at some new x_ν (may or may not be one of the original x's.)

The *true value* for the mean response is:

$$\mu_\nu = \beta_0 + \beta_1 x_\nu$$

What's the "obvious" best guess using the data?

$$\hat{\mu}_\nu = b_0 + b_1 x_\nu$$

estimate the mean response—with confidence

A confidence interval will be as usual based on:

$$\frac{\hat{\mu}_\nu - \mu_\nu}{s.e.(\hat{\mu}_\nu - \mu_\nu)} \sim t_{n-2}$$

estimate the mean response—with confidence

A confidence interval will be as usual based on:

$$\frac{\hat{\mu}_\nu - \mu_\nu}{s.e.(\hat{\mu}_\nu - \mu_\nu)} \sim t_{n-2}$$

The standard error of $\hat{\mu}_\nu - \mu_\nu$ is:

$$s_e \sqrt{\frac{1}{n} + \frac{(x_\nu - \overline{x})^2}{S_{xx}}}$$

estimate the mean response—with confidence

A confidence interval will be as usual based on:

$$\frac{\hat{\mu}_\nu - \mu_\nu}{s.e.(\hat{\mu}_\nu - \mu_\nu)} \sim t_{n-2}$$

The standard error of $\hat{\mu}_\nu - \mu_\nu$ is:

$$s_e \sqrt{\frac{1}{n} + \frac{(x_\nu - \overline{x})^2}{S_{xx}}}$$

So the 95% C.I. for the mean response at x_ν will be:

$$\hat{\mu}_\nu \pm t^*_{n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_\nu - \overline{x})^2}{S_{xx}}}$$

weight model example

Let's make a 95% CI for the mean response at a weight of $x_\nu = 200$ pounds. Here's the R output:

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -14.69314    2.76045  -5.323 2.29e-07 ***
## Weight        0.18938    0.01533  12.357  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.538 on 248 degrees of freedom
## Multiple R-squared:  0.3811, Adjusted R-squared:  0.3786
## F-statistic: 152.7 on 1 and 248 DF,  p-value: < 2.2e-16
```

Weight at 200 example

Weight at 200 example

The estimate will be -14.693 $+$ 0.189(200) $=$ 23.1821234.

Weight at 200 example

The estimate will be -14.693 + 0.189(200) = 23.1821234.

$s_e = 6.538$

Weight at 200 example

The estimate will be -14.693 $+$ 0.189(200) $=$ 23.1821234.

$s_e = 6.538$

We would need to be given $\overline{x}$, which in this case is the sample average of the `Weight` variable. This is: 178.0832.

Weight at 200 example

The estimate will be -14.693 + 0.189(200) = 23.1821234.

$s_e = 6.538$

We would need to be given $\overline{x}$, which in this case is the sample average of the `Weight` variable. This is: 178.0832.

We also need S_{xx}, which is 534240.6. Could we have determined that from the output given?

Weight at 200 example

The estimate will be -14.693 + 0.189(200) = 23.1821234.

$s_e = 6.538$

We would need to be given $\overline{x}$, which in this case is the sample average of the `Weight` variable. This is: 178.0832.

We also need S_{xx}, which is 534240.6. Could we have determined that from the output given?

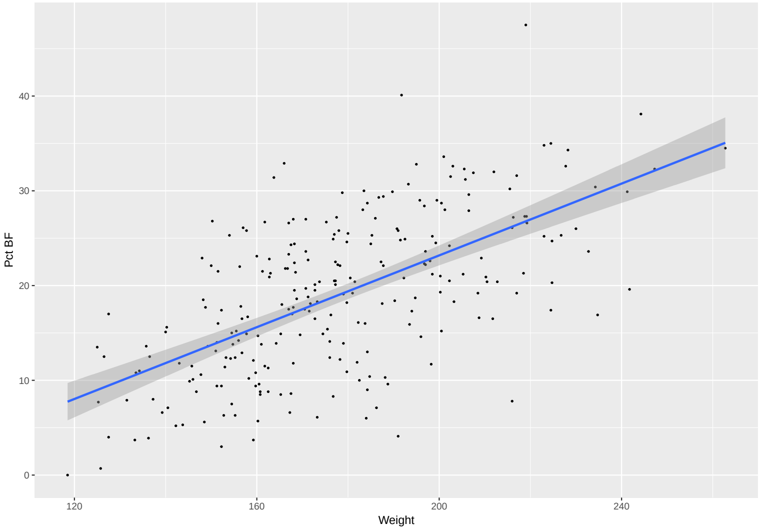
So the 95% CI for the mean `Pct BF` at `Weight`=200 is:

$$23.1821234 \pm 1.9695757 \cdot 6.538267\sqrt{\frac{1}{250} + \frac{(200 - 178.0832)^2}{1.8199849 \times 10^5}}$$

or:

$$(22.1328322, 24.2314145)$$

picture of 95% CI for mean response - weight model



predict y at a new x value

Suppose you want to predict what y might be at some new x_ν (may or may not be one of the original x's.)

predict y at a new x value

Suppose you want to predict what y might be at some new x_ν (may or may not be one of the original x's.)

There is no *true* value. We are predicting something random (and un-knowable)—not estimating something fixed (but unknown.)

predict y at a new x value

Suppose you want to predict what y might be at some new x_ν (may or may not be one of the original x's.)

There is no *true* value. We are predicting something random (and un-knowable)—not estimating something fixed (but unknown.)

What's the "obvious" best guess using the data?

$$\hat{y}_\nu = b_0 + b_1 x_\nu$$

The *same* guess as the estimate for μ_ν.

predict y at a new x value

Suppose you want to predict what y might be at some new x_ν (may or may not be one of the original x's.)

There is no *true* value. We are predicting something random (and un-knowable)—not estimating something fixed (but unknown.)

What's the "obvious" best guess using the data?

$$\hat{y}_\nu = b_0 + b_1 x_\nu$$

The *same* guess as the estimate for μ_ν.

The variation inherent in such a prediction is different.

predict a new value—with confidence

A prediction interval will be based on, similar to a confidence interval:

$$\frac{\hat{y}_\nu - y_\nu}{s.e.(\hat{y}_\nu - y_\nu)} \sim t_{n-2}$$

predict a new value—with confidence

A prediction interval will be based on, similar to a confidence interval:

$$\frac{\hat{y}_\nu - y_\nu}{s.e.(\hat{y}_\nu - y_\nu)} \sim t_{n-2}$$

The standard error of $\hat{y}_\nu - y_\nu$ is:

$$s_e\sqrt{1 + \frac{1}{n} + \frac{(x_\nu - \overline{x})^2}{S_{xx}}}$$

predict at `Weight` of 200

predict at `Weight` of 200

The prediction will be (also) -14.693 $+$ 0.189(200) $=$ 23.1821234.

predict at `Weight` of 200

The prediction will be (also) -14.693 + 0.189(200) = 23.1821234.

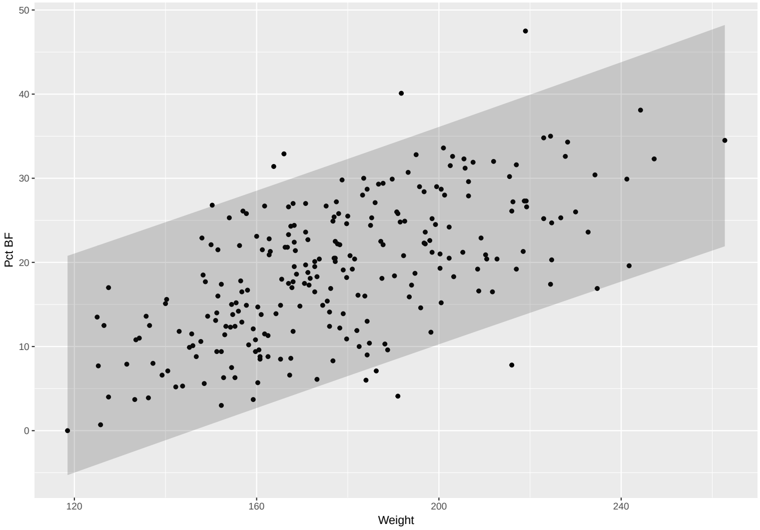
But the 95% "prediction interval" will be quite a bit wider:

$$23.1821234 \pm 1.9695757 \cdot 6.538267\sqrt{1 + \frac{1}{250} + \frac{(200 - 178.0832)^2}{1.8199849 \times 10^5}}$$

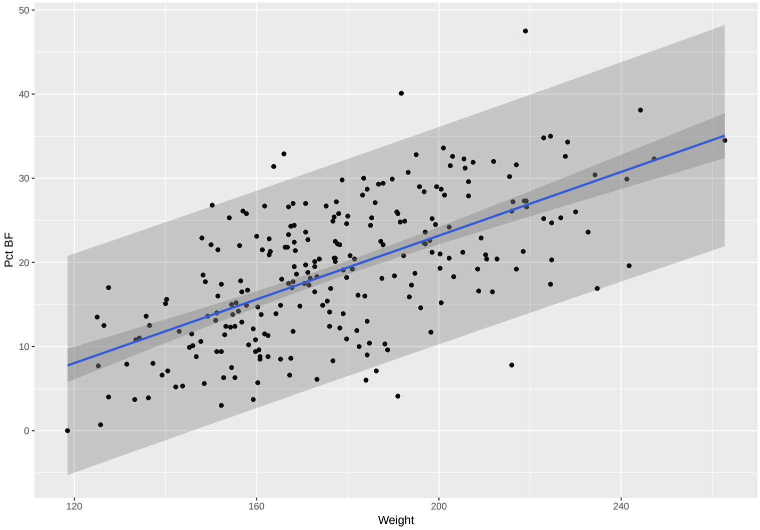
or:

$$(10.2618334, 36.1024133)$$

picture of 95% PI weight model



picture of both intervals



hmmm

```
## # A tibble: 250 × 15
##     `Pct BF`   Age Weight Height  Neck Chest Abdomen waist   Hip Thigh
##        <dbl> <int>  <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl>
## 1        0.0    40    118     68    34    79      69    27    85    47
## 2        0.7    35    126     66    34    91      75    30    89    50
## 3        3.0    35    152     68    37    92      82    32    93    55
## 4        3.7    27    159     72    36    90      80    31    96    55
## 5        3.7    27    133     65    36    94      74    29    88    50
## 6        3.9    42    136     68    38    88      78    31    89    52
## 7        4.0    47    128     67    34    83      70    28    87    51
## 8        4.1    25    191     74    38   101      82    32   100    63
## 9        5.2    55    142     67    35    93      83    33    92    54
## 10       5.3    25    144     72    35    92      76    30    92    52
## # ... with 240 more rows, and 5 more variables: Knee <dbl>,
## #   Ankle <dbl>, Bicep <dbl>, Forearm <dbl>, Wrist <dbl>
```