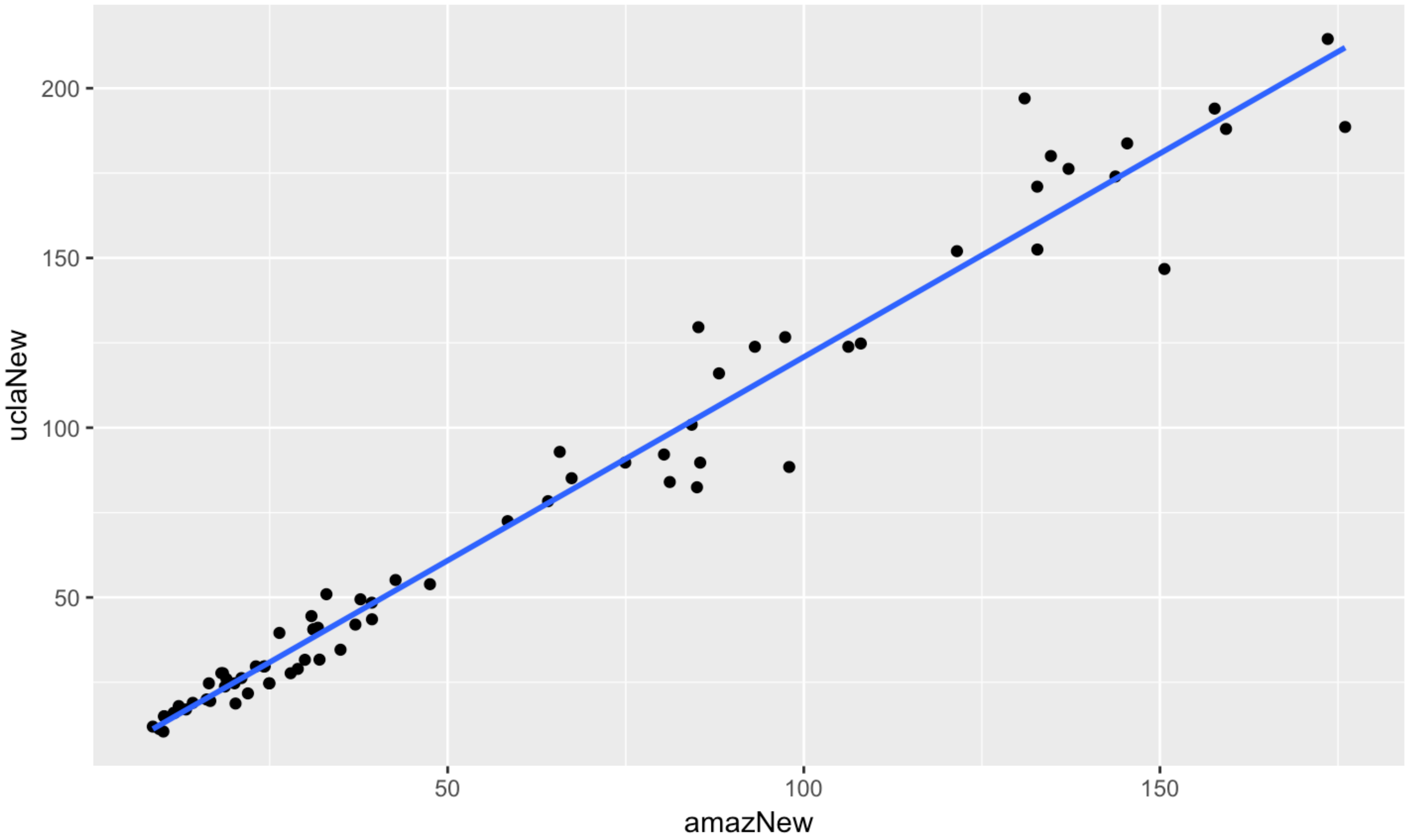
CORRELATION AND REGRESSION

Assessing model fit

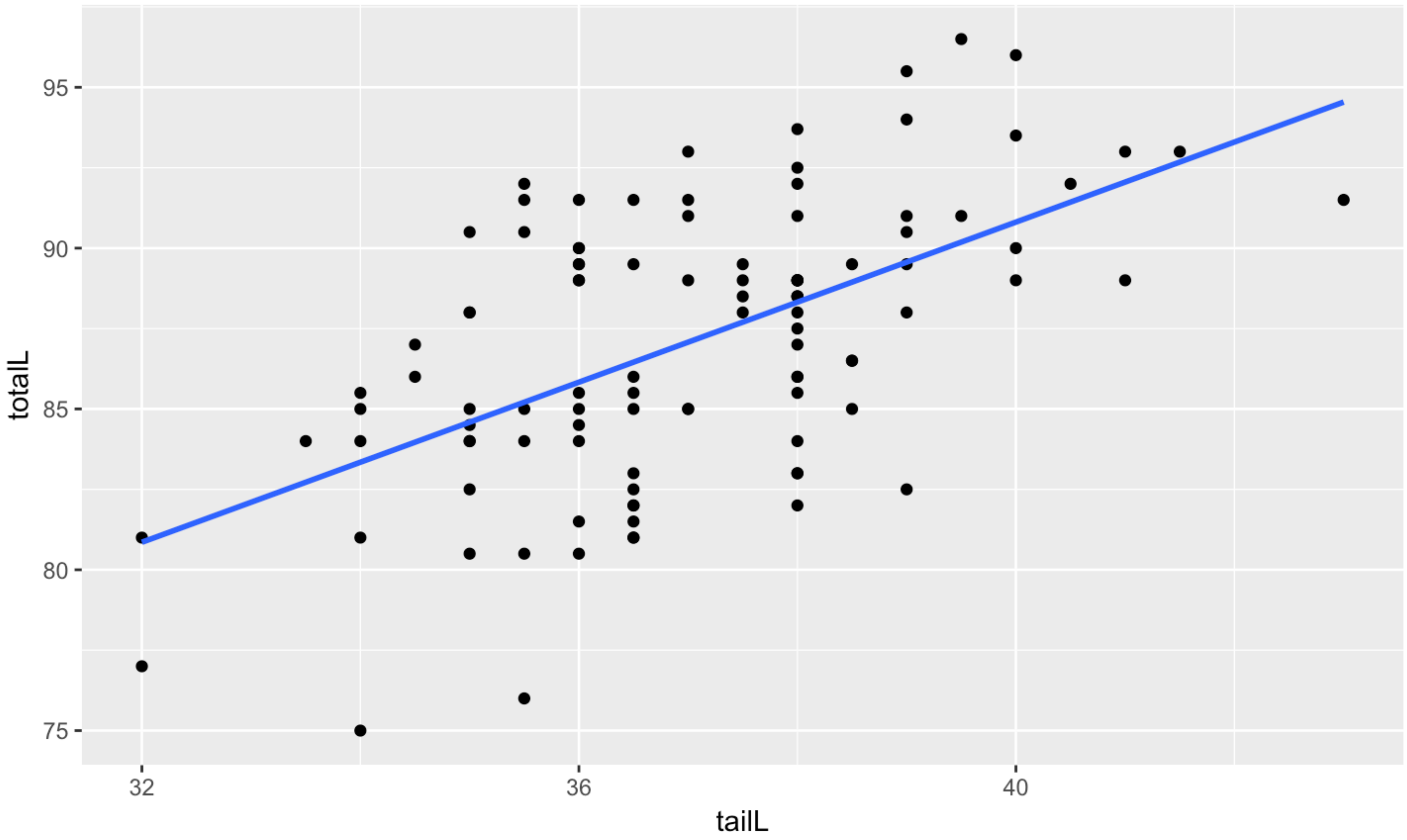
How well does our textbook model fit?

```
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```

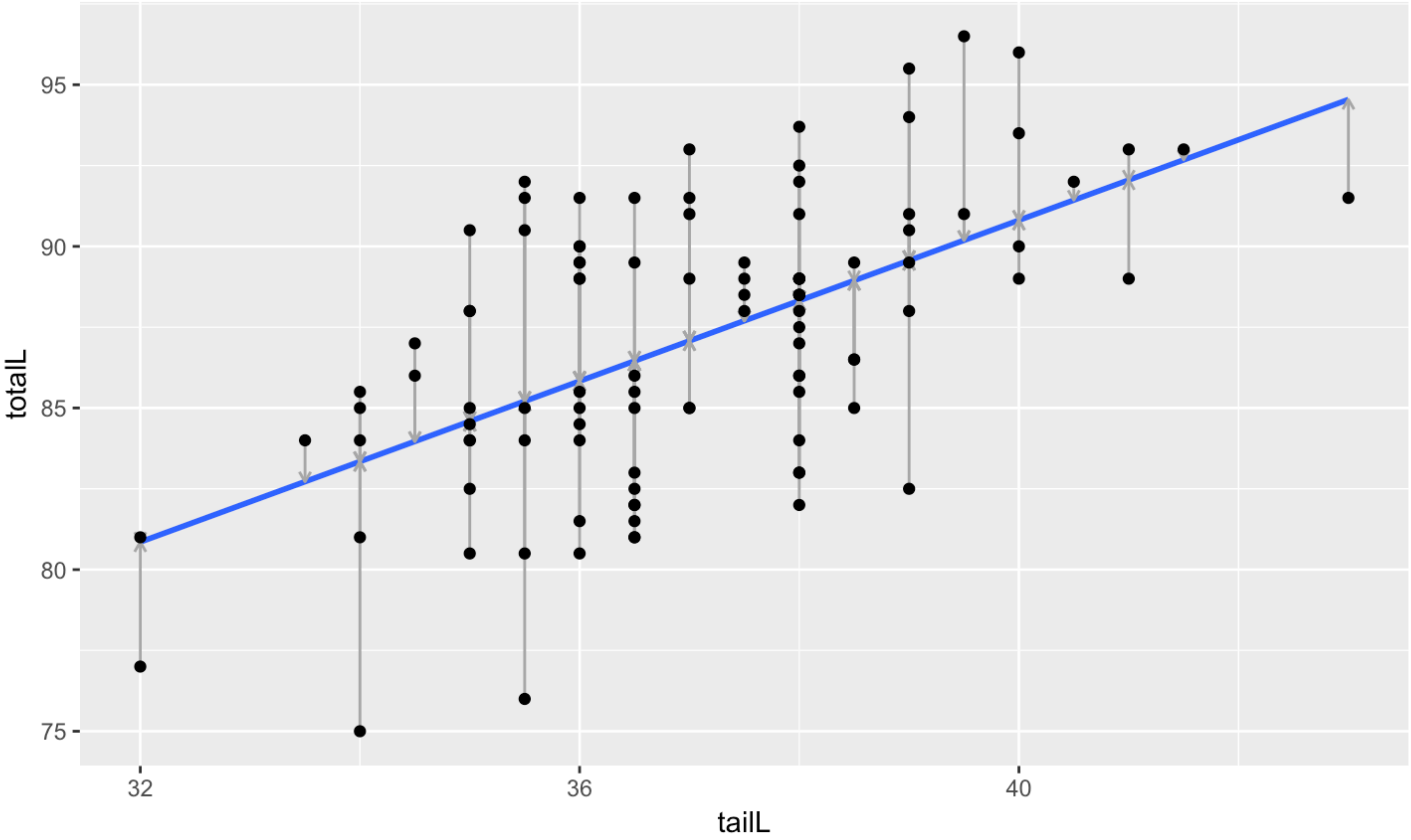


How well does our possum model fit?

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



Sums of squared deviations



SSE

```
> library(broom)
> mod_possum <- lm(totalL ~ tailL, data = possum)
> mod_possum %>%
    augment() %>%
    summarize(SSE = sum(.resid^2),
              SSE_also = (n() - 1) * var(.resid))
   SSE SSE_also
1 1301     1301
```

RMSE

$$RMSE = \sqrt{\frac{\sum_i e_i^2}{d.f}} = \sqrt{\frac{SSE}{n-2}}$$

Residual standard error (possums)

```
> summary(mod_possum)

Call:
lm(formula = totalL ~ tailL, data = possum)

Residuals:
   Min     1Q Median     3Q    Max
-9.210 -2.326  0.179  2.777  6.790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.04       6.66    6.16  1.4e-08
tailL           1.24       0.18    6.93  3.9e-10

Residual standard error: 3.57 on 102 degrees of freedom
Multiple R-squared:  0.32,   Adjusted R-squared:  0.313
F-statistic:   48 on 1 and 102 DF,  p-value: 3.94e-10
```

Residual standard error (textbooks)

```
> lm(uclaNew ~ amazNew, data = textbooks) %>%
    summary()

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Residuals:
   Min      1Q  Median      3Q     Max 
-34.78  -4.57   0.58   4.01  39.00 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9290     1.9354    0.48     0.63
amazNew       1.1990     0.0252   47.60   <2e-16

Residual standard error: 10.5 on 71 degrees of freedom
Multiple R-squared:  0.97,   Adjusted R-squared:  0.969 
F-statistic: 2.27e+03 on 1 and 71 DF,  p-value: <2e-16
```

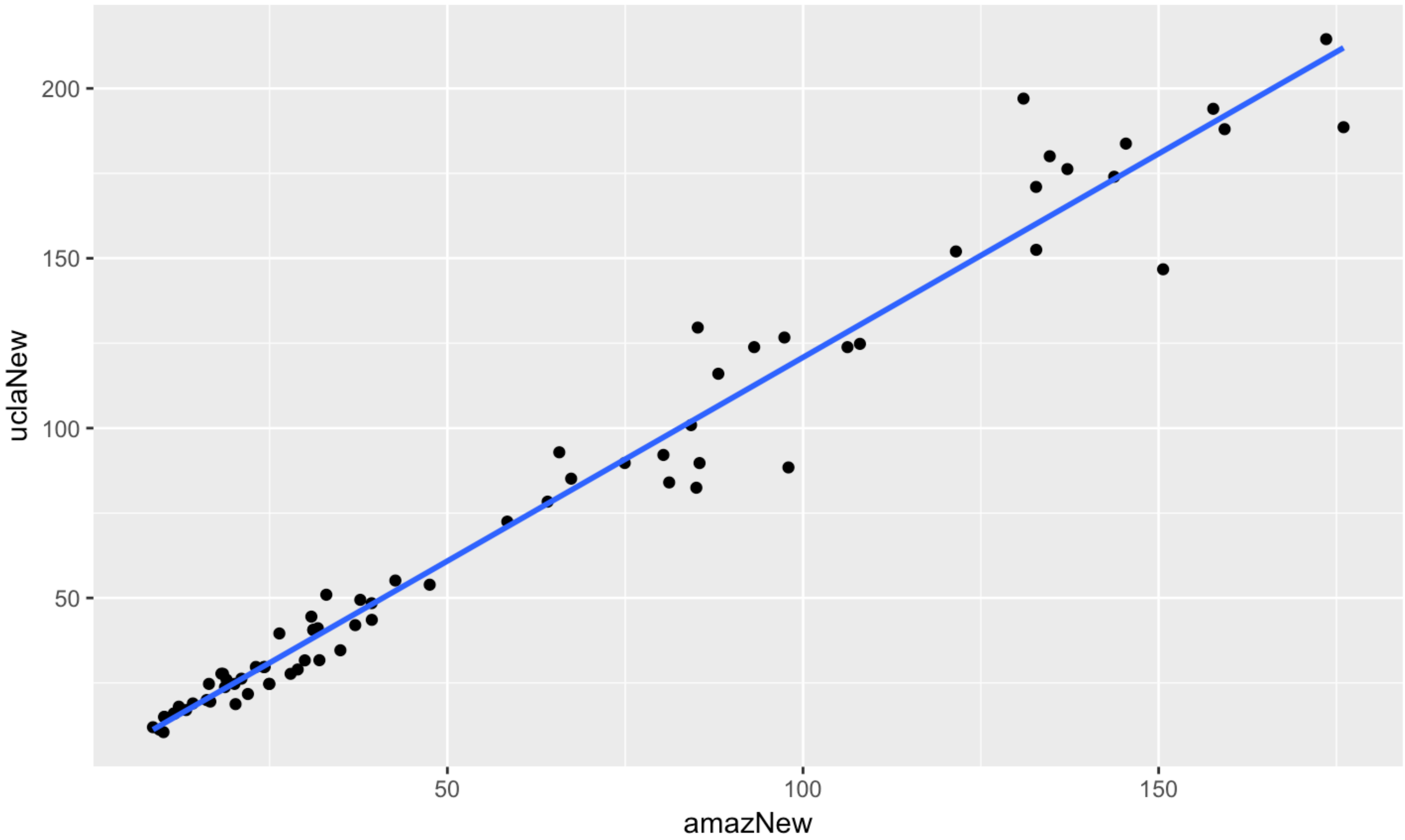
CORRELATION AND REGRESSION

Let’s practice!

Comparing model fits

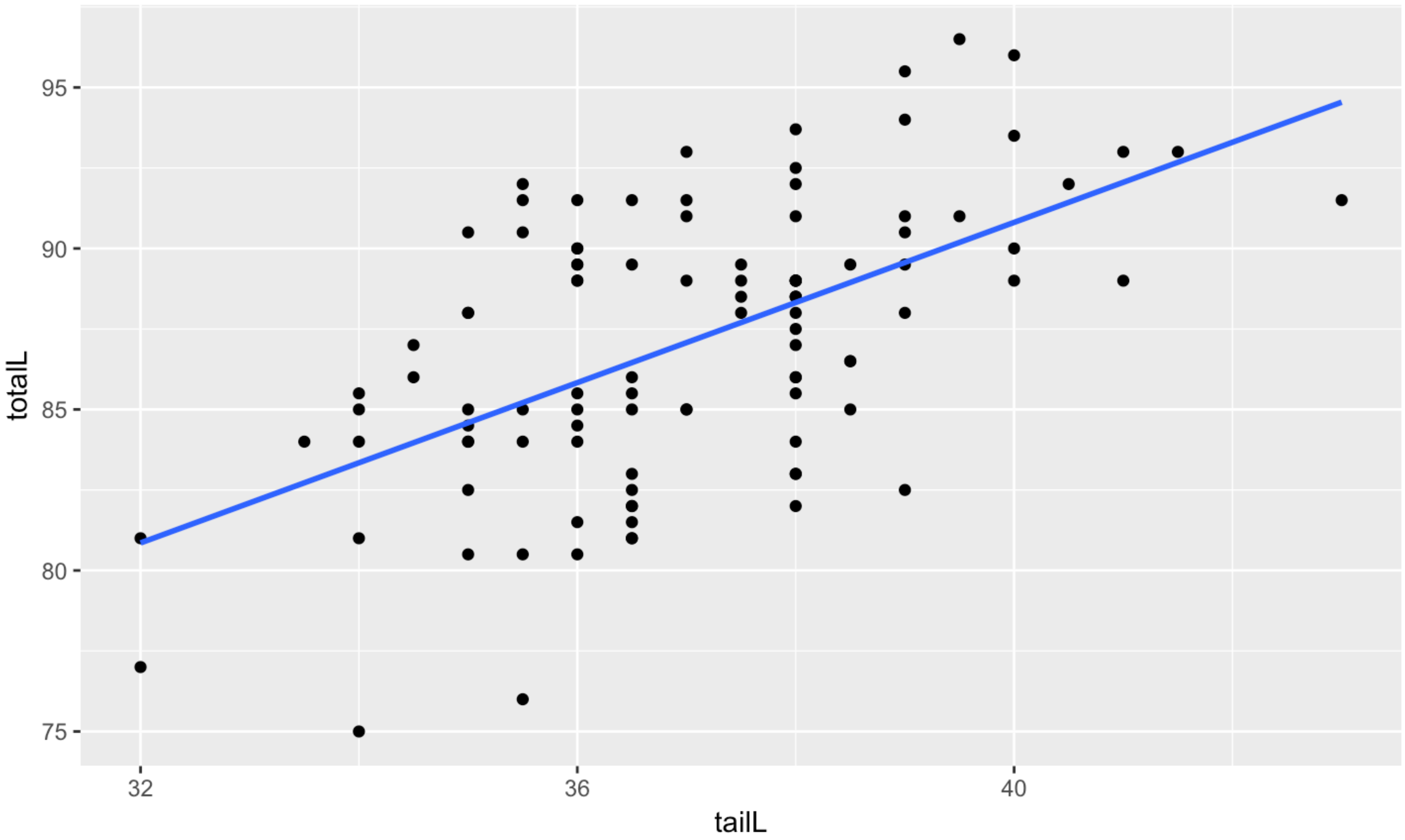
How well does our textbook model fit?

```
> ggplot(data = textbooks, aes(x = amazNew, y = uclaNew)) +
    geom_point() + geom_smooth(method = "lm", se = FALSE)
```



How well does our possum model fit?

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

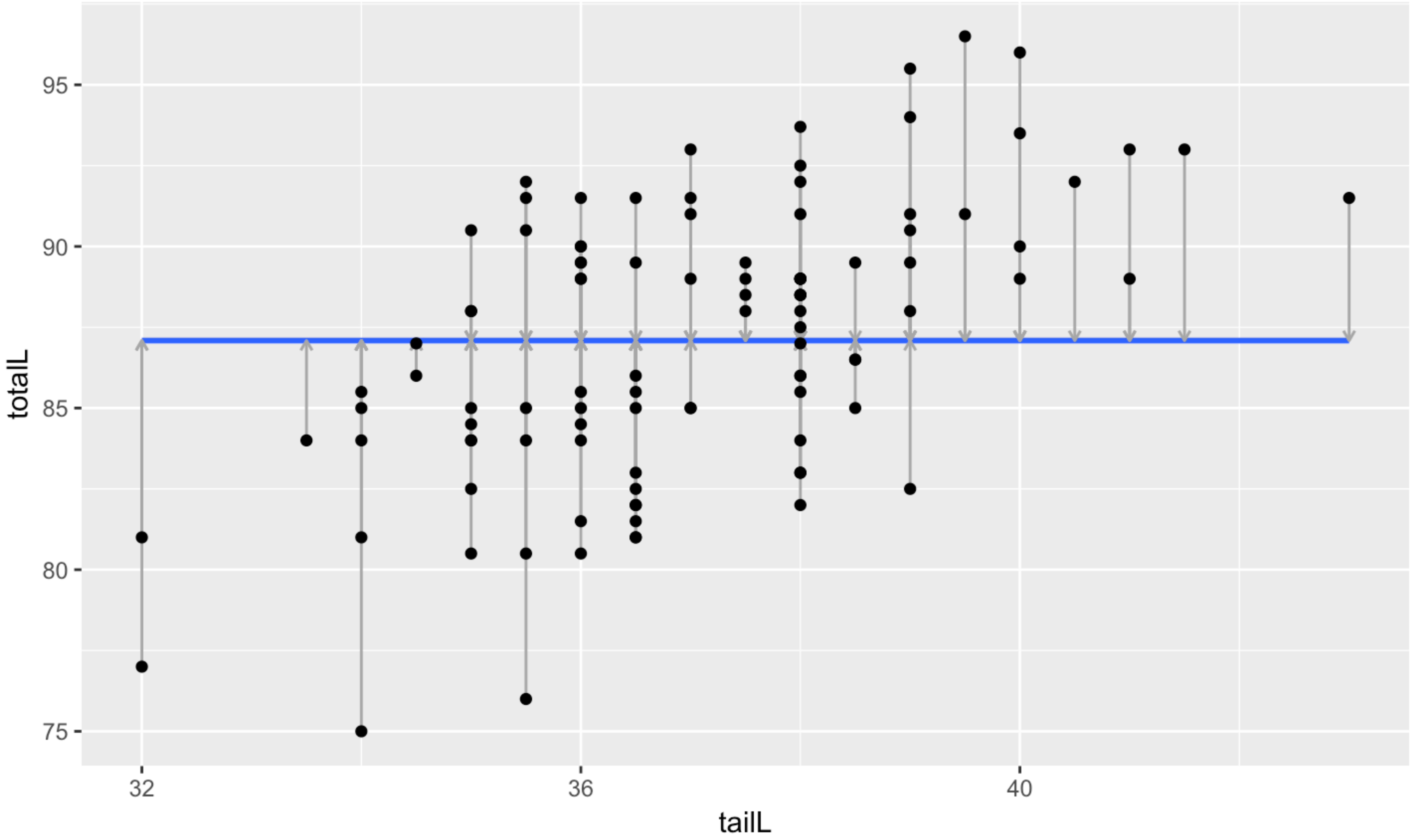


Null (average) model

- For all observations...

$$\hat{y} = \bar{y}$$

Visualization of null model



SSE, null model

```
> mod_null <- lm(totalL ~ 1, data = possum)
> mod_null %>%
    augment(possum) %>%
    summarize(SST = sum(.resid^2))
   SST
1 1914
```

SSE, our model

```
> mod_possum <- lm(totalL ~ tailL, data = possum)
> mod_possum %>%
    augment() %>%
    summarize(SSE = sum(.resid^2))
   SSE
1 1301
```

Coefficient of determination

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{Var(e)}{Var(y)}$$

Connection to correlation

- For simple linear regression...

$$r_{x,y}^2 = R^2$$

Summary

```
> summary(mod_possum)

Call:
lm(formula = totalL ~ tailL, data = possum)

Residuals:
   Min      1Q Median      3Q     Max
-9.210 -2.326  0.179  2.777  6.790

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.04       6.66    6.16  1.4e-08
tailL           1.24       0.18    6.93  3.9e-10

Residual standard error: 3.57 on 102 degrees of freedom
Multiple R-squared:  0.32,   Adjusted R-squared:  0.313
F-statistic:   48 on 1 and 102 DF,  p-value: 3.94e-10
```

Over-reliance on R-squared

> *"Essentially, all models are wrong, but some are useful."*
>
> - George Box

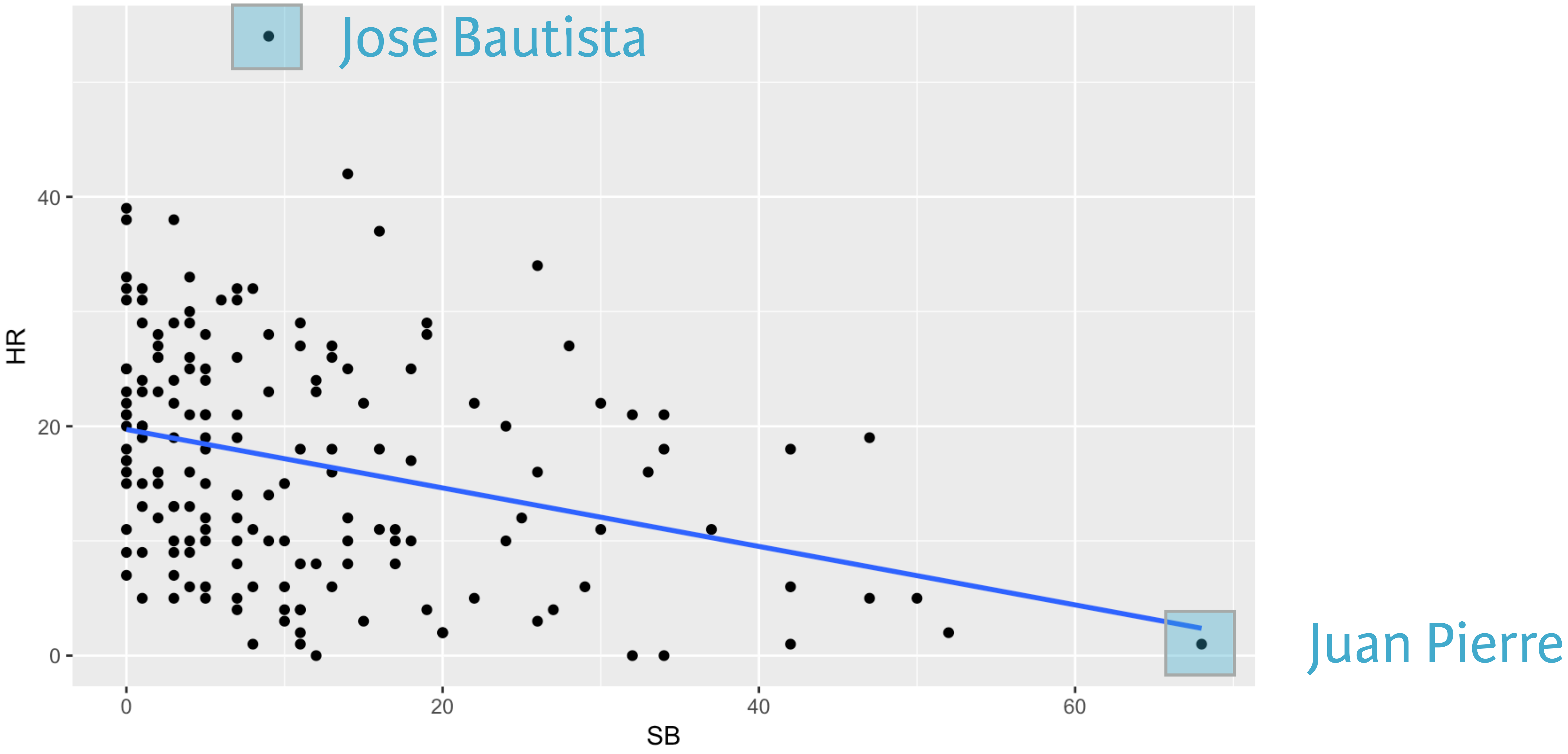
Let’s practice!

CORRELATION AND REGRESSION

Unusual points

Unusual points

```
> regulars <- mlbBat10 %>%
    filter(AB > 400)
> ggplot(data = regulars, aes(x = SB, y = HR)) +
    geom_point() +
    geom_smooth(method = "lm", se = 0)
```



Leverage

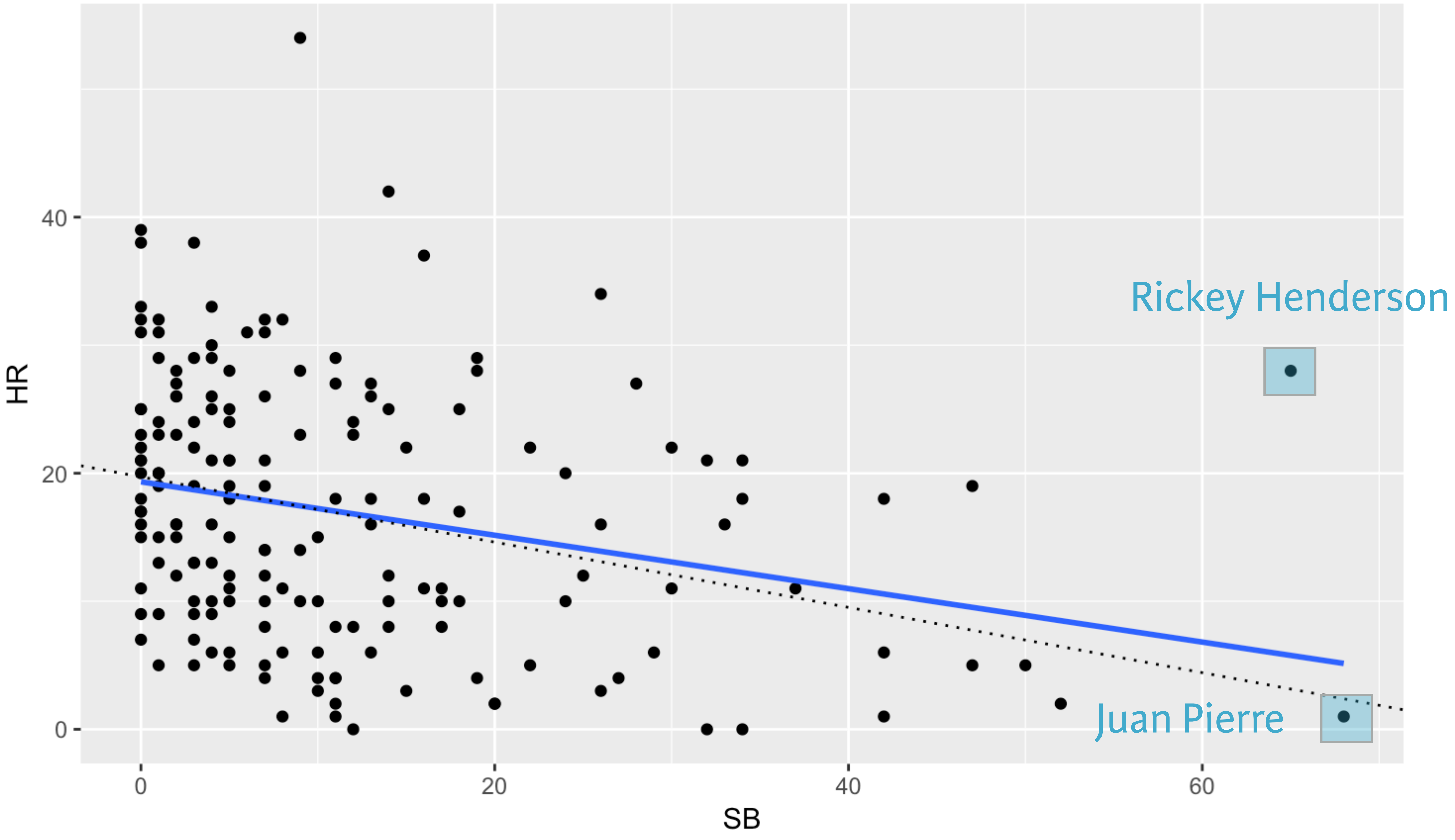
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Leverage computations

```
> library(broom)
> mod <- lm(HR ~ SB, data = regulars)
> mod %>%
    augment() %>%
    arrange(desc(.hat)) %>%
    select(HR, SB, .fitted, .resid, .hat) %>%
    head()
  HR SB .fitted .resid     .hat
1  1 68   2.383 -1.383 0.13082
2  2 52   6.461 -4.461 0.07034
3  5 50   6.971 -1.971 0.06417
4 19 47   7.736 11.264 0.05550
5  5 47   7.736 -2.736 0.05550
6  1 42   9.010 -8.010 0.04261
```

Juan Pierre

Consider Rickey Henderson...



Influence via Cook's distance

```
> mod <- lm(HR ~ SB, data = regulars_plus)
> mod %>%
    augment() %>%
    arrange(desc(.cooksd)) %>%
    select(HR, SB, .fitted, .resid, .hat, .cooksd) %>%
    head()
  HR SB .fitted .resid      .hat .cooksd
1 28 65   5.770 22.230 0.105519 0.33430
2 54  9  17.451 36.549 0.006070 0.04210
3 34 26  13.905 20.095 0.013150 0.02797
4 19 47   9.525  9.475 0.049711 0.02535
5 39  0  19.328 19.672 0.010479 0.02124
6 42 14  16.408 25.592 0.006061 0.02061
```

Henderson

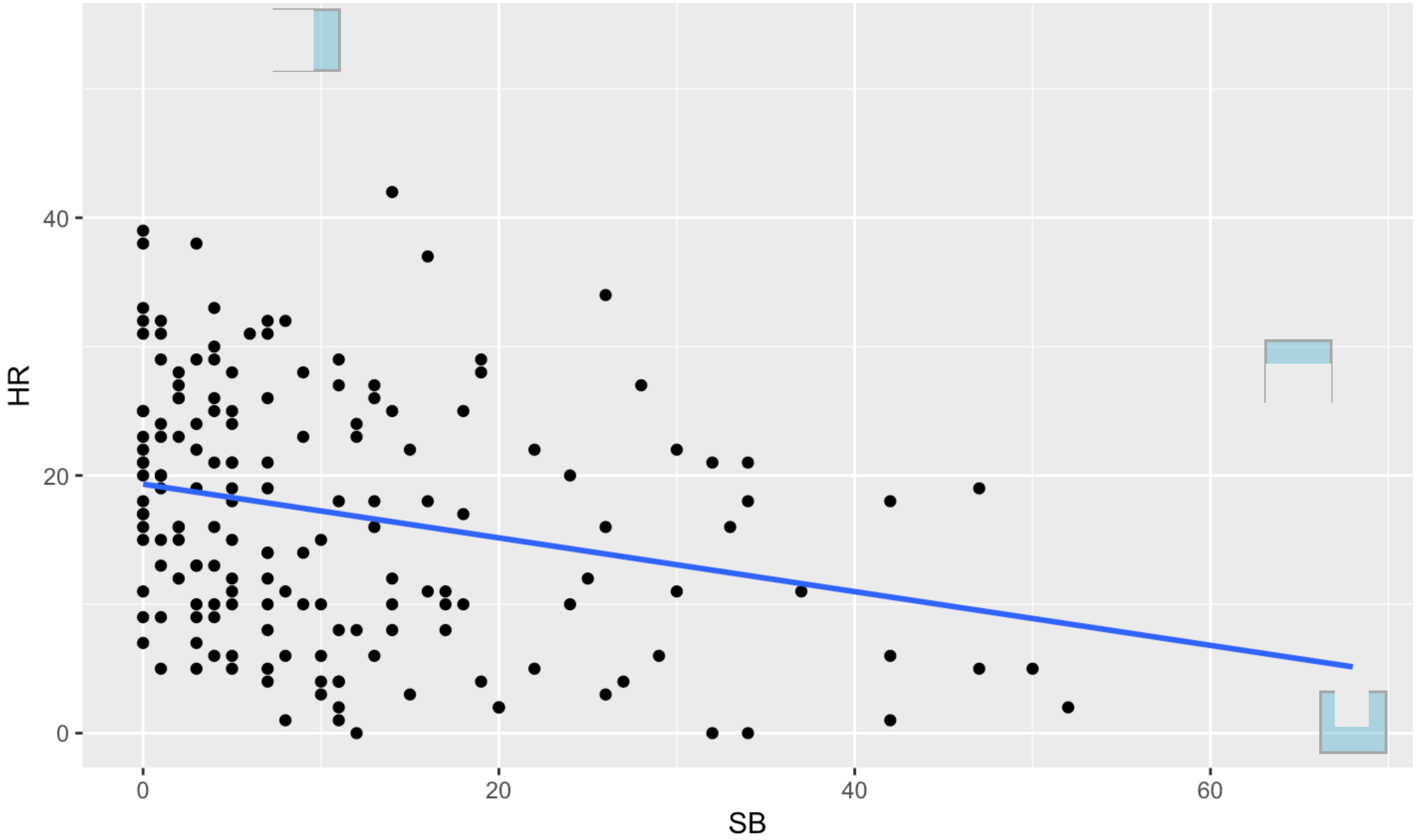
Let’s practice!

CORRELATION AND REGRESSION

Dealing with outliers

Dealing with outliers

```
> ggplot(data = regulars_plus, aes(x = SB, y = HR)) +
    geom_point() +
    geom_smooth(method = "lm", se = 0)
```



The full model

```
> coef(lm(HR ~ SB, data = regulars_plus))
(Intercept)          SB
    19.3282     -0.2086
```

Removing outliers that don't fit

```
> regulars <- regulars_plus %>%
    filter(!(SB > 60 & HR > 20)) # remove Henderson
> coef(lm(HR ~ SB, data = regulars))
(Intercept)          SB
    19.7169     -0.2549
```

- What is the justification?
- How does the scope of inference change?

Removing outliers that do fit

```
> regulars_new <- regulars %>%
    filter(SB < 60) # remove Pierre
> coef(lm(HR ~ SB, data = regulars_new))
(Intercept)          SB
    19.6870     -0.2514
```

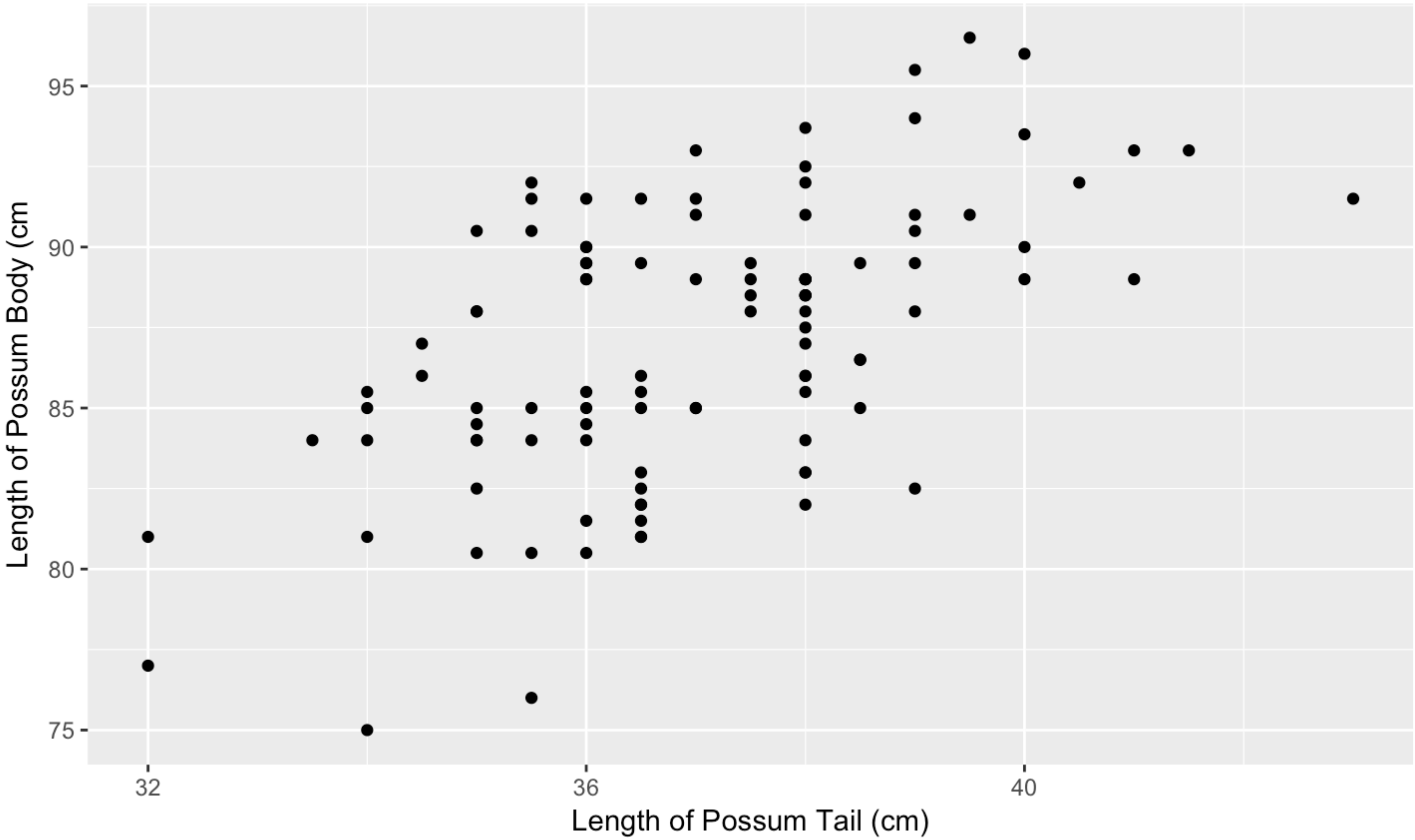
- What is the justification?
- How does the scope of inference change?

CORRELATION AND REGRESSION

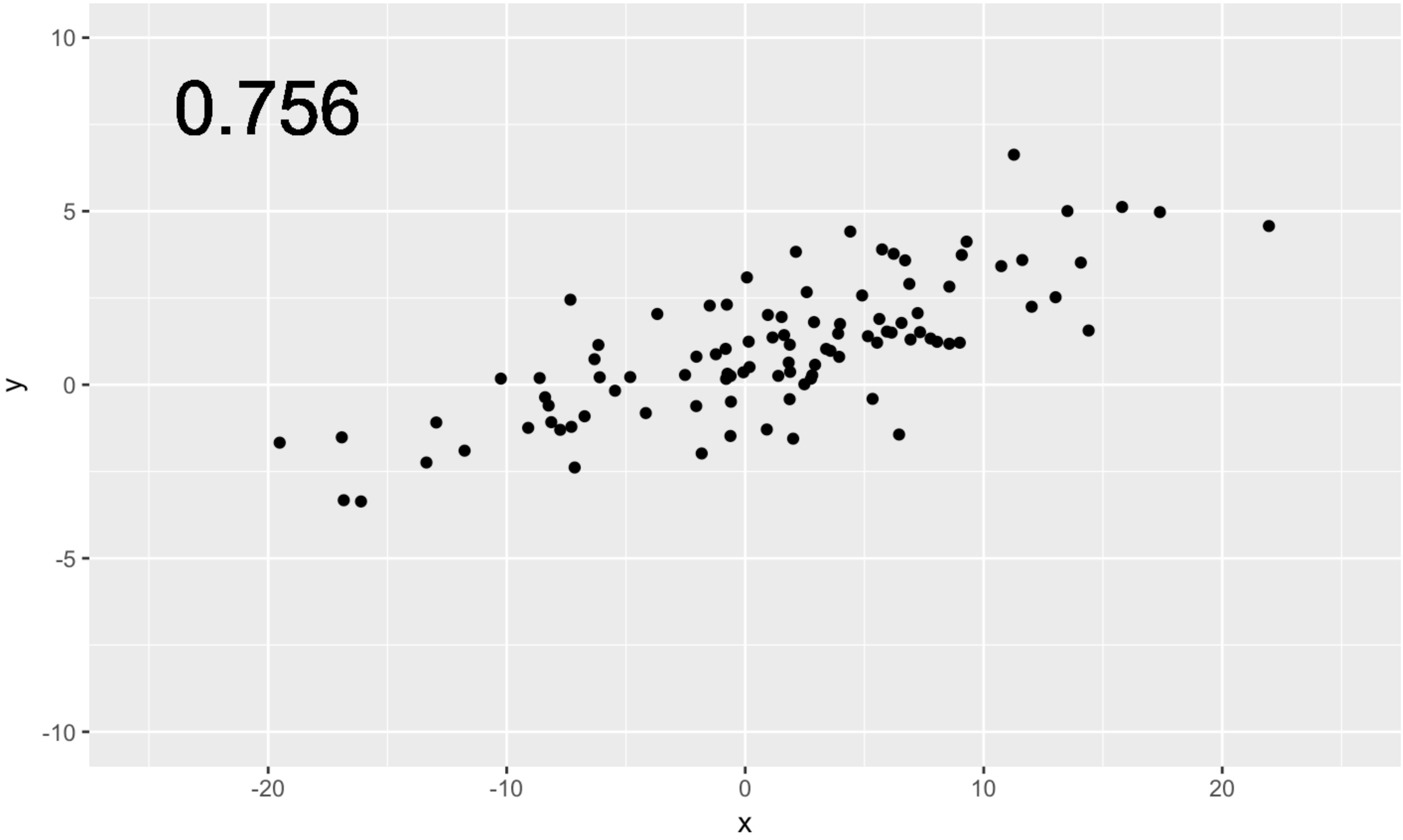
Let's practice!

Conclusion

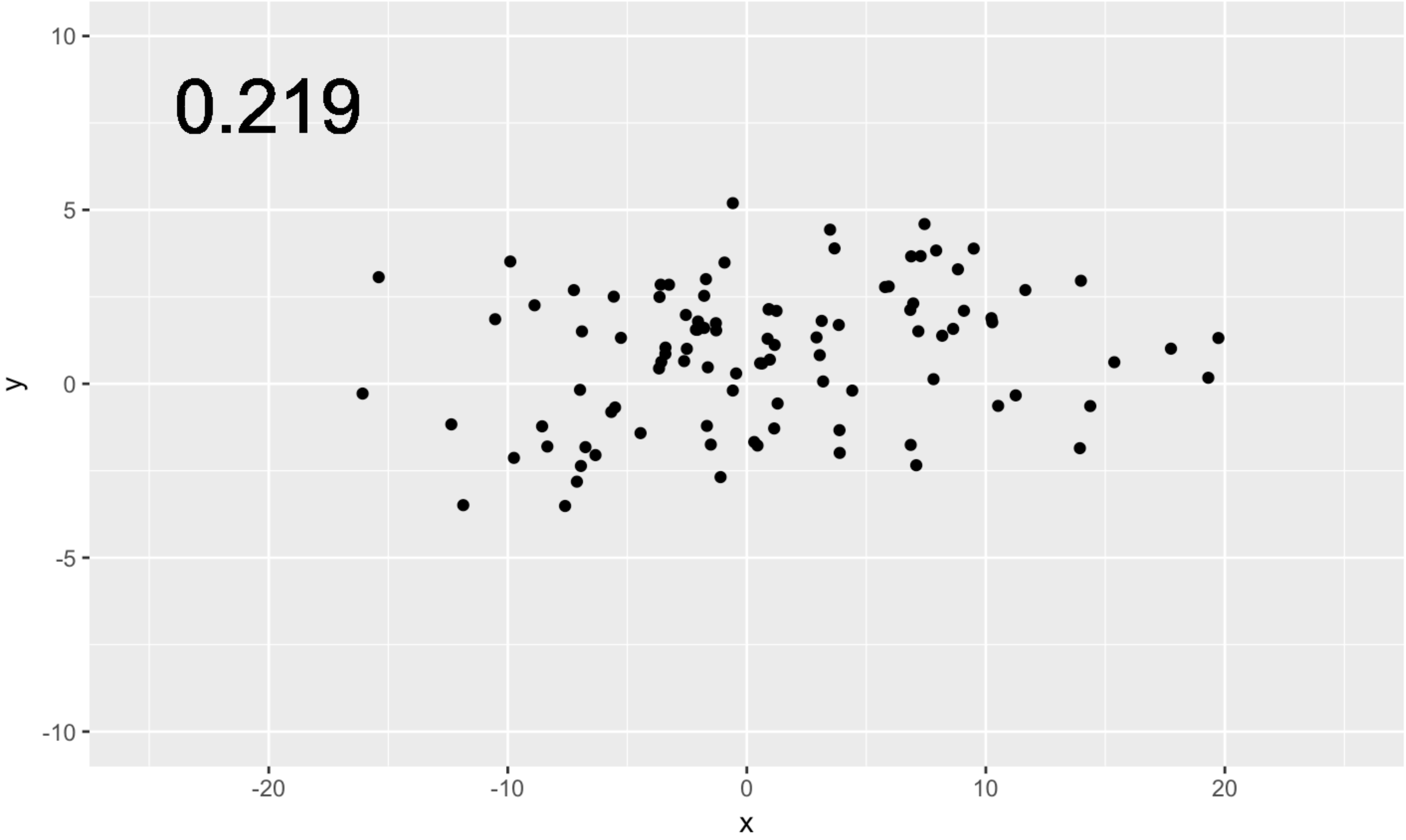
Graphical: scatterplots



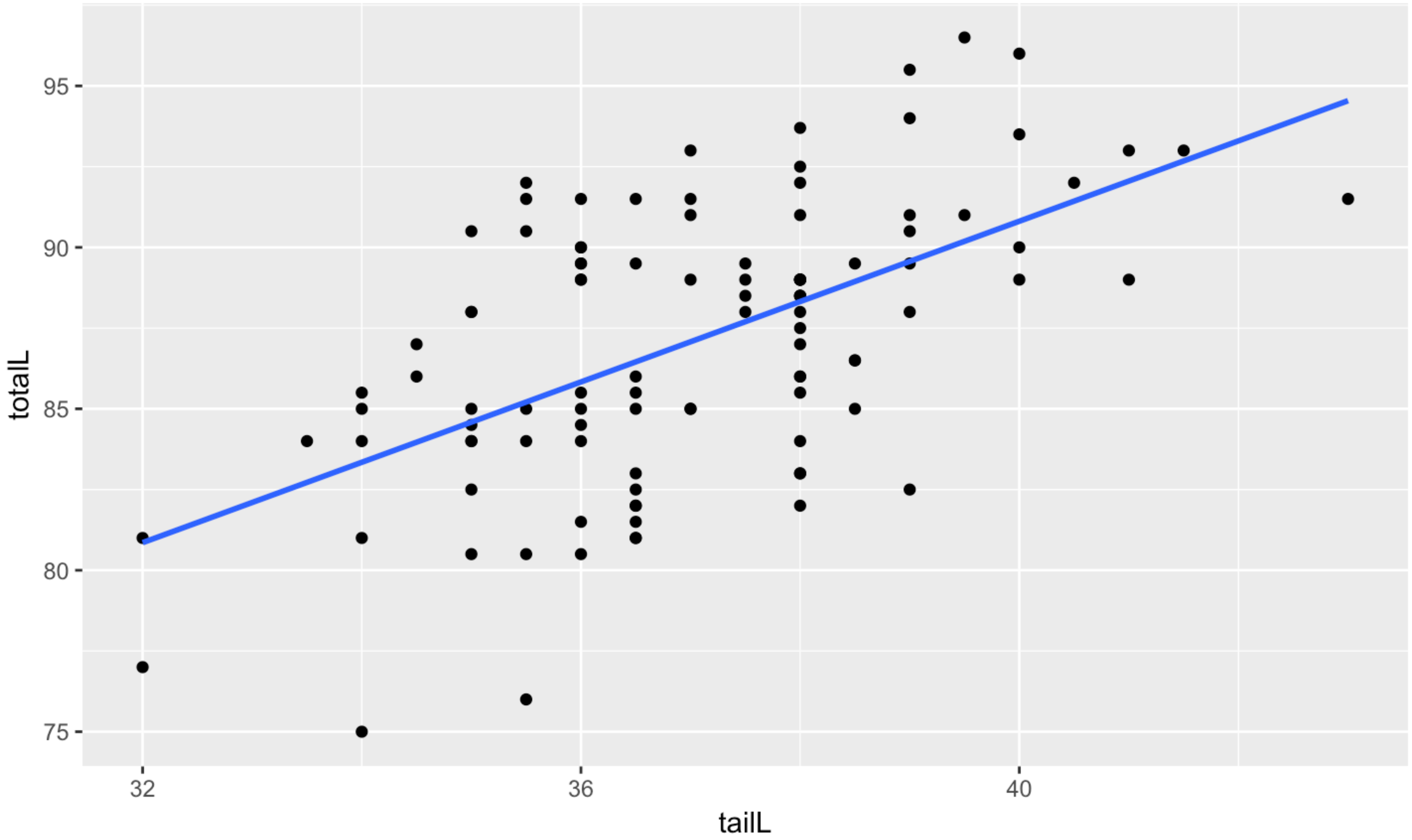
Numerical: correlation



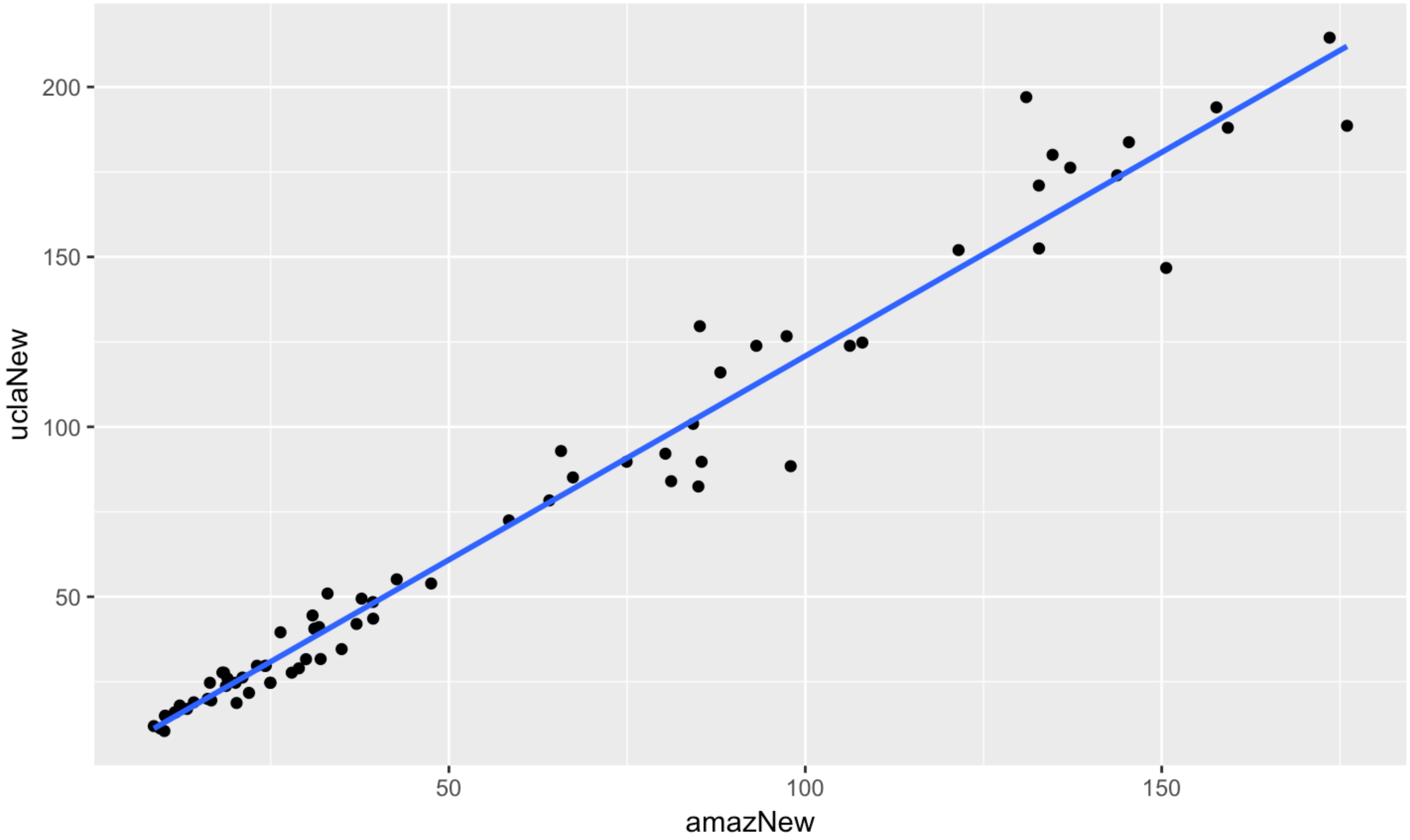
Numerical: correlation



Modular: linear regression



Focus on interpretation

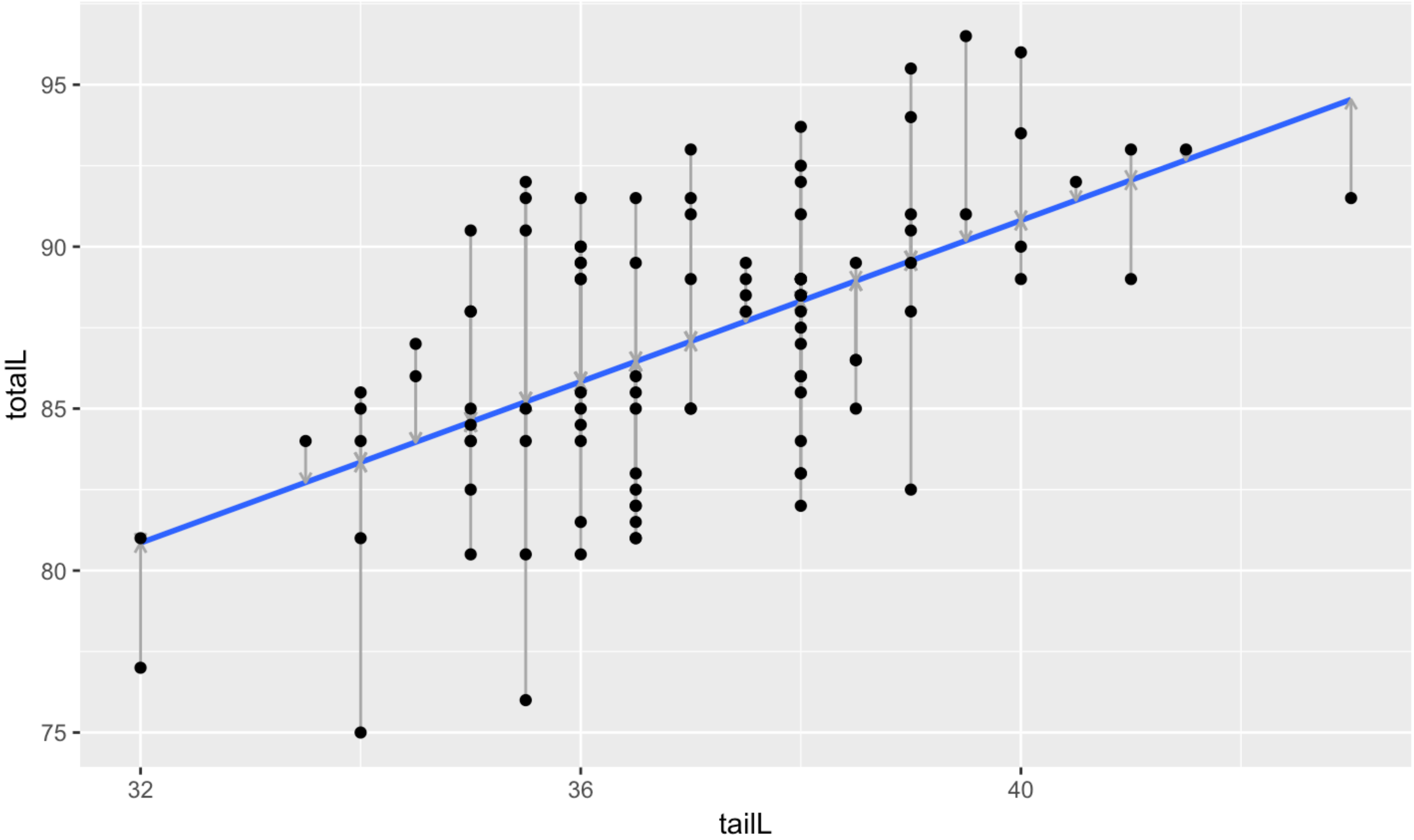


$$\widehat{uclaNew} = 0.929 + 1.199 \cdot amazNew$$

Objects and formulas

```
> summary(mod)

Call:
lm(formula = uclaNew ~ amazNew, data = textbooks)

Residuals:
   Min     1Q Median     3Q    Max
-34.78  -4.57   0.58   4.01  39.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9290     1.9354    0.48     0.63
amazNew       1.1990     0.0252   47.60   <2e-16

Residual standard error: 10.5 on 71 degrees of freedom
Multiple R-squared:  0.97,   Adjusted R-squared:  0.969
F-statistic: 2.27e+03 on 1 and 71 DF,  p-value: <2e-16
```

Model fit



Thanks!